Received: 1 November 2018 | Revised: 20 December 2019 | Accepted: 4 March 2020

DOI: 10.1002/sim.8534

RESEARCH ARTICLE

Statistics in Medicine WILEY

A new conditional performance score for the evaluation of adaptive group sequential designs with sample size recalculation

Carolin Herrmann[1,2] | **Maximilian Pilz**[3] | **Meinhard Kieser**[3] | **Geraldine Rauch**[1,2]

[1]Institute of Biometry and Clinical Epidemiology, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Berlin, Germany

[2]Berlin Institute of Health (BIH), Berlin, Germany

[3]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, Corporate Member of Freie Universität Berlin, Humboldt-Universität zu Berlin, Berlin Institute of Health, Charitéplatz 1, 10117 Berlin, Germany.
Email: carolin.herrmann@charite.de

In standard clinical trial designs, the required sample size is fixed in the planning stage based on initial parameter assumptions. It is intuitive that the correct choice of the sample size is of major importance for an ethical justification of the trial. The required parameter assumptions should be based on previously published results from the literature. In clinical practice, however, historical data often do not exist or show highly variable results. Adaptive group sequential designs allow a sample size recalculation after a planned unblinded interim analysis in order to adjust the sample size during the ongoing trial. So far, there exist no unique standards to assess the performance of sample size recalculation rules. Single performance criteria commonly reported are given by the power and the average sample size; the variability of the recalculated sample size and the conditional power distribution are usually ignored. Therefore, the need for an adequate performance score combining these relevant performance criteria is evident. To judge the performance of an adaptive design, there exist two possible perspectives, which might also be combined: Either the global performance of the design can be addressed, which averages over all possible interim results, or the conditional performance is addressed, which focuses on the remaining performance conditional on a *specific* interim result. In this work, we give a compact overview of sample size recalculation rules and performance measures. Moreover, we propose a new conditional performance score and apply it to various standard recalculation rules by means of Monte-Carlo simulations.

KEYWORDS
adaptive group-sequential design, clinical trial, performance score, sample size recalculation

1 | INTRODUCTION

In clinical trials in general, and in phase III efficacy trials in particular, a careful choice and a reliable justification of the sample size are very important for ethical and economical reasons. In an underpowered study, it is unlikely to gain enough evidence to demonstrate the research hypothesis and patients are thus unnecessarily exposed to study-specific

risks. In an overpowered study, the release of a new treatment is prolonged and late recruited patients may be allocated to a less effective treatment although there already may exist enough evidence to demonstrate the efficacy of the new treatment. In both cases, financial resources are wasted and a balanced benefit-risk assessment for the individual patient is no longer guaranteed.

For the determination of the sample size, different parameters are required, such as the expected treatment effect and its variance. In a classical one-stage clinical trial, the sample size is fixed in the planning stage even when the parameter assumptions could not reasonably be justified by the literature or by medical experience. A way to address this problem is the use of an adaptive group sequential design which is mentioned in the ICH E9 Guideline.[1] These study designs include one or more planned unblinded interim analyses. At the interim time points, the trial might be stopped either for efficacy or for futility. Otherwise, the trial is continued with possible adjustment of study design elements.[2] A commonly performed adjustment is sample size recalculation based on the re-estimated parameter values obtained from the interim data.

Early approaches for sample size recalculation were proposed, for example, by Cui et al,[3] by Bauer and Köhne[4] as well as by Lehmacher and Wassmer[5] in the 1990s. Most frequently, sample size recalculation rules are based on conditional power arguments. The conditional power is thereby defined as the probability that the null hypothesis is rejected at the final analysis given the observed value of the test statistic at interim. An easy recalculation strategy is to choose the second stage sample size such that the conditional power reaches a predefined boundary.[5] The conditional power approach is often criticized[6] as the available information at the interim stage is usually limited and thus the treatment effect estimate shows a rather high variability. A number of recalculation rules have been proposed which are based on conditional power arguments, but the adjustment of the sample size is performed in different ways, for example in a step-wise manner depending on predefined ranges of the conditional power.[7] Spiegelhalter et al[8,9] as well as Dmitrienko and Wang[6] proposed recalculation rules built upon a Bayesian conditional power approach (also known as the "predictive power") which is given by a weighted average of the conditional power for different treatment effects following a prespecified prior distribution. Jennison and Turnbull[10] proposed an optimization function relating the increase in sample size to the gain in conditional power. This optimization function is then used for sample size recalculation. Moreover, the authors proposed an approach to obtain globally optimal adaptive designs by applying variational techniques.[11]

General points of criticism for many existing adaptive group sequential designs are that the recalculated sample size can be large on average, that its variability is often high, and that the target power is often not met.[4,12] Comparing and judging different sample size recalculation rules is not an easy task and there exist no unique standards for performance assessment. Generally, within an adaptive design one needs to distinguish between conditional performance measures (when the interim data are already available) and unconditional performance measures (averaging over all interim results). In a general fixed design, the performance is naturally optimized in the planning stage before data collection is started. Therefore, the global, unconditional performance is of interest in this context. In an adaptive design, the perspective to optimize the design in the planning stage is still valid. However, there is an inherent need to investigate in addition the performance conditional on the interim result. As adaptive designs have the purpose of optimizing the second stage result based on the interim result, the conditional perspective seems natural in this context. However, global and conditional performance should not be considered as opposite criteria. Instead, both perspectives should be investigated when designing an adaptive trial.

As global performance measures, the global power of an adaptive design as well as the average sample size (under the null or alternative hypothesis) are two performance measures that are commonly reported. Liu et al[13] were the first who presented a global performance score for adaptive designs based on sample size and power criteria. This score compares the power and the average sample size of an adaptive design in relation to the "perfect" fixed design (under the true parameter setting) as a gold standard. Their performance score has the potential shortcoming that it does not take into account the variability of sample size and that it is not well-defined under the null hypothesis of the underlying test problem. It is also questionable whether the "perfect" fixed sample size design is really a sensible gold standard as a corresponding nonadaptive two-stage design has a smaller expected sample size. Thus, there is room for improvement in the definition of a global performance score.

Conditional performance measures are only insufficiently discussed in the literature. As conditional performance measures are also very important next to the unconditional ones and their target values are usually easier to define, we introduce a new conditional performance score in this work. With this new score, the conditional design properties can be reasonably judged. However, our new score should always be reported along with global, unconditional performance measures to assess both perspectives.

The objective of this paper is to critically review and compare different sample size recalculation rules as well as conditional and unconditional performance characteristics for adaptive designs. Subsequently, we discuss ways to combine the latter criteria within the new conditional performance score. This work is organized as follows: We introduce the study design and the test problem in Section 2. In Section 3, we present various established sample size recalculation rules. In Section 4, we give an overview of conditional and unconditional performance measures and scores for adaptive designs and present the new performance score. The performance of the different recalculation rules as well as the classical group sequential approach is compared by the new conditional score and the score by Liu et al[13] in a Monte-Carlo simulation study in Section 5. We conclude with a discussion in Section 6.

2 | THE STUDY DESIGN

2.1 | The test problem

Throughout this work, we consider the situation of a randomized, controlled trial comparing an intervention (I) with a control (C) based on a continuous, normally distributed outcome with common variance σ^2,

$$X_i^{\mathrm{I}} \overset{\text{iid}}{\sim} N(\mu^{\mathrm{I}}, \sigma^2),$$
$$X_i^{\mathrm{C}} \overset{\text{iid}}{\sim} N(\mu^{\mathrm{C}}, \sigma^2), \quad i = 1, \dots, n,$$

where n denotes the sample size per group which is assumed to be equal for the sake of simplicity. The underlying standardized effect is denoted by

$$\Delta := \frac{\mu^{\mathrm{I}} - \mu^{\mathrm{C}}}{\sigma}. \tag{1}$$

Without loss of generality, we assume further that large values of the primary endpoint are favorable. The hypotheses to be assessed in confirmatory analysis are thus given by

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

2.2 | Interim analysis, test statistics, and local significance levels

We assume now that the trial is planned with one interim analysis which takes place after $n_1 < n$ patients per group have been fully observed. The general idea of such a two-stage adaptive design is to define adequate test statistics Z_1 and Z_{1+2} for the interim analysis and for the final analysis, respectively, for which the joint distribution can be determined at the planning stage. The test statistic Z_{1+2} includes all data collected until the final analysis and is thus positively correlated to Z_1. For a normally distributed, continuous outcome, the test statistic at interim is given by

$$Z_1 := \frac{\overline{X}_1^{\mathrm{I}} - \overline{X}_1^{\mathrm{C}}}{S_{\text{pooled},1}} \cdot \sqrt{\frac{n_1}{2}}, \tag{3}$$

with means $\overline{X}_1^{\mathrm{I}}, \overline{X}_1^{\mathrm{C}}$ and pooled standard deviation $S_{\text{pooled},1}$. This corresponds to the standard two-sample t-test statistic including all data from the first stage. For reasonably large sample sizes, the t-test statistic will approach the z-test statistic. Therefore, in the following, we will assume that the sample sizes per group are sufficiently high and thus that the test statistics are normally distributed.[14] The trial is stopped at interim with rejection of H_0 if

$$Z_1 \geq q_{1-\alpha_1},$$

where α_1 denotes the corresponding local one-sided significance level for the interim analysis and $q_{1-\alpha_1}$ is the corresponding normal quantile. The trial is stopped for futility with maintenance of H_0 if

$$Z_1 < q_{1-\alpha_0},$$

where α_0 is an adequate stopping for futility bound. In the following, we denote the interval

$$\text{RA} := [q_{1-\alpha_0}; q_{1-\alpha_1}), \tag{4}$$

as the *recalculation area* (RA). Please also view our related comments in Section 4.5. If the trial is continued after the interim analysis, then additional n_2 patients per group are recruited. In principle, the second stage sample size can be chosen freely without any restriction. Common recalculation rules to determine the second stage sample size are provided in Section 3. The independent incremental test statistic including *exclusively* the data of the second stage is then given as

$$Z_2 := \frac{\overline{X}_2^{\text{I}} - \overline{X}_2^{\text{C}}}{S_{\text{pooled},2}} \cdot \sqrt{\frac{n_2}{2}},$$

with $\overline{X}_2^{\text{I}}, \overline{X}_2^{\text{C}}$ and $S_{\text{pooled},2}$ defined analogously as above. Note that Z_1 and Z_2 are stochastically independent by construction. The test statistic for the final analysis including *all* data is then given as

$$Z_{1+2} := \frac{w_1 \cdot Z_1 + w_2 \cdot Z_2}{\sqrt{w_1^2 + w_2^2}}, \tag{5}$$

where w_1, w_2 are predefined weights which must be fixed in the planning stage. This is also known as the inverse normal combination test.[4] A common way to choose these weights is to define

$$w_1 = w_2 = \sqrt{n_1}. \tag{6}$$

These weights are "optimal" in the case where the sample sizes per group for both stages are equally given by n_1, which relates to one half of the total sample size per group. If the sample size for the second stage n_2 is chosen larger than n_1, then the second stage data are down-weighted, whereas they are up-weighted if $n_2 < n_1$. The null hypothesis H_0 is rejected at the final analysis if

$$Z_{1+2} \geq q_{1-\alpha_{1+2}},$$

where α_{1+2} denotes the corresponding local one-sided significance level for the final analysis. It can easily be seen that for large sample sizes, approximately it holds that

$$\text{Cov}(Z_1, Z_{1+2}) = \frac{w_1}{\sqrt{w_1^2 + w_2^2}},$$

so the joint distribution of Z_1 and Z_{1+2} approximates a fully specified multivariate normal distribution, compare for example Reference 14. Using this joint distribution, local significance levels for the interim analysis and the final analysis can be specified. These local levels, denoted as α_1 and α_{1+2} are thereby defined such that the overall type I error is controlled by the global significance level α, that is

$$P_{H_0}\left(Z_1 \geq q_{1-\alpha_1} \vee \left(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1} \wedge Z_{1+2} \geq q_{1-\alpha_{1+2}}\right)\right) \leq \alpha.$$

There exist various possible ways to define the local levels by use of predefined alpha-spending functions, compare, for example References 15-18. Throughout this work, we focus on the most simple case given by equal local levels for each stage

$$\alpha_1 = \alpha_{1+2},$$

which correspond to the well-known Pocock boundaries.[19]

3 | SAMPLE SIZE RECALCULATION RULES

As specified above, the second stage sample size per group n_2 generally can be chosen freely as long as the weights w_1, w_2 remain fixed. Although it is possible to adapt the sample size fully flexible and not according to a prespecified algorithm, it seems more reasonable to define a sample size recalculation rule in advance in order to preserve the integrity of the study design and in order to be able to judge the design's performance.

A general restriction which is usually introduced within adaptive designs with sample size recalculation is the specification of a maximal sample size per group $n_{\max}$ which serves as an upper bound for the recalculated sample size n_2. This upper bound provides a guarantee that the trial's sample size is both flexible and feasible. Throughout this article, we use a maximal sample size per group given as

$$n_{\max} = f \cdot n_1, \tag{7}$$

with a constant sample size boundary factor f. The total sample size per group is a function of the observed value of the interim test statistic z_1 and is denoted by

$$N_{\text{total}}(z_1) := \begin{cases} n_1, & \text{if } z_1 \notin [q_{1-\alpha_0}; q_{1-\alpha_1}), \\ N_{\text{recalc}}(z_1), & \text{if } z_1 \in [q_{1-\alpha_0}; q_{1-\alpha_1}), \end{cases} \tag{8}$$

where $N_{\text{recalc}}(z_1)$ is specified for each recalculation design separately. There are a number of recalculation rules proposed in the literature of which we present in the following only a selection of the most common ones.

3.1 | Observed conditional power approach

The most common approach to define the second stage sample size is to choose n_2 such that the conditional power reaches a predefined boundary $1 - \beta$. The conditional power depends on the assumed underlying standardized treatment effect $\tilde{\Delta}$ (which does not necessarily coincide with the true standardized treatment effect Δ defined in (1)) and is a function of the total sample size per group n and the observed value of the test statistic at interim z_1. It is defined as the probability of rejecting the null hypothesis after inclusion of n patients per group given the observed value of the test statistic at interim z_1, and assuming a standardized treatment effect $\tilde{\Delta}$,

$$\text{CP}_{\tilde{\Delta}}(z_1, n) := \begin{cases} 0, & \text{if trial is stopped early for futility,} \\ 1 - \Phi\left(z_{1-\alpha_{1+2}} \cdot \frac{\sqrt{w_1^2 + w_2^2}}{w_2} - z_1 \cdot \frac{w_1}{w_2} - \tilde{\Delta} \cdot \sqrt{\frac{n_1}{2}} \cdot \sqrt{\frac{n - n_1}{n_1}} \right), & \text{if the sample size is recalculated,} \\ 1, & \text{if trial is stopped early for efficacy.} \end{cases} \tag{9}$$

There exist several approaches to define a good guess for the underlying standardized treatment effect $\tilde{\Delta}$. Within the *observed conditional power approach* presented here, the observed standardized treatment effect at interim is employed

$$\hat{\Delta}_1 := \frac{\overline{X}_1^{\text{I}} - \overline{X}_1^{\text{C}}}{S_{\text{pooled},1}}.$$

Hereby, it is implicitly assumed that the observed standardized treatment effect is equal to the true one. Remember that the observed standardized treatment effect at interim is related to the observed value of the interim test statistic by

$$\hat{\Delta}_1 = z_1 \sqrt{\frac{2}{n_1}}.$$

The required total sample size for the second stage is given by the smallest integer fulfilling

$$\tilde{n} \geq n_1 \cdot \left(1 + \left(\frac{z_\beta - z_{1-\alpha_{1+2}} \cdot \frac{\sqrt{w_1^2 + w_2^2}}{w_2}}{z_1} + \frac{w_1}{w_2} \right)^2 \right), \quad z_1 \in [q_{1-\alpha_0}; q_{1-\alpha_1}). \tag{10}$$

With this recalculated sample size, it holds that

$$\mathrm{CP}_{\hat{\Delta}_1}(z_1, \tilde{n}) \geq 1 - \beta, \quad z_1 \in [q_{1-\alpha_0}; q_{1-\alpha_1}).$$

The recalculated sample size per group within the recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$ is then given as the minimum of $n_{\max}$ and $\tilde{n}$ which is equivalent to

$$N_{\text{recalc}}^{\text{OCP}}(z_1) := \begin{cases} \tilde{n}, & \text{if } \mathrm{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) \geq 1 - \beta, \\ n_{\max}, & \text{if } \mathrm{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) < 1 - \beta. \end{cases} \tag{11}$$

Due to the imputation of the observed standardized treatment effect, this approach is referred to the *observed conditional power approach*.

3.2 | Restricted observed conditional power approach

When the observed conditional power approach is used for sample size recalculation, the limiting sample size $n_{\max}$ is used whenever the recalculated sample size $\tilde{n}$, which ensures a conditional power of $1 - \beta$, is larger than $n_{\max}$. This implies that in this case, the actual observed conditional power based on $n_{\max}$ can be considerably smaller than $1 - \beta$. Therefore, it might be reasonable to use the limiting sample size $n_{\max}$ only in case that this sample size ensures a predefined minimal conditional power level $1 - \beta_0$, for example, $1 - \beta_0 = 0.6$, and to stop the trial early else. This *restricted observed conditional power approach* is thus based on the following sample size within the recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$

$$N_{\text{recalc}}^{\text{restrOCP}}(z_1) := \begin{cases} n_1, & \text{if } \mathrm{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) < 1 - \beta_0, \\ \tilde{n}, & \text{if } \mathrm{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) \geq 1 - \beta, \\ n_{\max}, & \text{if } 1 - \beta_0 \leq \mathrm{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) < 1 - \beta, \end{cases} \tag{12}$$

where $\tilde{n}$ is defined as in Equation (10).

