Blinded sample size recalculation in multiple composite population designs with normal data and baseline adjustments

Roland G. Gera | Tim Friede

**Abstract**

The increasing interest in subpopulation analysis has led to the development of various new trial designs and analysis methods in the fields of personalized medicine and targeted therapies. In this paper, subpopulations are defined in terms of an accumulation of disjoint population subsets and will therefore be called composite populations. The proposed trial design is applicable to any set of composite populations, considering normally distributed endpoints and random baseline covariates. Treatment effects for composite populations are tested by combining \( p \)-values, calculated on the subset levels, using the inverse normal combination function to generate test statistics for those composite populations while the closed testing procedure accounts for multiple testing. Critical boundaries for intersection hypothesis tests are derived using multivariate normal distributions, reflecting the joint distribution of composite population test statistics given no treatment effect exists. For sample size calculation and sample size, recalculation multivariate normal distributions are derived which describe the joint distribution of composite population test statistics under an assumed alternative hypothesis. Simulations demonstrate the absence of any practical relevant inflation of the type I error rate. The target power after sample size recalculation is typically met or close to being met.

**KEYWORDS**

multiple testing, \( p \)-value combination, sample size calculation, sample size reestimation, subpopulation analysis

**INTRODUCTION**

In recent years, there has been an increased interest in evaluating the treatment effect across heterogeneous subpopulations (Basu et al., 2017; Kent et al., 2010; Varadhan et al., 2013). This approach has been coined **precision or personalized medicine**. One point where the pooling of treatment effects coming from diverse subpopulations might lead to issues is...
when several distinct treatment effects are averaged. The evaluation of such an averaged treatment effect might lead to a situation where a beneficial treatment effect for one subpopulation is masked by subpopulations in which the treatment benefit is not as high. This emphasizes the need for trial designs that aim to test treatment efficacy not only in the full population but in subpopulations as well.

One view of a subpopulation is to define it as an amalgamation of population subsets. These subsets are uniquely defined using biomarkers, resulting in numerous disjoint subsets. Since subsets are combined to form subpopulations, we call subpopulations which were derived by such an amalgamation “composite populations.” Common examples for biomarkers include genetic markers but also simple patient characteristics such as age or gender. In relation to the outcome, a biomarker can be prognostic and therefore describe natural differences in the overall outcome between patients in different subsets and/or predictive, where a biomarker indicates that a treatment has different efficacy in different subsets (Jenkins, Flynn, et al., 2011). If at the planning stage for a trial heterogeneous reactions in regard to the treatment are suspected, be it by subpopulation-specific efficacies, safety profiles, or any other clinically relevant behavior, it makes sense to divide the study population into subpopulations where similar responses towards the treatment are expected. These populations should be predefined prior to the study (EMA, 2019; Tanniou et al., 2016, 2019).

Incorporating heterogeneous study populations in a trial is not trivial. One challenge is the interpretation of study results. The Points to consider for multiplicity issues of clinical trials by the EMA (2016), for example, points out that, from a regulatory perspective, a positive result in the full/overall population does not necessarily mean that this positive result is applicable in all subpopulations present in the overall population. At least as long as there is reason to expect heterogeneity in the respective subpopulations. Another point to look out for when planning a study is how relevant covariates can be included in the study. Both, the Guideline on adjustment for baseline covariates in clinical trials (EMEA, 2003) as well as the ICH E9, support the inclusion of a priori defined baseline covariates in the analysis in cases where strong or moderate association between the covariates and the outcome is assumed.

Lastly, sample size calculation is an integral part of any trial planning phase. The sample size should be large enough to detect a certain relevant effect size while avoiding unnecessarily large sample sizes out of ethical, time, and economical considerations. However, the sample size is dependent on many parameters which are necessary to define the final sample size. Many of those parameters are often not of actual research interest and are therefore called nuisance parameters (Tavernier & Giraudeau, 2015). Estimates for any parameter are often based on either previous data or by making educated guesses about the magnitude of the true parameters. When considering a trial with multiple composite populations and population subsets where the effect of a new treatment has to be determined, common nuisance parameters are the covariances between the endpoint and each covariate as well as the variance of the endpoint itself. These parameters have to be defined for each subset. A misspecification can lead to an inadequately sized study, either being too small and therefore leading to a study with lacking power or being too large and wasting the time and money of researchers and subjects alike. A sample size review during an internal pilot study can provide a solution to this problem (Wittes & Britain, 1990). The basic principle of a sample size review is to look into the data at an earlier stage of the trial, to reestimate the nuisance parameters based on the available data and to recalculate the required sample size. Generally, regulatory agencies prefer a blinded (or noncomparative) sample size review, that is, without unveiling of the treatment allocation, over unblinded procedures (EMEA, 2007; ICH E-9 Expert Working Group, 1999). When treatment allocations remain concealed, the number of potential sources of bias is reduced. The editorial by Julious (2015) gives a broad overview of works in regard to pilot studies.

Already, methods accounting for some, but not all, highlighted challenges have been presented. An approach by Mehta et al. (2014) describes a trial design for survival endpoints in cancer trials. There only subsets were tested. Since test statistics are derived from disjoint subsets, those test statistics are independent from each other which makes controlling the family-wise type I error rate (FWER) a simple matter. While this approach is quite accessible, it is only appropriate for scenarios where testing is done on independent test statistics. In scenarios, however, where one or more subsets are part of two or more composite populations, this type of testing is not advised. Ignoring the correlation between test statistics, which occurs if the same subset is part of more than one subpopulation, will lead to less efficient tests. Using our method, trial designs with disjoint subsets as well as composite populations with any subset constellation can be considered.

Graf et al. (2019) investigated the issue of how to efficiently identify subpopulations in the context of a continuous biomarker. There, populations were defined using increasing biomarker thresholds to identify biomarker-positive subjects. Hence, subjects whose biomarker levels were below the threshold were grouped into the biomarker-positive subgroup and the treatment effect was then evaluated in the said subpopulation. With increasing thresholds, more and more subjects are selected for the biomarker-positive subpopulation. Each subpopulation encompasses all subjects already identified by previous thresholds. This population setting is called nested. Since typically several biomarker thresholds are tested,
the repeated testing raises the issue of multiple testing while at the same time resembling the structure of a group sequential trial. In the paper, various procedures which account for multiple testing were investigated and their operating characteristics were compared.

Placzek and Friede (2018) presented a trial design to evaluate treatment effects in nested subpopulations while also including a sample size calculation and blinded sample size recalculation scheme, considering continuous normally distributed endpoints. However, the trial design described in their paper does not apply to nonnested subpopulations and does not include the use of covariates.

In contrast, the method by Chiu et al. (2018) shows an approach for testing multiple subpopulations at once. Additionally, any type of population setting can be applied to this design. To achieve this flexibility, disjoint subsets are considered which are then combined into subpopulations using the inverse normal combination function. However, only normally distributed endpoints with equal and known variances in each subset are considered while also missing the use of covariates in the analysis. The testing approach presented in this paper is similar to the one by Chiu et al. but allows for varying variances and correlations in population subsets and includes covariates in the analysis while additionally showing how an effective sample size recalculation can be conducted.

On the topic of sample size calculation, Friede and Kieser (2011) presented a method to redetermine the nuisance parameter during an internal pilot study and use this to conduct a blinded sample size recalculation for analysis of covariance (ANCOVA) designs with one normally distributed covariate. But while this approach was later extended by Zimmermann et al. (2020) to account for multiple normally distributed covariates, it is not applicable for the subpopulation design. Here, we combine these ideas to blinded sample size recalculation within multiple composite population settings.

This paper is structured as follows. In Section 2, we motivate our design by an example on pulmonary arterial hypertension. Section 3 describes our approach for constructing the statistical model and hypothesis testing using any composite population structure, while also presenting the sample size calculation and recalculation schemes. The application of our method for testing and sample size estimation is shown using a numerical example, inspired by the motivating example. Simulation results are presented in Section 4. The paper closes with our findings, a brief discussion, and conclusions in Section 5.

## 2 MOTIVATING EXAMPLE

Pulmonary hypertension (PH) is a rare, chronic, and progressive disorder, which often occurs with nonspecific symptoms like shortness of breath, fatigue, swelling of the legs, swelling of the ankles, chest pain, and light-headedness. These symptoms are caused by an increased blood pressure in the pulmonary circulation system. Over time, this increased strain of the right heart chambers leads to a failure of the heart which ultimately leads to death if the illness remains untreated (Montani et al., 2013). While the early symptoms seem mild and often only occur during physically demanding exercise, it is important to treat PH as early as possible. New treatments for this illness are being developed frequently, requiring novel and efficient trial designs to accompany these trials. Grieve et al. (2014) summarize results of a workshop where novel designs and other design options for PH trials were discussed. Some of the highlighted methods include blinded sample size recalculation, subpopulation analysis, and population enrichment designs. Especially the last two options lent themselves very well to trials in PH, since currently five main subtypes of PH are recognized which may be divided further. Simonneau et al. (2019) describe these categories, which have been defined in more detail during the latest world symposium on PH. These main categories were (1) pulmonary arterial hypertension (PAH), (2) PH due to left heart disease, (3) PH due to lung diseases and/or hypoxia, (4) PH due to pulmonary artery obstructions, and, lastly, (5) PH with unclear and/or multifactorial mechanisms. Every single main category can be further partitioned into smaller population proportions which are demonstrated here for the first main category, (1) pulmonary arterial hypertension. Here, each partition is defined by the cause which resulted in the development of PAH. Those causes are identified as idiopathic PAH, heritable PAH, PAH linked to drugs or toxins, PAH associated with illnesses, PAH due to long-term responders to calcium channel blockers, PAH with overt features of venous/capillaries involvement, and persistent PH of the newborn syndrome. While most causes are specific and well defined, PAH associated with illnesses is a generic term encompassing five subordinate illnesses, like connective tissue disorder or HIV. For details on the population structure, see Table 1.

