Received: 12 June 2020 | Revised: 16 December 2020 | Accepted: 8 March 2021

DOI: 10.1002/pst.2122

MAIN PAPER

WILEY

Improving sample size recalculation in adaptive clinical trials by resampling

Carolin Herrmann[1] | **Corinna Kluge**[1] | **Maximilian Pilz**[2] | **Meinhard Kieser**[2] | **Geraldine Rauch**[1]

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Biometry and Clinical Epidemiology, Charitéplatz 1, 10117 Berlin, Germany

[2]Institute of Medical Biometry and Informatics, University Medical Center Ruprechts-Karls University Heidelberg, Heidelberg, Germany

Correspondence
Carolin Herrmann, Institute of Biometry and Clinical Epidemiology, Charité—Universitätsmedizin Berlin, Charitéplatz 1, Berlin 10117, Germany.
Email: carolin.herrmann@charite.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: KI 708/4-1, RA 2347/4-1; Charite Universitatsmedizin Berlin

Abstract

Sample size calculations in clinical trials need to be based on profound parameter assumptions. Wrong parameter choices may lead to too small or too high sample sizes and can have severe ethical and economical consequences. Adaptive group sequential study designs are one solution to deal with planning uncertainties. Here, the sample size can be updated during an ongoing trial based on the observed interim effect. However, the observed interim effect is a random variable and thus does not necessarily correspond to the true effect. One way of dealing with the uncertainty related to this random variable is to include resampling elements in the recalculation strategy. In this paper, we focus on clinical trials with a normally distributed endpoint. We consider resampling of the observed interim test statistic and apply this principle to several established sample size recalculation approaches. The resulting recalculation rules are smoother than the original ones and thus the variability in sample size is lower. In particular, we found that some resampling approaches mimic a group sequential design. In general, incorporating resampling of the interim test statistic in existing sample size recalculation rules results in a substantial performance improvement with respect to a recently published conditional performance score.

KEYWORDS

adaptive group sequential design, clinical trial, resampling, sample size recalculation

1 | INTRODUCTION

In clinical trials, a wrong specification of parameter values required for sample size calculations can have severe consequences: If the sample size is too small, an underlying relevant treatment effect cannot be detected. If the sample size is too large, patients recruited at later time points might be assigned to a treatment that is already known to be less efficient. Both scenarios are highly questionable from an ethical point of view.

Adaptive group sequential study designs are an attractive option to deal with planning uncertainties. At a pre-defined interim analysis, an unblinded data evaluation is performed. Based on the interim results, the trial may be stopped early or continued

Carolin Herrman and Corinna Kluge contributed equally to this article.

with a sample size adaptation. There exist different possibilities on how to adapt the sample size based on the interim results. Various advantages of adaptive designs are described in the recently published draft guideline by the Food and Drug Administration on adaptive designs,[1] yet no explicit advice is given on the choice of sample size recalculation rules. Likewise, the European Medicine Agency gives no clear advice on how to select a certain sample size recalculation rule in their reflection paper on adaptive designs.[2] The most common approach for adapting the sample size is based on conditional power arguments.[3] In this paper, the conditional power describes the probability of correctly rejecting the null hypothesis at the final analysis, given the observed interim data and the assumed true effect. Sample size recalculation rules based on conditional power arguments go back to Proschan and Hunsberger[4] as well as Lehmacher and Wassmer[5] in the 1990s. Nowadays, there exists a great variety of extended approaches based on these initial methods,[6-8] Another strategy is to consider the sample size update as the solution of an optimization problem in terms of a selected set of design features (e.g., conditional power, total sample size).[9,10] Although there thus exists a high number of strategies for recalculation, there are still major problems related to sample size recalculation. Those problems are, for example, not meeting the target power, large recalculated sample sizes and a high variability in the recalculated sample size.[11,12] Particularly, Bauer et al.[3] noted that the assumption that the observed interim effect is the true effect "may lead to a highly variable distribution of the conditional power resulting in highly variable sample sizes."

The uncertainty of the interim result is indeed one of the major problems with sample size recalculation. One way to tackle this problem is to use a Bayesian approach. Here, a certain probability distribution for the true effect is assumed. More precisely, the sample size calculation can be based on the *Bayesian predictive power*[13] instead of the conditional power. The underlying idea is to determine an average conditional power based on a prior distribution assumption for the true effect. The foundation for this approach was laid by Spiegelhalter et al.[14,15]

Another option to take account of the whole distribution of the interim test statistic is to implement resampling or bootstrapping approaches. For many applications it is known that the interim test statistic is approximately normally distributed where the expected value is just the test statistic for the true standardized effect and the variance equals 1. Naturally, the observed interim test statistic corresponds to the expectation of its bootstrapping distribution. By inserting the transformed interim test statistic as the expected value of the distribution, this allows us to draw random samples from this distribution directly instead of bootstrapping from the complete data set. This clearly reduces the computational effort. For this reason, we will use the term "resampling" in the following, although the procedure mimics a bootstrapping approach.

Preliminary work on the combination of bootstrapping with sample size recalculation was conducted by Hade et al.[16] For a blinded survival data setting, they proposed to use additional information from prior trials just before the ending of the planned accrual time. By repeatedly estimating the baseline survival function, the sample size was re-estimated. They determined the 80%-quantile of the bootstrapped distribution as the final sample size.

In this article, we propose to combine a resampling approach with a variety of existing sample size recalculation rules. The resampling approach can be applied to different types of endpoints with approximately normally distributed test statistics, however, we focus on a normally distributed endpoint for the sake of simplicity. By this approach, we treat the interim test statistic as a random variable. The idea is to repeatedly draw test statistics from the interim test statistic distribution, which is approximated by plugging in the observed interim test statistic as expectation. For each randomly drawn test statistic, we recalculate the sample size according to an established sample size recalculation rule. All these recalculated sample sizes are then combined into a single updated sample size for the second stage. We evaluate the resampling approach by means of performance characteristics from the *conditional perspective*. This refers to evaluating the sample size and power under the assumption of already knowing that the study is not stopped at interim, thus actually having the possibility to recalculate the sample size.

The article is structured as follows. In the following methods section, we describe the test problem, common recalculation strategies, and the new resampling approach that accounts for the uncertainty of the observed interim effect. Moreover, we give a short overview of the performance criteria used to qualify "good" recalculation strategies. The third section displays the simulation to evaluate the performance of the existing and new approach. The results are compared and discussed in detail in the last section. General recommendations for further applications are deduced.

2 | METHODS

2.1 | The test problem

We consider a randomized, controlled, two-armed clinical trial. The n observations of the intervention group I and control group C follow a normal distribution with means μ^I and μ^C and common variance σ^2, that is

$$X_j^I \sim \mathcal{N}\left(\mu^I, \sigma^2\right) \text{ and } X_j^C \sim \mathcal{N}\left(\mu^C, \sigma^2\right),\ j = 1, \ldots, n. \tag{1}$$

Throughout this work, we investigate the one-sided superiority test problem

$$H_0 : \mu^I - \mu^C \leq 0 \text{ versus } H_1 : \mu^I - \mu^C > 0, \tag{2}$$

hence referring to a setting where large values of the endpoint are considered as favorable. We consider an adaptive group sequential design with two stages, which is the simplest and most frequently applied adaptive group sequential design.[3] Hence, we have to define two independent test statistics,

$$T_i = \frac{\bar{X}_i^I - \bar{X}_i^C}{S_{\text{pooled},i}} \cdot \sqrt{\frac{n_i}{2}}, \tag{3}$$

where $i \in \{1, 2\}$ refers to the stages, $\bar{X}_i^I$ and $\bar{X}_i^C$ to the respective means, $S_{\text{pooled},i}$ to the pooled standard deviation as well as n_i to the sample size per group in stage i with $n_1 + n_2 = n$. Note that T_1 exclusively includes the data of the first and T_2 only those data of the second stage, both following approximately a normal distribution.

