Received: 30 November 2020 | Revised: 24 September 2021 | Accepted: 7 November 2021

DOI: 10.1002/bimj.202000345

Biometrical Journal

RESEARCH ARTICLE

Blinded sample size recalculation in adaptive enrichment designs

Marius Placzek[1] | **Tim Friede**[1,2]

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

[2] DZHK (German Center for Cardiovascular Research), partner site Göttingen, Göttingen, Germany

Correspondence
Marius Placzek, Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee 32, 37073 Göttingen, Germany.
Email: marius.placzek@med.uni-goettingen.de

Funding information
Deutsches Zentrum für Herz-Kreislaufforschung, Grant/Award Number: GOE S12 UMG; Bundesministerium für Bildung und Forschung, Grant/Award Number: 05M13MGE BIMIT

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

In the precision medicine era, (prespecified) subgroup analyses are an integral part of clinical trials. Incorporating multiple populations and hypotheses in the design and analysis plan, adaptive designs promise flexibility and efficiency in such trials. Adaptations include (unblinded) interim analyses (IAs) or blinded sample size reviews. An IA offers the possibility to select promising subgroups and reallocate sample size in further stages. Trials with these features are known as adaptive enrichment designs. Such complex designs comprise many nuisance parameters, such as prevalences of the subgroups and variances of the outcomes in the subgroups. Additionally, a number of design options including the timepoint of the sample size review and timepoint of the IA have to be selected. Here, for normally distributed endpoints, we propose a strategy combining blinded sample size recalculation and adaptive enrichment at an IA, that is, at an early timepoint nuisance parameters are reestimated and the sample size is adjusted while subgroup selection and enrichment is performed later. We discuss implications of different scenarios concerning the variances as well as the timepoints of blinded review and IA and investigate the design characteristics in simulations. The proposed method maintains the desired power if planning assumptions were inaccurate and reduces the sample size and variability of the final sample size when an enrichment is performed. Having two separate timepoints for blinded sample size review and IA improves the timing of the latter and increases the probability to correctly enrich a subgroup.

KEYWORDS
adaptive design, BSSR, enrichment, interim analysis, subgroup analysis

1 | INTRODUCTION

In the era of personalized medicine and targeted therapies, the statistical aspects of designing a clinical trial are also increasing in complexity. Subgroup analyses, treatment selection, and innovative designs rely on an increasing number of parameters, which are often unknown and have to be guessed, estimated, or determined by a rule of thumb in the planning stage of a trial. Ondra et al. (2016) give a systematic review about methods and statistical approaches investigating subgroups in clinical trials. In adaptive enrichment designs, interim analyses (IAs) offer the possibility to select populations that seemingly benefit the most from the treatment and alter recruitment in favor of the most promising subgroups. Since sample sizes are calculated prior to a study, often based on insufficient knowledge about nuisance parameters, it is plausible to additionally implement a reevaluation of those parameters during the ongoing study and adjusting the sample size if needed. The idea of a blinded sample size review is to reestimate the parameters at a prespecified timepoint in a blinded fashion, that is, without revealing treatment groups, and recalculating the sample size. Those sample size adjustments do not inflate the type I error probability. In this paper, we will combine both methods for blinded sample size recalculation (BSSR) and adaptive enrichment strategies in a design with multiple subgroups, normally distributed endpoints, and Wald-type test statistics. Hence, there are several nuisance parameters, which are crucial for the planning and execution of a trial, such as the variances in the (sub)populations, the prevalences of the subgroups, timepoint of the sample size review, and timepoint of the IA. We will handle these step by step: Approximations and exact procedures for sample size determination prior to and sample size recalculation within the trial as well as methods for analyzing a multiple nested subgroups design have been presented by Placzek and Friede (2018). For several scenarios concerning known or unknown variances in the populations, we gave exact or approximative multivariate t-distributions with a single degrees-of-freedom parameter to perform hypothesis testing and sample size (re)calculations, whereas Graf et al. (2019) use a multivariate normal distribution, which is then adjusted by quantile substitution based on univariate t-distributions. Since the nuisance parameters directly impact the sample size and therefore a missspecification of these parameters leads to incorrectly powered studies, it is vital to have proper estimates of those parameters. Pilot studies of sufficient size conducted in the same population can provide these estimates (Kieser & Wassmer, 1996). However, such pilot studies require resources and take time to conduct. Often it is more efficient to perform a pilot as part of the trial. In contrast to an external pilot study, the idea is to treat the first part of the trial as a source of information about those nuisance parameters. This first part is therefore referred to as internal pilot study (IPS) (Wittes & Brittain, 1990). Nuisance parameters are then reestimated based on the data of the internal pilot and the sample size adjusted accordingly. Such a recalculation is preferably done using blinded estimators as requested by regulatory guidelines, cf. European Medicines Agency (EMEA) (2007); International Conference on Harmonisation E9 Expert Working Group (1999). Adaptive designs have been widely discussed covering subgroup selection (Chiu et al., 2018; Jenkins et al., 2011; Stallard et al., 2014) or population enrichment (Mehta et al., 2014; Wassmer & Dragalin, 2015) and the combination of group-sequential designs and subpopulation enrichment (Rosenblum, Luber, et al., 2016; Rosenblum, Qian, et al., 2016). An important part of those designs is combining evidence from the different stages in the final analysis controlling the type I error rate. In the setting of adaptive seamless designs, Bretz et al. (2006) describe flexible test procedures with hypothesis selection at interim giving several applications in Schmidli et al. (2006). The combination test (CT) approach (Bauer & Koehne, 1994) was used in a one-subgroup design by Brannath et al. (2009) and Jenkins et al. (2011), while the conditional error function (CEF) approach (Mueller & Schaefer, 2001) was analyzed in a one-subgroup design by Friede et al. (2012) and Mehta et al. (2014). Sugitani et al. (2018) investigated weighted significance levels associated with the hypotheses of the overall and subgroup population in adaptive enrichment designs and improved the efficiency of several methods including the CT and CEF approach. In Placzek and Friede (2019), we extended the CEF approach providing the case of multiple subgroups with no restrictions to the variance in the populations. Considerations on the timepoint of the IA are given in Benner and Kieser (2017).

The paper is organized as follows. Section 2 gives a motivating example. In Section 3, the statistical model and hypothesis tests are introduced as well as BSSR and adaptive enrichment designs. In Section 4, we propose a strategy combining BSSR in an IPS and adaptive enrichment designs. This is followed by simulations in Section 5. We close with a brief discussion of the findings and limitations of our study (Section 6).

2 | A MOTIVATING EXAMPLE

To motivate an IPS and corresponding sample size review in an adaptive trial design, we take a look at a clinical trial from the field of precision medicine that was reported by Sorkness et al. (2019). The phenotype-stratified clinical trial SIENA (Steroids in Eosinophil Negative Asthma) aimed to check the paradigm that inhaled corticosteroids (ICS) are the appropriate first-line treatment for patients with mild persistent asthma. Two phenotypic strata were assigned based on an a priori defined extent of sputum eosinophilia (Eos Low vs. Eos High). Retrospective data from the National Heart, Lung, and Blood Institutes (NHLBI) Asthma Clinical Research Network (ACRN) suggested that approximately 50% of patients with mild-moderate asthma, not already treated with ICS, have < 2% eosinophils in induced sputum (Eos Low). Therefore it appeared reasonable to expect a 1:1 ratio of the two strata in the study and hence sample size and power calculations assumed such a distribution. Participants were treated with an ICS, LAMA (long-acting muscarinic antagonist) or placebo. The main objective focused on the different response to ICS and LAMA between the Eos Low and Eos High strata. However, after 8 months of recruitment, the observed ratio of Eos Low to Eos High was around 3:1 (79 patients, 60:19). This unexpected imbalance was monitored by the Steering Committee and different options and alternatives were considered. Since the initial sample size calculation assumed balanced strata, possible solutions included selectively enriching the Eos High stratum, closing the Eos Low stratum early or generally revising the analysis plan. Finally it was decided to change the statistical plan and focus on the originally secondary endpoints still testing treatment effects within both strata rather than between. Power and sample size calculation were revised as well and SIENA was successfully completed in a manner that maintained meaningful outcomes (221 Eos Low and 74 Eos High). Retrospectively, Sorkness et al. (2019) conclude that it is recommended at the planning stage, especially when examining phenotype- or biomarker-stratified populations, to incorporate plans for IAs or sample size reassessment leading to adaptive trial designs that, as they assume, will become even more relevant in the precision medicine era.