Using the PAH example, we can demonstrate the concepts of subsets and composite populations. A subset is defined as the smallest partitioning of a population, which is not subdivided further and where the subjects are expected to react to the treatment uniformly. In the present case, idiopathic PAH would be defined as a subset while PAH associated with
illness would not since the latter splits itself into five subsets. To what level of detail the consideration of a subset is sensible always depends on the context, of course.

A composite population, on the other hand, then describes a (sub-)population of interest on which a formal test is performed. Such a population may be defined as a single subset, or as the union of several subsets. Using this combination approach, it is possible to reflect essentially any kind of population setting; in particular, this allows the consideration of hierarchical or overlapping structures. Hence, a researcher is free to test any combination of subsets during a trial while using the trial methods we describe in this paper.

McLaughlin et al. (2009) demonstrate cases in the PAH context where the consideration of differential treatment effects in different subsets is in fact crucial. They conclude that within trials investigating the efficacy of interventions a majority of patients suffered from idiopathic PAH, and treatment effects appeared most pronounced in this prominent subset. For patients affected by other PAH subsets, the beneficial treatment effect appeared to be smaller, if present at all. In regard to the interpretation of treatment effects, it is hence crucial to also take subsets and composite population associations into consideration. For further information, see McLaughlin et al. (2009). Besides the problem of heterogeneous treatment effects, McLaughlin et al. also highlight that many commonly used primary endpoints considered during trials for PH, such as the 6-min walk test, are influenced by other factors such as age and height. Overall, composite population testing allows for flexible testing strategies when multiple subgroups are considered and often random covariates can be identified which are correlated to the study outcomes and should be included in the study.

3 | DESIGN

3.1 | Statistical model

We consider a two-arm randomized controlled trial design where an active treatment (T) is tested against a comparator treatment (C). The total sample size of subjects in the full study population is denoted as $N$, while the treatment allocation is denoted by $\kappa = N_T/N_C$, with $N_T$ and $N_C$ representing the sample sizes for the active and control treatments. We further assume that the overall population ($F$) consists of $J$ disjoint subsets $S_j$ with subset sample sizes given by $N_{T,j}$ and $N_{C,j}$, respectively. The subset prevalences are independent of the treatment allocation and are defined by $\tau_j = N_{T,j}/N_T = N_{C,j}/N_C$ for $j = 1, \ldots, J$.

Based on subsets $S_j$, we can define composite populations allowing for arbitrary population structures within the full population. Let $R$ denote the total number of composite populations which are defined by $G_r = \bigcup_{j \in I_r} S_j$ for a corresponding index set $I_r$ for $r = 1, \ldots, R$. For instance, the composite population PAH associated with illnesses is represented by $G_1$ defined by index set $I_1 = \{S_{connective tissue disease}, S_{HIV}, S_{portal hypertension}, S_{congenital heart disease}, S_{schistosomiasis}\}$ (see Table 1).

Then, for our study design we assume that normally distributed outcomes $Y_{ji}$, with $i = T, C$, as well as $D$ random covariates $X_{jd}, d = 1, \ldots, D$, were measured for each subject in each subset. With regard to the covariates, no assumptions about their distributions are made. Let $\sigma^2_{Y_{ji}}$ denote the variance of the outcome, $\Sigma_{X_{jd}}$ the covariance matrix among all covariates and $\sigma_{Y,X_{jd}}$ the vector of covariances between the outcome and each covariate. It is assumed that variances
and covariances are the same for the treatment and control groups within any given subset \( j \); however, variances and covariances can vary between subsets. Let \( \mu_{Tj} \) and \( \mu_{Cj} \) represent the true means of the covariate in treatment and of the control group. Naturally, due to randomization, it follows that \( \mu_{Tj} = \mu_{Cj} = \mu_{jd} \). Then, a linear model is evaluated in subset \( j \) so that

\[
Y_{ji} = \delta_j U_i + \beta_{0j} + \beta_{1j} X_{j1} + \cdots + \beta_{Dj} X_{jd} + \epsilon_j,
\]

where \( \delta_j \) denotes the benefit a treatment imparts to patients in the treatment arm on average in subset \( j \), while treatment indicator \( U_i \) equals 1 for \( i = T \) and takes the value of 0 otherwise. The error terms are independent from each other and follow the distribution \( \epsilon_j \sim N(0, \sigma^2_\epsilon) \) in any given subset \( j \).

In this paper, the one-sided subset null hypothesis \( H_{0j} : \delta_j \leq 0 \) is tested against its alternative \( H_{A_j} : \delta_j > 0 \). Therefore, it is assumed that greater values of \( \delta_j \) are associated with an increased treatment benefit. The null hypothesis \( H_{0j} \) for treatment benefit \( \delta_j \) is tested using the test statistic

\[
T_j = \frac{\hat{\delta}_j}{S(\hat{\delta}_j)},
\]

which is a well-known formula as noted by Zimmermann et al. (2020). The equation to calculate \( \hat{\rho}^2_{Y_j, X_{j1} \cdots X_{jd}} \) was also described by Zimmerman et al. and is given by

\[
\hat{\rho}^2_{Y_j, X_{j1} \cdots X_{jd}} = S'_{Y_j X_{j1} \cdots X_{jd}} S_{Y_j X_{j1} \cdots X_{jd}} / S^2_{Y_j}.
\]

This coefficient measures the linear association between the variables \( X_{j1} \cdots X_{jd} \) and the outcome \( Y_j \) and is also often denoted as \( R^2 \).

The \( D \times 1 \) column vector \( (\bar{X}_{Tj} - \bar{X}_{Cj})' \) represents the vector \( (\bar{X}_{Tj1} - \bar{X}_{Cj1}, \ldots, \bar{X}_{Tjd} - \bar{X}_{Cjd})' \). However, since it is assumed that the true covariate means are independent from treatment, the vector \( (\bar{X}_{Tj} - \bar{X}_{Cj}) \) converges to a \( \mathbf{0} \) vector. Hence, \( (\bar{X}_{Tj} - \bar{X}_{Cj})' ((N_{Tj} + N_{Cj} - 2) \hat{\Sigma}_{X_j})^{-1} (\bar{X}_{Tj} - \bar{X}_{Cj}) \) will be ignored during the sample size calculation and recalculation phase. Similar justification was also given in the papers by Friede and Kieser (2011) and Zimmermann et al. (2020).

Given \( \delta_j = 0 \), the subset test statistic \( T_j \) follows a central \( t \)-distribution with \( N_{Tj} + N_{Cj} - 2 - D \) degrees of freedom. A \( p \)-value testing the one-sided \( H_{0j} \) is calculated by \( p_j = 1 - \Psi_{N_{Tj} + N_{Cj} - 2 - D}(T_j) \), where \( \Psi_{df} \) denotes the cumulative distribution function for a \( t \)-distribution with \( df \) degrees of freedom with the sole restriction that for each subset \( df \geq 1 \) must be true.

To derive test statistics for composite populations, \( p \)-value combination functions are applied. To merge multiple independent subset \( p \)-values, we use the inverse normal combination function (Lehmacher & Wassmer, 1999). Hence, a test statistic is needed to test the composite population null hypothesis \( H_{0G} \), against its alternatives. Birnbaum (1954) defines the null hypothesis for such composite populations as

\[
H_{0G} : \text{all } \delta_j \leq 0, j \in I_r
\]

\[
H_{AG} : \text{one or more } \delta_j > 0, j \in I_r.
\]

Hence, the test statistics for these composite null hypotheses above are calculated as

\[
Z_{Gr} = \sum_{j \in I_r} \left( \frac{w_j}{\sum_{k \in I_r} w_k} \Phi^{-1}(1 - p_j) \right).
\]

Here, \( \Phi^{-1}(\cdot) \) denotes the quantile function of the standard normal distribution. Let \( w_j \) define prefixed weights for subset \( S_j \) with \( w_j > 0 \) for all \( j = 1, \ldots, J \) and \( \sum_{j=1}^{J} w_j = 1 \). These weights are often a function of sample size, preestimated standard errors, and, if it is assumed that the same treatment allocation is applied in all subsets, the subset prevalences \( \tau_j \). Since \( p \)-values are derived for each subset separately and combined afterwards, parameters like variances and correlations can vary between subsets without impacting other subset \( p \)-values.
Invoking the \textit{p-clud} property as described by Brannath et al. (2009), it can be shown that if \( H_{0G_r} \) is true and the \textit{p-clud} property is satisfied, then the \( p \)-value for \( H_{0G_r} \) given by \( 1 - \Phi(Z_{G_r}) \) is stochastically larger than or equal to the uniform distribution. The \textit{p-clud} property is fulfilled if the distribution of the \( p \)-value \( p_j \), \( j \in I_r \), and the conditional distributions of \( p_{j'} \) given \( p_j \), \( j' \in I_r \ \backslash \ j \), are stochastically larger than or equal to the uniform distribution on \([0,1]\). Here, using the \( t \)-test statistic, this condition is always satisfied, since the subset \( p \)-values emerge from disjoint subsets and are therefore always independent from each other.