The trial is continued with the second stage if the interim test statistic T_1 falls within the so called *recalculation area* (RA) given by the interval $\left[q_{1-\alpha_0}; q_{1-\alpha_1}\right)$, where α_0 refers to a futility stopping bound for the one-sided p-value of stage one, α_1 to the local one-sided significance level and q indicates the respective quantiles of the normal distribution. Hence, the trial is stopped after the first stage with an early rejection of the null hypothesis if $T_1 \geq q_{1-\alpha_1}$, or with acceptance of the null hypothesis if $T_1 < q_{1-\alpha_0}$.

All observed data over the two stages are combined by means of the inverse normal combination test[11] represented by the combined test statistic

$$T_{1+2} = \frac{w_1 \cdot T_1 + w_2 \cdot T_2}{\sqrt{w_1^2 + w_2^2}}, \tag{4}$$

consisting of the two stochastically independent test statistics T_1 and T_2, where we choose the weights $w_1 = \sqrt{n_1}$ and $w_2 = \sqrt{n_2}$. The null hypothesis is rejected at the final analysis if $T_{1+2} \geq q_{1-\alpha_{1+2}}$, where α_{1+2} corresponds to the local one-sided significance level for the final analysis. Local significance levels can, for example, be chosen according to the adjustments proposed by Pocock[17] or by O'Brien and Fleming.[18]

2.2 | Sample size recalculation approaches

There exist different ways for adapting the sample size during an ongoing trial. One of the simplest methods is the group sequential study design (GS), where the sample size for every stage equals a fixed pre-defined number. Group sequential designs may be considered as a special case of adaptive group sequential designs where the stage-wise sample sizes can be chosen in a data-dependent way.[5,19] Comparisons of the group sequential and the adaptive group sequential study designs can be found in References.[12,20-22] The most popular strategy in adaptive group sequential study designs is to update the sample size such that a certain pre-specified conditional power value is reached.[3] The *conditional power* describes the probability of correctly rejecting the null hypothesis, given the observed value of the test statistic at interim $T_1 = t_1$ and total sample size n per group. Moreover, it depends on the true standardized treatment effect $\Delta = (\mu^I - \mu^C)/\sigma$. The corresponding formula looks as follows:

$$CP_\Delta(t_1, n) := \begin{cases} 0, \text{ if trial is stopped early for futility,} \\ 1 - \Phi\left(q_{1-\alpha_{1+2}} \cdot \frac{\sqrt{w_1^2 + w_2^2}}{w_2} - t_1 \cdot \frac{w_1}{w_2} - \Delta \cdot \sqrt{\frac{n_1}{2}} \cdot \sqrt{\frac{n - n_1}{n_1}} \right), \\ \text{if the sample size is recalculated,} \\ 1, \text{ if trial is stopped early for efficacy.} \end{cases} \tag{5}$$

In the following, we describe three ways of recalculating the sample size based on the *observed conditional power* which means that in the above formula Δ is replaced by the observed interim effect $t_1 \cdot \sqrt{2/n_1}$. There are similar approaches which insert the assumed effect for Δ, also referred to *anticipated conditional power*. Since the focus of this work is on the effects of the proposed resampling tool, which can be combined with every existing recalculation rule, it is not so essential here to cover a wide range of recalculation rules differing just by the way which value to employ for Δ. Therefore, we only consider recalculation rules based on the observed conditional power here. For all investigated approaches, we limit the maximal total sample size per group to $n_{\max}$ for feasibility reasons. A more detailed description of the following three recalculation rules is given in Herrmann et al.[23]

2.2.1 | The observed conditional power approach

For observed interim test statistics t_1 falling in the recalculation area $\left[q_{1-\alpha_0}; q_{1-\alpha_1}\right)$, the sample size per group that ensures an observed conditional power of $1-\beta$ is determined by the smallest integer $\tilde{n}$ that fulfills the inequality

$$\tilde{n} \geq n_1 \cdot \left(1 + \left(\frac{q_\beta - q_{1-\alpha_{1+2}} \cdot \frac{\sqrt{w_1^2 + w_2^2}}{w_2} + t_1 \cdot \frac{w_1}{w_2}}{t_1}\right)^2\right). \tag{6}$$

The total sample size per group according to the *observed conditional power (OCP) approach* is given by

$$n_{\mathrm{OCP}}(t_1) = \begin{cases} \min(\tilde{n}(t_1), n_{\max}) & \text{if } t_1 \in \mathrm{RA}, \\ n_1 & \text{else.} \end{cases} \tag{7}$$

2.2.2 | The restricted observed conditional power approach

The *restricted observed conditional power (ROCP) approach* is very similar to the observed conditional power approach but, as the name suggests, it contains a restriction. One point of criticism regarding the observed conditional power approach is that when formula (6) demands higher sample sizes than the maximal sample size $n_{\max}$, the sample size is fixed to $n_{\max}$ irrespective of the conditional power that can be obtained by this highest sample size. Hence, it could be reasonable to augment the sample only if a certain minimal conditional power $1-\beta_{\mathrm{low}}^{\mathrm{ROCP}}$ can be attained. Consequently, the total sample size per group for the ROCP approach equals

$$n_{\mathrm{ROCP}}(t_1) = \begin{cases} \min(\tilde{n}(t_1), n_{\max}) & \text{if } t_1 \in \mathrm{RA} \\ & \text{and } CP(t_1, n_{\max}) \geq 1-\beta_{\mathrm{low}}^{\mathrm{ROCP}}, \\ n_1 & \text{else.} \end{cases} \tag{8}$$

2.2.3 | The promising zone approach

The *promising zone (PZ) approach* was proposed by Mehta and Pocock.[8] They specify an initial total sample size n_{ini} per group that is smaller than the maximally allowed total sample size $n_{\max}$ per group. Moreover, they define a lower bound for the conditional power $1-\beta_{\mathrm{low}}^{\mathrm{PZ}}$. Note that $1-\beta_{\mathrm{low}}^{\mathrm{PZ}}$ does not necessarily need to equal $1-\beta_{\mathrm{low}}^{\mathrm{ROCP}}$. Depending on the observed interim test statistic t_1, the updated sample size is determined either by the initially proposed total sample size n_{ini}, as $\tilde{n}$ according to (6), or as the limiting maximal sample size $n_{\max}$ per group. Explicitly, the total sample size per group according to the *PZ* approach is given by

$$n_{PZ}(t_1) = \begin{cases} \min(\tilde{n}(t_1), n_{max}) & \text{if } t_1 \in RA \text{ and } 1-\beta_{low}^{PZ} \leq CP(t_1, n_{ini}) < 1-\beta, \\ n_{ini} & \text{if } t_1 \in RA \text{ and } CP(t_1, n_{ini}) < 1-\beta_{low}^{PZ}, \\ & \text{or } t_1 \in RA \text{ and } CP(t_1, n_{ini}) \geq 1-\beta, \\ n_1 & \text{else.} \end{cases} \tag{9}$$

