The New England Journal of Statistics in Data Science Volume 0, 1–20 (2026)
DOI: https://doi.org/10.51387/26-NEJSDS103

Adaptive Sample Size Using a Totality of Evidence Approach in Rare Disease Clinical Trials

Lan SHI*, Yong LIN, Philip HE, and Di SHU

Abstract

Clinical trial design for rare diseases can be challenging due to limited data, heterogeneous clinical manifestations and progression, and a frequent lack of adequate knowledge about the disease. Multiple endpoints are usually used to collectively assess the effectiveness of the investigational drug on multiple aspects of the disease. Here we propose an adaptive design based on the promising zone framework, allowing for sample size re-estimation (SSR) using interim data for a clinical trial involving multiple endpoints. The proposed SSR procedure incorporates two global tests: the ordinary least squares (OLS) test and the nonparametric permutation test. We consider two SSR approaches: one is based on power (SSR-Power) and the other on conditional power (SSR-CP). Simulation results show that the adaptive design achieves type I error control and satisfactory power. Compared with the permutation test, the OLS test has improved type I error control when the sample size is small and the timing of the interim analysis is early; while the permutation test achieves slightly higher power in most scenarios. Regarding the SSR methods, SSR-CP consistently achieves higher power than SSR-Power but often requires a larger sample size and more frequently reaches the maximum allowable sample size. The proposed design is particularly useful when the trial has a small initial sample size and has opportunity to adjust the sample size at an interim analysis to achieve adequate power.

KEYWORDS AND PHRASES: Adaptive design, Clinical trials, Cohen's d, Conditional power, Global test, Group sequential design, Ordinary least squares, Permutation, Promising zone, Rare disease, Sample size re-estimation, Small population, Totality of evidence.

1. INTRODUCTION

The complex nature of rare diseases and limited data present unique challenges in the design of clinical trials [1, 2, 3, 4]. To better capture the intended effects of an investigational treatment, multiple endpoints, rather than a single endpoint, are commonly used in rare disease trials. For example, in the ENDEAR trial [5] that evaluated the efficacy and safety of nusinersen in infants with spinal muscular atrophy (SMA), both the motor milestone response and event-free survival were used as primary endpoints. Multiple endpoints allow for a more comprehensive assessment of treatment effects, particularly when there are no historical data available to prioritize one aspect of the disease over another. Furthermore, when the drug has an effect on each of the multiple endpoints, using multiple endpoints may improve statistical power and reduce the required sample size by aggregating information in different clinically significant outcomes [6, 7]. In the setting of dual primary endpoints, where statistical significance can be claimed if any endpoint is statistically significant, there is a multiplicity issue, and the family-wise type I error rate (FWER) must be strongly controlled in order to make valid endpoint-level claims [7, 8, 9]. In this setting, multiple multiplicity adjustment methods can be used to strongly control the FWER [7, 10, 11, 12, 13].

In the setting of rare disease trials, when a single endpoint cannot be a complete representation of the treatment effect, two common methods are used to control the type I error. One is using the composite endpoint, which combines multiple clinical results with the same type of data into a single variable [7, 14]. Another is the co-primary endpoint approach [15], which requires statistical significance of all co-primary endpoints [7, 16]. In this paper, we consider the global testing method, which aggregates multivariate summary statistics [17, 18, 19] such as z-scores [3, 20, 21] and p-values [22, 23] into a single univariate test framework to assess the totality of evidence. Sun et al. [20] proposed a method using the average z-scores as a univariate test statistic, coupled with the Wilcoxon rank-sum test [20]. Li et al. [21] introduced a nonparametric approach through a permutation test based on the average z-scores. Although this approach was originally applied to combine the same data type of endpoints, it can also be generalized to apply to mixed data types of endpoints. A modification of Sun et al.'s method was developed by replacing the Wilcoxon rank-sum test with an exact small sample ordinary least squares (OLS) test, called the z-score OLS, proposed by Zhang et al.

*Corresponding author.

[3]. They further extended this approach to the hybrid OLS, which combines individual endpoint's test statistics (e.g., t-test) into a univariate OLS test statistic, with correlations among test statistics estimated via permutation.