Since \( R \) composite populations are being tested, multiple testing needs to be accounted for by using the closed testing principle (Brannath et al., 2009; Marcus et al., 1976). In the context of the closed testing principle, \( H_{0G_r} \) is called an elementary null hypothesis. Starting from such elementary null hypotheses intersection, null hypotheses are defined by intersecting two or more elementary null hypotheses. As an example, the intersection hypothesis which includes all elementary null hypotheses is given by \( H_{0\cap \bigcap_{r=1}^{R} G_r} = \cap_{r=1}^{R} H_{0G_r} \). By hierarchically testing, each relevant intersection null hypotheses at level \( \alpha \) and testing the elementary null hypotheses only if all previous intersection null hypotheses have been rejected first, multiple testing can be accounted for. Hence, elementary null hypothesis \( H_{0G_r} \) can only be rejected at nominal level \( \alpha \), given that all intersection hypotheses containing the elementary null hypothesis \( H_{0G_r} \) were rejected at level \( \alpha \). Hence, for each (intersection) null hypothesis a new critical boundary is determined. To demonstrate the closed testing principle, we use an example where the elementary null hypotheses \( H_{0G_1}, H_{0G_2}, \) and \( H_{0G_3} \) are tested. Then, \( H_{0G_1} \) is only rejected at level \( \alpha \), if the intersection null hypotheses \( H_{0G_1 \cap G_2}, H_{0G_1 \cap G_3}, \) and \( H_{0G_1 \cap G_2 \cap G_3} \) have been rejected at level \( \alpha \) as well. During this procedure, each intersection hypothesis is tested with an affiliated critical boundary. The same approach is applied for testing \( H_{0G_2} \) and \( H_{0G_3} \).

One option for calculating the critical boundaries for each intersection null hypothesis is by using the Bonferroni or the Šidák correction (Bretz et al., 2006; Friede et al., 2012, 2020; Schmidli et al., 2006). These corrections, however, do not account for potential correlations between composite population test statistics \( Z_{G_r} \). The correlation between composite population test statistics occurs when one or more subsets are part of one or more composite populations. An approach that accounts for these correlations between completely nested, partially nested, or disjoint composite populations was presented by Spiessens and Debois (2010). There, a multivariate normal distribution is used to describe the joint distribution of the test statistics given a true intersection null hypothesis.

Using this approach, distributions for the test statistics under the null hypothesis and corresponding critical boundaries can be defined which fulfill any given \( \alpha \) level. By accounting for the covariances between test statistics, more efficient testing strategies emerge. For example, when testing intersection null hypothesis \( H_{0\cap \bigcap_{r=1}^{R} G_r} \) assuming that \( \delta_j = 0 \) for all \( j = 1, \ldots, J \), hence the edge of the intersection null hypothesis, the vector of composite population test statistics \( Z, Z = (Z_{G_1}, \ldots, Z_{G_R}) \) follows the distribution

\[
Z \sim \mathcal{N}_R(\mathbf{0}, \Sigma_0)
\]

with

\[
\Sigma_0 = \begin{pmatrix}
1 & \text{cov}(Z_{G_1}, Z_{G_2}) & \cdots & \text{cov}(Z_{G_1}, Z_{G_R}) \\
\text{cov}(Z_{G_1}, Z_{G_2}) & 1 & \ddots & \vdots \\
\vdots & \ddots & 1 & \text{cov}(Z_{G_{R-1}}, Z_{G_R}) \\
\text{cov}(Z_{G_1}, Z_{G_R}) & \cdots & \text{cov}(Z_{G_{R-1}}, Z_{G_R}) & 1
\end{pmatrix}
\]

Matrix \( \Sigma_0 \) denotes the covariance matrix between composite population test statistics. This distribution can then be used to calculate critical boundaries for \( H_{0\cap \bigcap_{r=1}^{R} G_r} \). The covariance between any two composite populations is solely dependent on the overlapping subsets that are present in both. To show the calculation of the covariances, let \( Z_{G_r} \) and \( Z_{G_{r'}} \) be two composite population test statistics. Then the covariance between these two test statistics is calculated as

\[
\text{cov}(Z_{G_r}, Z_{G_{r'}}) = \text{cov} \left( \sum_{j \in I_r} \sqrt{\frac{w_j}{\sum_{k \in I_r} w_k}} \Phi^{-1}(1 - p_j), \sum_{j' \in I_{r'}} \sqrt{\frac{w_{j'}}{\sum_{k' \in I_{r'}} w_{k'}}} \Phi^{-1}(1 - p_{j'}) \right).
\]
This expression can be simplified, however, by first separating the sums on both sides of the covariance into two proportions, the part where the same subset is shared by both composite populations and the part where subsets uniquely belong to only one composite population. Since the covariance between any two disjoint subsets or amalgamation of disjoint subsets is zero, using simple covariance rules we write

\[
\text{cov}(Z_{G_r}, Z_{G_{r'}}) = \sum_{l \in I_r \cap I_{r'}} \text{cov} \left( \sqrt{w_l \sum_{k \in I_r} w_k \Phi^{-1}(1-p_l)}, \sqrt{w_l \sum_{k' \in I_{r'}} w_{k'} \Phi^{-1}(1-p_l)} \right)
\]

\[
= \sum_{l \in I_r \cap I_{r'}} \sqrt{w_l \sum_{k \in I_r} w_k} \sqrt{w_l \sum_{k' \in I_{r'}} w_{k'}} \text{cov}(\Phi^{-1}(1-p_l), \Phi^{-1}(1-p_l))
\]

\[
= \sum_{l \in I_r \cap I_{r'}} \frac{w_l}{\sqrt{\sum_{k \in I_r} w_k \sum_{k' \in I_{r'}} w_{k'}}} \text{var}(\Phi^{-1}(1-p_l))
\]

\[
= \sum_{l \in I_r \cap I_{r'}} \frac{w_l}{\sqrt{\sum_{k \in I_r} w_k \sum_{k' \in I_{r'}} w_{k'}}}.
\]

The covariance between any two \(Z_{G_r}\) and \(Z_{G_{r'}}\) can therefore be calculated using (5). This in turn is then applied to calculate each entry in \(\Sigma_0\). This gives the covariance structure for the distribution of the intersection hypothesis. Then, given \(\alpha\) and the joint distribution under intersection null hypothesis \(R\) critical boundaries can be determined which fulfill

\[
\alpha \geq P_{H_{\omega_{\gamma=1}^R G_r}} \left( Z_{G_1} \geq c_1 \lor \cdots \lor Z_{G_{r'}} \geq c_{r'} \right).
\]  

(6)

There exists an infinite number of sets containing \(R\) critical boundaries which fulfill (6). For testing, the smallest equicoordinated critical boundary \(c_{\gamma=1}^{R G_r}\) is used to investigate \(H_{\omega_{\gamma=1}^R G_r}\). It can be found by solving

\[
\alpha \geq P_{H_{\omega_{\gamma=1}^R G_r}} \left( Z_{G_1} \geq c_{\gamma=1}^{R G_r} \lor \cdots \lor Z_{G_{r'}} \geq c_{\gamma=1}^{R G_r} \right).
\]  

(7)

This critical boundary is called equicoordinated, since the same quantile \(c_{\gamma=1}^{R G_r}\) is applied in (7). Naturally, for other intersection null hypotheses, new critical boundaries have to be derived following the same scheme.

Note that these critical boundaries are derived for the assumptions that \(\delta_j = 0\) \(\forall j = 1, \ldots, J\). This is called the edge of the null hypothesis since any other scenario covered by the null hypothesis is more extreme. Next, it has also to be shown that critical boundaries derived for \(\delta_j = 0\) \(\forall j = 1, \ldots, J\) also controls the type I error rate in cases where one or more subsets have treatment effects of \(\delta_j < 0\). Let \(\text{Var}(\Phi^{-1}(1-p_j))\) denote the variance of the transformed test statistic \(\Phi^{-1}(1-p_j)\) for subset \(j\). Then for \(\delta_j = 0\) it follows that \(\text{Var}(\Phi^{-1}(1-p_j)) = 1\), since then \(p_j\) follows a uniform distribution. However, if \(\delta_j < 0\), then \(\text{Var}(\Phi^{-1}(1-p_j)) \leq 1\). To show this inequality two scenarios must be considered. First, situations are considered with large sample sizes, that is, a large number of degrees of freedom (say at least 30 or more degrees of freedom), where asymptotic arguments apply. Hence, the \(t\)-distribution can be approximated by normal distribution \(N(\delta_j, 1)\). Here, \(\delta_j\) denotes the noncentrality parameter. Then, it follows that \(\text{Var}(\Phi^{-1}(1-p_j)) = \text{Var}(\Phi^{-1}(1-(1-\Phi(N(\delta_j, 1)))) = \text{Var}(N(\delta_j, 1)) = 1\). Regarding scenarios with smaller numbers of degrees of freedom, say \(DF < 30\) simulations have been conducted showing the \(\text{Var}(\Phi^{-1}(1-p_j)) \leq 1\). These simulations can be found in Figure 3 of the online supplement. With this result and Formula (3), it is possible to decompose the variance of a composite population \(\text{Var}(Z_{G_r})\) to

\[
\text{Var}(Z_{G_r}) = \text{Var} \left( \sum_{j \in I_r} \sqrt{w_j \sum_{k \in I_r} w_k \Phi^{-1}(1-p_j)} \right)
\]
\[ \sum_{j \in I_R} \frac{w_j}{\sum_{k \in I_R} w_k} \text{Var}(\Phi^{-1}(1 - p_j)). \]

This shows that \( \text{Var}(Z_{G_r}) \) equals one if and only if all \( \text{Var}(\Phi^{-1}(1 - p_j)) \) are equal to one for all \( j \in I_R \). Otherwise, it follows that \( \text{Var}(Z_{G_r}) < 1 \).