3.3 | Promising zone approach

Mehta and Pocock[7] proposed the so-called *promising zone approach* which is also based on the observed conditional power. This approach was criticised by Jennison and Turnbull[10] with respect to expected sample size, however they did not consider other performance measures. The design starts with a nonadaptive group sequential design with interim sample size per group given by n_1 and second stage sample size n_2 which sum up to an initial sample size per group of $n_{\text{ini}} = n_1 + n_2$, where $n_{\text{ini}} < n_{\max}$. The recalculated second stage sample size is then determined according to the following rule:

- If the conditional power based on the observed interim test statistic z_1 and the initial sample size n_{ini} falls below a predefined boundary $1 - \tilde{\beta}_0$, then the interim results are declared as *unfavorable* and an increase of the sample size according to the observed conditional power approach is considered as inadequate. Hence, the study continues with the originally planned second stage sample size n_{ini}.
- If the conditional power based on the observed interim test statistic z_1 and the initial sample size n_{ini} falls above this boundary but below the anticipated power of $1 - \beta$, this is called *the promising zone*. Within this area, the sample size for the second stage is increased according to the observed conditional power approach defined in Equation (11).

- If the conditional power based on the observed interim test statistic z_1 and the initial sample size n_{ini} is equal or above the anticipated power, the results are seen as *favorable* and the originally planned second stage sample size n_{ini} is considered to be sufficient.

The final recalculated sample size per group within the recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$ is thus given as

$$N_{\text{recalc}}^{\text{Prom}}(z_1) := \begin{cases} n_{\text{ini}}, & \text{if } \text{CP}_{\hat{\Delta}_1}(z_1, n_{\text{ini}}) < 1-\tilde{\beta}_0, \text{ (unfavorable zone)} \\ \tilde{n}, & \text{if } 1-\beta > \text{CP}_{\hat{\Delta}_1}(z_1, n_{\text{ini}}) \geq 1-\tilde{\beta}_0 \text{ and } \text{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) \geq 1-\beta, \text{ (promising zone)} \\ n_{\max}, & \text{if } 1-\beta > \text{CP}_{\hat{\Delta}_1}(z_1, n_{\text{ini}}) \geq 1-\tilde{\beta}_0 \text{ and } \text{CP}_{\hat{\Delta}_1}(z_1, n_{\max}) < 1-\beta, \text{ (promising zone)} \\ n_{\text{ini}}, & \text{if } \text{CP}_{\hat{\Delta}_1}(z_1, n_{\text{ini}}) \geq 1-\beta, \text{ (favourable zone).} \end{cases} \tag{13}$$

The choice of the lower power boundary $1-\tilde{\beta}_0$ limiting the promising zone is intensively discussed in Mehta and Pocock.[7] Note that $1-\beta_0$ as applied in the restricted observed conditional power approach given in Equation (12) also defines a lower bound for the conditional power and, therefore, we use a similar notation. However, the power bound in the promising zone approach (13) is not directly related to the power bound used in Equation (12) and is chosen differently in applications.

3.4 | Optimization function approach

Jennison and Turnbull[10] proposed an alternative approach based on the idea to choose the second stage's sample size such that an optimization function is maximized. In the remainder of this paper, this method is therefore referred to as the *optimization function approach*. Again, their design starts with a nonadaptive group sequential design with interim sample size per group given by n_1 and second stage sample size n_2 which sum up to an initial sample size per group of $n_{\text{ini}} = n_1 + n_2$, where $n_{\text{ini}} < n_{\max}$. The optimization function $f(n)$ then is a combination of the observed conditional power for a given total sample size per group n and the deviation from n to n_{ini} and is defined as

$$f_\gamma(z_1, n) := \text{CP}_{\hat{\Delta}_1}(z_1, n) - \gamma\,(n - n_{\text{ini}}), \quad \gamma > 0. \tag{14}$$

The recalculated sample size within the recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$ is then given as

$$N_{\text{recalc}}^{\text{OptFunc}}(z_1) := argmax_{n\in[0,n_{\max})} f_\gamma(z_1, n). \tag{15}$$

It is clear that the mathematical properties of $f_\gamma(z_1, n)$ depend on the underlying effect size or interim test statistic, respectively. Especially, the skewness and monotonicity of the function change in dependence of z_1. This explains constant sample size values in certain effect size regions (cf Figure 2). The choice of the tuning parameter γ is discussed and illustrated in examples by Jennison and Turnbull.[10] It is obvious that the choice of γ importantly influences the form of $f_\gamma(\cdot)$. The choice of γ in dependence of effect size and variability seems, however, to be difficult as it is just the motivation for an adaptive design that these parameters are unknown in the planning stage.

4 | EVALUATING THE PERFORMANCE OF AN ADAPTIVE DESIGN

The idea of an adaptive design with the option to change the sample size during the ongoing trial is to increase the efficiency of a study. However, it is not evident how to quantify this efficiency. The efficiency of a clinical trial with a fixed study design is usually measured by means of (a) the sample size and (b) the power of the trial. An intuitive approach would be to choose the same criteria to evaluate the performance properties of an adaptive design.

4.1 | Performance measures based on power concepts—the global (unconditional) power

The power of an adaptive design can be reported equivalently as for a fixed sample size design. Indeed, in the literature on adaptive designs, the power is commonly reported as a performance measure, compare, for example, References 5, 7,

13, 20,21. The power of an adaptive design for a standardized treatment effect Δ is given as the probability to stop the trial for efficacy either at the interim stage or at the final analysis,

$$\begin{aligned}\text{Pow}_\Delta &:= P_\Delta\left(\left(Z_1 \geq q_{1-\alpha_1}\right) \vee \left(Z_{1+2} \geq z_{1-\alpha_{1+2}} \wedge \left(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1}\right)\right)\right) \\ &= P_\Delta\left(Z_1 \geq q_{1-\alpha_1}\right) + P_\Delta(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1}) \cdot P\left(Z_{1+2} \geq z_{1-\alpha_{1+2}} \middle| q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1}\right). \end{aligned} \tag{16}$$

Alternatively, the power can also be written as

$$\text{Pow}_\Delta = \int_{-\infty}^{\infty} \text{CP}_\Delta(z_1)\, dP(z_1) = P_\Delta\left(Z_1 \geq q_{1-\alpha_1}\right) + \int_{q_{1-\alpha_0}}^{q_{1-\alpha_1}} \text{CP}_\Delta(z_1)\, dP(z_1), \tag{17}$$

where $\text{CP}_\Delta(z_1)$ corresponds to the conditional power as introduced in Equation (9). The dependence on the sample size is omitted as the conditional power is used here in the context of a predefined adaptive recalculation rule where the final sample size thus only depends on z_1. The power Pow_Δ refers to an unconditional performance measure because it is not conditional on the interim result. It should be noted that the global, unconditional power for an adaptive design with sample size recalculation is not completely comparable to the power of a fixed design, as there are two disjoint options to reject the null hypothesis, and the researcher is often interested in differentiating between these two options. Moreover, even if the unconditional power of the specific underlying design at hand is known to be $1-\beta$, the scientist would not consider the design as acceptable if the conditional power for the second stage is only 0.5 or lower. In the setting of an adaptive design, it seems therefore also natural to consider conditional performance measures.

4.2 | Performance measures based on power concepts—the conditional power

As outlined in Section 3, many sample size recalculation rules are based on criteria for the observed conditional power. The unconditional power presented before was criticized as a performance measure for adaptive designs for several reasons, in particular as there is no intuitive target value. When considering the conditional power within the recalculation area, $\text{RA} = [q_{1-\alpha_0}; q_{1-\alpha_1})$, the target value is given by $1-\beta$. In other words, whenever the sample size at interim is potentially increased, the recalculated sample size should guarantee that the resulting conditional power is close to $1-\beta$. Thereby, it must be kept in mind that the observed conditional power $\text{CP}_{\hat{\Delta}_1}$ depends on the observed value of the interim test statistic z_1. As a consequence, the observed conditional power within the recalculation area cannot be reported as a single performance measure but the distribution of the observed conditional power must be summarized. As a location measure, the expected observed conditional power under the assumption of being in the recalculation area (RA) can be reported as a performance measure

$$\mathbb{E}\left[\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right] := \mathbb{E}\left[\text{CP}_{\hat{\Delta}_1}(Z_1) | Z_1 \in \text{RA}\right] = \frac{1}{P(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1})} \cdot \int_{q_{1-\alpha_0}}^{q_{1-\alpha_1}} \text{CP}_{\hat{\Delta}_1}(z_1)\, dP(z_1). \tag{18}$$

The above expected observed conditional power naturally quantifies the location of the random distribution. However, it is intuitive that a sample size recalculation rule that reaches a conditional power near $1-\beta$ in the recalculation area on average can still not be considered as "good" if the variance of the distribution is large, that is, if the observed conditional power values importantly vary among each realization of the random experiment. It is astonishing that although basic statistical text books always suggest that any measure of location must be reported along with an adequate measure of variation, the empirical variance or standard deviation of the observed conditional power is hardly ever reported. We strongly recommend to overcome this obvious fault and to never judge the performance of an adaptive design only on performance measures for the location. Therefore, the observed conditional power distribution should also be reported along with its variance $\text{Var}(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)) := \text{Var}(\text{CP}_{\hat{\Delta}_1}(Z_1) | Z_1 \in \text{RA})$ or standard deviation, respectively.

4.3 | Performance measures based on the sample size—the total expected sample size

Evaluating the power or observed conditional power alone is not meaningful as a high (conditional) power can always be achieved if the sample size is increased to an arbitrarily large amount. However, it is evident that there is the intention to

keep the sample size as small as possible. Therefore, the gain in (conditional) power must always be outweighed against the required increase in sample size. For a fixed design without sample size recalculation, the required sample size for a given treatment effect can simply be calculated and reported. In an adaptive design, this is more difficult, as the sample size depends on the observed value of the interim test statistic given as $N_{\text{total}}(z_1)$, compare Equation (8).

Note that when defining the recalculated sample sizes for the different investigated designs, we omitted the dependence on Δ in the notation before. As the observed value of the interim test statistic comes from a normal distribution $N(\hat{\Delta}_1\sqrt{\frac{n_1}{2}}, 1)$, this dependence is, however, intuitively given.

To quantify the required sample size within an adaptive design, the expected total sample size per group $\mathbb{E}[N_{\text{total}}(Z_1)]$ of an adaptive design for a given treatment effect Δ is often reported. It should be noted that the expected sample size for an adaptive design with sample size recalculation is not directly comparable to the sample size of a fixed design, as the expected sample size depends not only on the interim sample size n_1 or the recalculated second stage sample size but also on the probability to enter stage two given by $P(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1})$. As the general intention of the sample size recalculation is to achieve a reasonable power about $1-\beta$ in case the study is not already stopped at interim and there is also the option to reject the null hypothesis already at interim, the expected sample size for an adaptive design is usually smaller than the sample size of a fixed design.

4.4 | Performance measures based on the sample size—the conditional expected sample size

If one focuses only on the recalculation area $[q_{1-\alpha_0}; q_{1-\alpha_1})$, that is, if we focus on the sample size given that the interim results suggest a second stage, the expected sample size per group is then given as

$$\mathbb{E}\left[N_{\text{total}}^{\text{RA}}(Z_1)\right] := \mathbb{E}\left[N_{\text{total}}(Z_1)|Z_1 \in \text{RA}\right] = \frac{1}{P(q_{1-\alpha_0} \leq Z_1 < q_{1-\alpha_1})} \cdot \int_{q_{1-\alpha_0}}^{q_{1-\alpha_1}} N_{\text{total}}(z_1)\, dP(z_1), \tag{19}$$

which should always be reported along with $\text{Var}(N_{\text{total}}^{\text{RA}}(Z_1)) := \text{Var}(N_{\text{total}}(Z_1)|Z_1 \in \text{RA})$. If the study is continued after the interim analysis, the total sample size should be slightly larger than the required sample size for a fixed design. This is due to the fact that the adaptive design includes adjustment for multiple testing and therefore the fixed sample size rather defines a "lower bound" for a target value.

4.5 | A note on conditional performance measures

So far, we have distinguished between conditional and unconditional performance measures, where conditional performance measures are based on the condition of entering the recalculation area $\text{RA} = [q_{1-\alpha_0}; q_{1-\alpha_1})$. When looking at the different sample size calculation rules, we see that there are some rules which suggest no increase of sample size within certain sub-regions of the interval $[q_{1-\alpha_0}; q_{1-\alpha_1})$. For example, the restricted observed conditional power approach suggests an increase of the sample size within the area $[q_{1-\alpha_0}; q_{1-\alpha_1})$ if and only if a conditional power of $1-\beta_0$ can be reached with the maximally allowed sample size. However, when comparing different recalculation rules, conditional performance measures must rely on the same condition. The interval limited by the local critical value and the futility bound seems a natural choice for such a condition.

4.6 | Performance scores

Usually, the performance of adaptive designs is summarized by providing several separate performance criteria based on the power and the sample size. It is intuitive that a good power performance can always be achieved to the cost of a high sample size (ie, a bad sample size performance) and vice versa, irrespective of the fact which power or sample size criteria, unconditional or conditional, are applied. Moreover, even if a specific sample size recalculation rule shows good power *and* sample size performance on average, it is still necessary to investigate the variance of each of the performance

criteria. Users are therefore confronted with the problem to outweigh several performance criteria which are negatively correlated. A direct and fair comparison between different sample size recalculation rules is thus difficult. This situation indicates the need for a performance score which includes all these aspects of performance and outweigh them formally in a predefined way.

4.7 | Performance score by Liu et al

Liu et al[13] were the first who proposed a performance score for adaptive designs based on the unconditional sample size and power. According to their definition, a "good" adaptive design is characterized by a sample size close to the optimal fixed design sample size and an unconditional power nearby $1-\beta$.

As the idea of sample size calculation is to determine the smallest sample size that ensures a prespecified power value, Liu et al[13] defined a relative oversizing and an underpowering function. Small values of these two performance functions for a given design relate to a well performing adaptive design.

Taking the optimal fixed design as the gold standard, a design is considered as oversized if the recalculated study sample size is larger than the "correct" fixed sample size. This is expressed by the following relative oversizing function

$$\mathrm{ROS}_{f_s,\beta}(\Delta) := \frac{\mathbb{E}\left[\frac{N_{\text{total}}(Z_1)}{n^{\text{fix}}_{\Delta,1-\beta}} - 1\right]_+}{f_s - 1}, \tag{20}$$

where Δ is the true underlying standardized effect and $n^{\text{fix}}_{\Delta,1-\beta}$ the "ideal" sample size per group in the fixed design for a target power of $1-\beta$ given as

$$n^{\text{fix}}_{\Delta,1-\beta} = \frac{2 \cdot \left(q_{1-\alpha} + q_{1-\beta}\right)^2}{\Delta^2}, \tag{21}$$

where $q_{1-\alpha}$ and $q_{1-\beta}$ correspond again to the respective normal quantiles. The parameter $f_s > 1$ corresponds to a constant scaling factor, which was suggested to be chosen as $f_s = 2$ in Liu et al.[13] The expression $[...]_+$ denotes the positive function, which equals the argument if it is positive and 0 otherwise. An adaptive design is thus defined to be 100% oversized if $N_{\text{total}} = f_s \cdot n^{\text{fix}}_{\Delta,1-\beta}$.

In order to define an underpowering function on the same scale, the amount of underpowering is also expressed in terms of sample size. A study design is considered to be underpowered if the power of the adaptive study design for the true standardized effect Δ is smaller than $1-\beta$. Again, let Pow_Δ denote the power of the adaptive design for a true effect Δ. The relative underpowering function is then given as

$$\mathrm{RUP}_{f_p,\beta}(\Delta) = \frac{\left[n^{\text{fix}}_{\Delta,1-\beta} - n^{\text{fix}}_{\Delta,\text{Pow}_\Delta}\right]_+}{n^{\text{fix}}_{\Delta,1-\beta} - n^{\text{fix}}_{\Delta,(1-f_p)\cdot(1-\beta)}}, \tag{22}$$

where $f_p < 1$ is again a constant scaling factor, which was suggested to be chosen as $f_p = 0.2$ in Liu et al.[13] A 100% underpowered design is thus given if $\text{Pow}_\Delta = (1-f_p)\cdot(1-\beta)$. The two performance criteria are combined within a single performance score as follows, either point-wise

$$S^{\text{Liu}}_{f_s f_p,\beta}(\Delta) := \mathrm{ROS}_{f_s,\beta}(\Delta) + \mathrm{RUP}_{f_p,\beta}(\Delta), \tag{23}$$

or as an average score over a predefined interval of plausible treatment effects $[\Delta_l, \Delta_u]$ as

$$\mathrm{AS}^{\text{Liu}}_{f_s f_p,\beta} = \int_{\Delta_l}^{\Delta_u} \mathrm{ROS}_{f_s,\beta}(\Delta) + \mathrm{RUP}_{f_p,\beta}(\Delta)\, d\Delta. \tag{24}$$

Liu et al[13] also discuss the option to define a weighted average score. We will not investigate this option for the sake of simplicity. The range of the point-wise and average Liu score is $[0, \infty)$. Thereby, a value close to 0 of the score relates to a good performance.

The above performance score is based on the idea that the global, unconditional power and sample size of an ideal sample size recalculation rule should only minimally deviate from the reference design.