2.3 | Performance evaluation for sample size recalculation rules

When aiming at improving sample size recalculation rules, appropriate performance evaluation criteria need to be prespecified. Typical evaluation criteria are the average sample size as well as global power. Both can be considered as random variables in the adaptive design setting. Herrmann et al.[23] pointed out that it is important to not only consider these location measures but to additionally provide variation measures. Generally, there are different perspectives for evaluating the performance of a sample size recalculation rule. The global perspective is the one before the trial is started and thus takes an "average" look on the two options to stop the trial early or to recalculate the sample size at interim. However, it comes with the difficulty that the mixture of performance features related to an early stop and performance features related to a sample size recalculation may be difficult to interpret. Another option is to consider the conditional perspective. Here, the researcher asks before the trial is started how the sample size should be recalculated if at interim the observed effect falls within the recalculation area. This means that we evaluate the recalculation rules under the assumption that we already know that the trial is not stopped at interim $\left(t_1 \in \left[q_{1-\alpha_0}; q_{1-\alpha_1}\right)\right)$ but we do *not* know the observed t_1-value yet. The word "conditional" thus refers to the recalculation area and not to a particular value of t_1. In this paper, we focus on this conditional perspective. Therefore, we investigate sample size recalculation possibilities under the following performance measures:

1. the expected conditional power $\mathbb{E}\left[CP_{\Delta}^{RA}\right]$,
2. the variance of the conditional power $\mathrm{Var}\left(CP_{\Delta}^{RA}\right)$,
3. the expected conditional total sample size per group $\mathbb{E}\left[CN_{\Delta}^{RA}\right]$, which is the average sample size per group conditional on having entered the recalculation area, and
4. the variance of the conditional total sample size per group $\mathrm{Var}\left(CN_{\Delta}^{RA}\right)$.

All performance measures are simulated under a range of true standardized effect sizes $\Delta = \frac{\mu^I - \mu^C}{\sigma}$. Moreover, all these evaluation criteria can be combined within a single performance value, the *conditional performance score CS* by Herrmann et al.[23] Here, we only describe the key features of this score. The conditional performance score consists of four components: a location and variation component for the conditional power ($e_{CP}(\Delta)$ and $v_{CP}(\Delta)$), as well as a location and variation component for the conditional sample size ($e_{CN}(\Delta)$ and $v_{CN}(\Delta)$). The underlying idea for the location components is to evaluate the expected values in relation to pre-defined target values. If the maximally allowed sample size is not greater than the related fixed sample size and the effect size does not equal zero, then the initially planned power value $1-\beta$ is taken as target value for the conditional power and the fixed sample size as target value for the conditional sample size. In the other cases, the first stage's sample size n_1 and the global one-sided significance level α are defined as target values since the trial may then be declared as not worth the effort to continue with the second stage.[23] Concerning the variation components, the observed variation is compared to the maximally possible variation in the respective setting. All four score components can take values between 0 and 1. Hence, they can all be evaluated separately and it is possible to combine them in two sub-scores, a *conditional power sub-score* SCP(Δ) and a *conditional sample size sub-score* SCN(Δ), or to a single performance value, the conditional performance score CS given by

$$CS(\Delta) = \frac{1}{2} \cdot (SCP(\Delta) + SCN(\Delta)). \tag{10}$$

For all (sub-)scores and components it holds that larger values correspond to a better performance. The components within the two sub-scores can be weighted in different ways by considering, for example for the conditional power sub-score

$$\mathrm{SCP}(\Delta) = \gamma_{\mathrm{loc}} \cdot e_{\mathrm{CP}}(\Delta) + \gamma_{\mathrm{var}} \cdot v_{\mathrm{CP}}(\Delta), \text{with } \gamma_{\mathrm{loc}} + \gamma_{\mathrm{var}} = 1, \tag{11}$$

where γ_{loc} and γ_{var} describe the respective weights for the location component e_{CP} as well as variation component v_{CP}. The same construction applies to the conditional sample size sub-score. In this paper, we consider an equal weighting of all components, that is, $\gamma_{\mathrm{loc}} = \gamma_{\mathrm{var}} = 0.5$. The detailed formulas for the conditional performance components and (sub-) scores can be found in the Appendix and were initially published by Herrmann et al.[23]

2.4 | Resampling approach for sample size recalculation

To incorporate the variability of the interim effect, resampling—as a tool to assess the variability of a random variable—may be an option worth to be considered. The resampling approach is only performed if the observed interim test statistic falls within the recalculation area, hence a second stage is suggested. Therefore, we resample B test statistics from a normal distribution with the observed interim test statistic as mean and a standard deviation of 1. All resampled test statistics, also the ones that do not fall into the RA, are included in the computation of the final value of the second stage's sample size as follows: For each of the B resampled test statistics, the second-stage sample size is recalculated resulting in a set of samples sizes $\tilde{n}_{(*),1}(t_1), \tilde{n}_{(*),2}(t_1), ..., \tilde{n}_{(*),B}(t_1)$, where (*) denotes the index for the initial sample size recalculation rule. Note that some of the "recalculated" sample sizes may thus correspond to the initial sample size n_1. Finally, a summary location measure combining all B sample sizes determines the final value of the second-stage sample size. Here, we distinguish between two approaches:

1. The simplest approach is to define the second stage sample size as the mean of all resampled sample sizes

$$n_{(*)}^{R1}(t_1) = \frac{1}{B}\sum_{b=1}^{B} \tilde{n}_{(*),i}(t_1). \tag{12}$$

We denote this resampling method as the *R1 approach*.

2. Since the first stage's sample size has a large influence on the resampled sample size, an alternative option is considered where we use the mean plus the standard deviation of the resampled sample sizes to obtain the final value for the second stage sample size

$$n_{(*)}^{R2}(t_1) = \frac{1}{B}\sum_{b=1}^{B} \tilde{n}_{(*),i}(t_1) + \frac{1}{B-1}\sqrt{\sum_{b=1}^{B}\left(\tilde{n}_{(*),i}(t_1) - \frac{1}{B}\sum_{b=1}^{B}\tilde{n}_{(*),i}(t_1)\right)^2}. \tag{13}$$

The standard deviation as an additional term implies that higher sample sizes are chosen. We call this method the *R2 approach*. Note that instead of adding the standard deviation, in principle also other measures of the resampled sample size distribution (e.g., predefined quantiles) could be used to achieve this effect. Therefore, the R2 approach is just to be seen as one exemplary option.

Figure 1 shows the recalculated sample sizes for the observed conditional power, restricted observed conditional power, and promising zone approach for the original recalculation rules (gray lines), the recalculation rules combined with the resampling method R1 (blue lines) and the recalculation rules with the resampling method R2 (green lines). The first stage's sample size was set to 50, the maximal sample size limited to 200, the futility stop bound α_0 was set to 0.5, and local significance levels were adjusted according to Pocock[17] to maintain the global significance level $\alpha = 0.025$. Note that the resampling approaches converge to a smooth line if the number B of resampling samples tends to infinity. Here, every single point of the blue and green lines in the figure represents the mean of $B = 5{,}000$ resampled sample sizes based on the observed interim test statistics.

3 | SIMULATION STUDY

To evaluate the performance of the different sample size recalculation approaches presented above, we conducted a simulation study[24] with the software R.[25] Random numbers were generated by the function rnorm. The resulting approaches were evaluated by means of specific performance characteristics and by the new conditional performance score (10) with parameter settings $\gamma_{\mathrm{loc}} = \gamma_{\mathrm{var}} = 0.5$.