Then, for testing \( H_{0 \cap R} \cap \bigcap_{r=1}^R G_r \), critical value \( c_{\cap R} \cap \bigcap_{r=1}^R G_r \) is determined assuming \( \delta_j \) is true. Therefore, the distribution of the joint composite population test statistic is described by \( \mathcal{N}_R(0, \Sigma_0) \) and the critical value \( c_{\cap R} \cap \bigcap_{r=1}^R G_r \) that solves (7). Let \( f_Z(z) \) denote the density function of \( \mathcal{N}_R(0, \Sigma_0) \). Then it is possible to determine \( c_{\cap R} \cap \bigcap_{r=1}^R G_r \) by finding the critical value that fulfills

\[
\alpha = 1 - \int_{-\infty}^{c_{\cap R} \cap \bigcap_{r=1}^R G_r} \cdots \int_{-\infty}^{c_{\cap R} \cap \bigcap_{r=1}^R G_r} f_Z((z_1, \ldots, z_R))dz_1 \cdots dz_R
\]

where the integral represents \( P_{H_{0 \cap R} \cap \bigcap_{r=1}^R G_r} (Z_{G_1} \leq c_{\cap R} \cap \bigcap_{r=1}^R G_r \wedge \cdots \wedge Z_{G_R} \leq c_{\cap R} \cap \bigcap_{r=1}^R G_r) \). If, however, one or more subsets have \( \delta_j < 0 \) then the joint distribution is no longer defined by \( \mathcal{N}_R(0, \Sigma_0) \) but by \( \mathcal{N}_R(\mu_{<0}, \Sigma_{<0}) \). The new mean vector \( \mu_{<0} \) has one or more entries which are smaller than zero, while \( \Sigma_{<0} \) represents the covariance matrix where one or more trace elements are smaller than one. Corresponding covariances will shrink according to each variance while the correlation stays the same. Then, the density function for the new distribution is denoted by \( f_{<0}(z) \). Using both densities, the inequality

\[
\alpha = 1 - \int_{-\infty}^{c_{\cap R} \cap \bigcap_{r=1}^R G_r} \cdots \int_{-\infty}^{c_{\cap R} \cap \bigcap_{r=1}^R G_r} f_Z((z_1, \ldots, z_R))dz_1 \cdots dz_R > 1 - \int_{-\infty}^{c_{<0}} \cdots \int_{-\infty}^{c_{<0}} f_{<0}((z_1, \ldots, z_R))dz_1 \cdots dz_R
\]

is determined. This shows that using \( c_{\cap R} \cap \bigcap_{r=1}^R G_r \) as the critical value will control the type I error rate not only for the edge of the null hypothesis but \( c_{\cap R} \cap \bigcap_{r=1}^R G_r \) becomes conservative for any other scenario covered by the null hypothesis. By using the closed testing principle, the same reasoning holds for any intersection and elementary null hypotheses. Simulations demonstrating the control of the type I error rate are given in the online supplement.

### 3.2 Sample size calculation

For the sample size calculation, several different power definitions could be used. Here, we opted to apply the so-called disjunctive power. The disjunctive power is the probability to reject at least one false null hypothesis, which is the probability to reject \( H_{0 \cap R} \cap \bigcap_{r=1}^R G_r \), when \( H_{A \cap R} \cap \bigcap_{r=1}^R G_r \) is true (Senn & Bretz, 2007). The required sample size for a study depends on subset prevalences, variances, and correlations. However, those parameters are typically not of research interest. For an initial sample size calculation, assumptions on parameters like the effect sizes \( \delta_j \), the outcome variance \( \sigma^2_{Y_j} \), and \( \sigma^2_{Y_j, X_j1 \cdots X_jD} \) have to be made. While typically a lower limit for the treatment benefit \( \delta_j \) can be defined, previous studies are typically used to obtain estimates for the nuisance parameters. If no such studies exist, values for the nuisance parameters have to be estimated without previous knowledge.

To identify parameters which are based on initial assumptions, those parameters are identified using *. Therefore, the assumed treatment benefit is written as \( \delta^*_j \), the assumed outcome variance is \( \sigma^2_{Y_j}^* \), and so on. Hence, composite population test statistics that are based on prior assumptions are denoted by \( Z_{G_r}^* (N) \), where each parameter except the sample size \( N \) are fixed values. The respective subset sample sizes for treatment and control groups are derived using assumed subset prevalences \( \tau^*_j \) and treatment allocation \( x^* \). Using those assumptions a joint test statistic under the alternative hypothesis is derived similar to the joint test statistic for the null hypothesis in (4). Since a treatment benefit is assumed to exist in at least one subset, the distribution of the joint test statistic consists of a nonzero mean vector \( (Z_{G_1}^*(N), \ldots, Z_{G_R}^*(N)) \) and
a covariance structure under the alternative hypothesis $\Sigma^*_A$. Let the joint cumulative distribution under the alternative be specified by $G^*(N) = \mathcal{N}_R((Z^*_{G_1}(N), \ldots, Z^*_{G_R}(N), \Sigma^*_A))$. Naturally, the covariance structure under the alternative hypothesis, $\Sigma^*_A$, is different from the covariance structure under null hypothesis $\Sigma_0$. This divergence increases with increasing treatment effect. Since no closed formula to determine $\Sigma^*_A$ is available, $\Sigma^*_A$ has to be approximated numerically using the following steps:

1. Simulate random subject data using $\tilde{N}, \tau_j, \kappa^*, Y_{j1}, X_{j1}^*, \ldots, X_{jd}^*, \Sigma^*_Y, \Sigma^*_X$, and $\sigma^*_{YX}$ for subsets $j = 1, \ldots, J$. Here, any $\tilde{N}$ is applicable as long as $\tilde{N} \cdot \tau_j^* - 2 - D > 1$ is true for all $j = 1, \ldots, J$. Otherwise, (1) cannot be evaluated due to a lack of subjects in at least one subset.

2. Evaluate $\delta_{ij}(\tilde{N})$ and $\alpha_{ij}(\tilde{N})$.

3. Using the method discussed in previous sections calculate $Z^*_{Gr}(\tilde{N})$ for $r = 1, \ldots, R$. Save each $Z^*_{Gr}(\tilde{N})$ in a respective vector.

4. Repeat steps 1–3 until a sufficiently large number of test statistics have been obtained to reliably calculate covariances between test statistics, say 10,000 to 15,000 repeats.

Afterwards, calculate the covariances between composite population test statistics $\tilde{S}_A$ using the simulated data. Note that $\Sigma^*_A$ and $\tilde{S}_A$ are only influenced by the expected effect sizes. For this reason, $\tilde{S}_A$ has only to be estimated once at the beginning of the sample size calculation and can even be re-used during the sample size recalculation, as long as the estimated effect sizes remain the same. For this, we simulate patient data with given prevalences, treatment allocation, treatment effects, covariance impact, and so on. Using these data, composite population test statistics are derived using the methods described in the previous section. Each of the $R$ composite population test statistics is saved in one of $R$ accompanying lists. After a sufficient number of simulations have been conducted, simply calculate the covariance between any two lists of composite population test statistics. With this approach, $\tilde{S}_A$ is determined and replaces $\Sigma^*_A$. Note that $\Sigma^*_A$ is only influenced by the expected effect sizes, $\tilde{S}_A$ has to be typically estimated only once at the beginning of the sample size calculation process.

For each simulation run, we advise to simulate at least enough subject data so test statistics can be evaluated in each subset but not simulate so much data that composite population test statistics converge to infinity due to numerical inaccuracies. Using covariance matrix $\tilde{S}_A$, a joint distribution for the composite population test statistics under the given $\delta_1 = \delta^*_1, \ldots, \delta_J = \delta^*_J$ is defined by $\tilde{G}(N) = \mathcal{N}_R((Z^*_{G_1}(N), \ldots, Z^*_{G_R}(N), \tilde{S}_A))$.

Using an approximation for $S(\tilde{\delta})^2$, so that $S^2(\tilde{\delta}) \approx (N_{Tj} + N_{Cj} - 2)/(N_{Tj} + N_{Cj} - 2 - D) \cdot (1 - \sigma^2_{Yj,Xj1\ldots Xjd})^2 \cdot (\frac{1}{N_{Tj}} + \frac{1}{N_{Cj}})$, the required sample size is calculated by determining the sample size $N$ which fulfills

$$1 - \beta \geq P_{\delta_1 = \delta^*_1, \ldots, \delta_J = \delta^*_J}(Z^*_{G_1}(N) \geq \gamma^*_1G_1, \ldots, Z^*_{G_R}(N) \geq \gamma^*_R G_R).$$

(8)

For each $N$, (8) can be evaluated by integrating over joint distribution $\tilde{G}(N)$ using $\gamma^*_r G_r$ as the lower limit for the integration. The smallest $N$ which satisfies (8) is denoted as $N_0$ and is in turn used for the study. Note, that during the sample size investigation lower bounds for $N_0$ do apply, namely that $N_0 \cdot \tau_j - 2 - D$ should be greater than one for any given subsets $j$. Here, the justification for this lower bound is the same as for $\tilde{N}$. The calculation of the critical boundaries can easily be done using R software packages like mvtnorm (Genz & Bretz, 2009; Genz et al., 2020) or by utilizing the cumulative density function for multivariate normal distributions implemented in SAS or MATLAB.

### 3.3 Sample size recalculation using the internal pilot study design

During the initial planning stage, researchers are basing their justification for the sample size on prior assumptions of the nuisance parameters which could be wrong, inaccurate, or not applicable to the current study population. Therefore, it seems sensible to stop the trial at a certain point during the recruitment phase and to reestimate nuisance parameters based on collected data. Such a procedure is executed during an internal pilot study (IPS) (Wittes & Brittain, 1990) and can be broken down into three steps. The first step is to calculate the initial global sample size $N_0$ based on initial assumptions about the nuisance parameters. For the second step, let $\nu$ denote a fraction of the initially planned sample size. After $\nu \cdot N_0$ subjects are recruited, an IPS is conducted wherein the nuisance parameters are reestimated based on the newly collected
data. The number of patients at which the pilot study takes place is defined as $N_1$, where $N_1 = \nu \cdot N_0$. An overview of how to size an internal pilot study can be found in Friede and Kieser (2006) as well as in Friede and Schmidli (2010). Based on the $N_1$ subjects parameters are reestimated without unveiling treatment affiliations and those parameters are then used to conduct a blinded sample size recalculation (BSSR). Throughout the sample size recalculation, we assume prevalences $\tau_j$ to be true and known and are therefore not reestimated.