Simulations comparing these methods with non-prioritized mixed type endpoints by Zhang et al. [3] suggest that the hybrid OLS provides type I error control and tends to be conservative, particularly in smaller sample sizes. Regarding power, both the z-score OLS and the hybrid OLS perform well across various scenarios for moderate to no correlations between endpoints.

A common challenge in the setting of rare diseases is that clinical data and scientific knowledge are limited. As a result, it may not be straightforward to pre-specify the endpoints and their magnitude of treatment effect to support regulatory approval when designing a clinical trial for development of a new therapy. Conventional fixed size designs can lead to the risk of designing an under-powered study or unnecessary resource waste in cost and timelines [3]. Adaptive sample size designs allowing modifications based on interim data that are not available at the time of the trial design stage are particularly appealing in rare disease settings. These designs can potentially increase the likelihood of trial success and optimize resource utilization, which are desirable for small patient populations [24, 25]. Among these, the promising zone design based on conditional power is a widely adopted approach [26, 27, 28, 29, 30]. The core concept is partitioning the interim conditional power into three zones: favorable, promising, and unfavorable, with sample size increases occurring only in the promising zone. Originally developed by Chen, DeMets, and Lan [27] and extended by Mehta and Pocock [26], this design defines the promising zone as having conditional power greater than or equal to a specified threshold (e.g., 50%) and utilizes the conventional test statistic to control type I error inflation associated with unblinded SSR, as an alternative to the CHW weighted test statistic [31]. A recent study [30] has shown that the CHW weighted test is uniformly more powerful than the method of Mehta and Pocock within the promising zone. Despite this, the promising zone concept, as a general framework, remains highly valuable and practical in adaptive sample size designs.

In this paper, we propose a study design that allows sample size adjustment at interim based on global testing of multiple endpoints when considering the totality of evidence in the context of rare diseases. Two global tests are considered including the OLS test and a nonparametric permutation test. Sample size re-estimation is conducted within a generalized promising zone framework, where the zones are determined based on the interim conditional power. In the promising zone, two distinct SSR approaches are considered: one is based on power (SSR-Power) and the other on conditional power (SSR-CP). For SSR-Power, the sample size is recalculated using the original power formula that is used for initial planning of the trial, but with updated estimates for the effect size and other nuisance parameters from interim data. In contrast, SSR-CP re-estimates the sample size based on the conditional power calculated using the interim data. We derive the power and conditional power formulas in the setting of multiple endpoints. The SSR-Power approach achieves power directly, irrespective of the timing of the interim analysis; while the SSR-CP approach achieves conditional power, i.e., the chance of statistical success, which incorporates the timing of the interim analysis. Additionally, we integrate the weighted combination test proposed by Cui et al. [31] and Lehmacher and Wassmer [32] to ensure rigorous type I error control. This integration allows greater flexibility in defining conditional power thresholds, which need not be identical to those proposed by Chen et al. (2004) [27] or Mehta and Pocock (2011) [26]. It is also straightforward to implement using traditional group sequential design boundaries and offers notable flexibility by allowing data from two stages to be combined without relying on interim decision criteria. We evaluate the performance of the proposed design through the simulation study under various scenarios, including settings with small sample sizes.

The remainder of the paper is organized as follows. Section 2 introduces the proposed study design, including two types of totality-of-evidence tests and two SSR procedures. Section 3 details the setup, scenarios and implementation of the simulation study. Section 4 reports the empirical results of the simulation study. Section 5 provides an example illustrating how to use the proposed design with our open-source `R` package `SSRTE`. Finally, Section 6 discusses limitations and practical considerations in implementing the method.