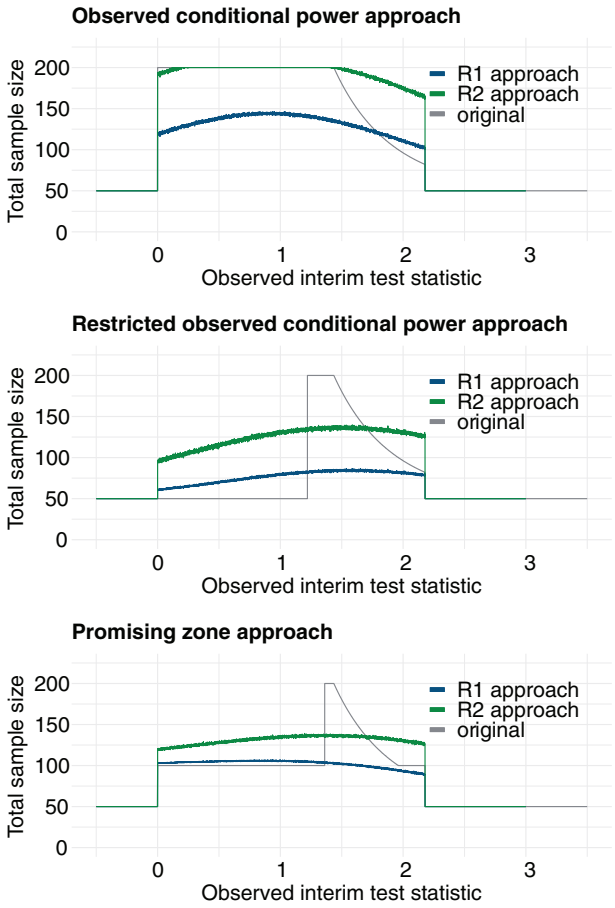

FIGURE 1 Sample size recalculation rules as functions of the observed interim test statistic. The gray solid lines present the original recalculation rules, the blue solid lines describe the resampling approach with the mean as summary location measure (R1 approach) and the green solid lines describe the resampling approach with the mean plus standard deviation as summary location measure (R2 approach)

3.1 | Simulation setup

For the simulation study, we rely on the design characteristics described in Section 2. Thus, we chose equally sized groups of $n_1 = 50$ subjects in the first stage. The initial second stage sample size per group was set to $n_2 = 50$, hence leading to an initial overall sample size per group of $n_{\text{ini}} = n_1 + n_2 = 100$, and the maximum possible sample size per group is set to four times the interim sample size n_1 by $n_{\max} = 200$. Accordingly, the weights for the inverse normal combination test were selected as $w_1 = w_2 = \sqrt{50}$. The global one-sided significance level was given by $\alpha = 0.025$ and the local significance levels were calculated according to Pocock,[17] that is, $\alpha_1 = \alpha_{1+2} = 0.0147$. Moreover, the futility bound was set to $\alpha_0 = 0.5$. The desired conditional power was chosen to be $1-\beta = 0.8$. The lower bound for the conditional power in the restricted observed conditional power approach (ROCP) was fixed to $1-\beta_{\text{low}}^{\text{ROCP}} = 0.6$. The lower bound for the promising zone (PZ) was set to $1-\beta_{\text{low}}^{\text{PZ}} = 0.36$, as applied by Mehta and Pocock.[8] We investigated the performance of the different designs for a variety of underlying true standardized treatment effect $\Delta \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For each scenario, 10,000 simulation samples were drawn. The number of samples for the resampling approaches was set to $B = 5{,}000$. For the sake of comparison, a group sequential design (GS) with $n_1 = n_2 = 50$ and the same decision boundaries as given above was also simulated and evaluated.

3.2 | Simulation results

Detailed simulation results can be found in Tables A1–A3 in the Appendix. In the following, we only present and discuss the main performance results with respect to the conditional performance score which are shown in Table 1.

In the standard sample size recalculation scenario without resampling, the group sequential study design is the clear performance winner except for a true underlying effect of $\Delta = 0.3$. This is mainly due to no variation in recalculated

sample sizes. The first resampling approach with the mean as summary location measure (R1 approach) performs better than the respective standard sample size recalculation rules without resampling with respect to the conditional performance score for all considered true standardized effect sizes Δ (Table 1, Columns 3 and 4). This is particularly due to the fact that the resampling approach reduces the variability in the recalculated sample sizes for all Δ. Moreover, the R1 approach does either perform even better than the group sequential approach or reveals a very similar total conditional performance score for the restricted observed conditional power and promising zone approach (Table 1, Columns 3 and 4). This stems mainly from a better performance with respect to the conditional power (Tables A1 and A2 in the Appendix). The performance of the observed conditional power approach is in most cases worse than the group sequential approach especially due to a worse conditional sample size performance (Table A2 in the Appendix). Moreover, the sample size recalculation approaches with resampling smooth the sample size curves and have the tendency to mimic the sample size shape of the group sequential design (cf. Figure 1) which is given by a horizontal line at $n = n_1 + n_2 = 100$ within the recalculation area. The shape of the sample size recalculation curve of the promising zone approach with resampling is here closest to the group sequential approach.

Note that there is a trend towards the first stage's sample size n_1 for the sample sizes recalculated with the R1 approach. This is due to the fact that test statistics outside the recalculation area might be resampled even if the observed interim test statistics falls within RA. To overcome this issue, resampling approach R2 is based on a different summary location measure which is given by the mean of the resampled sample sizes *plus* its standard deviation.

As it can be seen from the green lines in Figure 1, the observed conditional power approach with the R2 approach now has a stronger trend towards a group sequential design due to the sample size boundary $n_{\max}$. Compared to the

TABLE 1 Conditional performance scores for different sample size recalculation rules in their original version and combined with resampling approaches R1 and R2 for the design settings described in Section 3.1

		Conditional performance score $CS(\Delta)$		
Δ	Design	Original version	With R1 approach	With R2 approach
0.0	OCP	0.474	0.653	0.508
	ROCP	0.610	0.823	0.660
	PZ	0.651	0.762	0.668
	GS	0.776	–	–
0.1	OCP	0.430	0.616	0.465
	ROCP	0.540	0.791	0.617
	PZ	0.595	0.728	0.628
	GS	0.742	–	–
0.2	OCP	0.398	0.583	0.431
	ROCP	0.480	0.762	0.582
	PZ	0.549	0.697	0.594
	GS	0.710	–	–
0.3	OCP	0.621	0.633	0.692
	ROCP	0.390	0.557	0.623
	PZ	0.527	0.604	0.652
	GS	0.610	–	–
0.4	OCP	0.552	0.685	0.601
	ROCP	0.544	0.705	0.688
	PZ	0.622	0.746	0.700
	GS	0.756	–	–
0.5	OCP	0.541	0.660	0.584
	ROCP	0.522	0.733	0.664
	PZ	0.592	0.712	0.674
	GS	0.721	–	–

Abbreviations: OCP, observed conditional power approach; ROCP, restricted observed conditional power approach; PZ, promising zone approach; GS, group sequential approach.