Now, to estimate the nuisance parameters of the subset linear models without unveiling the treatment allocation of the subjects, we are building upon the framework of Zimmermann et al. (2020). They describe methods to blindly reestimate the variance for the adjusted treatment effects when testing a normally distributed outcome with an arbitrary number of normally distributed random baseline covariates. That approach is an extension of the work by Friede and Kieser (2011).

Hence, after $N_1$ subjects have been recruited the sample size review is conducted using the same linear regression model as in (1), but removing the first part with $\beta_j U_i$. This way, the treatment allocations of all subjects remain concealed. The model is evaluated in each subset, using all available data. However, similar to $N_0$, lower boundaries for $N_1$ do apply. $N_1$ has to be large enough so that $N_1 \cdot \tau_j - 1 - D > 1$ is fulfilled in any subset. Otherwise, one or more reduced linear models could not be analyzed due to a lack of subjects. Based on data from $N_1$ subjects, $J$ linear regression models are fitted for the blinded recalculation. The residuals $\hat{e}_{k,j}$ of these models are then utilized to approximate the variances for $\hat{\delta}_j$ by

$$\frac{N_{T_j} + N_{C_j} - 2}{N_{T_j} + N_{C_j} - 2 - D} \left(1 - \hat{\rho}_{Y_j,X_j1}^2 - \cdots - \hat{\rho}_{Y_j,X_jD}^2\right) S_{Y_j}^2 \approx \frac{\sum_{k=1}^{N_{T_j} + N_{C_j}} \hat{\epsilon}_{k,j}^2}{N_{T_j} + N_{C_j} - 1 - D}.$$

A justification for this approximation can be found in the online supplement. Then, this approximation is then applied to formula (2) and the sample size calculation procedure is repeated with the updated estimations resulting in a recalculated sample size $N_{\text{reest}}$. However, since the treatment allocation is ignored, a bias is introduced into the variance estimators leading to slightly inflated estimators. This is often not a problem since it has been shown for ANCOVA models with one random covariate, that the impact of this inflation is generally negligible (see Friede and Kieser, 2013).

The third and last step is to decide which recalculated sample size to use as the final sample size $N_{\text{final}}$. Wittes and Brittain (1990) proposed the so-called restricted design where the final sample size is defined as the maximum of the initially calculated sample size and the recalculated sample size, therefore $N_{\text{final}} = \max (N_0, N_{\text{reest}})$. Birkett and Day (1994), on the other hand, proposed the unrestricted design, where the sample size is the maximum of the sample size at the internal pilot study and the recalculated sample size, hence $N_{\text{final}} = \max (N_1, N_{\text{reest}})$. In this design, the recalculated sample size be lower than $N_1$ then the final analysis will be based on $N_1$ patients. In the following simulation study, we will use the unrestricted design to define $N_{\text{final}}$.

### 3.4 Numerical example

To the proposed approach, we provide a numerical example inspired by the motivating example. In this example, the determination of critical boundaries for intersection null hypothesis testing and the calculation of the initial sample size is shown. To keep the example concise, only the following PAH subsets are considered: idiopathic PAH, heritable PAH, PAH associated with connective tissue disease, and lastly PAH due to drugs or toxins. To keep in line with the notation introduced in previous sections, we equate idiopathic PAH to $S_1$, heritable PAH to $S_2$, PAH due to drugs or toxins to $S_3$, and PAH associated with connective tissue disease to $S_4$. Hence, a population consisting of four subsets is investigated. The subset null hypotheses are defined by $H_{0,j} : \delta_j \leq 0$ for $j = 1, \ldots, 4$.

The overall population is not considered for testing since it is expected that each subset might have a different reaction towards the intervention. Therefore, composite populations $G_1 = S_1, G_2 = S_2 \cup S_3 \cup S_4, \text{ and } G_3 = S_2 \cup S_3$ are tested.

In this example, prevalences are assumed to be $\tau_1^* = 0.5, \tau_2^* = 0.1, \tau_3^* = 0.2$, and $\tau_4^* = 0.2$. The outcome of the planned study is measured by the 6-minute walk test, which is defined as the distance a subject is able to walk within a timespan of 6 min. First, we consider data with one outcome and one covariate, where the squared correlation between the outcome and the covariate is expected to be $\hat{\rho}_{Y_1,X_{11}}^2 = \hat{\rho}_{Y_2,X_{21}}^2 = \hat{\rho}_{Y_3,X_{31}}^2 = \hat{\rho}_{Y_4,X_{41}}^2 = 0.2$. The effect of the treatment is set to be $\delta_1^* = 20, \delta_2^* = \delta_3^* = 10$, and $\delta_4^* = 0$, respectively. Setting the outcome standard deviation of each subset to $\sigma_{Y_1}^* = \sigma_{Y_2}^* = \sigma_{Y_3}^* = \sigma_{Y_4}^* = 25$, the variance after subtracting the variability explained by the covariate is

$$\left(1 - \hat{\rho}_{Y_1,X_{11}}^2\right) \sigma_{Y_1}^2 = \left(1 - \hat{\rho}_{Y_2,X_{21}}^2\right) \sigma_{Y_2}^2 = \left(1 - \hat{\rho}_{Y_3,X_{31}}^2\right) \sigma_{Y_3}^2 = \left(1 - \hat{\rho}_{Y_4,X_{41}}^2\right) \sigma_{Y_4}^2 = 500.$$
Using all these assumptions and using an equal treatment allocation with $\kappa^* = 1$, all subset test statistics are defined by

$$T_j^*(N) = \frac{\delta_j^*}{\sqrt{\frac{N \cdot \tau_{j}^{-2}}{N \cdot \tau_{j}^{-2} - 1} \cdot \text{500} \cdot \left( \frac{1}{N \cdot \tau_{j}^{-1/2}} + \frac{1}{N \cdot \tau_{j}^{1/2}} \right)}}.$$ 

From those test statistics, $p_j^*(N)$ values are generated using $t$-distributions with $N \cdot \tau_j^* - 3$ degrees of freedom. In the next step, composite test statistics are defined by

$$Z_{G_1}^*(N) = \Phi^{-1}(1 - p_j^*(N))$$

$$Z_{G_2}^*(N) = 4 \sum_{j=2}^{4} \frac{\omega_j^*}{\omega_2^* + \omega_3^* + \omega_4^*} \cdot \Phi^{-1}(1 - p_j^*(N))$$

$$Z_{G_3}^*(N) = 3 \sum_{j=2}^{3} \frac{\omega_j^*}{\omega_2^* + \omega_3^*} \cdot \Phi^{-1}(1 - p_j^*(N)).$$

The accompanying elementary null hypotheses for those composite population test statistics are $H_{0G_1}$, $H_{0G_2}$, and $H_{0G_3}$. Since the same treatment allocation is used in each subset, the weights are defined by the prevalence so that $\omega_j^* = \tau_j^*$. The next step is to determine the expected covariances between composite population test statistics. Given (5), the covariances are calculated as

$$\text{cov}(Z_{G_1}^*, Z_{G_2}^*) = 0,$$

$$\text{cov}(Z_{G_1}^*, Z_{G_3}^*) = 0,$$

$$\text{cov}(Z_{G_2}^*, Z_{G_3}^*) = \sum_{j=2}^{3} \frac{\omega_j^*}{(\omega_2^* + \omega_3^* + \omega_4^*) \cdot (\omega_2^* + \omega_3^*)} = \sqrt{0.6}.$$ 

Given previously defined elementary null hypotheses, the following intersection null hypotheses must be tested during the closed testing procedure. The intersection hypotheses are $H_{0G_1 \cap G_2 \cap G_3}$, $H_{0G_1 \cap G_2 \cap G_3}$, $H_{0G_1 \cap G_2}$, and $H_{0G_2 \cap G_3}$. For each intersection hypothesis, an efficient critical boundary can be derived by defining its multivariate normal distributions which describe the joint distribution of the composite population test statistics under the null hypothesis. Depending on the intersection null hypotheses, these normal distributions are denoted as

$$(Z_{G_1}, Z_{G_2}, Z_{G_3} | \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0) \sim N \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \sqrt{0.6} \\ 0 & \sqrt{0.6} & 1 \end{pmatrix} \right).$$

$$(Z_{G_1}, Z_{G_2} | \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

$$(Z_{G_1}, Z_{G_3} | \delta_1 = \delta_2 = \delta_3 = 0) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{0.6} \end{pmatrix} \right).$$

$$(Z_{G_2}, Z_{G_3} | \delta_2 = \delta_3 = \delta_4 = 0) \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sqrt{0.6} \\ \sqrt{0.6} & 1 \end{pmatrix} \right).$$
TABLE 2  Here, the required sample sizes are shown to achieve a trial with a targeted power of 90%. The effect sizes, variances, and critical values are taken from the previous example with one covariate. The squared correlation between the first covariate to the outcome was kept to 0.2 (first column), while the squared correlation of the second covariate to the outcome varied between 0, 0.1, and 0.2 (second column). The squared correlation among the covariates also varied and took on values of 0, 0.25, or 0.5 (third column). The results of the sample size calculations are shown in the last column.

| 𝜌₁,ₓ₁ | 𝜌₂,ₓ₂ | 𝜌₃,ₓ₃ | 𝜌₄,ₓ₄ | 𝑁₀  |
|-------|-------|-------|-------|-----|
| 0.2   | 0     | 0.25  | 0.5   | 135 |
| 0.2   | 0.1   | 0     | 0.5   | 119 |
| 0.2   | 0.2   | 0.25  | 0.5   | 124 |
| 0.2   | 0.2   | 0.25  | 0.5   | 129 |

By applying any given 𝛼 to each of those distributions, efficient critical boundaries are derived. Here, setting the one-sided 𝛼 level to 0.025, the critical boundaries derived for each intersection null hypotheses are 𝑐₁,₁ = 2.344, 𝑐₁,₂ = 2.236, and 𝑐₂,₂ = 2.161. In contrast, using the Bonferroni correction to derive critical boundaries for the same intersection, null hypotheses would result in 𝑐₁,₁ = 2.394 and 𝑐₁,₂ = 𝑐₂,₂ = 𝑐₁,₁ = 2.241 instead. Given these critical boundaries, using disjunctive power for the sample size calculation and setting the desired power to 90%, the required overall sample size 𝑁₀ would amount to 132 subjects.