"Underpowering" mainly occurs because of a limitation of the maximal sample size which is not related to the performance of the design. Similarly, the total unconditional expected sample size of a "correct" nonadaptive group sequential design is smaller than the corresponding sample size of a fixed design, in particular if the underlying effect is large.

Moreover, the performance score of Liu et al[13] does not take into account the variability of the random adaptive design. Another potential drawback is that the range of the score is not bounded from above so it is difficult to judge which value of the score defines a "bad" performance.

4.8 | A new conditional performance score

The need for a performance score that overcomes these shortcomings and considers conditional performance measures is evident. The basic ideas to construct such a score as proposed by Liu et al[13] are appealing; however, it has been discussed above that there is still room for improvement. Therefore, we propose a new conditional performance score within this section. Similarly to Liu et al,[13] we construct the global score based on two sub-scores, where one sub-score assesses the performance according to conditional sample size and the other sub-score measures the conditional power performance. The main arguments for this approach are as follows: (1) If the study results at interim neither suggest stopping for efficacy nor for futility, then the principal investigator is interested in optimizing the second stage performance. In contrast, the global performance is always influenced by the power and sample size at the interim stage and therefore a sufficient global performance does not necessarily guarantee an acceptable second stage performance. (2) Using conditional performance measures enables the definition of (at least rough) target values for the power and sample size. Note that the conditional perspective—and therefore also the following defined conditional performance score—naturally only refers to trial designs with interim analyses.

The sub-score assessing the performance according to sample size is thus based on assessing the expected sample size conditional on the recalculation area $\mathbb{E}\left[N_{\text{total}}^{\text{RA}}(Z_1)\right]$ as defined in Equation (19). Thereby, $\mathbb{E}\left[N_{\text{total}}^{\text{RA}}(Z_1)\right]$ is estimated as the average over all second stage sample sizes within the recalculation area. Equivalently, $\text{Var}(N_{\text{total}}^{\text{RA}}(Z_1))$ is estimated by the corresponding empirical variance.

As discussed in Section 4.4, it can be argued that if an increase in sample size is considered after the interim analysis, a sample size slightly larger than the required sample size for a fixed design $n_{\Delta,1-\beta}^{\text{fix}}$ with true standardized effect Δ and power $1-\beta$ can be interpreted as a reasonable target value. If the sample size required for the fixed design exceeds the maximally allowed sample size, then an increase of the sample size might be considered as "not worth the effort". We therefore suggest the following target values for the average conditional sample size for given Δ and β,

$$N_{\Delta,\beta}^{\text{target}} := \begin{cases} n_{\Delta,1-\beta}^{\text{fix}}, & \text{if } n_{\Delta,1-\beta}^{\text{fix}} \leq n_{\max} \text{ and } \Delta \neq 0, \\ n_1, & \text{if } n_{\Delta,1-\beta}^{\text{fix}} > n_{\max} \text{ or } \Delta = 0. \end{cases} \tag{25}$$

With this, the sub-score for the conditional sample size can be defined as

$$\text{SN}_\beta(\Delta) := \begin{cases} \frac{1}{2} \cdot \left(\underbrace{\left(1 - \frac{\left| \mathbb{E}\left[N_{\text{total}}^{\text{RA}}(Z_1)\right] - N_{\Delta,\beta}^{\text{target}} \right|}{n_{\max} - n_1} \right)}_{:=e_N(\Delta)} + \underbrace{\left(1 - \sqrt{\frac{\text{Var}\left(N_{\text{total}}^{\text{RA}}(Z_1)\right)}{\text{Var}_{\max}\left(N_{\text{total}}^{\text{RA}}(Z_1)\right)}} \right)}_{:=v_N(\Delta)} \right), & \text{if } \text{Var}_{\max}\left(N_{\text{total}}^{\text{RA}}(Z_1)\right) \neq 0, \\ \frac{1}{2} \cdot \left(\underbrace{\left(1 - \frac{\left| \mathbb{E}\left[N_{\text{total}}^{\text{RA}}(Z_1)\right] - N_{\Delta,\beta}^{\text{target}} \right|}{n_{\max} - n_1} \right)}_{:=e_N(\Delta)} + \underbrace{1}_{:=v_N(\Delta)} \right), & \text{if } \text{Var}_{\max}\left(N_{\text{total}}^{\text{RA}}(Z_1)\right) = 0, \end{cases} \tag{26}$$

where $e_N(\Delta) \in [0;1]$ is a parameter assessing the location of the conditional sample size and $v_N(\Delta) \in [0;1]$ is a parameter assessing the variance of the conditional sample size. Large values of $e_N(\Delta)$ and $v_N(\Delta)$ are favorable. Note that both parameters $e_N(\Delta)$ and $v_N(\Delta)$ depend on the true underlying effect Δ. This natural dependence was omitted in the single terms to ease readability. The sub-score for the conditional sample size (26) has a range between 0 and 1. The expression $\text{Var}_{\max}\left(N^{\text{RA}}_{\text{total}}\right)$ corresponds to the maximally possible variance for the conditional sample size which can be observed. The maximal value for the variance intuitively occurs when the conditional sample size is given by n_1 in 50% of all realizations and by $n_{\max}$ in the remaining 50% of all realizations. Hence, an upper limit for the maximally possible variance is given by

$$\text{Var}\left(N^{\text{RA}}_{\text{total}}(Z_1)\right) \leq \left(\frac{n_{\max} - n_1}{2}\right)^2 =: \text{Var}_{\max}\left(N^{\text{RA}}_{\text{total}}(Z_1)\right), \tag{27}$$

and used for the calculation of the sub-score for $\text{Var}_{\max}\left(N^{\text{RA}}_{\text{total}}(Z_1)\right)$. Note that by definition of Equation (26), the sub-score can also be applied to classical group sequential study designs where the variance of the recalculated sample size equals 0 and therefore corresponds to a perfect variation component $v_N(\Delta)$ of 1.

Similarly, the sub-score assessing the performance according to power is based on assessing the expected conditional power $\mathbb{E}\left[\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right]$ which is estimated as the average over all conditional power values within the recalculation area. The corresponding variance $\text{Var}(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1))$ is accordingly estimated by the empirical variance. As discussed in Section 4.2, in case the sample size is increased at interim, the conditional power should ideally reach a level of $1-\beta$. If the required sample size for the fixed design exceeds the maximally allowed sample size, an increase of the sample size is not recommended and the conditional power should then be close to the local one-sided significance level α. We therefore suggest the following target values for the average conditional sample size

$$\text{CP}^{\text{target}}_{\Delta,\beta} := \begin{cases} 1-\beta, & \text{if } n^{\text{fix}}_{\Delta,1-\beta} \leq n_{\max} \text{ and } \Delta \neq 0, \\ \alpha, & \text{if } n^{\text{fix}}_{\Delta,1-\beta} > n_{\max} \text{ or } \Delta = 0. \end{cases} \tag{28}$$

The sub-score for the conditional power can then be defined as

$$\text{SCP}_\beta(\Delta) := \frac{1}{2} \cdot \left(\underbrace{\left(1 - \frac{\left|\mathbb{E}\left[\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right] - \text{CP}^{\text{target}}_{\Delta,\beta}\right|}{1-\alpha}\right)}_{=:e_{\text{CP}}(\Delta)} + \underbrace{\left(1 - \sqrt{\frac{\text{Var}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right)}{\text{Var}_{\max}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right)}}\right)}_{=:v_{\text{CP}}(\Delta)} \right), \tag{29}$$

where $e_{\text{CP}}(\Delta) \in [0;1]$ is a parameter assessing the location of the conditional sample size and $v_{\text{CP}}(\Delta) \in [0;1]$ assessing the variance of the conditional sample size. As before, both parameters $e_{\text{CP}}(\Delta)$ and $v_{\text{CP}}(\Delta)$ depend on the true underlying effect Δ and large values of $e_{\text{CP}}(\Delta)$ and $v_{\text{CP}}(\Delta)$ are favorable. $\text{Var}_{\max}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}\right)$ corresponds to the maximally possible variance for the conditional power which can be observed. The maximal value for the variance intuitively occurs if the conditional power is given by 0 in 50% of all realizations and by 1 in the remaining 50% of all realizations,

$$\text{Var}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right) \leq \left(\frac{1-0}{2}\right)^2 = 0.25 =: \text{Var}_{\max}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right), \tag{30}$$

and used for the calculation of the conditional power sub-score for $\text{Var}_{\max}\left(\text{CP}^{\text{RA}}_{\hat{\Delta}_1}(Z_1)\right)$. Note that both conditional power values of 0 and 1 in the maximal variance scenario in Equation (30) are never observed in practice. However, based on these values an upper boundary for the maximal possible variance is obtained. Moreover, note that we do not need to consider two cases for the conditional power sub-score in Equation (29) as the maximal possible variance cannot become 0.

Both sub-scores are based on the idea of punishing deviations from the corresponding target value as well as large variances of the conditional sample size or power distributions. Both sub-scores have a range of $[0;1]$, where values close

to 1 indicate a good performance. It is, however, intuitive that $\mathrm{SN}_\beta(\Delta)$ and $\mathrm{SCP}_\beta(\Delta)$ are negatively correlated and therefore should always be investigated as a pair. We therefore suggest the following global point-wise performance score

$$S_\beta^{\mathrm{New}}(\Delta) := \frac{\mathrm{SN}_\beta(\Delta) + \mathrm{SCP}_\beta(\Delta)}{2}, \tag{31}$$

which allows to measure the performance with respect to sample size *in relation* to the performance with respect to power. Equivalently as for the score by Liu et al,[13] the new score can as well be extended to an average score over a predefined interval of plausible treatment effects $[\Delta_l, \Delta_u]$ as

$$\mathrm{AS}_\beta^{\mathrm{New}} = \int_{\Delta_l}^{\Delta_u} S_\beta^{\mathrm{New}}(\Delta)\, d\Delta. \tag{32}$$

The global point-wise and the average score both have a range of $[0; 1]$, where values close to 1 indicate a good performance.

By Equations (26), (29), and (31), the components of the conditional score are weighted equally. Note that these weights could also be chosen differently.

4.9 | Interpretation of the new conditional score

As noted above, the score has absolute values for best ($S_\beta^{\mathrm{New}}(\Delta) = 1$) and worst ($S_\beta^{\mathrm{New}}(\Delta) = 0$) performance. However, these two extremes can practically never be reached by adaptive group sequential designs. For a score value of 1, the expected value of the sample size $\mathbb{E}[N_{\mathrm{total}}^{\mathrm{RA}}(Z_1)]$ and conditional power $\mathbb{E}[\mathrm{CP}_{\hat{\Delta}_1}^{\mathrm{RA}}(Z_1)]$ must exactly meet the corresponding target values ($N_{\Delta,\beta}^{\mathrm{target}}, \mathrm{CP}_{\Delta,\beta}^{\mathrm{target}}$) and the two variances must be equal to 0. Similarly, for a performance score of 0, the two variances need to equal the maximally possible variance. Moreover, the absolute value of the difference between the expected sample size $\mathbb{E}[N_{\mathrm{total}}^{\mathrm{RA}}(Z_1)]$ and target sample size $N_{\Delta,\beta}^{\mathrm{target}}$ needs to equal $n_{\max} - n_1$. Similarly, the absolute value of the difference between the expected conditional power $\mathbb{E}[\mathrm{CP}_{\hat{\Delta}_1}^{\mathrm{RA}}(Z_1)]$ and target conditional power $\mathrm{CP}_{\Delta,\beta}^{\mathrm{target}}$ needs to equal $1 - \alpha$.

Hence, the question arises which score values indicate an observable high, medium, or low performance. A general recommendation cannot be given here, as this depends importantly on the underlying study-specific setting. However, in this section, we provide rules of thumb to derive score ranges for low, medium, and high performances.

For a high performance, we expect the observed variance of the total sample size conditional on entering the recalculation area to be at most 30% of the maximally possible variance of the sample size conditional on entering the recalculation area. Moreover, we anticipate a deviation of the observed sample size from the target sample size of at most 30% of the largest possible sample size deviation. The same considerations are made for the conditional power. This translates into the following conditions:

- $\mathrm{Var}\left(N_{\mathrm{total}}^{\mathrm{RA}}(Z_1)\right) \leq 0.3 \cdot \mathrm{Var}_{\max}\left(N_{\mathrm{total}}^{\mathrm{RA}}(Z_1)\right)$,
- $\mathrm{Var}\left(\mathrm{CP}_{\hat{\Delta}_1}^{\mathrm{RA}}(Z_1)\right) \leq 0.3 \cdot \mathrm{Var}_{\max}\left(\mathrm{CP}_{\hat{\Delta}_1}^{\mathrm{RA}}(Z_1)\right)$,
- $\left|\mathbb{E}\left[N_{\mathrm{total}}^{\mathrm{RA}}(Z_1)\right] - N_{\Delta,\beta}^{\mathrm{target}}\right| \leq 0.3 \cdot (n_{\max} - n_1)$ and
- $\left|\mathbb{E}\left[\mathrm{CP}_{\hat{\Delta}_1}^{\mathrm{RA}}(Z_1)\right] - \mathrm{CP}_{\Delta,\beta}^{\mathrm{target}}\right| \leq 0.3 \cdot (1 - \alpha)$.

Thus, a total performance score $S_\beta^{\mathrm{New}}(\Delta) = 0.5 \cdot \left((1 - 0.3) + \left(1 - \sqrt{0.3}\right)\right) \approx 0.576$ or higher corresponds to a high performance.

Accordingly, for a medium performance, we expect the observed variance of the total sample size conditional on entering the recalculation area to be at most 50% of the maximally possible variance of the sample size conditional on entering the recalculation area. Moreover, we anticipate a deviation of the observed sample size from the target sample size of at most 50% of the largest possible sample size deviation. The same considerations are made for the conditional

power. Hence, a performance score $S_\beta^{\text{New}} \in [0.396; 0.576)$ measures up to a medium performance and $S_\beta^{\text{New}}(\Delta) < 0.396$ corresponds consequently to a low performance.

These ranges can also be used for the sub-scores $\text{SN}_\beta(\Delta)$ and $\text{SCP}_\beta(\Delta)$ as well as the components for location and variation $(e_N(\Delta), v_N(\Delta), e_{\text{CP}}(\Delta), v_{\text{CP}}(\Delta))$. Obviously, different assumptions can be made as well as one could also distinguish between different values for the conditional power and sample size sub-scores. Note that a certain design can be classified differently if other target values are specified.

5 | SIMULATION STUDY

In order to compare our new conditional performance score to the performance score proposed by Liu et al,[13] we performed a Monte-Carlo simulation study with the software R,[22] where the sample size recalculation rules introduced in Section 3 were compared based on both performance scores. Moreover, we added a classical group sequential design for comparison purposes. Remember that Liu's score is a global, unconditional performance score and the new score is conditional on the interim results. Therefore, the performance rankings measure two different, but related, aspects and are thus not necessarily intended to deliver the same performance ranking.

5.1 | Simulation setup

Within our simulation setting, we considered the test problem as introduced in Section 2.1. The global one-sided significance level was set to $\alpha = 0.025$. The local significance levels were adjusted according to Pocock,[19] that is, $\alpha_1 = \alpha_{1+2} = 0.0147$. The futility bound was set to $\alpha_0 = 0.5$ meaning that observed treatment effects which point into the wrong direction lead to early stopping after the first stage, compare, for example, Reference 4. The anticipated power was given by $1 - \beta = 0.8$. We assumed equal sample sizes per group for the sake of simplicity. The optimal sample size for the fixed design was calculated with the help of the `power.t.test` function in R.[22] We considered three simulation settings differing in their interim and maximal sample sizes. Apart from the four presented adaptive group sequential sample size recalculation rules, we also calculated performance measures for the classical group sequential study design. For the classical group sequential scenario, we took the prespecified values for n_1 and n_2.

In the main scenario shown here, the initial sample size per group for the first stage was fixed to $n_1 = 50$. If the trial was not stopped at interim, additional patients were recruited. The initial value for the second stage sample size per group was fixed to $n_2 = 50$. This motivates the choice of the weights for the definition of the final test statistic given in Equation (5) which were fixed to $w_1 = w_2 = \sqrt{50}$. Moreover, we used a maximal bound for the sample size per group of $n_{\max} = 4 \cdot n_1 = 200$.

In the second setting, we assumed $n_1 = n_2 = 25$ with $w_1 = w_2 = \sqrt{25}$ and $n_{\max} = 8 \cdot n_1 = 200$. The latter setting was implemented to evaluate the influence of a different interim sample size, however, with a constant absolute maximal sample size of 200.

In the third setting, we assumed $n_1 = n_2 = 25$ with $w_1 = w_2 = \sqrt{25}$ and $n_{\max} = 4 \cdot n_1 = 100$. This setting was considered to assess the influence of the interim sample size with the same sample size boundary factor $f = 4$ as in the main setting. The second and third setting are addressed in the Appendix.