observed conditional power approach with resampling and the mean as summary location measure (R1 approach, blue lines), the step size at the upper edge of the recalculation area is higher. The restricted observed conditional power R2 approach looks similar to the restricted observed conditional power R1 approach (cf. Figure 1) but is located at a higher position. For the promising zone approach combined with the R2 approach, the shape of the sample size curve is now nearly horizontal and thus looks again very similar to a group sequential design with constant sample sizes per stage. Within the R2 method, the promising zone has the best overall conditional performance scores for most values of Δ (Table 1, Column 5) compared to the observed conditional and restricted observed conditional power R2 approaches. Overall, the R2 approaches perform again better than the original sample size recalculation approaches without resampling (Table 1, Columns 3 and 5). This is again mainly due to a reduction in variability of the conditional sample size (Tables A1 and A2 in the Appendix) since also here the resampling leads to more robust approaches. However, the group sequential approach outperforms the different R2 methods (Table 1, Columns 3 and 5) for $\Delta \in \{0.0,0.1,0.2,0.4,0.5\}$. Moreover, for the same true underlying standardized effect sizes, the R2 approach always has a worse conditional performance score than the R1 approach (Table 1, Columns 4 and 5). This is due to the fact that the sample size recalculation rules in the R2 approach do not well approach the target values for the sample size, which is supported by considerably worse conditional sample size sub-score SCN values (Tables A2 and A3 in the Appendix). Also, the conditional power location component is worse for $\Delta \in \{0.0,0.1,0.2\}$ and the conditional power variation component is also a little worse than the one of the R1 method (Tables A2 and A3 in the Appendix). Hence, the conditional performance score of the R1 and R2 approaches are rather similar for the true standardized effect sizes for $\Delta \geq 0.3$ but for $\Delta < 0.3$ the R1 approach is clearly better (Table 1, Columns 4 and 5). For $\Delta \geq 0.3$, the conditional sample size sub-scores SCN are better for the R1 method but in turn, the conditional power sub-scores SCP of the R2 method outperform the R1 method. Taking both sub-scores together, the overall performances are rather similar. For small true underlying effect sizes, where the first stage's sample size is chosen as a reference value for the conditional sample size component, the R1 method clearly outperforms the R2 method, which is mainly due to smaller expected conditional sample sizes.

3.3 | Clinical trial example

Based on a clinical trial of Bowden and Mander,[26] we consider a new and a standard treatment, N and S, for osteoarthritis patients with respect to pain relief after 2 weeks compared to baseline. Pain relief is measured on the McGill pain scale[27] where values range from 0 referring to no pain until 50 referring to maximally possible pain. The values are supposed to be normally distributed. For the sake of illustration, we adapt the initial clinical trial design to meet the design requirement of the methods proposed in here as already proposed by Herrmann and Rauch.[28] We assume that superiority of the new treatment is known from a pilot study but further evidence is required to quantify the effect size. Therefore, the hypotheses

$$H_0 : \left(\mu^N_{\text{baseline}} - \mu^N_{\text{2weeks}}\right) - \left(\mu^S_{\text{baseline}} - \mu^S_{\text{2weeks}}\right) \leq 0 \text{ versus } H_1 : \left(\mu^N_{\text{baseline}} - \mu^N_{\text{2weeks}}\right) - \left(\mu^S_{\text{baseline}} - \mu^S_{\text{2weeks}}\right) > 0, \tag{14}$$

where $\left(\mu^N_{\text{baseline}} - \mu^N_{\text{2weeks}}\right)$ describes the expected pain relief after 2 weeks for the new treatment and $\left(\mu^S_{\text{baseline}} - \mu^S_{\text{2weeks}}\right)$ for the standard treatment. The study is evaluated with an adaptive two-stage design and the possibility to recalculate the sample size at the interim analysis. More precisely, we choose $n_1 = n_2 = 50$ and $n_{\max} = 4 \cdot n_1 = 200$ as well as choose the inverse normal combination test[11] with weights $w_1 = w_2 = \sqrt{50}$. Furthermore, we decide on a binding futility stop bound $\alpha_0 = 0.5$, a global significance level $\alpha = 0.025$ and local significance levels adjusted according to Pocock.[17]

Suppose we observe an interim effect size $\Delta = 0.2$, referring to an interim test statistic of $T_1 = 1$, and we are interested in the conditional performance differences of the OCP, ROCP and PZ approaches with and without the R1 resampling approach. For the evaluation, we use an equal weighting of the conditional performance score components as a large recalculated sample size is only justifiable if it is not caused by random variation and if the sample size time meets the target value at the same. Therefore, variation and location components are considered as equally important. We primarily focus on the performance for $\Delta = 0.2$, which corresponds to the observed effect size, but also take the performance of the neighboring effect sizes $\Delta = 0.1$ and 0.3 into account. The performance values are given in Table 1 and Tables A1 and A2. Without resampling and an interim effect size of $\Delta = 0.2$, the OCP approach would suggest the maximal sample size of 200, whereas the ROCP approach suggests no increase of the sample size at all and thus no second

stage of the trial, and the PZ approach suggests the initially planned total sample size of 100. The resampling R1 approach suggests for all three approaches (OCP, ROCP, PZ) a trial continuation with total sample sizes varying between at least 75 and at most 150. The overall conditional performance measured by the conditional performance score is better for the R1 approach than for the original approach for all three recalculation rules and all three considered effect sizes (cf. Table 1). This comes mainly from the variance reduction of the conditional sample size and power by the resampling approach (cf. Tables A1 and A2). For $\Delta = 0.1$ and 0.2, the ROCP R1 resampling approach performs best whereas for 0.3, the OCP R1 resampling approach turns out to be the best (cf. Table 1). This change in the ranking is mainly due to the change in underlying target values for an effect of 0.3. If one is also interested in the global performance, the OCP R1 approach attains a higher global power than the other ROCP and PZ R1 approaches across the considered effect sizes due to higher sample sizes. As a general result it can be deduced that the resampling approaches flatten the shape of the sample size function and thus reduce the variability.

4 | DISCUSSION AND CONCLUSIONS

Incorporating resampling to the interim test statistics in established sample size recalculation rules leads to more robust recalculation approaches with a considerable performance improvement with respect to individual performance characteristics and a conditional performance score, mainly due to the reduced variance in the conditional sample size and conditional power. This was also seen in a fictitious clinical trial example. Note the weighting scheme of the conditional performance score and its reference values might also be chosen differently. Moreover, note that the observed performance jumps around $\Delta = 0.3$ for the conditional performance score are a general property of recalculation rules as for small effects no increase in sample size is favorable whereas from a certain medium effect an increase in sample size is reasonable. Irrespective of the performance score, the application of the proposed resampling tool resulted in a smoothing of the sample size curve. The form of the smoothed sample size function is concave where the kurtosis nearly vanishes for some scenarios. Thus, the sample size function approaches a constant line in some situations which in turn mimics a group sequential design. The concave form of the smoothed function means that within a certain interval of the recalculation area, the sample size increases with increasing observed interim effect. One might argue that an increase in sample size with increasing interim test statistic is not reasonable. However, despite their unintuitive character, concave sample size functions have shown to be optimal in some particular settings.[29] Furthermore, one could also argue that large "jumps" in the sample size function are also not reasonable as this implies that the sample size changes considerably if the observed test statistic is only minimally changed. Hence, they can be seen as two opposite points of view: On the one hand unintuitive "jumps" in sample size can be avoided with smoothed sample size curves by concave function shapes, and on the other hand this results in sample size functions which are no longer monotonically decreasing in the interim test statistic, which also is not intuitive. Note that these "jumps" are part of nearly all established sample size recalculation rules. In areas where conventional recalculation rules show these "jumps", the resampling approach defines a compromise between the extremes. One might also say that any sample size recalculation rule that includes large "jumps" is generally not reasonable and, as a consequence, the compromise proposed by the resampling approach cannot be optimal either. A general recommendation is thus to choose the design settings such that large jumps are omitted, for example, through a smaller maximal sample size $n_{\max}$ or a larger local significance level α_{1+2}. Even though the resampling approaches outperform the original sample size recalculation rules with respect to the conditional performance score, it does not mean that the resulting sample sizes are point-wise optimal. It rather reduces the average risk of choosing an entirely wrong sample size, which leads *on average* to good results. In the individual case, however, this can be fundamentally wrong. The latter is of course not a negative feature for the resampling approach but generally holds true for sample size recalculation rules. We believe that sample size recalculation rules with resampling are a good approach to take account of the cost–benefit ratio. Due to the characteristic of reducing the average distance to the ideal sample size, the method is suitable to balance the costs and the benefits of a study by choosing the best trade-off between both.