Further extending this example, it is demonstrated how the sample sizes change when including another covariate. Different correlation constellations are simulated. While critical values remain the same (no changes to the prevalences occurred), the degrees of freedom in each subset are changed from 𝑁 ∙ 𝜏 − 3 to 𝑁 ∙ 𝜏 − 4.

The first three rows of Table 2 provide insight into the efficiency of an investigation, where a second covariate with no correlation to the outcome was added. At first, the added covariate reduces the efficiency of the analysis. This is marked by an increase in the required sample size from 132 subjects to 135 subjects. However, if the second covariate has no correlation to the outcome but to the first covariate, an increase in efficiency is detected with the growing correlation between the covariates. The reverse trend is seen when a correlation between the outcome and the second covariate does exist. The peak efficiency of an investigation occurs when no correlation between covariates exists, while both covariates are correlated to the outcome. The efficiency of an analysis lessens with increasing correlations between covariates.

4 | SIMULATION STUDY

The performance of the proposed procedures was evaluated using simulation studies. For simplicity, a population consisting of a subset 𝑆₁ and its complement 𝑆₂ was considered. To keep the number of possible simulations reasonable, the data were simulated with one normally distributed outcome and one normally distributed covariate. A scenario investigating the performance with three composite populations can be found in the online supplement in Figure 6. In the simulations presented here, the data for the outcome and accompanying covariate were drawn from a bivariate normal distribution with mean vector (0, 0) for patients in the treatment arm in subset 𝑆₁ and from a bivariate normal distribution with mean vector (0, 0) for control subjects of subset 𝑆₁. All subjects for subset 𝑆₂ were drawn from a bivariate normal distribution with mean vector (0, 0). Therefore, it is assumed that the treatment does not benefit subjects in subset 𝑆₂. The respective covariance matrix for the bivariate distributions was defined by the corresponding scenario. The assumed prevalence 𝜏₁ always matched the true prevalence 𝜏₁. The effect of the intervention was tested in composite populations 𝐺₁ = 𝑆₁ and 𝐺₂ = 𝑆₁ ∪ 𝑆₂. Sample size calculation and recalculation were performed such that it matched a disjunctive power of 1 − 𝛽 = 0.9. For the final analysis, if a BSSR took place, the IPS data are pooled with the additional cumulated data after IPS. Results for scenarios corresponding to a disjunctive power of 1 − 𝛽 = 0.8 are presented in Tables 6 and 7 and Figures 4 and 5 of the online supplement. Further details on each scenario are shown in Table 3.
TABLE 3 Scenarios considered for the FWER and power simulation studies. The scenarios considered resulted from a combination of parameters present in this table. Therefore, during the investigation of the type I error rate, 54 scenarios were generated in which the ability to reject the intersection null hypothesis $H_{0,G_1 \cap G_2}$ for the settings: no sample size recalculation, earlier recalculation and later recalculation were investigated. Hence, 162 simulations were considered overall. For the power investigation, 90 scenarios were considered for the same three aforementioned simulations.

| Scenario | Under intersection null hypothesis $H_{0,G_1 \cap G_2}$ | Under intersection alternative hypothesis $H_{A,G_1 \cap G_2}$ |
|----------|-------------------------------------------------|-------------------------------------------------|
| Simulation runs | 10 000 | 10 000 |
| Significance level $\alpha$ | 0.025 (one-sided) | 0.025 (one-sided) |
| Allocation ratio $\kappa$ | 1 | 1 |
| Prevalence $\tau_1 = \tau_1$ | 0.25; 0.5; 0.75 | 0.25; 0.5; 0.75 |
| Proportion $\nu$ for IPS | 0.3; 0.5 | 0.3; 0.5 |
| Assumed treatment benefit $\delta^*_1$ in $S_1$ | 0.5; 1 | 0.5; 1 |
| Assumed treatment benefit $\delta^*_2$ in $S_2$ | 0 | 0 |
| Assumed variance $\sigma^2_Y$ in $S_1$ | 1 | 1 |
| Assumed variance $\sigma^2_Y$ in $S_2$ | 1 | 1 |
| Assumed squared correlation $\rho^*_{Y_1,Y_2}$ | 0.4 | 0.4 |
| True treatment benefit $\delta_1$ in $S_1$ | 0 | same as $\delta^*_1$ |
| True treatment benefit $\delta_2$ in $S_2$ | 0 | 0 |
| True variance $\sigma^2_Y$ in $S_1$ | 0.8; 1; 1.2 | 0.8; 1; 1.2 |
| True variance $\sigma^2_Y$ in $S_2$ | 1 | 1 |
| True squared correlation $\rho_{Y_1,Y_2}$ | 0.0; 0.4; 0.8 | 0.0; 0.2; 0.4; 0.6; 0.8 |

Abbreviation: IPS, internal pilot study.

First, the type I error rate is investigated by comparing the error rates of fixed designs and designs with earlier ($\nu = 0.3$) and later ($\nu = 0.5$) sample size recalculation. The sample sizes are calculated and recalculated using assumed treatment benefits $\delta^*_1 = 0.5$ and $\delta^*_1 = 1$. Table 4 provides the results for the simulations given the type I error rate for low and large sample sizes, considering a planned power of 90%. The expected Monte Carlo error for the type I error rate was calculated to be $\sqrt{0.025 \cdot 0.975 \cdot 10000} \approx 0.00156$. The error rates presented in Table 4 show the type I error rates for falsely rejecting the intersection null hypothesis. Error rates reflecting the probability to falsely reject the elementary null hypothesis $H_{0,G_1}$ are provided in Table 5 of the online supplement. Here, out of 162 simulations, only six exceeded the resulting 95% prediction interval which was true for fixed designs as well as recalculation designs.

Next, to evaluate the ability of the proposed blinded sample size recalculation to maintain the desired power, we performed simulations of the power, the mean recalculated sample size, and the variation of the recalculated sample size which are shown in Figures 1 and 2. The variation was represented in terms of the 10% as well as the 90% quantile. For the fixed designs (—) the parameters for the initial sample size calculation were assumed to be $\delta^*_1$, $\delta^*_2$, $\sigma^2_Y$, $\sigma^2_Y$, and $\rho^{*}_{Y_1,Y_2} = \rho^{*}_{Y_2,Y_2}$. In each scenario, initially assumed treatment benefits always matched the true treatment benefits, so that $\delta^*_1 = \delta_1$ and $\delta^*_2 = \delta_2$. The other two settings included a sample size recalculation at a later stage with $\nu = 0.5$ (---) or an earlier stage with $\nu = 0.3$ (—). During the recalculation process, the true prevalences values were used and no reestimation took place.

An improvement in the performance is seen when comparing the power of fixed designs ($\delta_1 = 0.5$; —, $\delta_1 = 1$; ---) to designs which conducted an earlier ($\delta_1 = 0.5$; —, $\delta_1 = 1$; ---) or later ($\delta_1 = 0.5$; ---, $\delta_1 = 1$; ---) sample size recalculation. (See Figure 1.) It stands out that in simulations where no IPS is conducted the power diverged widely from the desired power. However, in simulations where an IPS was conducted after $N_0 \cdot 0.5$ or $N_0 \cdot 0.3$ subjects were recruited, the power was met in most circumstances. An overpowered is observed in studies where a high correlation between the outcome and the covariate exists. In those scenarios, the required sample size to meet the desired power is rather low, so that the BSSR takes place at a point where already more than enough subjects have been recruited. This feature is also expressed in those simulations with an earlier IPS. Here, however, the overpowered is not as large, due to the earlier sample size recalculation.
**TABLE 4** Comparing type I error rates for falsely rejecting the intersection null hypotheses \( H_{G_1 \cap G_2} \) for fixed designs and designs with blinded sample size recalculation. \( N_0 \) denotes the initially calculated total sample size for power = 0.9 and using assumed variances and correlations but true prevalences. With 10,000 simulations per scenario and a simulated Type I error rate of 2.5%, the simulated error rate has a standard error of 0.00156 leading to a 95% prediction interval of [0.02194; 0.02806]. Values outside that interval are highlighted. Out of the 162 simulations, six were outside the interval.