Within each setting, we investigated eight scenarios for the underlying true standardized effect Δ ranging from 0.0 to 0.6 by steps of 0.1 (except for the interim step at 0.35). We decided for the extra effect size step at 0.35 as this is the effect area with the highest variability in sample size. For each of these eight scenarios, we drew 10 000 replications from a normal distribution $N(\Delta\sqrt{n_1/2}, 1)$ expressing the observed values of the interim test statistics. Based on this set of observed values for the interim test statistic, we applied the four different sample size recalculation rules presented in Section 3, that are the observed conditional power approach, the restricted observed conditional power approach, the promising zone approach as well as the optimization function approach. Moreover, we added also the classical group sequential design. For all resulting scenarios, Liu's score and the newly proposed conditional score were calculated—both point-wise for each Δ. Moreover, we provided Liu's measures for oversizing $\text{ROS}_{f_s,\beta}(\Delta)$ and underpowering $\text{RUP}_{f_p,\beta}(\Delta)$ as well as the conditional sub-scores for sample size $\text{SN}_\beta(\Delta)$ and power $\text{SCP}_\beta(\Delta)$ together with the location and variation components $e_N(\Delta), e_{\text{CP}}(\Delta), v_N(\Delta), v_{\text{CP}}(\Delta)$. For the main setting $(n_1 = n_2 = 50, f = 4)$, also averaged values and scores

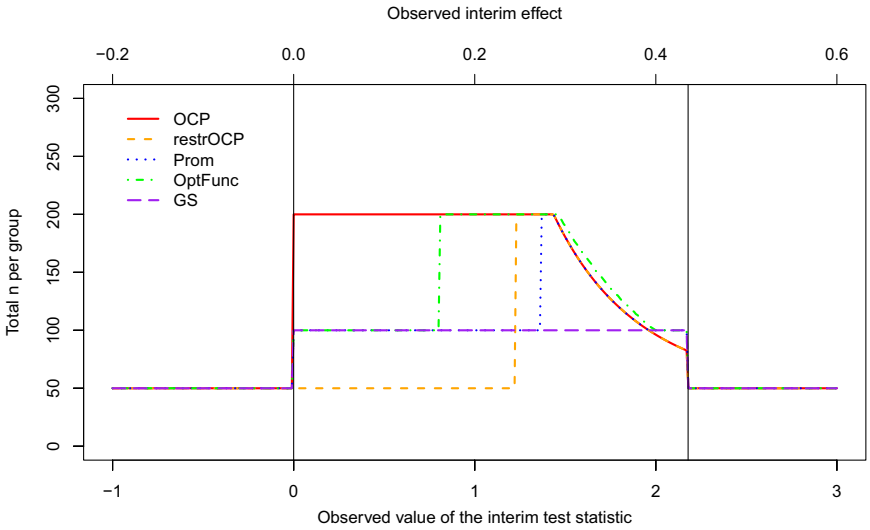

FIGURE 1 Sample size recalculation rules investigated within the Monte-Carlo simulation study for $n_1 = n_2 = 50$, $n_{\max} = 200$. GS, group sequential approach; OCP, observed conditional power approach; OptFunc, optimization function approach; Prom, promising zone approach; restrOCP, restricted observed conditional power approach [Colour figure can be viewed at wileyonlinelibrary.com]

over the range $[0.0; 0.6]$ as well as for the three effect ranges $[0.0; 0.2]$, $[0.3; 0.4]$, and $[0.5; 0.6]$ were calculated. Concerning specific score and design parameters, we used the same scaling factors for oversizing and underpowering as proposed by Liu et al[13] given as $f_s = 2$ and $f_p = 0.2$. Moreover, in the restricted observed conditional power design, the predefined minimal conditional power level was set to 0.6. Concerning the promising zone design, the boundary $1 - \tilde{\beta}_0$ for the unfavorable zone was given by 0.36 according to Reference 7. In the optimal function design, we chose $\gamma = 0.005/4$. Note that these parameter settings for the recalculation rules can be chosen differently. Figure 1 illustrates all investigated recalculation rules under the parameter settings used within this simulation as functions of the observed treatment effect at interim.

5.2 | Simulation results

The results of our main setting with $n_1 = n_2 = 50$ and $n_{\max} = 200$ are displayed in Figure 2 as well as in Tables 1-3. The second setting's ($n_1 = 25, n_{\max} = 200$) and third setting's ($n_1 = 25, n_{\max} = 100$) simulation results can be found in Tables A1-A4 in Appendix.

We focus on describing the simulation results of the main setting and state briefly a few points on the comparison with the other two settings. The specific results of the second and third setting are not described in detail and we refer the interested reader to Tables A1-A4 in Appendix. Concerning the main setting, we only analyze the global and conditional performance scores and their sub-scores with respect to the obtained sample size and conditional power values. If the reader is interested in a compact overview, whether a high or low conditional sub-score ($\mathrm{SN}_\beta(\Delta)$, $\mathrm{SCP}_\beta(\Delta)$) was mainly obtained due to the location ($e_N(\Delta), e_{\mathrm{CP}}(\Delta)$) or variation component ($v_N(\Delta), v_{\mathrm{CP}}(\Delta)$), we refer the reader to Table 2. The component values can also be found in Figure 2 above the respective boxplots. Figure 2 shows boxplots for the conditional sample size and the conditional power (conditional on having entered the recalculation area) for each design and selected standardized effects $\Delta = 0.0, 0.3, 0.6$. When looking at the boxplots describing the recalculated sample sizes per group for $\Delta = 0.3$ (Figure 2, second row), one expects the observed conditional power approach to perform best with respect to location and the classical group sequential study design together with the restricted observed conditional power approach to perform worst. This assumption can be verified by the values of the location component $e_N(0.3)$ (observed conditional power approach: 0.965; classical group sequential approach: 0.492; restricted observed conditional power approach: 0.493). Moreover, the boxplots show the restricted observed conditional power approach with the highest variability in sample size and the classical group sequential approach with the lowest. Also here, this goes in line with the respective variation components $v_N(0.3)$ (restricted observed conditional power approach: $v_N(0.3) = 0.247$; classical group sequential approach: $v_N(0.3) = 1.000$). Recall that a value of 0 refers to the worst and 1 to the best performance.

Table 1 presents the estimated point-wise performance scores and related conditional and unconditional performances measures for all investigated simulation scenarios. Column 1 displays the underlying true standardized treatment effect Δ and the corresponding sample size per group of the fixed design $n^{\mathrm{fix}}_{\Delta,0.8}$. The underlying sample size recalculation design is provided in Column 2. Column 3 shows the estimated total expected sample size over both stages $\mathbb{E}[N_{\mathrm{total},\Delta}]$ as introduced

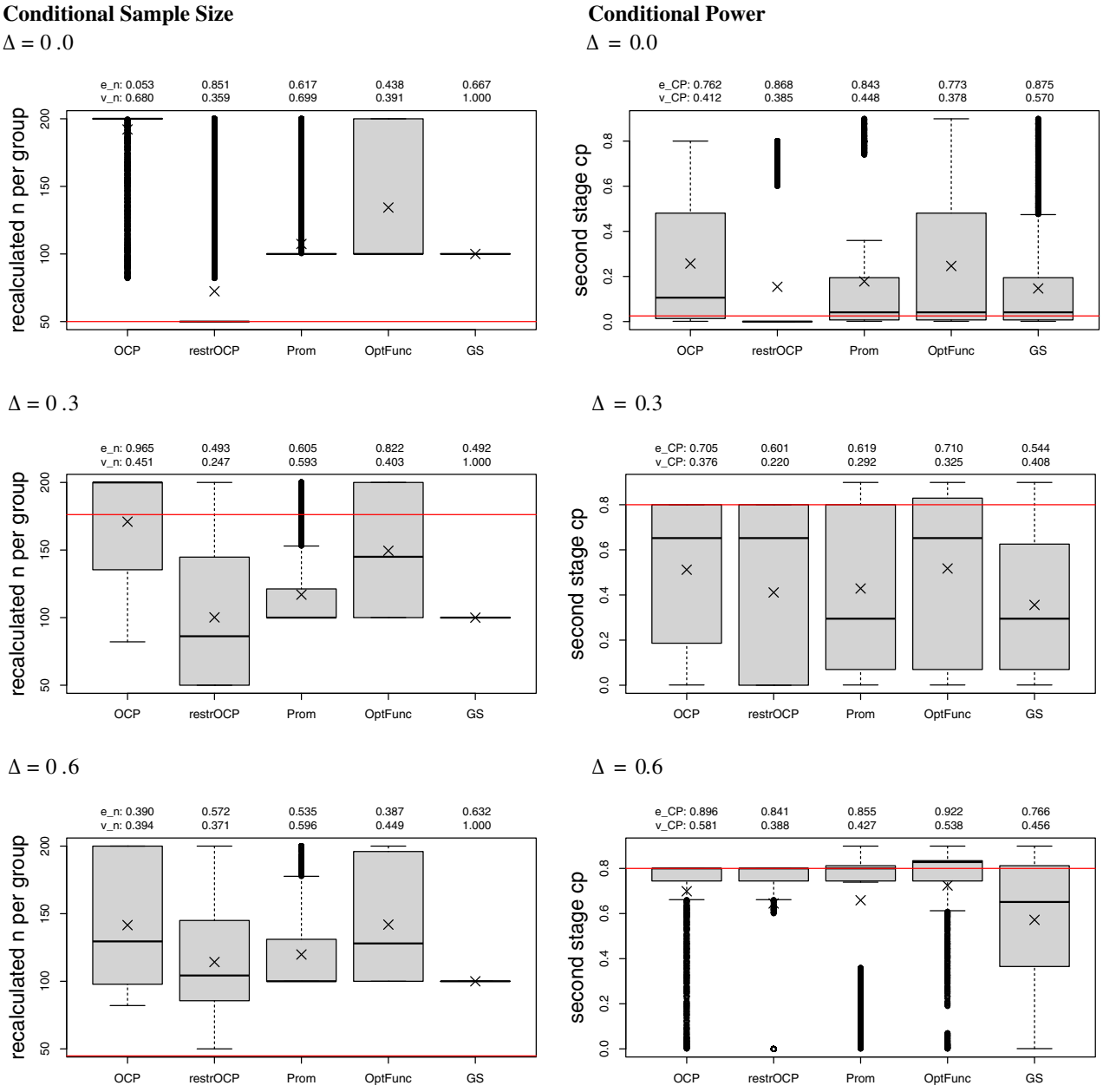

FIGURE 2 Boxplots for the conditional sample size and for the conditional power (conditional on having entered the recalculation area) for different recalculation rules and underlying effects Δ with $n_1 = n_2 = 50$, $n_{\max} = 200$. Red lines indicate the target values, $N^{\text{target}}_{\Delta,\beta}$ or $\text{CP}^{\text{target}}_{\Delta,\beta}$, respectively. Average values are represented by black crosses. e_N: location component for conditional sample size; v_N: variation component for conditional sample size; e_{CP}: location component for conditional power; v_{CP}: variation component for conditional power. GS, group sequential approach; OCP, observed conditional power approach; OptFunc, optimization function approach; Prom, promising zone approach; restrOCP, restricted observed conditional power approach [Colour figure can be viewed at wileyonlinelibrary.com]

in Equation (8). The index Δ is added here to explore the dependence on the true underlying treatment effect. As the total expected sample size is directly related to the relative oversizing function defined by Liu et al,[13] Column 4 subsequently displays the relative oversizing function $\text{ROS}_{f_s,\beta}(\Delta)$. Column 5 shows the global, unconditional power Pow_Δ based on the true underlying standardized treatment effect size Δ of the design. Again, the global power is directly related to the relative underpowering function defined by Liu et al[13] and therefore Column 6 displays the relative underpowering function $\text{RUP}_{f_p,\beta}(\Delta)$. Column 7 finally shows the total point-wise Liu performance score $S^{\text{Liu}}_{f_s f_p,\beta}(\Delta)$. Thereby, remember that Liu's point-wise performance score is not applicable for a true standardized effect size of $\Delta = 0.0$ as the score is not well-defined under the null hypothesis. As the newly proposed conditional performance score is based on conditional performance

measures, we report the estimated expected sample size conditional on having entered the recalculation area given by $\mathbb{E}\left[N^{\text{RA}}_{\text{total},\Delta}\right]$ in Column 8 and the related variance $\text{Var}(N^{\text{RA}}_{\text{total},\Delta})$ in Column 9. The new conditional sub-score assessing the sample size performance $\text{SN}_\beta(\Delta)$ is directly related to these two measures and is reported in Column 10. Columns 11 and 12 show the estimated expected conditional power conditional on entering the recalculation area given as $\mathbb{E}[\text{CP}^{\text{RA}}_\Delta]$ and the related variance $\text{Var}(\text{CP}^{\text{RA}}_\Delta)$. Again, the new conditional sub-score assessing the power performance $\text{SCP}_\beta(\Delta)$ is directly related to these two measures and is reported in Column 13. Finally, Column 14 shows the total point-wise new conditional performance score $S^{\text{New}}_\beta(\Delta)$. The numbers in square brackets after the performance scores show the ranking of the group sequential designs for each Δ according to the underlying performance score. Effects varying from $\Delta = 0.0$ until $\Delta = 0.6$ by steps of 0.1 are presented with the exception of the additional value $\Delta = 0.35$. Recall that the reason for this interim step is that the sample sizes for an effect size around $\Delta = 0.35$ show the highest variability as the sample size functions include a "jump" in this area.

Now, we come to the simulation results. The classical group sequential design plays a special role. Concerning the global score by Liu et al,[13] it is performing worse than the other designs with sample size recalculation for the underlying effect sizes since it tends to achieve a lower power. However, in terms of the conditional evaluation, the classical group sequential design is the clear winner for equally weighted components ($e_N(\Delta)$, $e_{\text{CP}}(\Delta)$, $v_N(\Delta)$, $v_{\text{CP}}(\Delta)$). This is due to the fact that particularly the component with respect to sample size variation attains always the perfect value of 1 paired with a high value for the conditional power variation component. In the following description, we compare the other four designs (observed conditional power, restricted observed conditional power, promising zone, and optimization function approach) such that we do not go in to further details for the classical group sequential design. Hence, the term sample size recalculation rules refers to the observed conditional power, restricted observed conditional power, promising zone, and optimization function approach in the following. Table 1 shows the following results. Under the null hypothesis (ie, $\Delta = 0.0$), the relative underpowering function, relative oversizing function, and Liu's score are not defined. The sub-score assessing the conditional sample size favors the promising zone approach (Row 3, $\text{SN}_\beta(0.0) = 0.658$). The observed conditional power approach (Row 1, $\text{SN}_\beta(0.0) = 0.366$) shows the worst result with respect to sample size which is due to hitting the target value worst of all four designs. The ranking with respect to the conditional power is similar. The best performance sub-score is also achieved by the promising zone design (Row 3, $\text{SCP}_\beta(0.0) = 0.646$), the optimization function approach shows the worst performance (Row 4, $\text{SCP}_\beta(0.0) = 0.576$). The total point-wise performance score thus shows the best performance for the promising zone approach (Row 3, $S^{\text{New}}_\beta(0.0) = 0.652$), whereas the observed conditional power approach (Row 1, $S^{\text{New}}_\beta(0.0) = 0.477$) and the optimization function approach (Row 4, $S^{\text{New}}_\beta(0.0) = 0.495$) show the worst results. This can also be verified graphically in Figure 2. The boxplots of the conditional sample sizes show that in the absence of a treatment effect the median conditional sample size of the observed conditional power approach is far from $N^{\text{target}}_{0.0,\beta} = n_1 = 50$ which would be optimal for $\Delta = 0.0$. The promising zone approach boxplot shows hitting the target conditional power of 0.025 together with a small variance. The new conditional score thus verifies the visual inspection of the boxplots, as the new conditional score shows the highest value for the promising zone design with $S^{\text{New}}_\beta(0.0) = 0.652$, followed by the restricted observed conditional power approach $S^{\text{New}}_\beta(0.0) = 0.615$, compare Column 14 of Table 1.

For underlying standardized effects of $\Delta = 0.1$ or $\Delta = 0.2$, the ranking with respect to the new score remains stable. However, now we can compare the ranking of the new conditional score to the ranking by Liu's score. Liu's score shows best performance values for the observed conditional power approach (Row 6, $S^{\text{Liu}}_{f_s f_p,\beta}(0.1) = 2.927$; Row 11, $S^{\text{Liu}}_{f_s f_p,\beta}(0.2) = 2.100$), and the optimization function approach (Row 9, $S^{\text{Liu}}_{f_s f_p,\beta}(0.1) = 2.944$; Row 14, $S^{\text{Liu}}_{f_s f_p,\beta}(0.2) = 2.198$). However, the Liu scores are all very similar across the different designs. Note that for these small values of Δ, Liu's score is exclusively based on the relative underpowering function as the maximal sample size of the adaptive design is smaller than the required sample size of the corresponding fixed design. Thus, in fact the Liu score only assesses the relative underpowering. It is evident that recalculation designs that augment the sample size only in case the observed effect suggests that a target conditional power value can be reached show a low power if the underlying effect is small.