2. METHODS

2.1 Notations

Consider a randomized clinical trial with L stages and K endpoints that are assumed to follow normal distributions. At stage ℓ ($\ell = 1, \ldots, L$), suppose the sample sizes for the treatment group and the control group are $n_{T,\ell} = rn_\ell$ and $n_{C,\ell} = n_\ell$ respectively. Without loss of generality, we assume that a higher value in each endpoint reflects an improvement in the participant's outcome. Denote $Y_{T,k,\ell,i}$ as the kth endpoint of participant i in the treatment group at stage ℓ and $Y_{C,k,\ell,j}$ as the kth endpoint of participant j in the control group at stage ℓ. Then,

$$Y_{T,k,\ell,i} \overset{i.i.d.}{\sim} N(\mu_{T,k}, \sigma^2_{T,k}),$$
$$Y_{C,k,\ell,j} \overset{i.i.d.}{\sim} N(\mu_{C,k}, \sigma^2_{C,k}).$$

Denote $\rho_{T,pq}$ and $\rho_{C,pq}$ as the correlations between different endpoints p and q for the same participant from the treatment group and from the control group, respectively.

$$\rho_{T,pq} = Corr(Y_{T,p,\ell,i}, Y_{T,q,\ell,i}),$$
$$\rho_{C,pq} = Corr(Y_{C,p,\ell,j}, Y_{C,q,\ell,j}),$$

for all $\ell = 1, \ldots, L$. For simplicity, we further assume a common variance and correlation for endpoint k as follows

$$\sigma^2_{T,k} = \sigma^2_{C,k} = \sigma^2_k, \qquad \forall k = 1, \ldots, K;$$
$$\rho_{T,pq} = \rho_{C,pq} = \rho_{pq}, \quad \forall 1 \le p < q \le K.$$

The assumption of common variances and correlations between treatment and control arms for a certain endpoint can be interpreted as assuming a shared population-level covariance structure, representing the pooled covariance across both groups. This assumption is reasonable when participants in the treatment and control arms are drawn from comparable populations, and when most variability arises from measurement error rather than true group differences. Furthermore, assuming equal variances is a common practice in sample size calculations for single endpoints, such as continuous or binary outcomes.

2.2 Standardized Effect Size

To have a standardized measurement of the effect size of continuous endpoints with varying scales, we adopt the widely used Cohen's d [33, 6, 34] as the effect size measure, defined as the difference between two means divided by the assumed common standard deviation of treatment groups. For endpoint k, Cohen's d is given by

$$\theta_k = \frac{\mu_{T,k} - \mu_{C,k}}{\sigma_k}.$$

Based on stage ℓ data, Cohen's d can be estimated as

$$d_{k,\ell} = \hat{\theta}_k = \frac{\bar{Y}_{T,k,\ell} - \bar{Y}_{C,k,\ell}}{\hat{\sigma}_{k,\ell}}$$

where $\hat{\sigma}_{k,\ell}$ is the pooled sample standard deviation obtained from the pooled sample covariance matrix (see (2.9)), and

$$\bar{Y}_{T,k,\ell} = \frac{1}{rn_\ell}\sum_{i=1}^{rn_\ell} Y_{T,k,\ell,i} \sim N(\mu_{T,k}, \frac{\sigma^2_k}{rn_\ell}),$$
$$\bar{Y}_{C,k,\ell} = \frac{1}{n_\ell}\sum_{j=1}^{n_\ell} Y_{C,k,\ell,j} \sim N(\mu_{C,k}, \frac{\sigma^2_k}{n_\ell}),$$
$$\bar{Y}_{T,k,\ell} - \bar{Y}_{C,k,\ell} \sim N(\mu_{T,k} - \mu_{C,k}, (\frac{1}{rn_\ell} + \frac{1}{n_\ell})\sigma^2_k).$$

For each pair of endpoints p and q where $1 \le p < q \le K$,

$$Cov(\bar{Y}_{T,p,\ell} - \bar{Y}_{C,p,\ell}, \bar{Y}_{T,q,\ell} - \bar{Y}_{C,q,\ell}) = (\frac{1}{rn_\ell} + \frac{1}{n_\ell})\rho_{pq}\sigma_p\sigma_q. \tag{2.1}$$

2.3 Totality of Evidence Tests

We use a global test to integrate information across endpoints. A natural and clinically meaningful approach is to test whether the mean Cohen's d across these endpoints exceeds zero. This can be interpreted as assessing whether the treatment has an overall positive effect on various aspects of disease manifestation.