3 | METHODS FOR ADAPTIVE ENRICHMENT DESIGNS

3.1 | Statistical model and hypothesis tests

In this paper, we consider a patient population, which is denoted by $F = S_0$, indicating that it represents the full population, and k nested subgroups $S_k \subset S_{k-1} \subset \cdots \subset S_1 \subset S_0 = F$ within. Let $\tau_1, \dots, \tau_k$ denote the corresponding subgroup prevalences, that is, τ_i denotes the proportion of subjects in S_i among all subjects in F. We assume that an individual observation is normally distributed and want to compare an experimental treatment to a control. For simplicity and without loss of generality the control group mean is set to 0 in all populations. Treatment effect sizes and hence treatment mean differences are denoted by $\theta_0, \dots, \theta_k$. We further assume that in each population S_i the variances $\sigma^2_{T,S_i} = \sigma^2_{C,S_i}$ are equal for the treatment and the control group. Therefore, only one variance $\sigma^2_{S_i}$ per population S_i is needed for the statistical model, $i = 0, \dots, k$. Let $n_T^{S_i}$ and $n_C^{S_i}$ be the number of subjects in the treatment and control group for each subpopulation, $i = 1, \dots, k$ and let $n_C^{S_0} = n_C$ and $n_T^{S_0} = n_T$ denote the number of subjects in the experimental treatment and control group in the whole patient population. For unbalanced sample sizes, we introduce an allocation parameter $a = n_T/n_C$. This means there are $n_T^{S_i} = a \cdot n_C^{S_i}$ subjects in the treatment group of population $S_i, i = 0, \dots, k$.

Hypotheses will be tested using standardized test statistics $Z^{\{F\}}, Z^{\{S_1\}}, \dots, Z^{\{S_k\}}$. In previous papers, we already considered the multiple nested subpopulations design, which implies a certain correlation structure on the test statistics (Placzek & Friede, 2018, 2019). We broke down how the treatment effects and variances from S_j are a combination of treatment effects and variances of $S_j \backslash S_{j+1}$ and $S_i, i > j$. These dependencies play a major role in determining the covariance matrix of the joint vector of the standardized test statistics. Hence, scenarios with less dependencies, for example, nonoverlapping subgroups, can easily be derived from this case by adjusting the covariance matrix (Placzek & Friede, 2019).

Here, we will stick to the multiple nested subgroups design. The null hypothesis of no treatment effect in the full population

$$H_0^{\{F\}}\colon\ \theta_0 = 0$$

will be tested using $Z^{\{F\}}$ while hypotheses

$$H_0^{\{S_i\}}\colon\ \theta_i = 0,\ i = 1, \dots, k$$

of no effect in an individual subpopulation are tested with test statistics $Z^{\{S_i\}}$, respectively. To control the familywise error rate (FWER) in the strong sense while testing these multiple hypotheses, we apply a closed testing procedure (Marcus et al., 1976). The intersection hypothesis

$$H_{\boldsymbol{\upsilon}} = H_0^{\bigcap_{\{i \in \boldsymbol{\upsilon}\}} S_i}\colon\ \theta_i = 0\ \forall i \in \boldsymbol{\upsilon} \subseteq \{0, \dots, k\}$$

is tested using the joint distribution of the standardized test statistics. To this end let

$$\boldsymbol{Z} = (Z^{\{F\}}, Z^{\{S_1\}}, \dots, Z^{\{S_k\}})'$$

and let $\boldsymbol{T}_1$ denote the set of hypotheses that are planned to be tested at the final analysis.

There are two important observations here. On the one hand, there are various nuisance parameters (prevalences, variances) whose missspecification at the planning stage of a trial can lead to an overpowered or underpowered study. On the other hand, there are multiple (sub)populations, hence multiple hypotheses, and we would like to focus on the most promising ones. We will address both issues beginning with a recap of methods for BSSR in an IPS design followed by subgroup selection and subgroup enrichment in adaptive designs with an IA. These two approaches are then combined to a joint novel strategy in Section 4.

3.2 | BSSR in an IPS

To implement a sample size review, we suggest using an IPS design, that is, at a prespecified timepoint a data look is performed in a blinded fashion, as preferred by, for example, regulatory authorities (European Medicines Agency (EMEA), 2007; Food and Drug Administration (FDA), 2019, 2006; International Conference on Harmonisation E9 Expert Working Group, 1999), in order to check assumptions that were made prior to the study by reestimating nuisance parameters, for example, the variances or prevalences of the (sub)populations. Let n_{IPS} denote the number of subjects available for the blinded sample size review. To reestimate the variances without breaking the blind we use so-called *lumped variance estimators*, cf. Zucker et al. (1999), that is, one sample variance estimators treating control group and treatment group observations as one group and combining disjunct variance estimates from each population. Let $n_{IPS}^{S_i}$ denote the sample size for $S_i \backslash S_{i+1}$ at the sample size review, then the variance estimators and prevalence estimators are given by

$$\hat{\sigma}_F^2 = \frac{1}{n_{IPS} - 1} \sum_{i=0}^{k} \sum_{j=1}^{n_{IPS}^{S_i}} (X_{ij} - \bar{X}_{i\cdot})^2,$$

$$\hat{\sigma}_{S_i}^2 = \frac{1}{n_{IPS}^{S_i} - 1} \sum_{s=i}^{k} \sum_{j=1}^{n_{IPS}^{S_i}} (X_{sj} - \bar{X}_{s\cdot})^2, \ i = 1, \dots, k,$$

$$\hat{\tau}_i = \frac{n_{IPS}^{S_i}}{n_{IPS}}, \ i = 1, \dots, k,$$

with

$$\bar{X}_{i\cdot} = \frac{1}{n_{IPS}^{S_i}} \sum_{j=1}^{n_{IPS}^{S_i}} X_{ij}, \ i = 0, \dots, k,$$

where X_{ij} denotes an individual observation in $S_i \backslash S_{i+1}$, $i = 0, \dots, k; \ j = 1, \dots, n_{IPS}^{S_i}$. Here $S_{k+1} = \emptyset$. These new estimates are then used to recalculate the sample size, usually by plugging in the new values into the sample size formula that was used at the planning stage of the trial. This will be showcased in Section 4 where the new procedure combining BSSR and adaptive enrichment designs is proposed. Methods for the analysis, sample size determination, as well as recalculation of the sample size in designs with nested subgroups have been discussed in Placzek and Friede (2018). There we considered several scenarios, namely, known variances, the case of unknown but equal variances across all populations, and the case of completely unknown variances. Hence, to calculate the test statistic we (1) do not have to estimate a variance component, (2) have to estimate one overall variance, or (3) have to estimate a variance for each population separately. Accordingly, the vector of test statistics is either multivariate normal distributed, multivariate t-distributed, or approximately multivariate t-distributed, that is, under the global null hypothesis $H_0^{\cap_{i \in \boldsymbol{\mathcal{U}}} S_i}$ for $\boldsymbol{Z} = (Z^{\{F\}}, Z^{\{S_1\}}, \dots, Z^{\{S_k\}})'$ it holds that

$$(a)\quad \boldsymbol{Z}_a = \left(\sqrt{\frac{n_C}{a^*}}\frac{\hat{\Delta}_F}{\sigma_F}, \sqrt{\frac{\tau_1 n_C}{a^*}}\frac{\hat{\Delta}_{S_1}}{\sigma_{S_1}}, \dots, \sqrt{\frac{\tau_k n_C}{a^*}}\frac{\hat{\Delta}_{S_k}}{\sigma_{S_k}}\right)' \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_a),$$

$$(b)\quad \boldsymbol{Z}_b = \left(\sqrt{\frac{n_C}{a^*}}\frac{\hat{\Delta}_F}{\hat{\sigma}}, \sqrt{\frac{\tau_1 n_C}{a^*}}\frac{\hat{\Delta}_{S_1}}{\hat{\sigma}}, \dots, \sqrt{\frac{\tau_k n_C}{a^*}}\frac{\hat{\Delta}_{S_k}}{\hat{\sigma}}\right)' \sim MT_{(a+1)n_C-2(k+1)}(\mathbf{0}, \boldsymbol{\Sigma}_b),$$

$$(c)\quad \boldsymbol{Z}_c = \left(\sqrt{\frac{n_C}{a^*}}\frac{\hat{\Delta}_F}{\hat{\sigma_F}}, \sqrt{\frac{\tau_1 n_C}{a^*}}\frac{\hat{\Delta}_{S_1}}{\hat{\sigma_{S_1}}}, \dots, \sqrt{\frac{\tau_k n_C}{a^*}}\frac{\hat{\Delta}_{S_k}}{\hat{\sigma_{S_k}}}\right)' MT_{df}(\mathbf{0}, \boldsymbol{\Sigma}_c).$$

Here a^* is defined as $a^* = 1 + 1/a$. In (b) the degrees of freedom depend on the number of subjects in the whole population $(a+1)n_C$ and the number of subgroups k. In (c) the degrees of freedom can be chosen depending on the number of subjects in the smallest or the largest population in order to get a conservative ($df = (a+1)n_C^{S_k} - 2$) or a liberal ($df = (a+1)n_C - 2(k+1)$) approximation. The covariance matrices depend on the variances, the estimators of the variances, and the prevalences of the subpopulations, respectively. Alternatively, in scenario (c) Graf et al. (2019) suggest approximating the joint distribution using a multivariate normal distribution, calculating an equicoordinate quantile c_α and transforming this to critical values c_i for each population by applying univariate t-distributions

$$c_i = \Psi^{-1}_{(a+1)n_C^{S_i}-2}(\Phi_{0,1}(c_\alpha)),\ i = 0, \dots, k. \tag{1}$$

Here Ψ_{df} denotes the distribution function of a univariate t-distribution and $\Phi_{0,1}$ of a standard normal distribution. In any case, these distributional properties of $\boldsymbol{Z}$ can be used to perform hypothesis testing and also to determine a sample size prior to the study or recalculate the sample size in an ongoing study. In Placzek and Friede (2018), we investigated the proposed approximations with and without BSSR and reported findings on type I error probability and power in several scenarios. The results are satisfactory for sufficiently large sample sizes; for small sample sizes, we compared and gave advice on different choices for the single degree-of-freedom parameter of the multivariate t-distribution as well as the optimal timepoint of the BSSR.