For an underlying standardized effect of $\Delta = 0.3$, a visual inspection of Figure 2 would suggest the observed conditional power approach as the clear "winner" of the sample size recalculation rules as it meets the target power and sample size best and shows the lowest variance. Here, the Liu score and the new conditional score show a similar ranking starting indeed with the observed conditional power approach as the best design whereas the restricted observed conditional power approach shows the worst performance. Again, Liu's score is exclusively based on the relative underpowering function. Therefore, a performance comparison with respect to sample size aspects alone is not possible for Liu's score. However, it is noticeable that the observed conditional power approach performs best now (Row 16, $S^{\text{New}}_\beta(0.3) = 0.624$) where it was the worst performing design for $\Delta = 0.2$ according to the new score (Row 11, $S^{\text{New}}_\beta(0.2) = 0.398$). This is due

TABLE 1 Estimated point-wise performance scores and related conditional and unconditional performances measures for all investigated simulation scenarios with $n_1 = n_2 = 50$, $n_{max} = 200$

		Parameters to be Estimated											
Δ ($n^{fix}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{total,\Delta}]$	$ROS_{f_s,\beta}(\Delta)$	Pow_Δ	$RUP_{f_p,\beta}(\Delta)$	$S^{Liu}_{f_s,f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$	$Var\left(N^{RA}_{total,\Delta}\right)$	$SN_\beta(\Delta)$	$\mathbb{E}\left[CP^{RA}_\Delta\right]$	$Var\left(CP^{RA}_\Delta\right)$	$SCP_\beta(\Delta)$	$S^{New}_\beta(\Delta)[\cdot]$
0.0	OCP	119.396	–	0.025	–	–	192.119	575.126	0.366	0.257	0.087	0.587	0.477[5]
(–)	restrOCP	60.941	–	0.022	–	–	72.407	2314.387	0.605	0.154	0.095	0.626	0.615[3]
	Prom	78.024	–	0.025	–	–	107.392	508.872	0.658	0.178	0.076	0.646	0.652[2]
	OptFunc	91.189	–	0.025	–	–	134.352	2088.885	0.414	0.246	0.097	0.576	0.495[4]
	GS	74.415	–	0.025	–	–	100.00	0.000	0.833	0.147	0.046	0.722	0.778[1]
0.1	OCP	138.768	0.000	0.119	2.927	2.927[1]	186.565	934.978	0.341	0.337	0.102	0.520	0.431[5]
(1571)	restrOCP	70.458	0.000	0.090	3.017	3.017[5]	81.473	2830.996	0.540	0.230	0.125	0.541	0.541[3]
	Prom	89.464	0.000	0.101	2.982	2.982[3]	110.714	686.199	0.623	0.253	0.102	0.564	0.593[2]
	OptFunc	108.898	0.000	0.114	2.944	2.944[2]	140.612	2144.090	0.389	0.331	0.116	0.503	0.446[4]
	GS	82.500	0.000	0.094	3.004	3.004[4]	100.000	0.000	0.833	0.208	0.064	0.652	0.743[1]
0.2	OCP	143.616	0.000	0.370	2.100	2.100[1]	179.358	1341.474	0.325	0.422	0.106	0.471	0.398[5]
(395)	restrOCP	79.521	0.000	0.261	2.469	2.469[4]	90.792	3126.289	0.491	0.317	0.146	0.468	0.480[3]
	Prom	96.006	0.000	0.287	2.384	2.384[3]	113.571	803.273	0.599	0.336	0.120	0.495	0.547[2]
	OptFunc	119.292	0.000	0.342	2.198	2.198[2]	145.747	2106.018	0.375	0.421	0.122	0.448	0.411[4]
	GS	86.185	0.000	0.257	2.480	2.480[5]	100.000	0.000	0.833	0.278	0.080	0.588	0.711[1]
0.3	OCP	131.934	0.000	0.702	0.667	0.667[1]	170.972	1694.175	0.708	0.512	0.097	0.540	0.624[1]
(177)	restrOCP	83.994	0.000	0.511	1.575	1.575[4]	100.191	3192.310	0.370	0.411	0.152	0.410	0.390[5]
	Prom	95.377	0.000	0.561	1.368	1.368[3]	116.997	933.024	0.599	0.429	0.125	0.456	0.527[4]
	OptFunc	117.361	0.000	0.651	0.942	0.942[2]	149.456	2007.543	0.612	0.517	0.114	0.518	0.565[3]
	GS	83.865	0.000	0.511	1.576	1.576[5]	100.000	0.000	0.746	0.356	0.088	0.476	0.611[2]
0.35	OCP	121.773	0.000	0.826	0.000	0.000[1]	165.726	1858.207	0.594	0.555	0.090	0.574	0.584[4]

(Continues)

TABLE 1 (Continued)

		Parameters to be Estimated											
Δ ($n^{fix}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{total,\Delta}]$	$ROS_{f_s,\beta}(\Delta)$	Pow_Δ	$RUP_{f_p,\beta}(\Delta)$	$S^{Liu}_{f_s f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$	$Var\left(N^{RA}_{total,\Delta}\right)$	$SN_\beta(\Delta)$	$\mathbb{E}\left[CP^{RA}_\Delta\right]$	$Var\left(CP^{RA}_\Delta\right)$	$SCP_\beta(\Delta)$	$S^{New}_\beta(\Delta)[\cdot]$
(131)	restrOCP	83.943	0.000	0.636	1.018	1.018[5]	104.729	3106.080	0.544	0.462	0.148	0.442	0.493[5]
	Prom	92.022	0.000	0.693	0.717	0.717[3]	117.755	924.542	0.756	0.475	0.123	0.483	0.620[2]
	OptFunc	111.628	0.000	0.778	0.171	0.171[2]	149.368	1941.097	0.642	0.564	0.106	0.553	0.598[3]
	GS	81.010	0.000	0.645	0.973	0.973[4]	100.000	0.000	0.900	0.399	0.089	0.495	0.698[1]
0.4	OCP	110.203	0.123	0.907	0.000	0.123[3]	160.343	1980.502	0.502	0.592	0.082	0.607	0.555[4]
(101)	restrOCP	81.508	0.000	0.741	0.427	0.427[5]	107.750	2932.263	0.613	0.509	0.141	0.475	0.544[5]
	Prom	88.009	0.000	0.805	0.000	0.000[1]	119.665	996.853	0.724	0.525	0.119	0.515	0.619[2]
	OptFunc	103.586	0.056	0.868	0.000	0.056[2]	148.215	1888.115	0.550	0.605	0.097	0.588	0.569[3]
	GS	77.280	0.000	0.763	0.275	0.275[4]	100.000	0.000	1.000	0.440	0.089	0.516	0.758[1]
0.5	OCP	86.306	0.374	0.977	0.000	0.374[5]	149.414	2120.409	0.410	0.654	0.062	0.676	0.543[4]
(65)	restrOCP	72.446	0.154	0.886	0.000	0.154[2]	111.462	2533.312	0.508	0.587	0.118	0.547	0.527[5]
	Prom	75.274	0.199	0.938	0.000	0.199[3]	119.205	926.352	0.614	0.600	0.102	0.579	0.597[2]
	OptFunc	84.508	0.346	0.963	0.000	0.346[4]	144.491	1796.357	0.450	0.674	0.075	0.661	0.556[3]
	GS	68.260	0.087	0.921	0.000	0.087[1]	100.000	0.000	0.881	0.517	0.085	0.564	0.722[1]
0.6	OCP	68.303	0.566	0.995	0.000	0.566[4]	141.560	2065.001	0.392	0.699	0.044	0.739	0.557[4]
(45)	restrOCP	62.843	0.441	0.958	0.000	0.441[2]	114.246	2223.650	0.471	0.645	0.094	0.614	0.534[5]
	Prom	63.954	0.467	0.987	0.000	0.467[3]	119.807	919.209	0.565	0.659	0.082	0.641	0.595[2]
	OptFunc	68.376	0.568	0.993	0.000	0.568[5]	141.924	1705.966	0.418	0.724	0.053	0.730	0.565[3]
	GS	59.995	0.376	0.982	0.000	0.376[1]	100.000	0.000	0.816	0.572	0.074	0.611	0.714[1]

Δ: true standardized effect; $n^{fix}_{\Delta,0.8}$: optimal sample size in the fixed design; $\mathbb{E}[N_{total,\Delta}]$: Expected sample size per group for both stages; Pow_Δ: power of the adaptive design for standardized treatment effect Δ;
$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$: expected sample size per group conditional on entering the recalculation area; $Var\left(N^{RA}_{total,\Delta}\right)$: variance of sample size per group conditional on entering the recalculation area;
$\mathbb{E}\left[CP^{RA}_\Delta\right]$: expected conditional power conditional on entering the recalculation area; $Var(CP^{RA}_\Delta)$: variance of conditional power conditional on entering the recalculation area;
$ROS_{f_s,\beta}(\Delta)$: relative oversizing function; $RUP_{f_p,\beta}(\Delta)$: relative underpowering function; $S^{Liu}_{f_s f_p,\beta}(\Delta)$: point-wise Liu score;
$SN_\beta(\Delta)$: conditional sample size sub-score; $SCP_\beta(\Delta)$: conditional power sub-score; $S^{New}_\beta(\Delta)$: point-wise new conditional score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach;
$[\cdot]$: numbers in square brackets present the ranking of the group sequential designs according to the corresponding point-wise performance score ([1]: best performance to [5]: worst performance).

TABLE 2 Estimated performance sub-scores and related location and variation components for all investigated designs and $n_1 = n_2 = 50$, $n_{max} = 200$

		Parameters to be Estimated					
Δ	**Design**	$e_N(\Delta)$	$v_N(\Delta)$	$SN_\beta(\Delta)$	$e_{CP}(\Delta)$	$v_{CP}(\Delta)$	$SCP_\beta(\Delta)$
0.0	OCP	0.053	0.680	0.366	0.762	0.412	0.587
	restrOCP	0.851	0.359	0.605	0.868	0.385	0.626
	Prom	0.617	0.699	0.658	0.843	0.448	0.846
	OptFunc	0.438	0.391	0.414	0.773	0.378	0.576
	GS	0.667	1.000	0.833	0.875	0.570	0.722
0.1	OCP	0.090	0.592	0.341	0.680	0.361	0.520
	restrOCP	0.790	0.291	0.540	0.789	0.292	0.541
	Prom	0.595	0.651	0.623	0.766	0.361	0.564
	OptFunc	0.396	0.383	0.389	0.686	0.320	0.503
	GS	0.667	1.000	0.833	0.813	0.492	0.652
0.2	OCP	0.138	0.512	0.325	0.593	0.349	0.471
	restrOCP	0.728	0.255	0.491	0.701	0.236	0.468
	Prom	0.576	0.622	0.599	0.681	0.309	0.495
	OptFunc	0.362	0.388	0.375	0.594	0.302	0.448
	GS	0.667	1.000	0.833	0.741	0.435	0.588
0.3	OCP	0.965	0.451	0.708	0.705	0.376	0.540
	restrOCP	0.493	0.247	0.370	0.601	0.220	0.410
	Prom	0.605	0.593	0.599	0.619	0.292	0.456
	OptFunc	0.822	0.403	0.612	0.710	0.325	0.518
	GS	0.492	1.000	0.746	0.544	0.408	0.476
0.35	OCP	0.762	0.425	0.594	0.749	0.400	0.574
	restrOCP	0.831	0.257	0.544	0.654	0.231	0.442
	Prom	0.918	0.595	0.756	0.667	0.299	0.483
	OptFunc	0.871	0.413	0.642	0.758	0.349	0.553
	GS	0.800	1.000	0.900	0.588	0.402	0.495
0.4	OCP	0.598	0.407	0.502	0.787	0.427	0.607
	restrOCP	0.949	0.278	0.613	0.701	0.249	0.475
	Prom	0.869	0.579	0.724	0.718	0.311	0.515
	OptFunc	0.679	0.421	0.550	0.800	0.376	0.588
	GS	1.000	1.000	1.000	0.631	0.402	0.516
0.5	OCP	0.433	0.386	0.410	0.850	0.503	0.676
	restrOCP	0.686	0.329	0.508	0.782	0.313	0.547
	Prom	0.635	0.594	0.614	0.795	0.362	0.579
	OptFunc	0.466	0.435	0.450	0.870	0.452	0.661
	GS	0.763	1.000	0.881	0.709	0.418	0.564
0.6	OCP	0.390	0.394	0.392	0.896	0.581	0.739
	restrOCP	0.572	0.371	0.471	0.841	0.388	0.614
	Prom	0.535	0.596	0.565	0.855	0.427	0.641
	OptFunc	0.387	0.449	0.418	0.922	0.538	0.730
	GS	0.632	1.000	0.816	0.766	0.456	0.611

Δ: true standardized effect; $e_N(\Delta)$: location component for conditional sample size; $v_N(\Delta)$: variation component for conditional sample size; $SN_\beta(\Delta)$: conditional sample size sub-score; $e_{CP}(\Delta)$: location component for conditional power; $v_{CP}(\Delta)$: variation component for conditional power; $SCP_\beta(\Delta)$: conditional power sub-score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach.

TABLE 3 Estimated average performance scores and related sub-scores for all investigated designs and $n_1 = n_2 = 50$, $n_{\max} = 200$

		Parameters to be Estimated					
Average Range	**Design**	$\overline{\text{ROS}}_{f_s,\beta}$	$\overline{\text{RUP}}_{f_p,\beta}$	$\text{AS}^{\text{Liu}}_{f_s,f_p,\beta}[\cdot]$	$\overline{\text{SN}}_{\beta}$	$\overline{\text{SCP}}_{\beta}$	$\text{AS}^{\text{New}}_{\beta}[\cdot]$
0.0-0.6	OCP	0.152	0.813	0.965[1]	0.455	0.589	0.522[4]
	restrOCP	0.085	1.215	1.300[5]	0.518	0.515	0.517[5]
	Prom	0.095	1.065	1.160[3]	0.642	0.547	0.595[2]
	OptFunc	0.139	0.894	1.033[2]	0.481	0.572	0.527[3]
	GS	0.066	1.187	1.253[4]	0.855	0.578	0.717[1]
0.0-0.2	OCP	0.000	2.514	2.514[1]	0.344	0.526	0.435[5]
	restrOCP	0.000	2.743	2.743[5]	0.545	0.545	0.545[3]
	Prom	0.000	2.683	2.683[3]	0.627	0.568	0.598[2]
	OptFunc	0.000	2.571	2.571[2]	0.393	0.509	0.451[4]
	GS	0.000	2.742	2.742[4]	0.833	0.654	0.744[1]
0.3-0.4	OCP	0.041	0.222	0.263[1]	0.601	0.574	0.588[3]
	restrOCP	0.000	1.007	1.007[5]	0.509	0.442	0.476[5]
	Prom	0.000	0.695	0.695[3]	0.693	0.485	0.589[2]
	OptFunc	0.019	0.371	0.390[2]	0.601	0.553	0.577[4]
	GS	0.000	0.941	0.941[4]	0.882	0.496	0.689[1]
0.5-0.6	OCP	0.470	0.000	0.470[5]	0.401	0.708	0.555[4]
	restrOCP	0.298	0.000	0.298[2]	0.490	0.581	0.536[5]
	Prom	0.323	0.000	0.333[3]	0.590	0.610	0.600[2]
	OptFunc	0.457	0.000	0.457[4]	0.434	0.696	0.565[3]
	GS	0.232	0.000	0.232[1]	0.849	0.588	0.719[1]

$\overline{\text{ROS}}_{f_s,\beta}$: average relative oversizing function; $\overline{\text{RUP}}_{f_p,\beta}$: average relative underpowering function; $\text{AS}^{\text{Liu}}_{f_s,f_p,\beta}$: average Liu score;
$\overline{\text{SN}}_{\beta}$: average conditional sample size sub-score; $\overline{\text{SCP}}_{\beta}$: average conditional power sub-score; $\text{AS}^{\text{New}}_{\beta}$: average new conditional score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach;
[·]: numbers in square brackets present the ranking of the group sequential designs according to the corresponding average performance score ([1]: best performance to [5]: worst performance).

to the fact that the recalculated sample sizes are very similar for $\Delta = 0.2$ and $\Delta = 0.3$ (Row 11, $\mathbb{E}[N^{\text{RA}}_{\text{total}}] = 179.358$; Row 16, $\mathbb{E}[N^{\text{RA}}_{\text{total}}] = 170.972$) but the target values are very different ($N^{\text{target}}_{0.2;\beta} = 50; N^{\text{target}}_{0.3;\beta} = 177$).

Considering an underlying standardized effect size of $\Delta = 0.35$, the ranking according to Liu's score for the sample size recalculation rules remains the same as for $\Delta = 0.1, 0.2$, and 0.3. All designs do not oversize according to Liu's definition as their expected sample sizes are not higher than the one of the fixed design. For $\Delta = 0.35$, the observed conditional power approach receives the optimal total score of 0 (Row 21, $S^{\text{Liu}}_{f_s,f_p,\beta}(0.35) = 0.000$) as the power is also higher than 0.8. Concerning the conditional score, the score values for the four sample size recalculation designs are all rather similar. However, the promising zone design is judged best (Row 23, $S^{\text{New}}_{\beta}(0.35) = 0.620$) due to an expected sample size close to the target sample size together with a small variance (Row 23, $\text{Var}\left(N^{\text{RA}}_{\text{total},\Delta}\right) = 924.542$).

For an underlying standardized effect of $\Delta = 0.4$, the rankings with respect to Liu's score and the new conditional score are the same for the sample size recalculation approaches. Both scores judge the promising zone approach as the best recalculation rule. Liu's score even judges that approach as optimal since both sub-scores take the value of 0.000. Hence, $S^{\text{Liu}}_{f_s,f_p,\beta}(0.4) = 0.000$ for the promising zone approach. This is again due to the fact that the expected sample size is smaller than the target sample size design and the power is higher than 0.8 such that no oversizing or underpowering

occurs. However, the variability of sample size and hence also of the conditional power are not taken into account. Note that the observed conditional power approach and the optimization function approach show some oversizing (Row 26, $\mathrm{ROS}_{f_s,\beta}(0.4) = 0.123$; Row 29, $\mathrm{ROS}_{f_s,\beta}(0.4) = 0.056$) and the restricted observed conditional power approach is the only sample size recalculation design that shows underpowering (Row 27, $\mathrm{RUP}_{f_p,\beta}(0.4) = 0.427$). With respect to the new conditional score, the promising zone design is also judged as best design (Row 28, $S_\beta^{\mathrm{New}}(0.4) = 0.619$) as the expected sample size conditional on entering the second stage meets the target value quite well and shows a rather small variability. However, in total the conditional performance scores of the four recalculation designs for $\Delta = 0.4$ lie all very close to each other.