Define a global treatment effect as the mean Cohen's d:

$$\bar{\theta} = \frac{1}{K}\sum_{k=1}^{K}\theta_k. \tag{2.2}$$

At each stage, we conduct a superiority test regarding the global treatment effect:

$$H_0 : \bar{\theta} = 0 \quad \text{vs.} \quad H_1 : \bar{\theta} > 0.$$

We consider two different choices of test statistics: one based on an exact small sample OLS test, and the other on a nonparametric permutation test. Both tests, however, share a common preliminary step: standardizing the observed group difference for each endpoint into a t-statistic, and then taking the mean of these t-statistics across all endpoints. Specifically, we first assume that the true σ_k is known and express the derivation in terms of z-scores. Denote the z-score for endpoint k at stage ℓ as $z_{k,\ell}$ given by

$$z_{k,\ell} = \frac{\bar{Y}_{T,k,\ell} - \bar{Y}_{C,k,\ell}}{\sigma_k\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}} \sim N(\mu_{z_k}, 1),$$

where

$$\mu_{z_k} = \frac{\mu_{T,k} - \mu_{C,k}}{\sigma_k\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}} = \frac{\theta_k}{\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}}.$$

Let $\boldsymbol{z}_\ell = \left(z_{1,\ell}, \ldots, z_{K,\ell}\right)^T$ denote the z-score vector for all endpoints at stage ℓ, and $\mathbf{1} = (1, \ldots, 1)^T$ denote a vector of K 1's, the mean z-score across all endpoints at stage ℓ is

$$\bar{z}_\ell = \frac{1}{K}\sum_{k=1}^{K} z_{k,\ell} = \frac{1}{K}\mathbf{1}^T\boldsymbol{z}_\ell \tag{2.3}$$

Therefore, the t-statistic of each endpoint at stage ℓ and the mean of these t-statistics among endpoints are given by

$$t_{k,\ell} = \frac{\bar{Y}_{T,k,\ell} - \bar{Y}_{C,k,\ell}}{\hat{\sigma}_{k,\ell}\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}} = \frac{d_{k,\ell}}{\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}},$$
$$\bar{t}_\ell = \frac{1}{K}\sum_{k=1}^{K} t_{k,\ell} = \frac{\bar{d}_\ell}{\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}}. \tag{2.4}$$

Here, $t_{k,\ell}$ and $\bar{t}_\ell$ are obtained by substituting σ_k with $\hat{\sigma}_{k,\ell}$ in $z_{k,\ell}$ and $\bar{z}_\ell$, respectively.

2.3.1 Exact Small Sample OLS Test

An exact student's t-test [3, 17, 18] gives

$$T_\ell = \frac{\bar{t}_\ell}{se(\bar{t}_\ell)} \sim t_{v_\ell}, \tag{2.5}$$

where $v_\ell = 0.5(n_{T,\ell} + n_{C,\ell} - 2)(1 + \frac{1}{K^2})$. The numerator of T_ℓ is more technically expressed as $\bar{t}_\ell - \bar{\theta}/\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}$, which reduces to $\bar{t}_\ell$ because $\bar{\theta} = 0$ under the null hypothesis. The approximate degrees of freedom v_ℓ, as proposed by Logan and Tamhane [18], provide strict type I error control for small populations.

To find the expression of $se(\bar{t}_\ell)$, we first derive the covariance matrix of z-scores at stage ℓ. For z-scores from two endpoints p and q, $\forall 1 \le p < q \le K$, it can be shown that their correlation is $Corr(z_{p,\ell}, z_{q,\ell}) = \rho_{pq}$ (see details in Appendix A). Thus, $\boldsymbol{z}_\ell \sim \text{MVN}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$, where

$$\boldsymbol{\mu}_z = \begin{bmatrix} \mu_{z_1} \\ \vdots \\ \mu_{z_k} \end{bmatrix}, \boldsymbol{\Sigma}_z = \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1K} \\ \rho_{12} & 1 & \dots & \rho_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K} & \rho_{2K} & \dots & 1 \end{bmatrix}.$$

By (2.3), mean z-score $\bar{z}_\ell \sim N(\mu_{\bar{z}}, \sigma^2_{\bar{z}})$, with

$$\mu_{\bar{z}} = \frac{1}{K}\mathbf{1}^T\boldsymbol{\mu}_z = \frac{1}{K}\sum_{k=1}^{K}\mu_{z_k} = \frac{\bar{\theta}}{\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}} \tag{2.6}$$

$$\sigma^2_{\bar{z}} = \frac{1}{K^2}\mathbf{1}^T\boldsymbol{\Sigma}_z\mathbf{1} = \frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\rho_{pq}). \tag{2.7}$$