3.3 | Adaptive enrichment designs

Extensions to the standard fixed-sample designs, especially when dealing with subgroups and an uncertainty about the possibly heterogeneous treatment effects in different (sub)populations, are adaptive enrichment designs. These are designs with two or more stages and preplanned IAs. At the interim looks decisions are made on whether to continue the trial unchanged or to drop populations, which seem not promising based on the data obtained up to that point. In the latter case, recruiting and hypotheses testing are restricted to subjects from populations that are carried to the next stage resulting in an enrichment of the most promising subgroup. A crucial part is the combination of the different stages of the trial for the final analysis while still controlling the type I error probability. In Placzek and Friede (2019), we used the so-called CEF approach in an adaptive enrichment design with multiple nested subgroups and normally distributed endpoints, accounting for uncertainty in variance estimation. The distributional properties used for testing were the same as described in Section 3.2.

Combining two stages the CEF approach can be applied testing each hypothesis $H_{\boldsymbol{\upsilon}}$ as follows. After all data of the first stage are collected, the conditional error can be calculated as the probability to reject $H_{\boldsymbol{\upsilon}}$ after the second stage given the data collected so far, that is,

$$CE_{\boldsymbol{\upsilon}} = P_{H_{\boldsymbol{\upsilon}}}(\max_{i \in \boldsymbol{\upsilon}} Z_i \geq d_s | z_i^{(1)}, i \in \upsilon). \tag{2}$$

Here Z_i denotes the standardized test statistic corresponding to (sub)population i based on the accumulated data from both stages. Note that this is a Dunnett-type test since we are using a maximally selected test statistic. The critical boundary d_s depends on s, the number of populations involved in testing $H_{\boldsymbol{\upsilon}}$. $z_i^{(1)}$ is the first-stage standardized test statistic, which is not a random variable since stage 1 data are already available when calculating $CE_{\boldsymbol{\upsilon}}$. Let $\boldsymbol{T}_2$ denote the set of hypotheses corresponding to populations carried to the second stage and $\boldsymbol{I}_2$ the set of indices corresponding to those populations. After the second stage a stage 2 p-value $q_{\boldsymbol{\upsilon}}$ is calculated under the conditional distribution given the observed stage 1 test statistics,

$$q_{\boldsymbol{\upsilon}} = P_{H_{\boldsymbol{\upsilon}}}(\max_{i\in\boldsymbol{\upsilon}\cap\boldsymbol{I}_2} Z_i \geq z^{\max}_{\boldsymbol{\upsilon}\cap\boldsymbol{I}_2} | z_i^{(1)}, i \in \boldsymbol{\upsilon} \cap \boldsymbol{I}_2), \tag{3}$$

where $z^{\max}_{\boldsymbol{\upsilon}\cap\boldsymbol{I}_2}$ denotes the actually observed value of $\max_{i\in\boldsymbol{\upsilon}\cap\boldsymbol{I}_2} Z_i$. The hypothesis $H_{\boldsymbol{\upsilon}}$ is then rejected if $q_{\boldsymbol{\upsilon}} < CE_{\boldsymbol{\upsilon}}$. For normally distributed observations and known variances across the populations, these are trivial calculations using multivariate normal distributions, but in Placzek and Friede (2019), we also gave suggestions for the more complex and practical scenarios in which the variances are unknown and either equal or unequal across the subgroups (Placzek & Friede, 2019). Since it can be difficult to determine the conditional distributions analytically, in some cases it might be easier to simulate the conditional distributions in order to calculate conditional rejection probabilities when determining the conditional error. Note that this would also be helpful when recalculating the sample size based on unblinded estimates of the nuisance parameters conditioned on the observed treatment effects at interim. Since the trial is unblinded at the IA anyway, such an unblinded sample size recalculation can be integrated at this stage. It does not inflate the type I error rate since we are using the CEF approach here. However, in this paper, we focus on BSSR in an IPS and will not include unblinded SSR here. We will discuss it in Section 5 in more detail.

We compared this procedure to the traditional combination function approach. The principle of this approach is simple: For each hypothesis, stage 1 data are used to calculate stage 1 p-values and stage 2 data to calculate stage 2 p-values. First- and second-stage p-values are then combined using a combination function such as the weighted inverse normal combination function, which is given by

$$C(p_1, p_2) = 1 - \Phi(w_1\Phi^{-1}(1 - p_1) + w_2\Phi^{-1}(1 - p_2)), \tag{4}$$

where w_i, $i = 1, 2$ are prespecified weights with $0 \leq w_i \leq 1$ and $w_1^2 + w_2^2 = 1$. Weights are commonly determined proportional to the sample sizes per stage.

At the IA between the two stages it is decided which (sub)populations are carried to the next stage. There are several ways to do so, which do not impact the FWER control. For the simulations in Section 5, we chose the so-called ϵ rule. Here, the population with the maximum test statistic and every population with a test statistic within ϵ range of the maximum test statistic is carried forward to the second stage. For $\epsilon = 0$ and $\epsilon \to \infty$ this includes both extreme cases, namely, continuing only with the best or with all populations. Dropping not promising populations and shifting the now freed sample size toward the remaining populations when continuing recruitment leads to an enrichment. Note that the difficult part of the adaptive enrichment design is not the enrichment itself but the combination of data from the different stages of the trial, that is, pre- and postadaptation, in the final confirmatory analysis. The CEF approach and the combination function approach are two solutions to this problem.

In Placzek and Friede (2019), we conducted a simulation study and found that the best performing combination of methods is the CEF approach with the univariate t-approximation by Graf et al. (2018) outperforming the other approximations and the CT approach. It is therefore recommended to use the CEF approach, especially in the case of unknown and unequal variances or when there is little knowledge about the variances. Therefore, we will use the CEF approach rather than the CT approach in the following sections as our main focus is on the combination of the BSSR and the adaptive enrichment design.

4 | PROPOSED PROCEDURE FOR BSSR IN ADAPTIVE ENRICHMENT DESIGNS

After this recap of the adaptive tools, we will now present the proposed procedure incorporating both BSSR in an IPS and adaptive enrichment in a design with an IA. Combining the advantageous properties of both approaches creates a robust, efficient, and flexible testing strategy as we will see in the simulations. Figure 1 illustrates such a procedure for a simple one-subgroup design and one IA. We will describe it in detail for a design with multiple nested subgroups.

Prior to the trial, assumptions are made on the treatment effects and the nuisance parameters and a sample size N_0 is calculated: Assume we have no prior knowledge about the variances and decide to use the methods for a design with

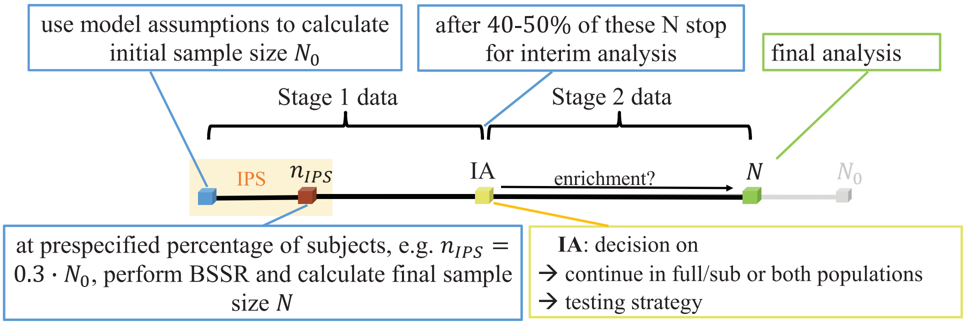
use model assumptions to calculate initial sample size N_0
after 40-50% of these N stop for interim analysis
final analysis
Stage 1 data
Stage 2 data
IPS
n_{IPS}
IA
enrichment?
N
N_0
at prespecified percentage of subjects, e.g. $n_{IPS} = 0.3 \cdot N_0$, perform BSSR and calculate final sample size N
IA: decision on
→ continue in full/sub or both populations
→ testing strategy