For increasing effect sizes, the ranking according to the new conditional performance score remains stable with the promising zone design showing the best performance (Row 33, $S_\beta^{\mathrm{New}}(0.5) = 0.597$; Row 38, $S_\beta^{\mathrm{New}}(0.6) = 0.595$). However, all designs achieve very similar score values. Liu's ranking is nearly the same for underlying effect sizes of 0.5 and 0.6. In both scenarios, the restricted observed conditional power approach performs best and the promising zone approach second best regarding the sample size recalculation rules. Only the observed conditional power approach and the optimization function approach change their positions as the observed conditional power design returns sample sizes that are minimally closer to the target sample size for $\Delta = 0.6$ than the optimization function design (Row 36, $\mathbb{E}[N_{\mathrm{total},\Delta}] = 68.303$; Row 39, $\mathbb{E}[N_{\mathrm{total},\Delta}] = 68.376$) where it is the other way around for $\Delta = 0.5$ (Row 31, $\mathbb{E}[N_{\mathrm{total},\Delta}] = 86.306$; Row 34, $\mathbb{E}[N_{\mathrm{total},\Delta}] = 84.508$). For these higher effects, the power of the four recalculation designs is higher than 0.8 and the sample sizes are higher than the target sample size. Hence, Liu's score is only based on the relative oversizing function.

From Table 1 as well as Table 2, we see that a change in expected conditional power conditional on entering the recalculation area (Table 1, Column 11) translates to a similar change in the location component for the conditional power of the conditional performance score (Table 2, Column 6). A change in the location component of the sample size $e_N(\Delta)$ (Table 2, Column 3) corresponds approximately to a hundredfold change in expected sample size conditional on entering the recalculation area (Table 1, Column 8).

Table 3 presents the estimated average performance scores over all eight effect sizes as well as three effect ranges ($[0.0; 0.2]$, $[0.3; 0.4]$, $[0.5; 0.6]$) together with the related average sub-scores for all investigated simulation scenarios. Column 1 of Table 3 displays the considered effect range and Column 2 the underlying sample size recalculation design. Column 3 shows the relative oversizing function $\mathrm{ROS}_{f_s,\beta}(\Delta)$ averaged over the respective effects, where the average is denoted by $\overline{\mathrm{ROS}}_{f_s,\beta}$; Column 4 with $\overline{\mathrm{RUP}}_{f_p,\beta}$ is denoted analogously. Column 5 finally shows the total estimated averaged Liu performance score $\mathrm{AS}^{\mathrm{Liu}}_{f_s f_p,\beta}$. The new conditional sub-score assessing the sample size performance averaged over the considered Δ is denoted by $\overline{\mathrm{SN}}_\beta$ and reported in Column 6. The average new conditional sub-score assessing the power performance $\overline{\mathrm{SCP}}_\beta$ is given in Column 7. Finally, Column 8 shows the estimated averaged new conditional performance score $\mathrm{AS}^{\mathrm{New}}_\beta$. As before, numbers in square brackets after the performance scores represent the ranking of the designs according to the underlying performance score and we only describe the performance of the four sample size recalculation rules (observed conditional power, restricted observed conditional power, promising zone, and optimization function approach).

Table 3 shows the following results. Generally, we can observe that the estimated average performance scores for the sample size recalculation rules lie all very close to each other, both for the average Liu score and the average new conditional score. However, the rankings with respect to the two average scores are quite different. The average Liu score over all eight effect sizes rates the observed conditional power approach as best (Row 1, $\mathrm{AS}^{\mathrm{Liu}}_{f_s f_p,\beta} = 0.965$) and the restricted conditional power approach as worst (Row 2, $\mathrm{AS}^{\mathrm{Liu}}_{f_s f_p,\beta} = 1.300$). This is due to the fact that the average of the relative underpowering function is rather small for the observed conditional power design (Row 1, $\overline{\mathrm{RUP}}_{f_p,\beta} = 0.813$) and relatively large for the restricted conditional power approach (Row 2, $\overline{\mathrm{RUP}}_{f_p,\beta} = 1.215$). The relative oversizing values lie all very close to each other and are incorporated in the average Liu score with the same weight as the average values of the relative underpowering function. Consequently, they hardly affect the average Liu score.

When looking at the average values of the new conditional score, the promising zone approach is the best (Row 3, $\mathrm{AS}^{\mathrm{New}}_\beta = 0.595$) and the restricted observed conditional power approach is the worst (Row 2, $\mathrm{AS}^{\mathrm{New}}_\beta = 0.517$). The conditional power sub-scores are relatively similar across the designs whereas the sample size sub-scores show larger differences. The promising zone approach has the best sample size sub-score (Row 3, $\overline{\mathrm{SN}}_\beta = 0.642$) as its overall variability in calculating the sample size is rather low compared to the other designs. For the observed conditional power approach, the sample size sub-score $\overline{\mathrm{SN}}_\beta$ equals 0.455 (Row 1) as the expected sample size per group conditional on entering the recalculation area is often far away from meeting the target value.

Concerning the Liu score for the three effect ranges $[0.0; 0.2]$, $[0.3; 0.4]$, and $[0.5; 0.6]$, the ranking for the first two ranges is equivalent to the overall ranking, for the reasons specified above. The ranking of the third range, $[0.5; 0.6]$, is completely opposite with the restricted observed conditional power design showing the best performance (Row 17, $AS^{Liu}_{f_s f_p, \beta} = 0.298$) and the observed conditional power approach showing the worst performance (Row 16, $AS^{Liu}_{f_s f_p, \beta} = 0.470$). This is due to the fact that only the relative oversizing sub-scores determine the total scores and the recalculated sample sizes of the latter design are relatively large.

The rankings of the new conditional performance score are mostly different from the overall ranking. However, the promising zone approach remains the best performing recalculation design for all other three effect ranges $[0.0; 0.2]$, $[0.3; 0.4]$, and $[0.5; 0.6]$. The other three designs change positions. For $[0.0; 0.2]$, the observed conditional power approach performs worst (Row 6, $AS^{New}_{\beta} = 0.435$). For $[0.3; 0.4]$ and $[0.5; 0.6]$, the restricted observed conditional power approach performs worst (Row 12, $AS^{New}_{\beta} = 0.476$; Row 17, $AS^{New}_{\beta} = 0.536$). Especially for the effect range $[0.5; 0.6]$, the performance values of the recalculation rules are all very close to each other.

After having had a closer look at the performance of the main simulation setting ($n_1 = n_2 = 50, f = 4$), it is also interesting to compare the performance of the sample size recalculation rules with different initial and maximal sample size assumptions. Here, we only give an overview of the different rankings with respect to the conditional performance score. Detailed information can be found in Tables A1-A4 in the Appendix.

In the second simulation setting described above, the same absolute maximal sample size as in the main setting is considered ($n_{max} = 200$) but with $n_1 = n_2 = 25$. The ranking of this new simulation setting with respect to the new conditional performance score is very similar to the one of the main simulation setting for the different underlying effect sizes. However, for larger effect sizes ($\Delta = 0.5, \Delta = 0.6$), the values of S^{New}_{β} are higher in the second setting. This is due to the fact that the sample size sub-score returns better values for the second setting as the sample size is smaller on average and hence closer to the target values.

In the third simulation setting described in the simulation setup, the same relation as in the main setting between maximal and initial sample size, $f = 4$, is considered but with $n_1 = n_2 = 25$ (instead of $n_1 = n_2 = 50$). For effect sizes $\Delta = 0.0, 0.1$, and 0.2, the rankings of the main and the third simulation setting with respect to the conditional performance score are very similar with the promising zone approach as best performing recalculation design (cf Tables 1 and A3, Rows 3, 8, 13). For $\Delta = 0.3$, the rankings differ considerably. In the third setting, the ranking remains the same as for the smaller effect sizes but for the main setting it changes completely with the observed conditional power approach performing best (Table 1, Row 16, $S^{New}_{\beta}(0.3) = 0.624$) and restricted observed conditional power approach performing worst (Table 1, Row 17, $S^{New}_{\beta}(0.3) = 0.390$). The performance of the observed conditional power approach is due to an average sample size very close to the target sample size $N^{target}_{0.3;\beta} = 177$ in the main setting, whereas theoretically that sample size can never be reached in the third setting as the maximal sample size is bounded by 100 and thus $n_1 = 25$ is taken as target sample size for the third setting. For effect sizes of $\Delta = 0.35$, 0.4, 0.5, and 0.6, the rankings are again rather similar. Especially for $\Delta = 0.5$ and $\Delta = 0.6$, the score values of the four sample size recalculation rules are also very close to each other. For effect sizes Δ of 0.35 and 0.4, the sample size and conditional power sub-scores are always better in the main setting than in the third setting.

Briefly, the maximal sample size has a great influence on the fact if and how well the target sample size can be reached. Smaller sample sizes n_1 for the first stage favor small target sample sizes, which is reflected by the conditional performance score.

5.3 | A clinical trial example

The ChroPac multicenter, randomized, controlled trial[23,24] investigated the quality of life after surgical treatment for chronic pancreatitis patients. Two groups, one receiving a duodenum-preserving surgery and one with resection of the duodenum, were compared with regard to physical functioning. The endpoint was measured by a score ranging from 0 until 100 obtained by the EORTC QLQ-C30 questionnaire 9 months after surgery. High score values indicate a good performance. A score difference of 10 points was supposed to be clinically relevant and a standard deviation of 20 points was assumed from previous trials (cf References 23,24) resulting in an assumed standardized effect of $\Delta = 0.5$.

The original trial was not planned with the option to recalculate the sample size. For our illustrative example, we thus make further assumptions. We set the one-sided significance level to 2.5% and the desired power to 80%. The score values are supposed to be normally distributed with known common variance for both groups. With these assumptions

and $\Delta = 0.5$, $n^{\text{fix}}_{0.5,0.8} = 63$ patients are required per group according to Equation (21). After the inclusion of 32 patients ($= \lceil 0.5 \cdot n^{\text{fix}}_{0.5,0.8} \rceil$) per group, an interim analysis is performed with the possibility to recalculate the sample size. The maximal sample size per group is limited to $2 \cdot n^{\text{fix}}_{0.5,0.8} = 126$.

A standardized interim effect of $\hat{\Delta}_1 = 0.4$ is observed, hence, falling in the recalculation area. Now the question arises which recalculation rule is the best performing one. We evaluate this with regard to the new conditional performance score as well as the overall power and total sample size. For the inverse normal combination test, we take weights $w_1 = w_2 = \sqrt{32}$. We performed a simulation study where 10 000 observations were drawn from a normal distribution with an expectation equal to $0.4 \cdot \sqrt{32/2}$ and a standard deviation of 1. Note that the boundaries for the interpretation of the conditional score remain the same as in Section 4.9.

According to the corresponding simulation results presented in Table 4, the observed conditional power approach and the optimization function approach are proposed by the new conditional score (Row 1, $S^{\text{New}}_{\beta}(0.4) = 0.619$; Row 4: $S^{\text{New}}_{\beta}(0.4) = 0.617$) next to the classical group sequential approach (Row 5, $S^{\text{New}}_{\beta}(0.4) = 0.646$). This goes in line with an overall power close to 80% (Row 1, $\text{Pow}_{0.4} = 0.753$; Row 4, $\text{Pow}_{0.4} = 0.729$) for the observed conditional power and optimization function approach but not for the group sequential approach (Row 5, $\text{Pow}_{0.4} = 0.565$). The higher power of the observed conditional power and optimization function approach is obtained at the cost of higher total sample sizes (Row 1, $\mathbb{E}[N_{\text{total};0.4}] \approx 82$; Row 4, $\mathbb{E}[N_{\text{total};0.4}] \approx 78$) than for the other three designs with a smaller overall power. However, all the sample sizes are much smaller than the maximally allowed sample size $n_{\max} = 126$. All three designs show a high performance with respect to the boundaries of the score interpretation from Section 4.9. The observed conditional power approach and the optimization function approach perform also reasonably well (mostly with a medium performance) if the true effect size Δ is larger than 0.4 (Row 6, $S^{\text{New}}_{\beta}(0.45) = 0.578$; Row 9, $S^{\text{New}}_{\beta}(0.45) = 0.577$; Row 11, $S^{\text{New}}_{\beta}(0.5) = 0.556$; Row 14, $S^{\text{New}}_{\beta}(0.5) = 0.556$; Row 16, $S^{\text{New}}_{\beta}(0.55) = 0.544$; Row 19, $S^{\text{New}}_{\beta}(0.55) = 0.543$; Row 21, $S^{\text{New}}_{\beta}(0.6) = 0.541$; Row 24, $S^{\text{New}}_{\beta}(0.6) = 0.540$). The classical group sequential design is the only design that shows a high performance for all considered standardized treatment effect sizes. However, it does only reach a global power of 0.8 for effect sizes $\Delta = 0.5$ and $\Delta = 0.6$.

As an effect size of $\Delta = 0.5$ was initially defined to be clinically relevant, the simulation results for smaller effect sizes than 0.4 should not guide the choice of the recalculation rule. Hence, these are not listed in Table 4. While considering a high power and small sample size as equally important, one might however decide for the optimization function approach as it shows a high or medium performance and almost the same performance as for the observed conditional power approach can be achieved at the cost of fewer patients for standardized effect sizes of 0.4, 0.45, and 0.5. Moreover, the design performs also reasonably good in comparison to the others for larger effect sizes than $\Delta_1 = 0.4$. The classical group sequential design is another option since it shows a high conditional performance for all considered effect sizes. However, an overall power of 0.8 is only attained for $\Delta = 0.5$ and $\Delta = 0.6$.

6 | DISCUSSION

Adaptive group-sequential designs generally allow adapting the sample size during the ongoing trial and thereby account for planning uncertainties. So far, there exist a number of different performance measures to assess the quality of such designs. However, existing performance measures ignore important features of performance (eg, the variability of the recalculated sample size) and there exist no unique standards to assess the global performance within a related score. In this work, we contributed to overcome this shortcoming by presenting a new conditional performance score for the evaluation of adaptive designs, where the term "conditional" refers to the condition of already knowing the interim result. We present different unconditional and conditional performance measures and discuss their intuitive target values. In an adaptive setting, global performance measures and conditional performance measures are two important perspectives which should both be assessed. As conditional performance measures have a very natural interpretation in adaptive designs and their target values are easier to define, we introduced a conditional performance score. As a new aspect, our new score does not only include measures of location but also of variation. The new score is based on a sub-score assessing the sample size and a sub-score assessing the power which can both be interpreted separately. Moreover, the two sub-scores can be split into a location and variation component. We applied the performance score proposed by Liu et al[13] as well as the new score to four different sample size recalculation approaches that were proposed in the literature.

Liu's score is a global, unconditional score that focuses on penalizing high sample sizes as well as low power values compared to the fixed sample size design. This concept nicely adapts the idea of calculating the smallest sample size that

TABLE 4 Estimated point-wise performance scores and related conditional score components as well as unconditional performance measures for all investigated simulation scenarios for the ChroPac trial ($n_1 = n_2 = 32$, $n_{max} = 126$)

		Parameters to be Estimated								
Δ	**Design**	$\mathbb{E}[N_{total,\Delta}]$	Pow_Δ	$e_N(\Delta)$	$v_N(\Delta)$	$SN_\beta(\Delta)$	$e_{CP}(\Delta)$	$v_{CP}(\Delta))$	$SCP_\beta(\Delta)$	$S_\beta^{New}(\Delta)[\cdot]$
0.4	OCP	81.337	0.753	0.926	0.447	0.687	0.719	0.383	0.551	0.619[2]
	restrOCP	53.316	0.558	0.621	0.249	0.435	0.616	0.221	0.419	0.427[5]
	Prom	60.211	0.615	0.732	0.594	0.663	0.637	0.295	0.466	0.564[4]
	OptFunc	77.849	0.729	0.983	0.426	0.704	0.749	0.309	0.529	0.617[3]
	GS	53.056	0.565	0.616	1.000	0.808	0.561	0.405	0.483	0.646[1]
0.45	OCP	75.955	0.843	0.727	0.427	0.577	0.753	0.403	0.578	0.578[3]
	restrOCP	53.291	0.657	0.876	0.254	0.565	0.661	0.234	0.448	0.506[5]
	Prom	58.377	0.717	0.965	0.589	0.777	0.676	0.302	0.489	0.633[2]
	OptFunc	74.342	0.820	0.755	0.436	0.596	0.788	0.328	0.558	0.577[4]
	GS	51.424	0.670	0.843	1.000	0.921	0.596	0.402	0.499	0.710[1]
0.5	OCP	70.014	0.905	0.603	0.412	0.508	0.784	0.425	0.605	0.556[3]
	restrOCP	51.872	0.739	0.957	0.273	0.615	0.698	0.248	0.473	0.544[5]
	Prom	56.270	0.805	0.871	0.576	0.724	0.717	0.312	0.515	0.619[2]
	OptFunc	69.877	0.763	0.606	0.442	0.524	0.824	0.351	0.588	0.556[3]
	GS	49.459	0.670	0.996	1.000	0.998	0.631	0.402	0.516	0.757[1]
0.55	OCP	63.815	0.944	0.512	0.401	0.457	0.811	0.452	0.631	0.544[3]
	restrOCP	49.921	0.808	0.823	0.296	0.559	0.733	0.272	0.503	0.531[5]
	Prom	52.982	0.870	0.754	0.593	0.674	0.748	0.328	0.538	0.606[2]
	OptFunc	64.615	0.930	0.495	0.447	0.471	0.855	0.377	0.616	0.543[4]
	GS	47.245	0.840	0.882	1.000	0.941	0.663	0.405	0.534	0.738[1]
0.6	OCP	57.723	0.969	0.451	0.395	0.423	0.836	0.484	0.660	0.541[3]
	restrOCP	47.374	0.862	0.727	0.314	0.520	0.763	0.295	0.529	0.525[5]
	Prom	49.637	0.919	0.666	0.592	0.629	0.778	0.351	0.564	0.597[2]
	OptFunc	59.106	0.960	0.414	0.454	0.434	0.884	0.410	0.647	0.540[4]
	GS	44.781	0.889	0.796	1.000	0.898	0.692	0.415	0.554	0.726[1]