Therefore, $se(\bar{t}_\ell)$ can be derived by replacing the unknown parameter ρ_{pq} with its estimate $\hat{\rho}_{pq,\ell}$:

$$se(\bar{t}_\ell) = \hat{\sigma}_{\bar{z}} = \sqrt{\frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\hat{\rho}_{pq,\ell})}. \tag{2.8}$$

Similar to $\hat{\sigma}_{k,\ell}$, $\hat{\rho}_{pq,\ell}$ can also be obtained from the pooled sample covariance matrix $\boldsymbol{S}_{\ell,pooled} \in \mathbb{R}^{K\times K}$, estimated using observations from stage ℓ. Specifically,

$$\boldsymbol{S}_{\ell,pooled} = \frac{(n_{T,\ell} - 1)\mathbf{S}_{T,\ell} + (n_{C,\ell} - 1)\mathbf{S}_{C,\ell}}{n_{T,\ell} + n_{C,\ell} - 2}, \tag{2.9}$$

$\mathbf{S}_{T,\ell}$ and $\mathbf{S}_{C,\ell}$ are the sample covariance matrix for $\boldsymbol{Y}_{T,\ell} \in \mathbb{R}^{n_{T,\ell}\times K}$ and $\boldsymbol{Y}_{C,\ell} \in \mathbb{R}^{n_{C,\ell}\times K}$.

Finally, by (2.4) and (2.8), we arrive at

$$T_\ell = \frac{\bar{d}_\ell/\sqrt{\frac{1}{rn_\ell} + \frac{1}{n_\ell}}}{\sqrt{\frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\hat{\rho}_{pq,\ell})}}.$$

Denote the one-sided p-value of T_ℓ as p_{T_ℓ}, the corresponding critical value in the z-test is

$$z^*_\ell = \Phi^{-1}(1 - p_{T_\ell}),$$

where Φ^{-1} denotes the inverse of the cumulative distribution function (CDF) of the standard normal distribution.

2.3.2 Permutation Test

One advantage of the permutation test [21] is that it does not require the derivation of $\bar{t}_\ell$'s distribution or the analytic form of $se(\bar{t}_\ell)$. Instead, at stage ℓ, we randomly shuffle the treatment and control groups for a large number of permutations to generate the empirical null distribution of the mean t-statistic, denoted by $\{\bar{t}_{\ell,0}\}$. Then we estimate the chance that the observed $\bar{t}_\ell$ is greater than or equal to $\{\bar{t}_{\ell,0}\}$ under the null hypothesis of no overall treatment effect. This probability is the one-sided p-value from the permutation test, p_{P_ℓ}.

Lastly, as in the exact OLS test, we convert this p-value into the corresponding critical value in the z-test as follows:

$$z^*_\ell = \Phi^{-1}(1 - p_{P_\ell}).$$

2.4 Sample Size Re-Estimation

In the context of multiple endpoints, we consider the promising zone design framework and explore two SSR approaches. In addition, we derive the adaptive sample size procedure using the weighted combination test procedures from Cui et al. [31] and Lehmacher and Wassmer [32] to maintain strong control of type I error. Although these procedures can be applicable to adaptive trials with multiple interim stages, for simplicity, we consider a practical setting that the SSR is conducted only once at a given interim analysis. Therefore, mathematically, this adaptive trial can simply be modeled as a two-stage design ($L = 2; \ell = 1, 2$), with an interim analysis conducted immediately following stage 1 and a final analysis after stage 2. Suppose that the target power is $1 - \beta$.