FIGURE 1 Combining blinded sample size recalculation in an internal pilot study design and an adaptive enrichment procedure with an interim analysis. In this example, there is one subpopulation inside a full population and a two-stage design shall reveal whether there is an increased treatment benefit in the subpopulation while still simultaneously examining the full population.

unknown and potentially unequal variances across the populations. For simplicity, we consider a balanced design, that is, a 1:1 allocation, hence $a = 1$, $a^* = 2$, and $n_T = n_C = n$. This can easily be modified to an arbitrary allocation ratio, cf. Placzek and Friede (2019). Best guesses for the treatment effects and the variances are denoted by $\theta_0^*, \dots, \theta_k^*$ and $\sigma_{S_0}^*, \dots, \sigma_{S_k}^*$. Sample size calculation is based on the power to reject the global intersection hypothesis, which is tested using the joint distribution of the vector of standardized test statistics $\boldsymbol{Z}$. As described in Section 3.2, this joint distribution can be approximated by a multivariate t-distribution. For the planning purpose we choose the conservative approximation with $df = 2n_{S_k} - 2$ degrees of freedom. Accordingly, let $\boldsymbol{t}_{\mathbf{0},\boldsymbol{\Sigma},df,1-\alpha}$ denote the $(1-\alpha)$-equicoordinate quantile of the distribution $MT_{df}(\mathbf{0}, \boldsymbol{\Sigma})$ under the null hypothesis $H_0^{\{\cap_{i=0}^k S_i\}}$. We define $\boldsymbol{G}_{\boldsymbol{\delta},\tilde{\boldsymbol{\Sigma}},df}$ as the distribution function of $MT_{df}(\boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}})$ under the alternative. Here, $\boldsymbol{\delta}$ is the noncentrality parameter

$$\boldsymbol{\delta} = (\delta_0, \dots, \delta_k)' = \left(\sqrt{\frac{n}{2}} \frac{\theta_0^*}{\sigma_F^*}, \sqrt{\frac{n_{S_1}}{2}} \frac{\theta_1^*}{\sigma_{S_1}^*}, \dots, \sqrt{\frac{n_{S_k}}{2}} \frac{\theta_k^*}{\sigma_{S_k}^*} \right)', \tag{5}$$

and $\tilde{\boldsymbol{\Sigma}}$ a slightly shifted version of $\boldsymbol{\Sigma}$ under the alternative (Placzek & Friede, 2018). Through an iterative search algorithm, we can find the initial sample size N_0 required to achieve a power of $(1-\beta)$ via

$$N_0/2 = \min n \ \text{ s.t. } \ 1 - \boldsymbol{G}_{\boldsymbol{\delta},\tilde{\boldsymbol{\Sigma}},df}(\boldsymbol{t}_{\mathbf{0},\tilde{\boldsymbol{\Sigma}},df,1-\alpha}) \geq 1 - \beta. \tag{6}$$

We start recruitment and after a prespecified number of subjects, for example, $t_{IPS} = 30\%$, a sample size review with $n_{IPS} = t_{IPS} \cdot N_0$ is performed. Here, variances and prevalences are recalculated in a blinded fashion and a new final sample size N is determined: To reestimate the variances without breaking the blind, we use the lumped variance estimators as described in Section 3.2. We obtain estimators $\hat{\sigma}_F, \hat{\sigma}_{S_i}$ and $\hat{\tau}_i$, $i = 1, \dots, k$. These reestimated nuisance parameters are then plugged in the previously described sample size determination algorithm, cf. (5) and (6), replacing the original guesses $\sigma_{S_0}^*, \dots, \sigma_{S_k}^*$ not only in the noncentrality parameter vector $\boldsymbol{\delta}$ but also in the covariance matrix $\tilde{\boldsymbol{\Sigma}}$. The final sample size N is recalculated. Note that this calculation of the final sample size is still based on rejection of the global intersection hypothesis assuming all populations are kept in the trial until the end. This means it does not anticipate an enrichment at an IA. Consequently, if in a one-subgroup design the subpopulation is selected at the IA, recruiting only from the subgroup until N is reached will substantially increase the power above 80%. Therefore, in the IPS analysis, we can calculate a final sample size N_S replacing N specifically for the case of dropping the full population $F \backslash S$ and only continuing with the subgroup at the IA. Calculation can be done using the same estimates from the IPS and formulas as before but additionally adjusting the numbers of patients $n, n_{S_1}, \dots, n_{S_k}$, and the guesses of $\theta_0^*, \dots, \theta_k^*$ accordingly, for the noncentrality parameter $\boldsymbol{\delta}$ in (5). At the IA, if the subpopulation is selected, we can choose to recruit from the subgroup until N_S is reached keeping the power at 80%. This completes the IPS stage of the trial. The idea of updating the information on the prevalence has previously been discussed by Gurka et al. (2010) in the context of epidemiological studies.

Recruiting is continued until the IA takes place. The timing of this IA depends on the adjusted sample size N and a prespecified portion t_1. For example, it might be scheduled midway of the trial, that is, $t_1 = 0.5$. At the IA the $N_1 = t_1 \cdot N$ observations available are used to decide in which populations testing should be continued. To do so, a selection rule, for

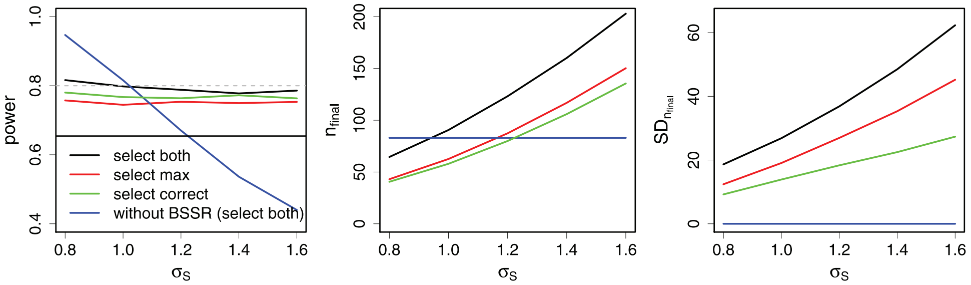

FIGURE 2 Comparison between methods combining BSSR and adaptive testing strategies (black, red, green) and a strategy without BSSR (blue). Adaptive testing strategy is the conditional error function approach deciding at an interim analysis whether to continue testing in both populations (black, blue) or only in the population with the maximum test statistic at interim (red) or only in the subpopulation (green). Prevalence $\tau = 0.4$, $t_{IPS} = 0.3$, $t_1 = 0.5$, $\Delta_S = 0.75$. Assumed variance in the subpopulation is $\sigma_S^* = 1$ while the true variance varies on the x-axis.

example, the ϵ rule, is applied to the test statistics calculated at interim in an unblinded fashion. One can also decide on the further testing strategy. Assume we stick to the CEF approach and decide to perform testing using the approximation by Graf et al (2019). For each hypothesis $H_{\boldsymbol{v}}$ we calculate the conditional error as the probability to reject the hypothesis at the final analysis using all N observations given this particular stage 1 data (N_1 observations). We demonstrate this with the global intersection hypothesis H_{IS}: First, we have to calculate the critical values c_i for testing at the final analysis. To do so, we approximate the joint distribution of $\boldsymbol{Z}$ by a multivariate normal distribution, compute an equicoordinate $(1-\alpha)$-quantile, and apply Equation (1). Now we determine

$$CE_{IS} = P_{H_{IS}}\left(\exists i \text{ s.t. } Z^{\{S_i\}} > c_i | z_1^{\{S_i\}}, i = 0, \dots, k\right),$$

which requires the multivariate conditional distribution of $\boldsymbol{Z}|\boldsymbol{Z}_1$. We suggest simulating this distribution to obtain the conditional error. Recruiting is then continued in the second stage according to the decisions made at interim, that is, if it is decided to further test only in certain subpopulations, only subjects from these populations are recruited and hence there is an enrichment of the design. When the final sample size N is reached the final analysis is carried out. This means calculating stage 2 p-values $q_{\boldsymbol{v}}$, for example, for the intersection hypothesis:

$$q_{IS} = P_{H_{IS}}(\max_{i \in \boldsymbol{I}_2} Z^{\{S_i\}} \geq z_{\boldsymbol{I}_2}^{\max} | z_1^{\{S_i\}}, i \in \boldsymbol{I}_2).$$

Here, $\boldsymbol{I}_2$ is the set of indices corresponding to populations carried to the second stage and $z_{\boldsymbol{I}_2}^{\max}$ the actual observed value of $\max_{i \in \boldsymbol{I}_2} Z^{\{S_i\}}$. A hypothesis is then rejected if its stage 2 p-value is smaller than its conditional error at the IA, for example, H_{IS} is rejected if $q_{IS} < CE_{IS}$.