Δ: true standardized effect; $\mathbb{E}[N_{total,\Delta}]$: expected sample size per group for both stages; Pow_Δ: power of the adaptive design for standardized treatment effect Δ;
$e_N(\Delta)$: location component for conditional sample size; $v_N(\Delta)$: variation component for conditional sample size;
$e_{CP}(\Delta)$: location component for conditional power; $v_{CP}(\Delta)$: variation component for conditional power;
$SN_\beta(\Delta)$: conditional sample size sub-score; $SCP_\beta(\Delta)$: conditional power sub-score; $S_\beta^{New}(\Delta)$: point-wise new conditional score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach;
[·]: numbers in square brackets present the ranking of the group sequential designs according to the corresponding point-wise performance score ([1]: best performance to [5]: worst performance).

ensures a certain power value yet it does not directly consider "overpowering" or "undersizing." While in a fixed design, an overpowered trial is naturally oversized and vice versa an underpowered trial is undersized, this cannot directly be transferred to adaptive designs where there no longer exist a unique sample size but average sample sizes are considered. A potential criticism of Liu's score is that it is highly questionable whether the "perfect" fixed sample size design is really a valid reference. In fact, the expected sample size of an adaptive design is not directly comparable to the required sample size of a fixed design. Moreover, under the null hypothesis, there is no reference sample size of the fixed design and therefore Liu's score is not well-defined under the null hypothesis. However, a nice feature of Liu's score is that the relative underpowering function and the relative oversizing function can be interpreted as sub-scores and may be

evaluated separately. Both sub-scores are measured on the same scale which is achieved by a transformation to a "sample size scale", which is appealing for comparative investigations. Nevertheless, this sample size scale can also be questioned. The power sub-score does not directly compare the achieved power with the power of the fixed design but the power values are transformed to a sample size scale. This transformation is clearly nonlinear and therefore expectations are not maintained. The score values range from 0 to ∞ and it needs to be discussed which multiple of oversizing or underpowering can be regarded as decent performance of a design. Moreover, one has to keep in mind that a "perfect" performance with a Liu score value of 0.000 (eg, as in the promising zone approach for $\Delta = 0.4$ in the main setting) does not fit the intuitive view of "optimal performance" as Liu's score does not consider measures of variability.

The above considerations led us to the definition of the new conditional score that also incorporates the idea of two sub-scores relating to sample size and power. The concept of adequate target values is easier in this setting. An important advantage is that our new score (and both sub-scores) incorporate parameters penalizing a high variability in sample size and conditional power. Moreover, the new score has the potential benefit that it ranges from 0 to 1 such that absolute values for "worst" and "best" performance are defined. This concept applies also to the two sub-scores that are measured on the same scale. The four location and variation parameters ($e_N(\Delta)$, $v_N(\Delta)$, $e_{CP}(\Delta)$, $v_{CP}(\Delta)$) allow to specify the performance of an approach in relation to a design with respect to location and variation of conditional power and sample size, respectively, and range also from 0 to 1. Even though the performance score was primarily developed to compare different recalculation rules, one might also be interested in the influence of other design parameters (like different n_1 or n_{max}) on a certain recalculation rule. Therefore, we gave guidance on how to determine score values indicating good or bad performance. However, the judgment of the performance of a certain design as high, medium, or low, or the interpretation of score differences depend importantly on the chosen target values. Note that the variance of the score depends on the underlying effect size. For example, small differences of the observed interim effect do not make a big difference for interim effects close to $\hat{\Delta}_1 = 0.1$ but they do for observed interim effects close to $\hat{\Delta}_1 = 0.3$ (cf Figure 1). The reason for the latter observation is that a slightly different observed effect size leads to a very different recalculated sample size. As a potential limitation but also as potential room for extending our new score, it should be noted that target values for the conditional power and the sample size are not necessarily "unique". In the definition of the target sample size in Equation (25), one could for instance also apply n_{max} instead of n_1 whenever $n^{fix}_{\Delta,1-\beta} > n_{max}$. However, our score can also easily be applied when the target values are modified. Generally, it must be noted that any combination of different performance measures within a single performance score always corresponds to some arbitrariness. For example, we combine the two sub-scores with equal weights in the final conditional performance score (which favors classical group sequential designs). A different weighting can easily be implemented by means of a weighted average of the conditional power and sample size sub-score. Generally, the sub-scores for sample size and power must not necessarily be combined by a simple linear combination.

The two performance scores were applied to four different sample size recalculation designs and a classical group sequential study design. With respect to certain effect sizes and overall performance, the designs are performing differently in terms of conditional and unconditional measures (cf Tables 1-3 and A1-A4). The performance score values also depend on the interim and maximal sample size. Both performance scores do not have the same ranking for all values of Δ. Therefore, comparing unweighted average performance scores over a wide range for Δ is not recommended. We suggest to investigate the point-wise scores separately or only for small effect ranges. Moreover, the score ranking with respect to Liu's score and the new conditional performance score are very different. This is due to the fact that the new score penalizes large variabilities in sample size and conditional power. Another point is that the target values for conditional power and sample size in the new score only suggest a sample size increase if this seems "worth the effort" with respect to the maximally allowed sample size. However, note that most investigated sample size recalculation approaches rely on specific parameter assumptions, which we did not try to optimize. Especially for the optimization function approach, we did not search for the optimal parameter γ and therefore the design might behave differently under other parameters. Similarly, the predefined minimal conditional power $1 - \beta_0$ in the restricted observed conditional power approach can be chosen differently. Thus, the performance rankings should not be over-interpreted. Moreover, note that the conditional perspective of our score is motivated by an interim look. For fixed sample size designs, we recommend the global evaluation perspective.

We generally propose the following performance assessment when planning an adaptive trial:

1. Investigate global unconditional performance measures for sample size and power or a related global performance score under various underlying effect scenarios.
2. Determine the probability to enter the recalculation area (given by the interim critical value and the futility bound) under various underlying effect scenarios.

3. In particular for those effect sizes, where it is likely to reach the recalculation area: Investigate conditional performance measures and our new conditional performance score.

The choice of other local significance levels (eg, the design by O'Brien and Fleming[18]) changes the performance of the sample size recalculation rules and is therefore also interesting for future research. Another task for future work is the development of optimized sample size recalculation rules based on the new conditional performance score. An R package implementing the new score and related performance measures is currently created. The simulation programs underlying this work are added as supplemental material to this paper.

ACKNOWLEDGEMENT

This work was supported by the German Research Foundation (grants RA 2347/4-1 and KI 708/4-1).

DATA AVAILABILITY

Original data were not analyzed. The R code that was used for producing simulated data and analyzing them is available as supplemental material.

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. The International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. Statistical Principles for Clinical Trials-E9. ICH. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Published 1998. Accessed December 17, 2019.
2. Müller HH, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*. 2001;57(3):886-891.
3. Cui L, Hung HM, Wang S. Modification of sample size in group sequential clinical trials. *Biometrics*. 1999;55(3):853-857.
4. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics*. 1994;50(4):1029-1041.
5. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics*. 1999;55(4):1286-1290.
6. Dmitrienko A, Wang M. Bayesian predictive approach to interim monitoring in clinical trials. *Stat Med*. 2006;25(13):2178-2195.
7. Mehta C, Pocock SJ. Adaptive increase in sample size when the results are promising: a practical guide with examples. *Stat Med*. 2011;30(28):3267-3284.
8. 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.
9. Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: conditional or predictive power? *Control Clin Trials*. 1986;7(1):8-17.
10. Jennison C, Turnbull BW. Adaptive sample size modification in clinical trials: start small then ask for more? *Stat Med*. 2015;34(29):3793-3810.
11. 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.
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. Liu GF, Zhu GR, Cui L. Evaluating the adaptive performance of flexible sample size designs with treatment difference in an interval. *Stat Med*. 2008;27(4):584-596.
14. Placzek M, Friede T. Clinical trials with nested subgroups: analysis, sample size determination and internal pilot studies. *Stat Methods Med Res*. 2018;27(11):3286-3303.
15. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. *Stat Med*. 1990;9(12):1439-1445.
16. Kim K, Demets DL. Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika*. 1987;74(1):149-154.
17. Lan KG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika*. 1983;70(3):659-663.
18. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35(3):549-556.
19. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64(2):191-199.
20. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Berlin, Heidelberg / Germany: Springer; 2016.

21. Wassmer G. Planning and analyzing adaptive group sequential survival trials. *Biom J*. 2006;48(4):714-729.
22. R development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.
23. Diener MK et al. ChroPac-trial: duodenum-preserving pancreatic head resection versus pancreatoduodenectomy for chronic pancreatitis Trial protocol of a randomised controlled multicentre trial. *Trials*. 2010;11(1):47.
24. Diener MK et al. Partial pancreatoduodenectomy versus duodenum-preserving pancreatic head resection in chronic pancreatitis: the multicentre, randomized, controlled, double-blind ChroPac trial. *Lancet*. 2017;390(10099):1027-1037.

How to cite this article: 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. *Statistics in Medicine*. 2020;39:2067–2100. https://doi.org/10.1002/sim.8534

Appendix A

TABLE A1 Estimated point-wise performance scores and related conditional and unconditional performance measures for all investigated simulation scenarios with $n_1 = n_2 = 25$, $n_{max} = 200$

		Parameters to be Estimated											
Δ ($n^{fix}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{total,\Delta}]$	$ROS_{f_s,\beta}(\Delta)$	Pow_Δ	$RUP_{f_p,\beta}(\Delta)$	$S^{Liu}_{f_s f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$	$Var\left(N^{RA}_{total,\Delta}\right)$	$SN_\beta(\Delta)$	$\mathbb{E}\left[CP^{RA}_\Delta\right]$	$Var\left(CP^{RA}_\Delta\right)$	$SCP_\beta(\Delta)$	$S^{New}_\beta(\Delta)[\cdot]$
0.0	OCP	98.627	–	0.025	–	–	175.782	2157.566	0.304	0.364	0.111	0.493	0.398[5]
(–)	restrOCP	43.064	–	0.023	–	–	61.995	3611.718	0.551	0.272	0.138	0.502	0.527[3]
	Prom	39.082	–	0.025	–	–	53.838	141.638	0.850	0.179	0.077	0.644	0.747[2]
	OptFunc	67.678	–	0.025	–	–	112.400	4623.551	0.362	0.372	0.147	0.440	0.401[4]
	GS	37.208	–	0.025	–	–	50.000	0.000	0.929	0.147	0.046	0.722	0.825[1]
0.1	OCP	111.623	0.000	0.092	3.011	3.011[1]	167.472	2734.251	0.294	0.421	0.115	0.458	0.376[5]
(1571)	restrOCP	50.548	0.000	0.075	3.064	3.064[3]	67.020	3687.930	0.533	0.333	0.150	0.455	0.494[3]
	Prom	43.271	0.000	0.069	3.080	3.080[4]	55.051	178.473	0.838	0.229	0.095	0.587	0.712[2]
	OptFunc	80.359	0.000	0.091	3.014	3.014[2]	116.050	4433.345	0.359	0.438	0.155	0.395	0.377[4]
	GS	40.200	0.000	0.066	3.090	3.090[5]	50.000	0.000	0.929	0.187	0.058	0.676	0.802[1]
0.2	OCP	117.346	0.000	0.279	2.411	2.411[1]	157.948	3284.462	0.293	0.478	0.112	0.434	0.363[4]
(395)	restrOCP	57.146	0.000	0.197	2.676	2.676[3]	71.280	3650.945	0.523	0.395	0.154	0.417	0.470[3]
	Prom	46.730	0.000	0.164	2.785	2.785[4]	56.285	211.108	0.828	0.286	0.111	0.532	0.680[2]
	OptFunc	90.343	0.000	0.262	2.465	2.465[2]	119.073	4227.353	0.360	0.505	0.154	0.362	0.361[5]
	GS	42.365	0.000	0.148	2.834	2.834[5]	50.000	0.000	0.929	0.234	0.071	0.627	0.778[1]
0.3	OCP	113.526	0.000	0.562	1.365	1.365[1]	147.307	3721.682	0.569	0.540	0.102	0.548	0.558[1]
(177)	restrOCP	61.988	0.000	0.389	2.035	2.035[3]	76.103	3608.416	0.371	0.465	0.150	0.442	0.406[5]
	Prom	48.262	0.000	0.317	2.282	2.282[4]	57.139	224.455	0.574	0.347	0.121	0.419	0.497[3]
	OptFunc	94.530	0.000	0.513	1.567	1.567[2]	121.062	3991.146	0.481	0.578	0.144	0.507	0.494[4]
	GS	43.095	0.000	0.284	2.392	2.392[5]	50.000	0.000	0.639	0.286	0.081	0.452	0.546[2]
0.35	OCP	108.477	0.000	0.687	0.750	0.750[1]	141.491	3862.307	0.612	0.570	0.094	0.575	0.594[3]
(131)	restrOCP	62.155	0.000	0.487	1.671	1.671[3]	76.849	3431.734	0.513	0.497	0.145	0.463	0.488[5]
	Prom	48.407	0.000	0.412	1.951	1.951[4]	57.664	234.554	0.706	0.380	0.125	0.432	0.596[2]

(Continues)

TABLE A1 (Continued)

		Parameters to be Estimated											
Δ ($n^{\text{fix}}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{\text{total},\Delta}]$	$\text{ROS}_{f_s,\beta}(\Delta)$	Pow_Δ	$\text{RUP}_{f_p,\beta}(\Delta)$	$S^{\text{Liu}}_{f_s,f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{\text{RA}}_{\text{total},\Delta}\right]$	$\text{Var}\left(N^{\text{RA}}_{\text{total},\Delta}\right)$	$\text{SN}_\beta(\Delta)$	$\mathbb{E}\left[\text{CP}^{\text{RA}}_\Delta\right]$	$\text{Var}\left(\text{CP}^{\text{RA}}_\Delta\right)$	$\text{SCP}_\beta(\Delta)$	$S^{\text{New}}_\beta(\Delta)[\cdot]$
	OptFunc	94.707	0.000	0.633	1.033	1.033[2]	122.274	3870.848	0.622	0.615	0.134	0.539	0.580[4]
	GS	42.915	0.000	0.371	2.099	2.099[5]	50.000	0.000	0.771	0.315	0.085	0.460	0.616[1]
0.4	OCP	101.789	0.037	0.783	0.134	0.172[1]	135.551	3962.147	0.539	0.600	0.087	0.603	0.571[4]
(101)	restrOCP	62.239	0.000	0.584	1.267	1.267[3]	78.612	3345.005	0.608	0.532	0.137	0.492	0.550[5]
	Prom	48.196	0.000	0.513	1.566	1.566[4]	58.395	252.683	0.790	0.415	0.126	0.447	0.619[2]
	OptFunc	91.937	0.000	0.728	0.507	0.507[2]	121.368	3716.994	0.591	0.650	0.125	0.570	0.580[3]
	GS	42.365	0.000	0.465	1.756	1.756[5]	50.000	0.000	0.857	0.344	0.087	0.470	0.664[1]
0.5	OCP	85.529	0.362	0.895	0.000	0.362[2]	123.469	3937.376	0.473	0.647	0.072	0.654	0.563[5]
(65)	restrOCP	58.561	0.000	0.742	0.419	0.419[3]	79.598	2974.982	0.645	0.593	0.118	0.550	0.597[3]
	Prom	46.086	0.000	0.702	0.665	0.665[4]	59.304	261.932	0.893	0.480	0.123	0.485	0.689[2]
	OptFunc	82.629	0.316	0.860	0.000	0.316[1]	118.751	3429.464	0.510	0.710	0.105	0.630	0.570[4]
	GS	40.367	0.000	0.654	0.929	0.929[5]	50.000	0.000	0.959	0.400	0.089	0.497	0.728[1]
0.6	OCP	68.774	0.577	0.949	0.000	0.577[4]	112.027	3814.709	0.455	0.684	0.057	0.701	0.578[5]
(45)	restrOCP	51.817	0.188	0.852	0.000	0.188[3]	78.314	2646.007	0.610	0.640	0.098	0.604	0.607[3]
	Prom	42.476	0.000	0.846	0.000	0.000[1]	59.743	253.384	0.866	0.541	0.116	0.527	0.697[2]
	OptFunc	69.476	0.593	0.933	0.000	0.593[5]	113.421	3171.296	0.482	0.756	0.085	0.685	0.584[4]
	GS	37.575	0.000	0.811	0.000	0.000[1]	50.000	0.000	0.985	0.457	0.089	0.526	0.756[1]

Δ: true standardized effect; $n^{\text{fix}}_{\Delta,0.8}$: optimal sample size in the fixed design; $\mathbb{E}[N_{\text{total},\Delta}]$: expected sample size per group for both stages; Pow_Δ: power of the adaptive design for standardized treatment effect Δ;
$\mathbb{E}\left[N^{\text{RA}}_{\text{total},\Delta}\right]$: expected sample size per group conditional on entering the recalculation area; $\text{Var}\left(N^{\text{RA}}_{\text{total},\Delta}\right)$: variance of sample size per group conditional on entering the recalculation area;
$\mathbb{E}\left[\text{CP}^{\text{RA}}_\Delta\right]$: expected conditional power conditional on entering the recalculation area; $\text{Var}(\text{CP}^{\text{RA}}_\Delta)$: variance of conditional power conditional on entering the recalculation area;
$\text{ROS}_{f_s,\beta}(\Delta)$: relative oversizing function; $\text{RUP}_{f_p,\beta}(\Delta)$: relative underpowering function; $S^{\text{Liu}}_{f_s,f_p,\beta}(\Delta)$: point-wise Liu score;
$\text{SN}_\beta(\Delta)$: conditional sample size sub-score; $\text{SCP}_\beta(\Delta)$: conditional power sub-score; $S^{\text{New}}_\beta(\Delta)$: point-wise new conditional score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach;
[·]: numbers in square brackets present the ranking of the group sequential designs according to the corresponding point-wise performance score ([1]: best performance to [5]: worst performance).