2.4.1 Efficacy Boundaries

Consider the Lan-DeMets O'Brien-Fleming-type alpha spending function [35] as an illustrative example to determine the efficacy rejection boundaries, which can be calculated using `R` package `gsDesign` [36]. These boundaries are independent of the sample size, but are determined by the desired significance level α, the number of interim stages, the timing of the interim analyses, and the type of tests. Here, the term "timing" refers to the information fraction (IF), defined as the proportion of participants with the results observed in the interim analysis relative to the total number of participants planned for the study. Figure 1 displays the efficacy boundaries for a two-stage study with $\alpha = 2.5\%$, under three different timings of interim analysis. Specifically, these timings correspond to when 33.3%, 50%, or 66.7% of the planned participants have been recruited, that is, at an information fraction of $\frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$, respectively. There is a clear decrease in the efficacy boundary at interim, as indicated by the first dot on each line, when the interim analysis occurs later in the study.

The alpha spending function approach allows interim analyses to be conducted flexibly, without the need for prespecified time points [35, 37, 38, 39], although this is not an

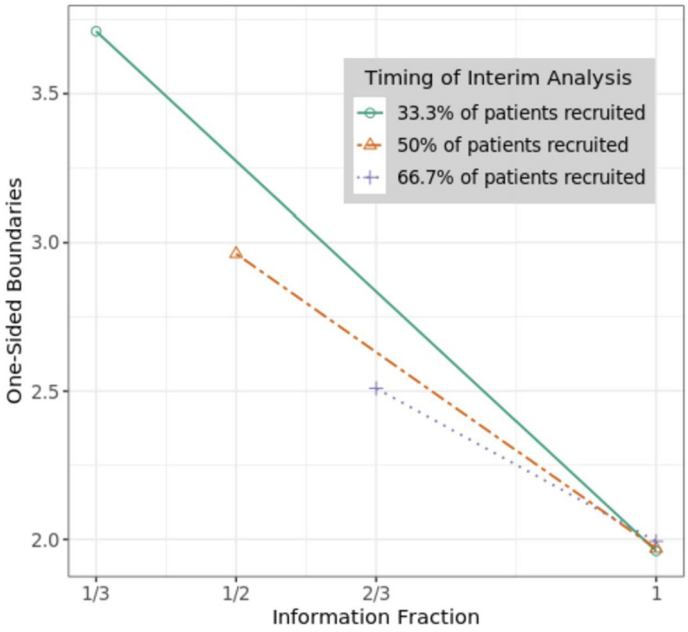

Figure 1: One-sided efficacy boundaries using the Lan-DeMets O'Brien-Fleming-type alpha spending function for a two-stage study with $\alpha = 2.5\%$, under three different timings of interim analysis. Specifically, these timings correspond to when 33.3%, 50%, or 66.7% of the planned participants have been recruited, that is, at an information fraction of $\frac{1}{3}$, $\frac{1}{2}$ or $\frac{2}{3}$, respectively.

issue in our settings. This flexibility contrasts with methods that require the number and timings of interim analyses be predetermined, such as the original Pocock and O'Brien-Fleming boundaries [40, 41].

2.4.2 Promising Zone Framework and Conditional Power

When the trial is not stopped early in the interim, we need to determine if the sample size should be increased. This decision is made based on the value of interim conditional power. Let the original total sample size planned (i.e., the combination of both stages) be defined as $N_{total} = N_C + N_T = (1+r)n = (1+r)(n_1+n_2)$, where $N_C = n = n_1 + n_2$ and $N_T = rN_C$ are the total sample sizes for the control group and treatment group, respectively. Then, the interim conditional power can be estimated as

$$CP(\bar{d}_1) = \Phi\left(\frac{\frac{\sqrt{n_1}\bar{t}_1 - \sqrt{n}Z_{\alpha_2}}{\sqrt{n_2}} + \frac{\sqrt{n_2}\bar{t}_1}{\sqrt{n_1}}}{\sqrt{\frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1})}}\right), \qquad (2.10)$$

where $\bar{d}_1$ is the observed mean Cohen's d at interim, contained in $\bar{t}_1 = \frac{\bar{d}_1\sqrt{n_1}}{\sqrt{\frac{1}{r}+1}}$. By this formula, we can see that conditional power depends on the timing of the interim analysis, the estimated mean Cohen's d at the interim, and the estimated sum of correlations across endpoints (i.e. $2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1}$). See the detailed derivation in Appendix A.