5 | SIMULATIONS

To investigate the properties of the procedure proposed in Section 4, we conduct a simulation study, which will focus on three main aspects: First, we examine power, sample sizes and variability of final sample sizes of the proposed method combining BSSR and adaptive testing strategies for different selection rules. We include the comparison to a design without BSSR (Figure 2). Additionally, we show some type I error rates for the corresponding scenarios under the global null hypothesis. Next, we investigate the impact of different timepoints for the BSSR and the IA. We provide simulations suggesting an optimal timepoint of the IA (Figure 5) and finally analyze the impact of performing the BSSR prior to the unblinded IA in comparison to performing both stops at the same timepoint (Figure 6).

We start by comparing a one subgroup adaptive enrichment design with BSSR and such a design without recalculation in terms of power and sample size. Therefore, we simulate a subgroup with prevalence $\tau = 0.4$ and an effect of $\Delta_S = 0.75$ while there is no effect in the rest of the full population $F\backslash S$. The true variance in the complement of the subgroup is $\sigma_{F\backslash S} = 1$ and the true variance σ_S varies on the x-axis from 0.8 to 1.6. When calculating the initial sample size, it is always assumed that $\sigma^*_{F\setminus S} = \sigma^*_S = 1$, that is, the variance is equal and 1. Sample size calculation is done assuming unequal and unknown variances. Therefore, we use the conservative t-approximation for the distribution of $\boldsymbol{Z}$. We aim for a trial with 80% power to reject the intersection hypothesis. The size of the IPS is 30% of the initial sample size ($t_{IPS} = 0.3$) while the IA is performed after 50% of the final sample size is observed ($t_1 = 0.5$). As an adaptive method to combine the two stages of the trial, we choose the CEF approach applying the univariate t-approximation by Graf et al. (2019). Note that in all simulations in this section we assume a known, fixed prevalence τ. This means it is not estimated during the BSSR or the final analysis. In a previous paper, we found that additionally estimating this parameter does not notably change the simulation results, cf. Placzek and Friede (2018). Therefore, we focused on a known, fixed prevalence here for the ease of presentation. The number of simulation runs is 10,000.

Figure 2 (left panel) shows the simulated power of four different strategies: The black line shows rejection rates for a selection rule at interim that always selects all populations to continue in the second stage ($\epsilon = \inf$ in the context of the epsilon selection rule) and the red line corresponding rates for a rule that always selects only the population with the maximum test statistic at interim ($\epsilon = 0$). The green line depicts a theoretical method that always correctly picks the population that actually benefits most. Here, this is the subpopulation since the true effect is generated only in the subpopulation in this simulation. Hence, the green line shows rejection rates for a strategy that always selects the subpopulation at interim. In these two strategies (red, green) there is an enrichment in terms of recruiting only patients from the subgroup if it is selected at interim. In that case keeping the recalculated sample size and only reallocating N would increase the power considerably above 80%. Therefore, we adjust the sample size once again with an estimate from the IPS at the sample size review, which we calculated anticipating the selection of the subgroup at interim as described in Section 4.

The last method (blue line) depicts the CEF approach without a BSSR in an IPS. The central panel shows the corresponding final sample sizes while the right panel presents the standard deviation of the final recalculated sample sizes.

The benefit of having an IPS to adjust the sample in case of bad initial assumptions is obvious. As expected, in the scenario without the option for a sample size adjustment (blue line) there are either too many subjects included in the trial or too few subjects depending on whether the true variance in the subpopulation was lower or higher than assumed and thus leading to an underpowered or overpowered study. Only if the assumption was correct ($\sigma_S = 1$), the nominal power of 80% is obtained. In the scenarios with an IPS, the BSSR can make up for misspecifications of the nuisance parameters at the planning stage and the new final sample sizes lead to trials containing the desired power over the range of of σ_S. Not surprisingly the method that carries all populations to the next stage (black) achieves the power best since the sample size calculation and recalculation is powered for the rejection of the intersection hypothesis. However, a slight decrease in power can be observed for a larger variability in the subpopulation. The strategy that only selects the most promising population to be continued (red) has a lower power across all values of σ_S. This is due to the fact that we stick to the adjusted sample size in case of selecting the full population and only adjust a second time if the subpopulation is selected. Naturally, testing only in the full population at the final analysis would require a much higher sample size since the effect is actually generated in the subpopulation. Consequently the overall power is a bit decreased due to the decreased mean recalculated final sample sizes. Here (central panel) the advantage of the enrichment is visible. Only recruiting from the seemingly more favorable population lets us decrease the final sample size while still while still maintaining the preplanned power. For the theoretical approach that always chooses the subgroup (green), the power is slightly better and the recalculated sample sizes even a bit lower. It also does not completely achieve the nominal power of 80% since the readjustment is based on rejecting the intersection hypothesis and not on the rejection of the subgroup hypothesis. Final sample sizes and the standard deviations of the final sample sizes (right panel) increase with increasing variance in the subpopulation for the designs, which perform a BSSR. The design without sample size adjustment (blue) always sticks to its initial sample size, hence there is no variability. Strategies with enrichment need the lowest sample sizes and have lower SDs of the final sample sizes compared to the strategy that always continues with all populations.

Figure 3 complements these simulations of power and sample size with scenarios where there is an effect in both populations, that is, $\Delta_F = 0.5$, $\Delta_S = 0.5$. Methods and settings remain the same. The number of simulation runs is 10,000. The results show that the desired power of 80% is attained in all settings. The final sample size and the variability of the final sample size behave similar to Figure 2. Compared to the scenarios with an effect only in the subpopulation, there is almost no difference between the two selection rules.

Since type I error rate control of the combined procedure follows directly from FWER control of both components, that is, BSSR and adaptive enrichment design, cf. Placzek and Friede (2018, 2019), we only briefly show type I error rates here. BSSR and adaptive enrichments methods are applied independently, hence FWER control follows. This statement

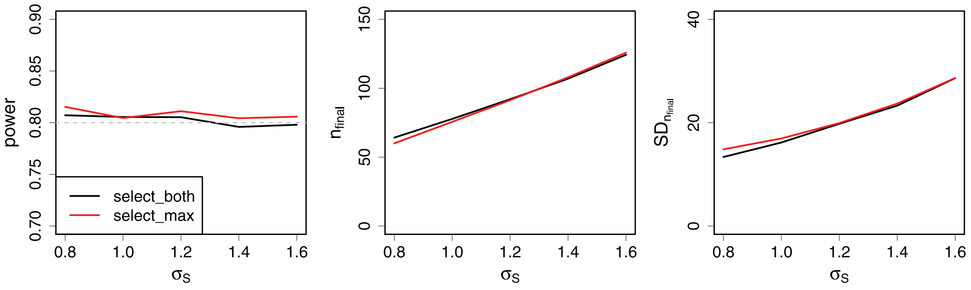

FIGURE 3 Power, final sample size, and variability of final sample size of the method combining BSSR and adaptive enrichment with effects in both populations. Adaptive testing strategy is the conditional error function approach deciding at an interim analysis whether to continue testing in both populations (black) or only in the population with the maximum test statistic at interim (red). Prevalence $\tau = 0.4$, $t_{IPS} = 0.3$, $t_1 = 0.5$, $\Delta_F = 0.5$, $\Delta_S = 0.5$. Assumed variance in the subpopulation is $\sigma_S^* = 1$ while the true variance varies on the x-axis.

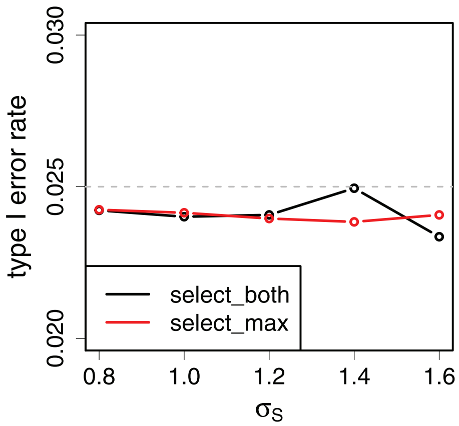

FIGURE 4 Type I error rates for combining BSSR with an adaptive enrichment design. Adaptive testing strategy is the conditional error function approach deciding at an interim analysis whether to continue testing in both populations (black) or only in the population with the maximum test statistic at interim (red). Prevalence $\tau = 0.4$, $t_{IPS} = 0.3$, $t_1 = 0.5$. Assumed effect in the subpopulation is $\Delta_S^* = 0.75$. Assumed variance in the subpopulation is $\sigma_S^* = 1$ while the true variance varies on the x-axis.

is supported by the simulation results presented in Figure 4. For a nominal level of $\alpha = 0.025$ we simulated the type I error rate under the global null hypothesis of no effect in any population for the same scenarios as just presented for the power and sample size, that is, a BSSR at $t_{IPS} = 0.3$ and a possible enrichment at an IA at $t_1 = 0.5$. Methods for BSSR and final analysis remain the same. The number of these simulation runs is 100,000. The results show that the type I error rate is controlled throughout the different scenarios for both selection rules. In all cases it is a bit below the nominal level of 0.025. This minimal conservatism is inherited from the applied conservative approach for BSSR, that is, choosing the degrees of freedom for the multivariate t-distribution in a conservative way.