| \( \tau_1 \) | \( \sigma^2_{Y_1} \) | \( \rho^2_{Y_1 X_1, X_2} = \sigma^2_{Y_2 X_1} \) | \( \delta^*_1 = 0.5 \) | \( N_0 \) | FWER no IPS | FWER \( \gamma = 0.3 \) | FWER \( \gamma = 0.5 \) | \( \delta^*_1 = 1 \) | \( N_0 \) | FWER no IPS | FWER \( \gamma = 0.3 \) | FWER \( \gamma = 0.5 \) |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.25 | 0.8 | 0.0 | 652 | 0.0276 | 0.0256 | 0.0252 | 173 | 0.0238 | 0.0232 | 0.0245 |
| 0.25 | 0.8 | 0.4 | 652 | 0.0234 | 0.0260 | 0.0281 | 173 | 0.0274 | 0.0268 | 0.0284 |
| 0.25 | 0.8 | 0.8 | 652 | 0.0262 | 0.0252 | 0.0266 | 173 | 0.0260 | 0.0251 | 0.0237 |
| 0.25 | 1.0 | 0.0 | 652 | 0.0257 | 0.0238 | 0.0238 | 173 | 0.0263 | 0.0251 | 0.0264 |
| 0.25 | 1.0 | 0.4 | 652 | 0.0249 | 0.0269 | 0.024 | 173 | 0.0260 | 0.0246 | 0.0234 |
| 0.25 | 1.0 | 0.8 | 652 | 0.0264 | 0.0266 | 0.0258 | 173 | 0.0267 | 0.0297 | 0.0245 |
| 0.25 | 1.2 | 0.0 | 652 | 0.0253 | 0.0235 | 0.0213 | 173 | 0.0268 | 0.0245 | 0.0258 |
| 0.25 | 1.2 | 0.4 | 652 | 0.0264 | 0.0278 | 0.0222 | 173 | 0.0249 | 0.0233 | 0.0263 |
| 0.25 | 1.2 | 0.8 | 652 | 0.0233 | 0.0256 | 0.0262 | 173 | 0.0249 | 0.0245 | 0.0262 |
| 0.50 | 0.8 | 0.0 | 314 | 0.0265 | 0.0231 | 0.0264 | 84 | 0.0262 | 0.0237 | 0.0246 |
| 0.50 | 0.8 | 0.4 | 314 | 0.0248 | 0.0252 | 0.025 | 84 | 0.0249 | 0.0261 | 0.0254 |
| 0.50 | 0.8 | 0.8 | 314 | 0.0230 | 0.0245 | 0.022 | 84 | 0.0250 | 0.0256 | 0.0256 |
| 0.50 | 1.0 | 0.0 | 314 | 0.0246 | 0.0228 | 0.0252 | 84 | 0.0241 | 0.0233 | 0.028 |
| 0.50 | 1.0 | 0.4 | 314 | 0.0253 | 0.0253 | 0.0245 | 84 | 0.0229 | 0.0241 | 0.0245 |
| 0.50 | 1.0 | 0.8 | 314 | 0.0256 | 0.0258 | 0.0263 | 84 | 0.0242 | 0.0229 | 0.026 |
| 0.50 | 1.2 | 0.0 | 314 | 0.0243 | 0.0244 | 0.0229 | 84 | 0.0249 | 0.0272 | 0.0241 |
| 0.50 | 1.2 | 0.4 | 314 | 0.0266 | 0.0242 | 0.0257 | 84 | 0.0252 | 0.0266 | 0.0234 |
| 0.50 | 1.2 | 0.8 | 314 | 0.0237 | 0.0258 | 0.0255 | 84 | 0.0264 | 0.0279 | 0.0252 |
| 0.75 | 0.8 | 0.0 | 203 | 0.0246 | 0.0256 | 0.0231 | 54 | 0.0242 | 0.0222 | 0.0213 |
| 0.75 | 0.8 | 0.4 | 203 | 0.0248 | 0.0241 | 0.0259 | 54 | 0.0256 | 0.0253 | 0.0194 |
| 0.75 | 0.8 | 0.8 | 203 | 0.0266 | 0.0273 | 0.0224 | 54 | 0.0255 | 0.0227 | 0.0222 |
| 0.75 | 1.0 | 0.0 | 203 | 0.0240 | 0.0223 | 0.0247 | 54 | 0.0243 | 0.025 | 0.0253 |
| 0.75 | 1.0 | 0.4 | 203 | 0.0242 | 0.0242 | 0.027 | 54 | 0.0254 | 0.0274 | 0.0227 |
| 0.75 | 1.0 | 0.8 | 203 | 0.0239 | 0.0270 | 0.0271 | 54 | 0.0244 | 0.0269 | 0.0256 |
| 0.75 | 1.2 | 0.0 | 203 | 0.0254 | 0.0230 | 0.0236 | 54 | 0.0262 | 0.0227 | 0.0269 |
| 0.75 | 1.2 | 0.4 | 203 | 0.0251 | 0.0244 | 0.0239 | 54 | 0.0258 | 0.0239 | 0.0246 |
| 0.75 | 1.2 | 0.8 | 203 | 0.0271 | 0.0258 | 0.0247 | 54 | 0.0249 | 0.0266 | 0.0243 |

Abbreviations: FWER, family-wise type I error rate; IPS, internal pilot study.

When comparing the simulations with earlier sample size recalculations to those with later sample size recalculations, we see a loss of power of 1–2% on average. This power loss becomes more severe the fewer subjects are present during the IPS. For scenarios where the IPS was conducted with only 13 subjects in \( S_1 \), the loss of power was 3–4%.

The panels of Figure 2 show the sample size when all parameters are known from the get go (---), the average sample sizes, and the spread of the recalculated sample sizes. Three points stand out. First, on average earlier BSSR (---) and later BSSR (-----) yield the same average recalculated sample size. Second, while the average sample sizes are the same, the spread of the sample size is always higher in earlier IPS (-----) compared to a later IPS (-----). And third, by comparing the recalculated sample sizes with the optimal sample size, the average sample size was always higher than the optimal sample size. This increase in sample size naturally follows when estimating the variance in a blinded fashion in the presence of a positive treatment effect. Through this variance inflation, the recalculated sample size is increased. In our simulations, this inflation of the variance amounted to an increase in the sample size of 20 subjects per simulation run on average. This increase in the average sample sizes is, however, independent of the required sample sizes and has properties making them preferable to methods without these inflations as noted by Friede and Kieser (2001).
The planned power in these scenarios was set to 0.9. The power characteristics are plotted against varying values of true correlation ($\rho_{Y_1, X_1}$), true variance of subset $S_1$ ($\sigma^2_{Y_1}$), and varying prevalences $\tau_1$. Given a desired power of 90% (---), the simulation results for treatment benefit $\delta_1 = 0.5$ are presented as follows: no IPS (---), IPS at $\nu = 0.3$ (---), and IPS at $\nu = 0.5$ (---). The results for a treatment benefit of $\delta_1 = 1$ are presented as follows: no IPS (---), IPS at $\nu = 0.3$ (---) and an IPS at $\nu = 0.5$ (---).

5 DISCUSSION AND CONCLUSION

In the present investigation, we propose procedures for hypothesis testing, sample size calculation, and blinded sample size recalculation for trials with normally distributed outcomes and an arbitrary number of baseline covariates. The properties of this procedure in terms of type I error rate, power, and the distribution of the sample sizes are investigated in the context of no IPS and with earlier and later BSSR. The advantage of the presented approach is the ability to form any subpopulation constellation by first defining the smallest disjoint population denominators, which we call subsets, and form any population using these subsets, which we define as composite populations. In the simulations, the pre-BSSR and post-BSSR data were pooled for the final analysis. As noted by Posch et al. (2018), this could lead to an inflation of the type I error rate. However, this inflation will be minuscule in all but extreme sample size cases. One method to circumvent this inflation would be to analyze pre- and post-BSSR data separately and to combine them for the final analysis using a combination function. We investigate scenarios, in which no relevant inflation of the type I error rate could be detected. The desired power is achieved in scenarios where the initial assumptions about nuisance parameters were met or in scenarios in which an IPS occurred with at least 20 subjects in the smallest subset. The power is slightly below the desired power in scenarios where the IPS is conducted with low sample sizes. This finding is in line with Sandvik et al. (1996) and Birkett and Day (1994) who conclude that an internal pilot study should only be conducted with 20 or more degrees of freedom.

With regard to a sample size recalculation with few subjects, Zucker et al. (1999) addressed this issue for the simple setting of a two-sample $t$-test in one population by multiplying the recalculated sample size by an inflation factor calculated...
FIGURE 2  Calculated and recalculated sample size for $\delta_1 = 0.5$ (a) and $\delta_1 = 1$ (b). The planned power in these scenarios was set to 0.9. The initially calculated sample sizes when using assumed parameters were $N_0 = 652,314,203$ and $N_0 = 173,84,54$ for (a) and (b), read from top to bottom, respectively. The optimal sample size when using the unknown true values for each scenario is given by (—). The average recalculated sample sizes for $\nu = 0.3$ (—) and for $\nu = 0.5$ (—). The spread using the 10% and 90% quantile sample sizes are highlighted by (⋯⋯⋯) and (⋯⋯⋯), respectively.

as $(t_{\alpha,\eta_1} + t_{\beta,\eta_1})^2/\left(\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\right)$. Here, $t_{\alpha,\eta_1}$ and $t_{\beta,\eta_1}$ denote the $t$-quantiles given $\alpha, \beta$ while $\eta_1$ denotes the degrees of freedom during the IPS. Placzek and Friede (2018) extended that idea for a multiple nested subpopulation setting. In their work, the intersection hypothesis is tested using a multivariate $t$-distribution with degrees of freedom according to either the full population or the smallest subgroup. To account for uncertainties during the IPS, degrees of freedom according to the full population or the smallest subgroup at the IPS are used. Both procedures improve the results of the power analysis at the cost of increased sample sizes. An alternative approach to the inflation factor was discussed by Kieser and Friede (2000) and Friede and Kieser (2006), where instead of using the point estimate of the residual variance, an upper confidence bound is used.

Further contrasting the fixed design with the BSSR design, the adaptation stage comes at the cost of an increased (average) total sample size. This inflation of sample size occurs due to an inflation of the residual variance by ignoring the treatment allocation. This reestimation of the sample size has no influence on the type I error rate. However, a blinded sample size recalculation is favored by many regulatory agencies because fewer sources for bias can affect the sample size (EMEA, 2007; FDA, 2010, 2016; ICH E-9 Expert Working Group, 1999).