TABLE A2 Estimated performance sub-scores and related location and variation components for all investigated designs and $n_1 = n_2 = 25$, $n_{max} = 200$

		Parameters to be Estimated					
Δ	**Design**	$e_N(\Delta)$	$v_N(\Delta)$	$SN_\beta(\Delta)$	$e_{CP}(\Delta)$	$v_{CP}(\Delta)$	$SCP_\beta(\Delta)$
0.0	OCP	0.138	0.469	0.304	0.652	0.334	0.493
	restrOCP	0.789	0.313	0.551	0.747	0.257	0.502
	Prom	0.835	0.864	0.850	0.842	0.446	0.644
	OptFunc	0.501	0.223	0.362	0.645	0.234	0.440
	GS	0.857	1.000	0.929	0.875	0.570	0.722
0.1	OCP	0.186	0.402	0.294	0.594	0.323	0.458
	restrOCP	0.760	0.306	0.533	0.684	0.225	0.455
	Prom	0.828	0.847	0.838	0.791	0.384	0.587
	OptFunc	0.480	0.239	0.359	0.577	0.214	0.395
	GS	0.857	1.000	0.929	0.834	0.518	0.676
0.2	OCP	0.240	0.345	0.293	0.535	0.332	0.434
	restrOCP	0.736	0.309	0.523	0.620	0.214	0.417
	Prom	0.821	0.834	0.828	0.732	0.332	0.532
	OptFunc	0.462	0.257	0.360	0.508	0.216	0.362
	GS	0.857	1.000	0.929	0.786	0.468	0.627
0.3	OCP	0.835	0.303	0.569	0.733	0.362	0.548
	restrOCP	0.428	0.314	0.371	0.657	0.226	0.442
	Prom	0.320	0.829	0.574	0.535	0.303	0.419
	OptFunc	0.685	0.278	0.481	0.772	0.242	0.507
	GS	0.279	1.000	0.639	0.473	0.431	0.452
0.35	OCP	0.935	0.290	0.612	0.765	0.385	0.575
	restrOCP	0.696	0.331	0.513	0.689	0.237	0.463
	Prom	0.586	0.825	0.706	0.569	0.294	0.432
	OptFunc	0.955	0.289	0.622	0.810	0.267	0.539
	GS	0.542	1.000	0.771	0.502	0.417	0.460
0.4	OCP	0.797	0.281	0.539	0.795	0.411	0.603
	restrOCP	0.877	0.339	0.608	0.725	0.259	0.492
	Prom	0.762	0.818	0.790	0.605	0.289	0.447
	OptFunc	0.878	0.303	0.591	0.846	0.294	0.570
	GS	0.714	1.000	0.857	0.532	0.409	0.470
0.5	OCP	0.662	0.283	0.473	0.843	0.465	0.654
	restrOCP	0.913	0.377	0.645	0.787	0.312	0.550
	Prom	0.971	0.815	0.893	0.671	0.298	0.485
	OptFunc	0.689	0.331	0.510	0.907	0.352	0.630
	GS	0.918	1.000	0.959	0.590	0.403	0.497
0.6	OCP	0.616	0.294	0.455	0.881	0.521	0.701
	restrOCP	0.809	0.412	0.610	0.836	0.372	0.604
	Prom	0.915	0.818	0.866	0.735	0.319	0.527
	OptFunc	0.608	0.356	0.482	0.955	0.415	0.685
	GS	0.970	1.000	0.985	0.649	0.404	0.526

Δ: true standardized effect; $e_N(\Delta)$: location component for conditional sample size; $v_N(\Delta)$: variation component for conditional sample size; $SN_\beta(\Delta)$: conditional sample size sub-score; $e_{CP}(\Delta)$: location component for conditional power; $v_{CP}(\Delta)$: variation component for conditional power; $SCP_\beta(\Delta)$: conditional power sub-score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach.

TABLE A3 Estimated point-wise performance scores and related conditional and unconditional performance measures for all investigated simulation scenarios with $n_1 = n_2 = 25$, $n_{max} = 100$

		Parameters to be Estimated											
Δ ($n^{fix}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{total,\Delta}]$	$ROS_{f_s,\beta}(\Delta)$	Pow_Δ	$RUP_{f_p,\beta}(\Delta)$	$S^{Liu}_{f_s f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$	$Var\left(N^{RA}_{total,\Delta}\right)$	$SN_\beta(\Delta)$	$\mathbb{E}\left[CP^{RA}_\Delta\right]$	$Var\left(CP^{RA}_\Delta\right)$	$SCP_\beta(\Delta)$	$S^{New}_\beta(\Delta)[\cdot]$
0.0	OCP	59.698	–	0.025	–	–	96.059	143.781	0.366	0.257	0.087	0.587	0.477[4]
(–)	restrOCP	30.471	–	0.022	–	–	36.204	578.597	0.605	0.154	0.095	0.626	0.615[3]
	Prom	39.012	–	0.025	–	–	53.696	127.218	0.658	0.178	0.076	0.646	0.652[2]
	OptFunc	48.929	–	0.025	–	–	74.004	595.033	0.348	0.263	0.103	0.557	0.453[5]
	GS	37.208	–	0.025	–	–	50.000	0.000	0.833	0.147	0.046	0.722	0.778[1]
0.1	OCP	67.136	0.000	0.078	3.055	3.055[2]	94.302	199.470	0.350	0.312	0.098	0.539	0.444[4]
(1571)	restrOCP	33.843	0.000	0.061	3.103	3.103[5]	39.545	683.960	0.554	0.207	0.117	0.565	0.560[3]
	Prom	43.148	0.000	0.069	3.080	3.080[3]	54.848	158.371	0.633	0.228	0.094	0.589	0.611[2]
	OptFunc	56.241	0.000	0.078	3.054	3.054[1]	76.383	576.661	0.337	0.323	0.119	0.502	0.420[5]
	GS	40.200	0.000	0.066	3.090	3.090[4]	50.000	0.000	0.833	0.187	0.058	0.676	0.754[1]
0.2	OCP	71.543	0.000	0.202	2.662	2.662[1]	92.007	270.240	0.334	0.370	0.105	0.499	0.416[4]
(395)	restrOCP	37.208	0.000	0.147	2.839	2.839[5]	42.576	743.982	0.519	0.263	0.135	0.510	0.515[3]
	Prom	46.561	0.000	0.163	2.786	2.786[3]	56.042	187.658	0.610	0.285	0.110	0.535	0.572[2]
	OptFunc	62.259	0.000	0.202	2.662	2.662[1]	78.642	546.963	0.331	0.388	0.130	0.453	0.392[5]
	GS	42.365	0.000	0.148	2.834	2.834[4]	50.000	0.000	0.833	0.234	0.071	0.627	0.730[1]
0.3	OCP	71.488	0.000	0.411	1.958	1.958[1]	89.228	345.692	0.324	0.433	0.105	0.466	0.395[4]
(177)	restrOCP	40.063	0.000	0.288	2.380	2.380[4]	45.811	781.616	0.489	0.326	0.147	0.462	0.475[3]
	Prom	48.087	0.000	0.317	2.284	2.284[3]	56.897	201.521	0.598	0.346	0.120	0.489	0.543[2]
	OptFunc	65.315	0.000	0.405	1.976	1.976[2]	80.699	504.956	0.329	0.458	0.133	0.413	0.371[5]
	GS	43.095	0.000	0.284	2.392	2.392[5]	50.000	0.000	0.833	0.286	0.081	0.581	0.707[1]
0.35	OCP	69.929	0.000	0.532	1.490	1.490[1]	87.698	381.997	0.321	0.465	0.103	0.453	0.387[4]

(Continues)

TABLE A3 (Continued)

		Parameters to be Estimated											
Δ ($n^{fix}_{\Delta,0.8}$)	Design	$\mathbb{E}[N_{total,\Delta}]$	$ROS_{f_s,\beta}(\Delta)$	Pow_Δ	$RUP_{f_p,\beta}(\Delta)$	$S^{Liu}_{f_s f_p,\beta}(\Delta)[\cdot]$	$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$	$Var\left(N^{RA}_{total,\Delta}\right)$	$SN_\beta(\Delta)$	$\mathbb{E}\left[CP^{RA}_\Delta\right]$	$Var\left(CP^{RA}_\Delta\right)$	$SCP_\beta(\Delta)$	$S^{New}_\beta(\Delta)[\cdot]$
(131)	restrOCP	41.062	0.000	0.374	2.086	2.086[4]	47.414	789.390	0.476	0.359	0.151	0.440	0.458[3]
	Prom	48.224	0.000	0.411	1.955	1.955[3]	57.408	210.633	0.590	0.378	0.124	0.467	0.529[2]
	OptFunc	65.714	0.000	0.525	1.520	1.520[2]	81.815	480.959	0.329	0.495	0.131	0.397	0.363[5]
	GS	42.915	0.000	0.371	2.099	2.099[5]	50.000	0.000	0.833	0.315	0.085	0.560	0.697[1]
0.4	OCP	67.450	0.000	0.650	0.951	0.951[1]	86.114	414.734	0.321	0.498	0.099	0.442	0.382[4]
(101)	restrOCP	42.032	0.000	0.469	1.741	1.741[4]	49.521	803.971	0.458	0.398	0.152	0.419	0.439[3]
	Prom	47.984	0.000	0.512	1.570	1.570[3]	58.090	224.414	0.580	0.413	0.125	0.447	0.513[2]
	OptFunc	64.857	0.000	0.639	1.004	1.004[2]	82.381	461.120	0.331	0.532	0.128	0.383	0.357[5]
	GS	42.365	0.000	0.465	1.756	1.756[5]	50.000	0.000	0.833	0.344	0.087	0.541	0.687[1]
0.5	OCP	60.506	0.000	0.833	0.000	0.000[1]	82.762	465.046	0.590	0.557	0.090	0.576	0.583[3]
(65)	restrOCP	42.004	0.000	0.645	0.973	0.973[5]	52.662	778.871	0.550	0.466	0.147	0.445	0.497[5]
	Prom	45.899	0.000	0.702	0.668	0.668[3]	58.999	234.290	0.760	0.478	0.123	0.485	0.622[2]
	OptFunc	60.586	0.000	0.820	0.000	0.000[1]	82.891	425.434	0.602	0.600	0.118	0.555	0.578[4]
	GS	40.367	0.000	0.654	0.929	0.929[4]	50.000	0.000	0.904	0.400	0.089	0.497	0.700[1]
0.6	OCP	52.096	0.195	0.932	0.000	0.195[4]	78.869	501.309	0.474	0.607	0.078	0.622	0.548[3]
(45)	restrOCP	39.736	0.000	0.783	0.134	0.134[3]	54.296	701.996	0.583	0.527	0.137	0.490	0.537[5]
	Prom	42.339	0.000	0.846	0.000	0.000[1]	59.472	229.043	0.700	0.540	0.115	0.527	0.614[2]
	OptFunc	53.846	0.235	0.923	0.000	0.235[5]	82.348	406.502	0.481	0.660	0.104	0.605	0.543[4]
	GS	37.575	0.000	0.811	0.000	0.000[1]	50.000	0.000	0.965	0.4457	0.089	0.526	0.746[1]

Δ: true standardized effect; $n^{fix}_{\Delta,0.8}$: optimal sample size in the fixed design; $\mathbb{E}[N_{total,\Delta}]$: expected sample size per group for both stages; Pow_Δ: power of the adaptive design for standardized treatment effect Δ;
$\mathbb{E}\left[N^{RA}_{total,\Delta}\right]$: expected sample size per group conditional on entering the recalculation area; $Var\left(N^{RA}_{total,\Delta}\right)$: variance of sample size per group conditional on entering the recalculation area;
$\mathbb{E}\left[CP^{RA}_\Delta\right]$: expected conditional power conditional on entering the recalculation area; $Var(CP^{RA}_\Delta)$: variance of conditional power conditional on entering the recalculation area;
$ROS_{f_s,\beta}(\Delta)$: relative oversizing function; $RUP_{f_p,\beta}(\Delta)$: relative underpowering function; $S^{Liu}_{f_s f_p,\beta}(\Delta)$: point-wise Liu score;
$SN_\beta(\Delta)$: conditional sample size sub-score; $SCP_\beta(\Delta)$: conditional power sub-score; $S^{New}_\beta(\Delta)$: point-wise new conditional score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach;
[·]: numbers in square brackets present the ranking of the group sequential designs according to the corresponding point-wise performance score ([1]: best performance to [5]: worst performance).

TABLE A4 Estimated performance sub-scores and related location and variation components for all investigated designs and $n_1 = n_2 = 25$, $n_{max} = 100$

		Parameters to be Estimated					
Δ	**Design**	$e_N(\Delta)$	$v_N(\Delta)$	$SN_\beta(\Delta)$	$e_{CP}(\Delta)$	$v_{CP}(\Delta)$	$SCP_\beta(\Delta)$
0.0	OCP	0.053	0.680	0.366	0.762	0.412	0.587
	restrOCP	0.851	0.359	0.605	0.868	0.385	0.626
	Prom	0.617	0.699	0.658	0.843	0.448	0.646
	OptFunc	0.347	0.350	0.348	0.756	0.359	0.557
	GS	0.667	1.000	0.833	0.875	0.570	0.722
0.1	OCP	0.076	0.623	0.350	0.705	0.373	0.539
	restrOCP	0.806	0.303	0.554	0.813	0.317	0.565
	Prom	0.602	0.664	0.633	0.791	0.387	0.589
	OptFunc	0.315	0.360	0.337	0.694	0.310	0.502
	GS	0.667	1.000	0.833	0.834	0.518	0.676
0.2	OCP	0.107	0.562	0.334	0.646	0.351	0.499
	restrOCP	0.766	0.273	0.519	0.756	0.265	0.510
	Prom	0.586	0.635	0.610	0.733	0.336	0.535
	OptFunc	0.285	0.376	0.331	0.628	0.278	0.453
	GS	0.667	1.000	0.833	0.786	0.468	0.627
0.3	OCP	0.144	0.504	0.324	0.582	0.350	0.466
	restrOCP	0.723	0.255	0.489	0.691	0.232	0.462
	Prom	0.575	0.621	0.598	0.671	0.306	0.489
	OptFunc	0.257	0.401	0.329	0.556	0.270	0.413
	GS	0.667	1.000	0.833	0.732	0.431	0.581
0.35	OCP	0.164	0.479	0.321	0.548	0.358	0.453
	restrOCP	0.701	0.251	0.476	0.657	0.223	0.440
	Prom	0.568	0.613	0.590	0.638	0.297	0.467
	OptFunc	0.242	0.415	0.329	0.518	0.275	0.397
	GS	0.667	1.000	0.833	0.703	0.417	0.560
0.4	OCP	0.185	0.457	0.321	0.515	0.369	0.442
	restrOCP	0.673	0.244	0.458	0.618	0.220	0.419
	Prom	0.559	0.601	0.580	0.602	0.292	0.447
	OptFunc	0.235	0.427	0.331	0.480	0.285	0.383
	GS	0.667	1.000	0.833	0.673	0.409	0.541
0.5	OCP	0.755	0.425	0.590	0.751	0.401	0.576
	restrOCP	0.844	0.256	0.550	0.658	0.233	0.445
	Prom	0.928	0.592	0.760	0.670	0.300	0.485
	OptFunc	0.753	0.450	0.602	0.795	0.314	0.555
	GS	0.803	1.000	0.904	0.590	0.403	0.497
0.6	OCP	0.546	0.403	0.474	0.802	0.442	0.622
	restrOCP	0.873	0.293	0.583	0.720	0.261	0.490
	Prom	0.804	0.596	0.700	0.734	0.321	0.527
	OptFunc	0.499	0.462	0.481	0.856	0.354	0.605
	GS	0.931	1.000	0.965	0.649	0.404	0.526

Δ: true standardized effect; $e_N(\Delta)$: location component for conditional sample size; $v_N(\Delta)$: variation component for conditional sample size; $SN_\beta(\Delta)$: conditional sample size sub-score; $e_{CP}(\Delta)$: location component for conditional power; $v_{CP}(\Delta)$: variation component for conditional power; $SCP_\beta(\Delta)$: conditional power sub-score;
GS: classical group sequential approach; OptFunc: optimization function approach; Prom: promising zone approach;OCP: observed conditional power approach; restrOCP: restricted observed conditional power approach.