Next, we focus on the timepoints of the sample size review and the IA. Previously we chose, as rule of thumb, to perform the BSSR after 30% of the initially planned patients have entered the study. Obviously the earlier a sample size recalculation is performed to correct misspecified assumptions the better. However, in Placzek and Friede (2018) we showed that there should be at least 20–25 subjects in the smallest subgroup when recalculating the sample size. Otherwise either the desired power will not be achieved or an adjustment based on the small number of subjects in the IPS has to be used at the cost of a notably increased expected final sample size. So, one has to check the initial sample size and determine the timepoint t_{IPS} for an early sample size review such that there are enough patients to recalculate the different nuisance parameters even in the smallest population.

Concerning the timepoint t_1 of the IA we simulated the power of the CEF approach for different timepoints of the IA. This means for a fixed sample size we performed hypothesis testing in a one subgroup design using the above-mentioned selection rule to continue always in the population with the maximum test statistic. We varied the subgroup size ($\tau = 0.2, 0.3, 0.4$) and the portion of the sample size used to perform the IA ($N_1/N = 0.1, 0.2, \dots, 0.9$).

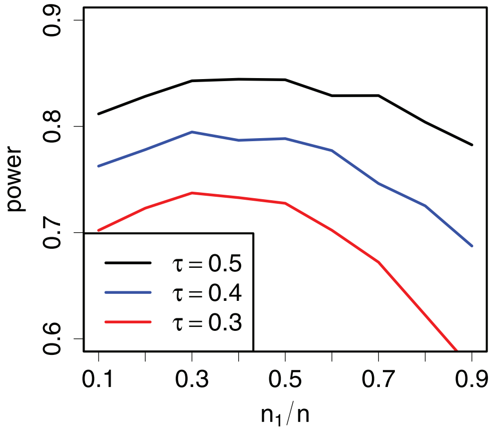

FIGURE 5 Comparing the power of the conditional error function approach for different timepoints of the interim analysis (x-axis) varying the subgroup prevalence $\tau = 0.3, 0.4, 0.5$ (red, blue, black). At interim always the best test statistic is selected ($\epsilon = 0$). Fixed sample size of $n = 200$ per group and an effect of $\Delta_S = 0.4$ in the subgroup.

Figure 5 shows the results. For all three values of τ there is a peak in the power curves with maxima at portions of 0.4–0.5 of the final sample size. Hence a timepoint for the IA efficiently taking advantage of the enrichment part of the trial is indeed after about half of the patients are observed. This is in line with findings of Benner and Kieser (2017), who analyzed the optimal timepoint of an IA in extensive simulations. It gives another reason for executing the blinded sample size review as early as possible, because determining the correct final sample size early makes sure that the optimal timepoint for the IA can be set properly. Imagine a late sample size review reducing the final sample size in such a way that at that point there are already 60–70% of the patients recruited having passed the best timepoint for the IA.

There is another variation related to the timing of the BSSR and the IA. One might be tempted to perform the sample size review and the IA at the same timepoint, for example, after half of the trial is completed. Instead of two stops there would only be one stop for the IA and a sample size reassessment. To take a look at this, we simulated and compared these two competing testing strategies (Figure 6). We remain in the one subgroup setting with prevalence of the subgroup $\tau = 0.4$. There is only an effect of $\Delta_S = 0.75$ in the subgroup. For initial sample size calculation, it is assumed that $\sigma^*_{F\setminus S} = \sigma^*_S = 1$ while the true variance in the subgroup varies on the x-axis. Sample size calculation and recalculation is performed assuming unequal and unknown variances, hence using the CEF approach with a suitable approximation as before. The nominal power is 80% and the number of simulation runs is 10,000.

On the one hand, a BSSR is performed after 30% of the initial sample size is recruited and then an IA after 50% of the final recalculated sample size is observed (black lines). We will refer to that as the two-stop strategy. On the other hand, both sample size adjustment and IA take place after 50% of the initial sample size (red lines). We call this the one-stop strategy. In the top left panel of Figure 6, the results for the power to reject the intersection hypothesis are shown. We still present both selection rules selecting both (dashed lines) or only one (solid lines) population at the IA. The method performing the sample size review and the IA at the same timepoint has a slightly higher power than the corresponding method with two separate stops, especially for scenarios where the mean recalculated sample size is small ($\sigma_S = 0.8$). This occurs because with only one stop midway the BSSR cannot correct the timing of the IA for the initially too large calculated sample size (perform it earlier). Hence, it might happen, that at the interim stop, there are already more subjects recruited than finally needed. Though it might not be very likely that a sample size adjustment in a real trial would reduce the planned sample size by large amounts, these cases emphasize the difference between the two approaches, that is, being able or not being able to change the timepoint of the IA based on early data. This reflects in the mean recalculated sample sizes (top right panel), which are larger than in the case of only one stop. On the other side, if the initial assumption of σ_S was too low ($\sigma^*_S = 1$ vs. $\sigma_S = 1.2, 1.4, 1.6$), an early BSSR would increase the final sample size resulting in a later IA. The strategy with only one fixed stop cannot postpone the IA to a later timepoint and therefore performs subgroup selection earlier than the two-stop strategy. Since the true effect is actually in the subgroup, this results in the same power with slightly lower mean recalculated sample sizes when always continuing with only one population (earlier opportunity to enrich). The bottom panels show the variability of the recalculated final sample sizes and the probabilities to (correctly) select the subpopulation and enrich the trial. The lower SDs of the one-stop strategy when

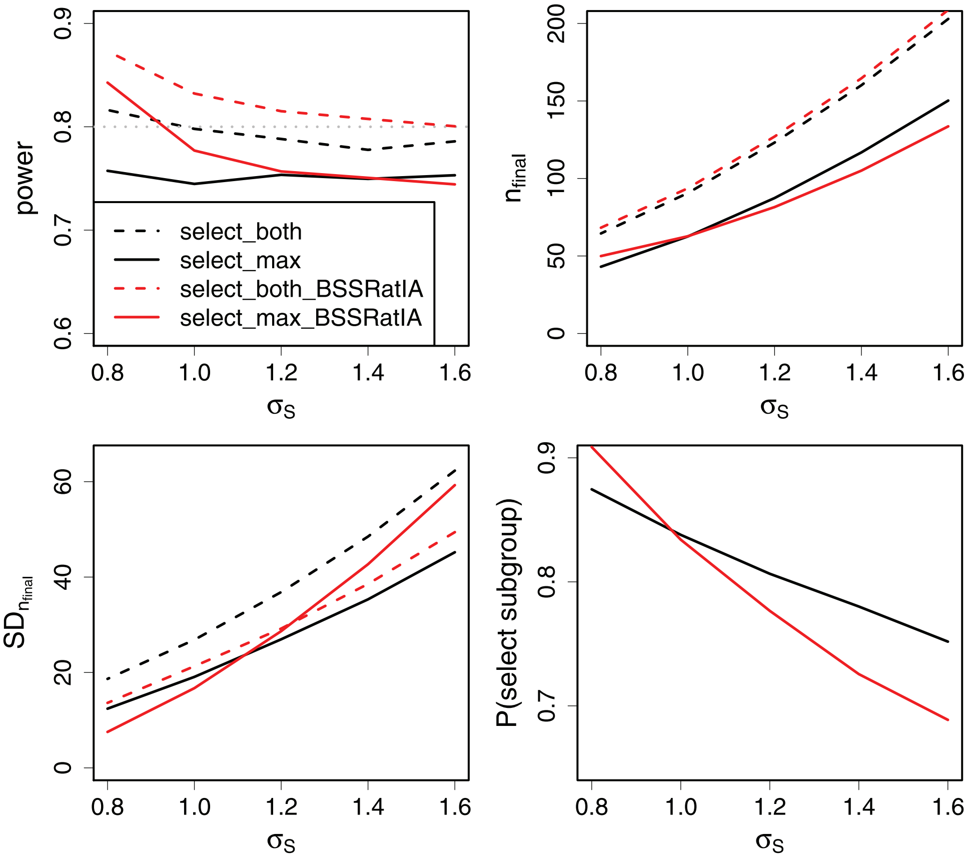

FIGURE 6 Comparison between two testing strategies: Perform the interim analysis (IA) and BSSR at the same time point ($t_1 = 0.5$) (red) or perform the IA after $t_1 = 0.5$ of the recalculated sample size obtained from a BSSR at $t_{IPS} = 0.3 \cdot N_0$ (black). Prevalence $\tau = 0.4$, $\Delta_S = 0.75$. Assumed variance in the subpopulation is $\sigma_S^* = 1$ while the true variance varies on the x-axis.

always continuing with both populations are due to the fact that the sample size recalculation in the two-stop design is performed earlier resulting in worse variance estimators. The biggest difference can be seen between the two strategies when always choosing the population with the maximum test statistic at interim. Performing sample size review and IA at the same fixed timepoint results in a notably larger variability in the final recalculated sample size as well as lower chances to select the correct population at interim, especially in case of a larger than expected variability in the subgroup. From that point of view the strategy performing an early BSSR and adjusting the timepoint of the IA is favorable. Therefore, someone who is quite confident about the nuisance parameters prior to the trial might do well using the strategy, which performs sample size reassessment and the IA at a fixed timepoint midway of the initially calculated sample size. If there is a lot of uncertainty about these parameters, one might prefer timing the IA depending on an IPS. We have seen that efficiency gains using the approach with two different timepoints for BSSR and IA are subtle but can be observed in realistic settings, for example, scenarios in which the BSSR suggests a larger sample size as preplanned and hence the IA is postponed until 50% of the recalculated sample size is available. This is reasonable since we have seen in Figure 5 that the optimal timepoint for the IA in terms of power is at 40–50% of the final sample size. Therefore, adjusting the timepoint of the IA based on the recalculated sample size is beneficial. We also presented other metrics to measure efficiency like mean recalculated sample size, variability of the recalculated sample size, and probability to choose the correct population at the IA. Especially when having to increase the sample size the latter is favorable for the two-stop approach.