The visual similarity of the procedures with resampling to sample size recalculation for group sequential designs is remarkable. Especially the promising zone approach combined with resampling approximates a standard group sequential design. This is because the promising zone approach includes large sample size jumps in a very small range of observed interim effects and this small range of large sample sizes has a low impact on the smoothed sample size curve. This further supports the thesis that group sequential designs might have an exceptional position among designs with sample size recalculation (cf. also References[12,20,21]). However, while sample size recalculation based on group

sequential designs does only depend on the interim test statistic with respect to stopping the trial early or not, the incorporation of resampling to sample size recalculation rules offers the possibility to base sample size recalculation on conditional power considerations and still avoid severe fluctuations in sample size. Hence, resampling makes sample size recalculation rules more robust and addresses obviously the randomness of the observed interim test statistic. Combined with the simulation code supplied, this may add to an appealing possibility of improving sample size recalculation rules in adaptive study designs.

Note that the resampling approaches described in formulas (12) and (13) may also be applied to studies with other types of endpoints as long as the test statistics are approximately normally distributed. The application of the resampling approach to binary endpoints is straightforward using the normal approximation test for rates. For time-to-event endpoints, the logrank test also allows to be applied in an adaptive design setting.[30]

As an alternative to the resampling approach proposed in here, a more direct approach to improve the performance of sample size recalculation could be to define a sample size recalculation function that optimizes the conditional performance score. This idea can be based on a numerical constrained optimization framework. The implementation of this alternative approach and the comparison to the resampling approach will be the task of future work.

ACKNOWLEDGMENT

This work was supported by the German Research Foundation (grants RA 2347/4-1 and KI 708/4-1). Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

Original data were not analyzed. The R code that was used for producing simulated data and analyzing them is available at https://github.com/shareCH/SSR-with-resampling.

ORCID

Carolin Herrmann https://orcid.org/0000-0003-2384-7303
Maximilian Pilz https://orcid.org/0000-0002-9685-1613
Meinhard Kieser https://orcid.org/0000-0003-2402-4333
Geraldine Rauch https://orcid.org/0000-0002-2451-1660

REFERENCES

1. U.S. Department of Health and Human Services, Food and Drug Administration. Adaptive Designs for Clinical Trials of Drugs and Biologics Guidance for Industry. 2019. https://www.fda.gov/media/78495/download
2. Committee for Medicinal Products for Human Use, European Medicines Agency. Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. *London*: *EMEA*. 2007. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf
3. Bauer P, Bretz F, Dragalin V, König F, Wassmer G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med*. 2016;35(3):325-347.
4. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics*. 1995;51:1315-1324.
5. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286-1290.
6. Denne JS. Sample size recalculation using conditional power. *Stat Med*. 2001;20(17–18):2645-2660.
7. Posch M, Bauer P, Brannath W. Issues in designing flexible trials. *Stat Med*. 2003;22(6):953-969.
8. Mehta CR, Pocock SJ. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Stat Med*. 2011;30(28):3267-3284.
9. Jennison C, Turnbull BW. Adaptive sample size modification in clinical trials: start small then ask for more? *Stat Med*. 2015;34(29):3793-3810.
10. Pilz M, Kunzmann K, Herrmann C, Rauch G, Kieser M. A variational approach to optimal two-stage designs. *Stat Med*. 2019;38(21):4159-4171.
11. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50(4):1029-1041.
12. Levin GP, Emerson SC, Emerson SS. Adaptive clinical trial designs with pre-specified rules for modifying the sample size: understanding efficient types of adaptation. *Stat Med*. 2013;32(8):1259-1275.
13. Dmitrienko A, Wang M-D. Bayesian predictive approach to interim monitoring in clinical trials. *Stat Med*. 2006;25(13):2178-2195.

14. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Control Clin Trials*. 1986;7(1):8-17.
15. Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med*. 1986;5(1):1-13.
16. Hade EM, Jarjoura D, Wei L. Sample size re-estimation in a breast cancer trial. *Clin Trials*. 2010;7(3):219-226.
17. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64(2):191-199.
18. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549-556.
19. Cui L, Hung HJ, Wang SJ. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999;55(3):853-857.
20. Jennison C, Turnbull BW. Efficient group sequential designs when there are several effect sizes under consideration. *Stat Med*. 2006;25(6):917-932.
21. Tsiatis AA, Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*. 2003;90(2):367-378.
22. Shih WJ. Sample size re-estimation—journey for a decade. *Stat Med*. 2001;20(4):515-518.
23. Herrmann C, Pilz M, Kieser M, Rauch G. A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation. *Stat Med*. 2020;39(15):2067–2100.
24. Kluge C, Herrmann C, Rauch G. Simulation code for sample size recalculation with resampling. GitHub. 2020. https://github.com/shareCH/SSR-with-resampling
25. Development Core Team R. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. 2017.
26. Bowden J, Mander A. A review and re-interpretation of a group-sequential approach to sample size re-estimation in two-stage trials. *Pharm Stat*. 2014;13(3):163-172.
27. Torgerson WS, Melzack R. On the language of pain. *Anesthesiology*. 1971;34:50-59.
28. Herrmann C, Rauch G. Smoothing corrections for improving sample size recalculation rules in adaptive group sequential study designs. *Methods Inf Med*. 2021;1-7.
29. Pilz M, Kilian S, Kieser M. A note on the shape of sample size functions of optimal adaptive two-stage designs. *Commun Stat Theor Method*. 2020;1-8.
30. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biom J*. 2006;48(4):714-729.

How to cite this article: Herrmann C, Kluge C, Pilz M, Kieser M, Rauch G. Improving sample size recalculation in adaptive clinical trials by resampling. *Pharmaceutical Statistics*. 2021;20:1035–1050. https://doi.org/10.1002/pst.2122

APPENDIX A

Formulas for the conditional performance score

In the following, the formulas for the conditional performance score together with its sub-scores are presented. More details for the motivation of the score can be found in Herrmann et al.[23] First, we describe the four components (e_{CN}, v_{CN}, e_{CP}, v_{CP}) and the two sub-scores (SCN, SCP) which can be composed to define the total score CS. The underlying idea of the two location components, e_{CP} and e_{CN}, is to compare the evaluated average conditional power and average conditional sample size with predefined target values. We specified the target values for the sample size as

$$\mathrm{CN}_{\mathrm{target}} := \begin{cases} n_{\Delta}^{\mathrm{fix}}, & \text{if } n_{\Delta}^{\mathrm{fix}} \leq n_{\max} \text{ and } \Delta \neq 0, \\ n_1, & \text{if } n_{\Delta}^{\mathrm{fix}} > n_{\max} \text{ or } \Delta = 0, \end{cases} \tag{A1}$$

where $n_{\Delta}^{\mathrm{fix}}$ refers to the required sample size in a fixed study design. The target values for the conditional power are given as

TABLE A1 Performance summary for the sample size recalculation rules without resampling for the design settings described in Section 3.1