Throughout the simulation, we assume that prevalences to be correctly specified and are therefore never reestimated. For the scenarios we considered, this was true. Naturally, if the prevalences are misspecified during the initial stage, the sample size calculation will be offset correspondingly. To this point, however, we revere to Placzek and Friede (2018), who concluded that the inclusion of a prevalence reestimation impacts the power only slightly. Moreover, in the present simulations, we only consider scenarios with two subsets and one baseline covariate. For more complex settings with more covariates, we revere to Zimmermann et al. (2020) where recalculation scenarios with more covariates are considered. They conclude that settings with more covariates still achieve the desired power, conditional that enough subjects are
present in during the IPS to accurately determine $\hat{\rho}_{Y, X_{1}, \ldots, X_{D}}^2$. But even with few subjects, the results of power simulations show increased efficiency compared to simulations with misspecified parameters and no sample size recalculation.

An alternative design where treatment effects are evaluated on a subpopulation level is to use a multiple linear model. In such model parameters like the treatment allocation, subpopulation affiliation, and subpopulation–treatment interaction term are included additionally to any other variables like age or gender. Using this design, it is possible to test a treatment in several subpopulations at once, while the covariates are estimated using the full dataset. However, some assumptions have to be met. For one, homoscedasticity must be assumed. Also, while the impact of covariates is estimated using the whole dataset, this makes only sense if it is assumed that the impact of the covariate is uniform in all subpopulations. In contrast, our approach is free from such assumptions and only requires the effects of any explanatory variable and correlations between variables to be constant within subsets. Another difference is the evaluation of the treatment effect. While using linear models, an average treatment effect is evaluated. Hence, a positive treatment effect might get cancelled out by a negative treatment effect. Our approach, on the other hand, tests if at least one subset in the population exhibits a positive treatment effect. For further information on using multivariate linear models for subpopulation analysis, see, for example, Cleophas and Zwinderman (2012) (Ch. 15).

A future route to extend the methods provided here is to incorporate the adaptive enrichment design (Brannath et al., 2009; Friede et al., 2012; Jenkins, Stone, et al., 2011; Placzek & Friede, 2019; Wang et al., 2007). Adaptive trials are made up of two or more stages where the efficacy of a treatment is evaluated at interim analyses between stages. During the first stage, subjects from the whole population are recruited but a change in the recruitment plan can be initiated during an interim analysis when this change was predefined at the start of the study.

ACKNOWLEDGMENTS
Open access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST
The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT
Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

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
Roland G. Gera 🐠 https://orcid.org/0000-0001-8896-3749
Tim Friede 🐠 https://orcid.org/0000-0001-5347-7441

REFERENCES
Basu, S., Sussman, J. B., & Hayward, R. A. (2017). Detecting heterogeneous treatment effects to guide personalized blood pressure treatment. *Annals of Internal Medicine, 166*, 354–360.

Birkett, M. A., & Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine, 13*, 2455–2463.

Birnbaum, A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association, 49*, 559–574.

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., König, 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.-D., König, F., Posch, M., & Jaki, T. (2018). Design and estimation in clinical trials with subpopulation selection. *Statistics in Medicine, 37*, 4335–4352.

Cleophas, T. J., & Zwinderman, A. H. (2012). *Statistics applied to clinical trials* (5th ed., pp. 177–186). Springer Netherlands.

EMA. (2016). *Guideline on multiplicity issues in clinical trials* (Draft. EMA/CHMP/44762/2017), https://www.ema.europa.eu/en/documents/scientific-guideline/draft-guideline-multiplicity-issues-clinical-trials_en.pdf
EMA. (2019). Guideline on the investigation of subgroups in confirmatory clinical trials (EMA/CHMP/539146/2013), https://www.ema.europa.eu/en/documents/scientific-guideline/guideline-investigation-subgroups-confirmatory-clinical-trials_en.pdf

EMEA. (2003). Points to consider on adjustment for baseline covariates (CPMP/EWP/2663/99), https://www.ema.europa.eu/en/documents/scientific-guideline/points-consider-adjustment-baseline-covariates_en.pdf. Accessed on August 23, 2018.

EMEA. (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design (CHMP/EWP/2459/02), https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf. Accessed August 23, 2018.

FDA. (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics (Draft). https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-design-clinical-trials-drugs-and-biologics-guidance-industry

FDA. (2016). Guidance for industry and Food and Drug Administration staff: Adaptive designs for medical device clinical studies. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adaptive-designs-medical-device-clinical-studies

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

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*, 48, 537–555.

Friede, T., Kieser, M. (2011). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics*, 10, 8–13.

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., 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., & Schmidli, H. (2010). Blinded sample size reestimation with count data: Methods and applications in multiple sclerosis. *Statistics in Medicine*, 29, 1145–1156.

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.

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Lecture Notes in Statistics. Springer-Verlag.

Genz, A., Bretz, F., Miwa, T., Mi, X., Scheipl, F., & Hothorn, T. (2021). *mvtnorm: Multivariate normal and t distributions*, https://CRAN.R-project.org/package=mvtnorm

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

Grieve, A. P., Chow, S.-C., Curram, J., Dawe, S., Harnisch, L. O., Henig, N. R., Hung, H. M. J., Ivy, D. D., Kawut, S. M., Rahbar, M. H., Xiao, S., & Wilkins, M. R. (2014). Advancing clinical trial design in pulmonary hypertension. *Pulmonary Circulation*, 3, 217–225.

ICH E-9 Expert Working Group. (1999). ICH harmonized tripartite guideline: Statistical principles for clinical trials. *Statistics in Medicine*, 18, 1905–1942.

Jenkins, M., Flynn, A., Smart, T., Harbron, C., Sabin, T., Ratnayake, J., Delmar, P., Herath, A., Jarvis, P., & Matcham, J. (2011). A statistician’s perspective on biomarkers in drug development. *Pharmaceutical Statistics*, 10, 494–507.

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.

Julious, S. A. (2015). Pilot studies in clinical research. *Statistical Methods in Medical Research*, 25, 995–996.

Kent, D. M., Rothwell, P. M., Ioannidis, J. P. A., Altman, D. G., & Hayward, R. A. (2010). Assessing and reporting heterogeneity in treatment effects in clinical trials: A proposal. *Trials*, 11, 85.

Kieser, M., & Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19, 901–911.

Lehmacher, W., & Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55, 1286–1290.

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

McLaughlin, V. V., Badesch, D. B., Delcroix, M., Fleming, T. R., Gaine, S. P., Galié, N., Gibbs, J. S. R., Kim, N. H., Oudiz, R. J., Peacock, A., Provencher, S., Sitbon, O., Tapson, V. F., & Seeger, W. (2009). End points and clinical trial design in pulmonary arterial hypertension. *Journal of the American College of Cardiology*, 54, 97–107.

Mehta, C., Schäfer, 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.

Montani, D., Günther, S., Dorfmüller, P., Perros, F., Girerd, B., Garcia, G., Jaïs, X., Savale, L., Artaud-Macari, E., Price, L. C., Humbert, M., Simonneau, G., & Sitbon, O. (2013). Pulmonary arterial hypertension. *Orphanet Journal of Rare Diseases*, 8, 97.

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.

Posch, M., Klinglmueller, F., König, F., & Miller, F. (2018). Estimation after blinded sample size reassessment. *Statistical Methods in Medical Research*, 27, 1830–1846.

Sandvik, L., Eriksen, J., Mowinckel, P., & Rodland, E. A. (1996). A method for determining the size of internal pilot studies. *Statistics in Medicine*, 15, 1587–1590.
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.

Senn, S., & Bretz, F. (2007). Power and sample size when multiple endpoints are considered. *Pharmaceutical Statistics, 6*, 161–170.

Simonneau, G., Montani, D., Celermajer, D. S., Denton, C. P., Gatzoulis, M. A., Krowka, M., Williams, P. G., & Souza, R. (2019). Haemodynamic definitions and updated clinical classification of pulmonary hypertension. *European Respiratory Journal, 53*, 1801913.

Spiessens, B., & Debois, M. (2010). Adjusted significance levels for subgroup analyses in clinical trials. *Contemporary Clinical Trials, 31*, 647–656.

Tanniou, J., Smid, S. C., van der Tweel, I., Teerenstra, S., & Roes, K. C. B. (2019). Level of evidence for promising subgroup findings: The case of trends and multiple subgroups. *Statistics in Medicine, 38*, 2561–2572.

Tanniou, J., Van Der Tweel, I., Teerenstra, S., & Roes, K. C. B. (2016). Subgroup analyses in confirmatory clinical trials: Time to be specific about their purposes. *BMC Medical Research Methodology, 16*, 20.

Tavernier, E., & Giraudou, B. (2015). Sample size calculation: Inaccurate a priori assumptions for nuisance parameters can greatly affect the power of a randomized controlled trial. *PLoS ONE, 10*, e0132578.

Varadhan, R., Segal, J. B., Boyd, C. M., Wu, A. W., & Weiss, C. O. (2013). A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. *Journal of Clinical Epidemiology, 66*, 818–825.

Wang, S.-J., O’Neill, R. T., & Hung, H. M. J. (2007). Approaches to evaluation of treatment effect in randomized clinical trials with genomic subset. *Pharmaceutical Statistics, 6*, 227–244.

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

Zimmermann, G., Kieser, M., & Bathke, A. C. (2020). Sample size calculation and blinded recalculation for analysis of covariance models with multiple random covariates. *Journal of Biopharmaceutical Statistics, 30*, 143–159.

Zucker, D. M., Wittes, J. T., 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:** Gera, R. G., & Friede, T. (2023). Blinded sample size recalculation in multiple composite population designs with normal data and baseline adjustments. *Biometrical Journal, 65*, 2000326. https://doi.org/10.1002/bimj.202000326