6 | DISCUSSION

In this paper, we presented an adaptive testing strategy that incorporates BSSR into adaptive enrichment designs. This does not only lead to a robust design against missspecifications of nuisance parameters at the planning stage of the trial but also improves the optimal timing of the IA during the study. In the framework of normally distributed endpoints and nested subgroups, potentially inhomogeneous with regard to variability and treatment effect, we gave estimators, approximations, and an algorithm for BSSR as well as a procedure applying the CEF approach in a two-stage design with an IA. We assessed the performance by simulations in terms of power, type I error rates, sample size, and variability of the final sample size for different selection rules and included a comparison with a design without BSSR. Familywise type I error rate control in the strong sense is given since both components of the suggested procedure, BSSR and adaptive enrichment, are independent and each independently controls the FWER in the strong sense as we have shown in previous papers (Placzek & Friede, 2018, 2019). The proposed method maintains the desired power if planning assumptions were inaccurate and reduces the sample size and variability of the final sample size when an enrichment is performed. Obviously, the greatest benefit can be seen if the true treatment effect is indeed in a subpopulation, which is then enriched over the course of the trial. Furthermore, we investigated the optimal timepoint of the IA and found it is around 40–50% of the final sample size. We then pointed out the benefits of having two separate timepoints for blinded sample review and IA improving the timing of the latter and increasing the probability to correctly select and enrich a subgroup. That way the BSSR may prevent an early or late IA in terms of maximizing power.

Here, to simplify notation, we considered nested subgroups in a balanced design. A generalization to unbalanced designs and nonnested subgroups should be fairly straigthforward, since the used CEF approach by Placzek and Friede (2019) has been described more generally and the BSSR procedure transfers easily to these settings.

Performance of the designs was comparatively assessed in different simulation scenarios considering type I error rate, power, expected total sample size, and variability of the sample size. We chose these performance indicators, since they are common metrics for adaptive designs. However, trial duration might be an additional metric of interest, especially in this setting, since both adaptations, BSSR and enrichment, can increase or decrease the trial duration. Naturally, there is a linear relationship between total sample size and trial duration. Hence, BSSR increasing the final sample size will result in an increased trial duration while the same is true when decreasing the sample size. The enrichment aspect of the trial can introduce another variability in trial duration. If it is decided to continue the second stage of the trial only with patients from a particular population, for example, a promising subgroup, subsequent recruitment is most likely slowed down increasing the trial duration. This may be counterbalanced by fewer patients needed, attaining the same statistical power due to the enrichment, and therefore decreasing the trial duration. Trial duration as a metric is more common in event-driven trials, for example, time to first analysis was considered and discussed by Asikanius et al. (2016) in an event-driven trial comparing different strategies to decide on the set of final hypotheses.

Friede et al. (2019) assess the operating characteristics, including trial duration, of a BSSR procedure in an event-driven trial and compare them with those of a fixed sample size design. There are additional benefits from repeated interim looks, although one might argue performing multiple looks would increase costs in terms of time and additional work for the statistician. However, experience shows that cleaning the data two or more times in preparation of the BSSR or the IA actually benefits the trial, since the trial statistician gets familiar with the data and a more continuous monitoring of data quality is stimulated. This improves data quality throughout the trial and saves time at the actual IA or final analysis. Hence, potential cost or time concerns are at least partially compensated. The idea of multiple BSSRs, since type I error rate is not affected, was taken to an extreme by Friede and Miller (2012), who considered blinded continuous monitoring and found that the sample size variability is reduced compared to a sample size recalculation in a single or repeated interim look while the expected sample size stays the same. Of course this advantage has to be balanced with the more complex implementation of such an approach.

Since data are unblinded at the IA, it is tempting to perform additional sample size adjustments at that timepoint in an unblinded fashion. There are two kinds of those recalculations. On the one hand, they can be based only on unblinded estimates of nuisance parameters not taking into account the observed treatment effects at the IA. In this context, Friede and Kieser (2013) compared sample size reestimation based on blinded versus unblinded variance estimators and showed that the unblinded method does not guarantee that the desired power is achieved. Especially in the case of a small IPS, that is, an early sample size review, a high variability in the unblinded estimators can lead to a power lower than the nominal level. Intuitively, the bias of blinded estimators in case of treatment group differences should be disadvantageous. Here, however, the one-sample variance estimator that we were using overestimates the within-group variance. This leads to larger recalculated sample sizes, which is actually beneficial compensating the small power loss that may be introduced by recalculating the sample size during an ongoing trial as mentioned above, in particular with high variability early in the trial. The amount of excess in sample size compared to the required sample size in a fixed design does not depend on the fixed sample size as we have shown in Placzek and Friede (2018), which confirmed the findings of Friede and Kieser (2001). It rather depends on the prevalence of the subgroup or the timing of the IPS, for example, ranging from

30 more subjects per treatment group to only eight more subjects varying the prevalence of the subgroup from 0.2 to 0.8 in a one-subgroup design. Furthermore, in contrast to the unblinded method the blinded method does not inflate the type I error rate. Therefore, we went with the authors recommendation to use BSSR throughout this paper. On the other hand, unblinded sample size recalculation can be based on both unblinded estimates of the nuisance parameters as well as observed treatment effects. For example, if there is only weak evidence of a positive treatment effect at interim in a trial where the sample size was set to detect a treatment effect, which is larger than the clinically relevant effect, it might be reasonable to increase the sample size to detect at least a clinically relevant treatment effect. Note that increasing the sample size to detect an effect smaller than that would undermine the credibility of the trial. In any way, if the sample size recalculation is based on the observed treatment effects, the type I error rate can be inflated to more than two times the size than initially planned as demonstrated by Proschan and Hunsberger (1995). Appropriate statistical adjustment is needed. Wassmer (2000) summarizes and reviews publications in which the design is changed in response to interim results according to either prespecified rules or in an unplanned way. Those include variance spending (Jennison & Turnbull, 2003) as well as alpha spending approaches. Naturally, the CEF approach discussed here enables effect-based sample size adjustments while controlling the type I error rate. Such sample size adjustments based on conditional power arguments go back to Proschan and Hunsberger (1995) and are explored in a great variety nowadays (Denne, 2001; Kieser, 2020).

Promising zone designs are trials with a prespecified zone for the interim test statistic in combination with a decision rule for increasing the sample size in case the interim test statistic lies within this promising zone. Example of such designs were given by Mehta and Pocock (2011) along with strategies for preserving the type I error rate. Choosing the promising zone and the corresponding sample size adjustment rule in an optimal way was discussed by Hsiao et al. (2019).

There are still major problems related to effect-driven sample size recalculations. Those problems include not achieving the desired power, large recalculated sample sizes, and a high variability in the recalculated sample size due to large variability of the observed interim effect (Bauer & Koehne, 1994; Levin et al., 2013).

Adaptive designs, and BSSR procedures in particular, are still a hot topic and consequently subject to recently published and active research with focus on a variety of different design aspects. For example, a recent work by Friede et al. (2020) presents a framework on adaptive seamless designs, designs that, for example, combine phase II and phase III characteristics such as treatment or subgroup selection and confirmatory testing (Friede et al., 2020). They provide methods with IAs informed by either the primary outcome or an early outcome and highlight an extension of the `R` package `asd` to include adaptive enrichment designs (Parsons et al., 2012).

Methods for BSSR in more complex designs that were recently considered include multitreatment crossover trials (Grayling et al., 2018a), stepped-wedge cluster randomized trials (CRTs) (Grayling et al., 2018b), and multicenter randomized controlled clinical trials based on noncomparative data (Harden & Friede, 2020).