Δ	Design	$\mathbb{E}[CN_\Delta^{RA}]$	$\mathrm{Var}(CN_\Delta^{RA})$	$e_{CN}(\Delta)$	$\nu_{CN}(\Delta)$	SCN(Δ)	$\mathbb{E}[CP_\Delta^{RA}]$	$\mathrm{Var}(CP_\Delta^{RA})$	$e_{CP}(\Delta)$	$\nu_{CP}(\Delta)$	SCP(Δ)	CS(Δ)
0.0	OCP	192.001	592.170	0.053	0.676	0.364	0.261	0.088	0.758	0.408	0.583	0.474
(50)	ROCP	73.107	2388.850	0.846	0.348	0.597	0.158	0.096	0.864	0.380	0.622	0.610
	PZ	107.346	513.570	0.618	0.698	0.658	0.180	0.076	0.841	0.448	0.644	0.651
	GS	100	0	0.667	1	0.833	0.149	0.047	0.873	0.567	0.720	0.776
0.1	OCP	186.297	940.290	0.091	0.591	0.341	0.340	0.102	0.677	0.360	0.519	0.430
(50)	ROCP	81.552	2810.183	0.790	0.293	0.541	0.232	0.126	0.788	0.291	0.539	0.540
	PZ	110.479	657.421	0.597	0.658	0.627	0.254	0.102	0.765	0.362	0.563	0.595
	GS	100	0	0.667	1	0.833	0.210	0.065	0.811	0.491	0.651	0.742
0.2	OCP	179.231	1335.010	0.138	0.513	0.326	0.423	0.106	0.592	0.350	0.471	0.398
(50)	ROCP	90.885	3112.382	0.727	0.256	0.492	0.318	0.146	0.700	0.236	0.468	0.480
	PZ	113.401	775.117	0.577	0.629	0.603	0.336	0.119	0.681	0.309	0.495	0.549
	GS	100	0	0.667	1	0.833	0.278	0.080	0.740	0.435	0.588	0.710
0.3	OCP	170.744	1713.040	0.964	0.448	0.706	0.509	0.099	0.702	0.370	0.536	0.621
(177)	ROCP	99.723	3166.808	0.490	0.250	0.370	0.410	0.152	0.600	0.219	0.410	0.390
	PZ	116.563	903.008	0.602	0.599	0.601	0.427	0.126	0.617	0.290	0.454	0.527
	GS	100	0	0.492	1	0.746	0.356	0.089	0.544	0.404	0.474	0.610
0.4	OCP	160.431	1987.207	0.597	0.406	0.502	0.587	0.083	0.781	0.424	0.603	0.552
(101)	ROCP	106.806	2911.707	0.955	0.281	0.618	0.503	0.142	0.695	0.247	0.471	0.544
	PZ	118.477	931.146	0.877	0.593	0.735	0.517	0.119	0.710	0.309	0.509	0.622
	GS	100	0	1	1	1	0.437	0.091	0.627	0.398	0.513	0.756
0.5	OCP	150.651	2120.43	0.425	0.386	0.406	0.652	0.062	0.849	0.503	0.676	0.541
(65)	ROCP	111.97	2608.735	0.683	0.319	0.501	0.583	0.119	0.778	0.309	0.543	0.522
	PZ	119.858	961.328	0.630	0.587	0.608	0.597	0.102	0.792	0.360	0.576	0.592
	GS	100	0	0.763	1	0.881	0.511	0.084	0.703	0.419	0.561	0.721

Note: Numbers in brackets below the standardized treatment effects represent the target values for the sample size. Performance measure abbreviations are stated in the beginning of Section "Detailed performance evaluation".

Abbreviations: OCP, observed conditional power approach; ROCP, restricted observed conditional power approach; PZ, promising zone approach; GS, group sequential approach.

TABLE A2 Performance summary for the sample size recalculation rules with resampling (R1 approach) for the design settings described in Section 3.1

Δ	Design	$\mathbb{E}\left[CN_\Delta^{RA}\right]$	$\mathrm{Var}\left(CN_\Delta^{RA}\right)$	$e_{CN}(\Delta)$	$\nu_{CN}(\Delta)$	SCN(Δ)	$\mathbb{E}\left[CP_\Delta^{RA}\right]$	$\mathrm{Var}\left(CP_\Delta^{RA}\right)$	$e_{CP}(\Delta)$	$\nu_{CP}(\Delta)$	SCP(Δ)	CS(Δ)
0.0	OCP	134.489	80.492	0.437	0.880	0.659	0.204	0.068	0.817	0.480	0.648	0.653
(50)	ROCP	73.390	58.926	0.844	0.898	0.871	0.115	0.032	0.908	0.643	0.776	0.823
	PZ	103.876	7.332	0.641	0.964	0.802	0.153	0.045	0.869	0.573	0.721	0.762
0.1	OCP	134.066	92.665	0.440	0.872	0.656	0.277	0.087	0.742	0.410	0.576	0.616
(50)	ROCP	75.260	57.174	0.832	0.899	0.865	0.164	0.045	0.858	0.576	0.717	0.791
	PZ	103.364	11.249	0.644	0.955	0.800	0.212	0.062	0.808	0.504	0.656	0.728
0.2	OCP	132.999	118.443	0.447	0.855	0.651	0.356	0.099	0.660	0.371	0.515	0.583
(50)	ROCP	77.044	50.109	0.820	0.906	0.863	0.221	0.057	0.799	0.524	0.662	0.762
	PZ	102.631	16.484	0.649	0.946	0.797	0.280	0.074	0.739	0.455	0.597	0.697
0.3	OCP	131.15	144.300	0.700	0.840	0.770	0.443	0.102	0.634	0.361	0.497	0.633
(177)	ROCP	78.639	40.646	0.350	0.915	0.632	0.285	0.065	0.472	0.492	0.482	0.557
	PZ	101.688	21.572	0.503	0.938	0.721	0.354	0.081	0.543	0.430	0.487	0.604
0.4	OCP	128.761	165.727	0.809	0.828	0.818	0.529	0.096	0.722	0.380	0.551	0.685
(101)	ROCP	79.973	28.828	0.866	0.928	0.897	0.354	0.067	0.542	0.482	0.512	0.705
	PZ	100.591	25.473	0.996	0.933	0.965	0.432	0.081	0.622	0.431	0.527	0.746
0.5	OCP	126.055	180.075	0.589	0.821	0.705	0.605	0.081	0.800	0.430	0.615	0.660
(65)	ROCP	80.861	19.339	0.890	0.941	0.916	0.417	0.064	0.607	0.494	0.551	0.733
	PZ	99.456	27.914	0.766	0.930	0.848	0.502	0.074	0.694	0.457	0.576	0.712

Note: Numbers in brackets below the standardized treatment effects represent the target values for the sample size. Performance measure abbreviations are stated in the beginning of Section "Detailed performance evaluation".

Abbreviations: OCP, observed conditional power approach; ROCP, restricted observed conditional power approach; PZ, promising zone approach.

TABLE A3 Performance summary for the sample size recalculation rules with resampling (R2 approach) for the design settings described in Section 3.1

Δ	Design	$\mathbb{E}\left[\mathrm{CN}_\Delta^{\mathrm{RA}}\right]$	$\mathrm{Var}\left(\mathrm{CN}_\Delta^{\mathrm{RA}}\right)$	$e_{\mathrm{CN}}(\Delta)$	$\nu_{\mathrm{CN}}(\Delta)$	SCN(Δ)	$\mathbb{E}\left[\mathrm{CP}_\Delta^{\mathrm{RA}}\right]$	$\mathrm{Var}\left(\mathrm{CP}_\Delta^{\mathrm{RA}}\right)$	$e_{\mathrm{CP}}(\Delta)$	$\nu_{\mathrm{CP}}(\Delta)$	SCP(Δ)	CS(Δ)
0	OCP	197.825	30.734	0.015	0.926	0.470	0.276	0.106	0.743	0.349	0.546	0.508
(50)	ROCP	119.568	167.081	0.536	0.828	0.682	0.199	0.074	0.821	0.456	0.639	0.660
	PZ	129.578	28.731	0.469	0.929	0.699	0.203	0.074	0.818	0.458	0.638	0.668
0.1	OCP	196.941	46.810	0.020	0.909	0.465	0.366	0.129	0.651	0.282	0.466	0.465
(50)	ROCP	122.562	154.592	0.516	0.834	0.675	0.275	0.097	0.743	0.376	0.559	0.617
	PZ	130.672	25.902	0.462	0.932	0.697	0.279	0.097	0.740	0.379	0.559	0.628
0.2	OCP	195.510	72.712	0.030	0.886	0.458	0.462	0.139	0.552	0.254	0.403	0.431
(50)	ROCP	125.351	130.921	0.498	0.847	0.673	0.359	0.114	0.657	0.326	0.491	0.582
	PZ	131.568	22.166	0.456	0.937	0.697	0.363	0.112	0.654	0.330	0.492	0.594
0.3	OCP	193.695	100.321	0.883	0.866	0.875	0.563	0.136	0.757	0.263	0.510	0.692
(177)	ROCP	127.744	102.595	0.677	0.865	0.771	0.452	0.120	0.643	0.308	0.475	0.623
	PZ	132.229	18.510	0.707	0.943	0.825	0.454	0.118	0.646	0.313	0.479	0.652
0.4	OCP	191.455	122.846	0.391	0.852	0.621	0.659	0.120	0.855	0.308	0.582	0.601
(101)	ROCP	129.669	70.143	0.803	0.888	0.845	0.545	0.115	0.738	0.322	0.530	0.688
	PZ	132.653	14.512	0.783	0.949	0.866	0.547	0.113	0.741	0.328	0.534	0.700
0.5	OCP	189.069	143.898	0.169	0.840	0.504	0.742	0.095	0.940	0.385	0.663	0.584
(65)	ROCP	130.861	46.995	0.557	0.909	0.733	0.628	0.100	0.823	0.369	0.596	0.664
	PZ	132.741	13.133	0.544	0.952	0.748	0.629	0.098	0.825	0.375	0.600	0.674

Note: Numbers in brackets below the standardized treatment effects represent the target values for the sample size. Performance measure abbreviations are stated in the beginning of Section "Detailed performance evaluation".

Abbreviations: OCP, observed conditional power approach; ROCP, restricted observed conditional power approach; PZ, promising zone approach.

$$\mathrm{CP}_{\mathrm{target}} := \begin{cases} 1-\beta, & \text{if } n_{\Delta}^{\mathrm{fix}} \leq n_{\max} \text{ and } \Delta \neq 0, \\ \alpha, & \text{if } n_{\Delta}^{\mathrm{fix}} > n_{\max} \text{ or } \Delta = 0, \end{cases} \tag{A2}$$

where α refers to the global one-sided significance level. With these notations, the sub-score for the conditional sample size is given by

$$\mathrm{SCN}(\Delta) := \gamma_{\mathrm{loc}} \cdot \underbrace{\left(1 - \frac{\left|\mathbb{E}\left[\mathrm{CN}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right] - \mathrm{CN}_{\mathrm{target}}\right|}{n_{\max} - n_1}\right)}_{:= e_{\mathrm{CN}}(\Delta)} + \gamma_{\mathrm{var}} \cdot \underbrace{\left(1 - \sqrt{\frac{\mathrm{Var}\left(\mathrm{CN}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right)}{\mathrm{Var}_{\max}\left(\mathrm{CN}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right)}}\right)}_{:= v_{\mathrm{CN}}(\Delta)}, \tag{A3}$$

with $\gamma_{\mathrm{loc}} + \gamma_{\mathrm{var}} = 1$ and $\mathrm{Var}_{\max}\left(\mathrm{CN}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right) := \left(\frac{n_{\max} - n_1}{2}\right)^2$. Note that the maximally possible variance of the conditional sample size that can be observed in a group sequential design with constant sample sizes equals 0 such that v_{CN} reduces to 1 in that case. Moreover, due to the conditional perspective, the score is defined for (adaptive) group sequential designs only and not for fixed sample size designs. The sub-score for the conditional power is given by

$$\mathrm{SCP}(\Delta) := \gamma_{\mathrm{loc}} \cdot \underbrace{\left(1 - \frac{\left|\mathbb{E}\left[\mathrm{CP}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right] - \mathrm{CP}_{\mathrm{target}}\right|}{1-\alpha}\right)}_{=: e_{\mathrm{CP}}(\Delta)} + \gamma_{\mathrm{var}} \cdot \underbrace{\left(1 - \sqrt{\frac{\mathrm{Var}\left(\mathrm{CP}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right)}{\mathrm{Var}_{\max}\left(\mathrm{CP}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right)}}\right)}_{=: v_{\mathrm{CP}}(\Delta)}, \tag{A4}$$

with $\gamma_{\mathrm{loc}} + \gamma_{\mathrm{var}} = 1$ and $\mathrm{Var}_{\max}\left(\mathrm{CP}_{\Delta}^{\mathrm{RA}}(\mathrm{T}_1)\right) := \left(\frac{1-0}{2}\right)^2 = 0.25$. The sub-scores, SCN and SCP, reach larger values if the respective variation is small and if the pre-defined target values are closely approached. Both sub-scores have a range of [0; 1]. Thus, the point-wise total conditional performance score CS, can be defined as

$$\mathrm{CS}(\Delta) := \frac{1}{2} \cdot (\mathrm{SCP}(\Delta) + \mathrm{SCN}(\Delta)). \tag{A5}$$

Detailed performance evaluation

This section provides the detailed performance summaries including all individual performance measures entering the conditional score. The reported performance criteria are:

- $\mathbb{E}\left[\mathrm{CN}_{\Delta}^{\mathrm{RA}}\right]$: Expected value of the conditional sample size in the RA.
- $\mathrm{Var}\left(\mathrm{CN}_{\Delta}^{\mathrm{RA}}\right)$: Variance of the conditional sample size in the RA.
- $e_{\mathrm{CN}}(\Delta)$: Location component of the conditional sample size sub-score.
- $v_{\mathrm{CN}}(\Delta)$: Variation component of the conditional sample size sub-score.
- $\mathrm{SCN}(\Delta)$: Conditional sample size sub-score.
- $\mathbb{E}\left[\mathrm{CP}_{\Delta}^{\mathrm{RA}}\right]$: Expected value of the conditional power in the RA.
- $\mathrm{Var}\left(\mathrm{CP}_{\Delta}^{\mathrm{RA}}\right)$: Variance of the conditional power in the RA.
- $e_{\mathrm{CP}}(\Delta)$: Location component of the conditional power sub-score.
- $v_{\mathrm{CP}}(\Delta)$: Variation component of the conditional power sub-score.
- $\mathrm{SCP}(\Delta)$: Conditional power sub-score.
- $\mathrm{CS}(\Delta)$: Final point-wise conditional score.