In this paper, we considered normally distributed endpoints. However, the ideas presented for those endpoints can be transferred to other endpoints, for example, binary, survival, or other event-based outcomes. Depending on the type of outcome the nuisance parameters involved in the model change (overall proportion, event rates, censoring rate). For example, Asendorf et al. (2019) consider BSSR in clinical trials with longitudinal negative binomial counts. Here the nuisance parameters are the overall rate and the shape parameter of the negative binomial distribution. Concerning the adaptive enrichment part of the procedure, the analysis methods have to be adjusted according to the outcome. The CEF principle does not depend on a particular distribution and can still be applied.

ACKNOWLEDGMENTS

This work was supported by the BMBF (Federal Ministry of Education and Research) under grant 05M13MGE BIMIT and the DZHK (German Center for Cardiovascular Research) under grant GOE SI2 UMG Information and Data Management. We would like to thank the associate editor and the two reviewers for their constructive comments that helped to improve the paper. Furthermore, we would like to thank Roland Gera for carefully proofreading the paper.

Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Not applicable

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

Marius Placzek https://orcid.org/0000-0002-8663-5378
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

Asendorf, T., Henderson, R., Schmidli, H., & Friede, T. (2019). Sample size re-estimation for clinical trials with longitudinal negative binomial counts including time trends. *Statistics in Medicine*, *38*, 1503–1528.

Asikanius, E., Rufibach, K., Bahlo, J., Bieska, G., & Burger, H. (2016). Comparison of design strategies for a three-arm clinical trial with time-to-event endpoint: Power, time-to-analysis, and operational aspects. *Biometrical Journal*, *58*, 1295–1310.

Bauer, P., & Koehne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, *50*, 1029–1041.

Benner, L., & Kieser, M. (2017). Timing of the interim analysis in adaptive enrichment designs. *Journal of Biopharmaceutical Statistics*, *28*, 622–632.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., & Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, *28*, 1445–1463.

Bretz, F., Schmidli, H., Koenig, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, *48*, 623–634.

Chiu, Y., Koenig, F., Posch, M., & Jaki, T. (2018). Design and estimation in clinical trials with subpopulation selection. *Statistics in Medicine*, *0*, 1–18.

Denne, J. (2001). Sample size recalculation using conditional power. *Statistics in Medicine*, *20*, 2645–2660.

European Medicines Agency (EMEA) (2007). *Reflection paper on methodological issues in conformatory clinical trials planned with an adaptive designs*. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf. Accessed: 2020-03-11.

Food and Drug Administration (FDA) (2006). *Adaptive designs for medical device clinical studies*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-designs-medical-device-clinical-studies. Accessed: 2020-09-20.

Food and Drug Administration (FDA) (2019). *Adaptive designs for clinical trials of drugs and biologics: Guidance for industry*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry. Accessed: 2020-03-11.

Friede, T., & Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, *20*, 3861–3873.

Friede, T., & Kieser, M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics*, *12*, 141–146.

Friede, T., & Miller, F. (2012). Blinded continuous monitoring of nuisance parameters in clinical trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*, 601–618.

Friede, T., Parsons, N., & Stallard, N. (2012). A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine*, *31*, 4309–4320.

Friede, T., Pohlmann, H., & Schmidli, H. (2019). Blinded sample size reestimation in event-driven clinical trials: Methods and an application in multiple sclerosis. *Pharmaceutical Statistics*, *18*, 351–365.

Friede, T., Stallard, N., & Parsons, N. (2020). Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. *Biometrical Journal*, *62*, 1264–1283.

Graf, A., Wassmer, G., Friede, T., Gera, R., & Posch, M. (2019). Robustness of testing procedures for confirmatory subpopulation analyses based on a continuous biomarker. *Statistical Methods in Medical Research*, *28*, 1879–1892.

Grayling, M., Mander, A., & Wason, J. (2018a). Blinded and unblinded sample size reestimation in crossover trials balanced for period. *Biometrical Journal*, *60*, 917–933.

Grayling, M., Mander, A., & Wason, J. (2018b). Blinded and unblinded sample size reestimation procedures for stepped-wedge cluster randomized trials. *Biometrical Journal*, *60*, 903–916.

Gurka, M., Coffey, C., & Gurka, K. (2010). Internal pilots for observational studies. *Biometrical Journal*, *52*, 590–603.

Harden, M., & Friede, T. (2020). Sample size recalculation in multicenter randomized controlled clinical trials based on noncomparative data. *Biometrical Journal*, *62*, 1284–1299.

Hsiao, S., Liu, L., & Mehta, C. (2019). Optimal promising zone designs. *Biometrical Journal*, *61*, 1175–1186.

International Conference on Harmonisation E9 Expert Working Group. (1999). ICH harmonised tripartite guideline. statistical principles for clinical trials. *Journal of Neurology*, *18*, 1905–1942.

Jenkins, M., Stone, A., & Jennison, C. (2011). An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics*, *10*, 347–356.

Jennison, C., & Turnbull, B. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, *22*, 971–993.

Kieser, M. (2020). Sample size recalculation based on conditional power. In *Methods and applications of sample size calculation and Recalculation in clinical trials*, Springer International Publishing (Ch. 28, pp. 271–277).

Kieser, M., & Wassmer, G. (1996). On the use of the upper confidence limit for the variance from a pilot sample for sample size determination. *Biometrical Journal*, *38*, 941–949.

Levin, G., Emerson, S., & Emerson, S. (2013). Adaptive clinical trial designs with pre-specified rules for modifying the sample size: Understanding efficient types of adaptation. *Statistics in Medicine*, *32*, 1259–1275.

Marcus, R., Eric, P., & Gabriel, K. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, *63*, 655–660.

Mehta, C., & Pocock, S. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, *30*, 3267–3284.

Mehta, C., Schafer, H., Daniel, H., & Irle, S. (2014). Biomarker driven population enrichment for adaptive oncology trials with time to event endpoints. *Statistics in Medicine*, *33*, 4515–4531.

Mueller, H., & Schaefer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, *57*, 886–891.

Ondra, T., Dmitrienko, A., Friede, T., Graf, A., Miller, F., Stallard, N., & Posch, M. (2016). Methods for identification and confirmation of targeted subgroups in clinical trials: A systematic review. *Journal of Biopharmaceutical Statistics*, *26*, 99–119.

Parsons, N., Friede, T., Todd, S., Marquez, E., Chataway, J., Nicholas, R., & Stallard, N. (2012). An R package for implementing simulations for seamless Phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics & Data Analysis*, *56*, 1150–1160.

Placzek, M., & Friede, T. (2018). Clinical trials with nested subgroups: Analysis, sample size determination and internal pilot studies. *Statistical Methods in Medical Research*, *27*, 3286–3303.

Placzek, M., & Friede, T. (2019). A conditional error function approach for adaptive enrichment designs with continuous endpoints. *Statistics in Medicine*, *38*, 3105–3122.

Proschan, M., & Hunsberger, S. (1995). Designed extension of studies based on conditional power. *Biometrics*, *51*, 1315–1324.

Rosenblum, M., Luber, B., Thompson, R., & Hanley, D. (2016). Group sequential designs with prospectively planned rules for subpopulation enrichment. *Statistics in Medicine*, *35*, 3776–3791.

Rosenblum, M., Qian, T., Du, Y., Qiu, H., & Fisher, A. (2016). Multiple testing procedures for adaptive enrichment designs: Combining group sequential and reallocation approaches. *Biostatistics*, *17*, 650–662.

Schmidli, H., Bretz, F., Racine, A., & Maurer, W. (2006). Confirmatory seamless Phase II/III clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biometrical Journal*, *48*, 635–643.

Sorkness, C., King, T., Dyer, A., Chinchilli, V., Mauger, D., Krishnan, J., Blake, K., Castro, M., Covar, R., Israel, E., Kraft, M., Lang, J., Lugogo, N., Peters, S., Wechsler, M., Wenzel, S., & Lazarus, S. (2019). Adapting clinical trial design to maintain meaningful outcomes during a multicenter asthma trial in the precision medicine era. *Contemporary Clinical Trials*, *77*, 98–103.

Stallard, N., Hamborg, T., Parsons, N., & Friede, T. (2014). Adaptive designs for confirmatory clinical trials with subgroup selection. *Journal of Biopharmaceutical Statistics*, *24*, 168–187.

Sugitani, T., Posch, M., Bretz, F., & Koenig, F. (2018). Flexible alpha allocation strategies for confirmatory adaptive enrichment clinical trials with a prespecified subgroup. *Statistics in Medicine*, *37*, 3387–3402.

Wassmer, G. (2000). Basic concepts of group sequential and adaptive group sequential test procedures. *Statistical Papers*, *41*, 253–279.

Wassmer, G., & Dragalin, V. (2015). Designing issues in confirmatory adaptive population enrichment trials. *Journal of Biopharmaceutical Statistics*, *25*, 651–669.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*, 65–72.

Zucker, D., Wittes, J., Schabenberger, O., & Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, *18*, 3493–3509.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Placzek, M., & Friede, T. (2023). Blinded sample size recalculation in adaptive enrichment designs. *Biometrical Journal*, *65,* 2000345. https://doi.org/10.1002/bimj.202000345