Three zones according to the conditional power are defined as

- Unfavorable: $CP(\bar{d}_1) \le CP_{min}$, do not increase the sample size;
- Promising: $CP_{min} < CP(\bar{d}_1) < 1-\beta$, increase the sample size;
- Favorable: $CP(\bar{d}_1) \ge 1-\beta$, do not increase the sample size.

In this study, we consider a fixed value for CP_{min} to illustrate the design. In practical applications, users can determine this value individually based on specific contextual considerations.

2.4.3 Sample Size Re-Estimation Approaches

If the estimated interim conditional power falls within the promising zone, the sample size will be re-estimated. It may then be increased, or maintained at its current level if the re-estimated sample size is smaller than initially planned. Moreover, we impose a fixed upper limit for the new sample size, which is twice the original sample size, that is, $N_{max,total} = 2N_{total}$, or equivalently $N_{max,C} = 2N_C$. Two distinct SSR methods are proposed, each using a different benchmark.

Method 1: SSR-Power

The first method is named SSR-Power because it uses the target power as benchmark. It is calculated with the original sample size formula used in the study planning stage to obtain the initially planned sample size. The sample size formula can be derived as follows.

Denote the true $\bar{\theta}$ under the alternative hypothesis H_1 as $\bar{\theta} = \bar{\theta}^*$. Based on (2.6), we can denote the true $\mu_{\bar{z}}$ under H_1 as

$$\mu^*_{\bar{z}} = \frac{\bar{\theta}^*}{\sqrt{\frac{1}{rN_C} + \frac{1}{N_C}}} = \frac{\sqrt{N_{total}}}{\sqrt{(\frac{1}{r}+1)(r+1)}}\bar{\theta}^*.$$

Recall $\bar{z}_\ell \sim N(\mu_{\bar{z}}, \sigma_{\bar{z}})$, so for the mean z-score at interim stage (i.e., $\bar{z}_1$) under H_1, we have $\bar{z}_1 \sim N(\mu^*_{\bar{z}}, \sigma_{\bar{z}})$. By the definition of power, i.e., $P\left\{\frac{\bar{z}_1}{\sigma_{\bar{z}}} > Z_\alpha \middle| \frac{\mu^*_{\bar{z}}}{\sigma_{\bar{z}}}\right\} = 1-\beta$, we derive the sample size formula as

$$N_{total} = \frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\rho_{pq})\left(\frac{Z_\alpha + Z_\beta}{\frac{\bar{\theta}^*}{\sqrt{(\frac{1}{r}+1)(1+r)}}}\right)^2. \qquad (2.11)$$

With interim data, we can estimate $\bar{\theta}^*$ and ρ_{pq} using $\bar{d}_1$ and $\hat{\rho}_{pq,1}$, separately. The resulting re-estimated total sam-

ple size using interim data can be expressed as:

$$\hat{N}_{total} = \frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1})\left(\frac{Z_\alpha + Z_\beta}{\frac{\bar{d}_1}{\sqrt{(\frac{1}{r}+1)(1+r)}}}\right)^2. \tag{2.12}$$

Therefore, the re-estimated total (two-stage) sample sizes for the control and treatment groups are given by:

$$M_C = min\left\{max\left\{\left\lceil\frac{\hat{N}_{total}}{1+r}\right\rceil, N_C\right\}, N_{max,C}\right\}, \tag{2.13}$$

$$M_T = \lceil rM_C\rceil. \tag{2.14}$$

Method 2: SSR-CP

The other method is named SSR-CP, where the total sample size of the control group is re-estimated based on the conditional power formula (2.10):

$$\hat{N}_C = \hat{n} \quad \text{such that}$$
$$\Phi\left(\frac{\frac{\sqrt{n_1}\bar{t}_1 - \sqrt{\hat{n}}Z_{\alpha_2}}{\sqrt{\hat{n}-n_1}} + \frac{\sqrt{\hat{n}-n_1}\bar{t}_1}{\sqrt{n_1}}}{\sqrt{\frac{1}{K^2}(K + 2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1})}}\right) = 1-\beta, \tag{2.15}$$

Similarly, the re-estimated total (two-stage) sample size for the control group is

$$M_C = min\{max\{\hat{N}_C, N_C\}, N_{max,C}\}. \tag{2.16}$$

As for the treatment group, M_T can be calculated by (2.14).

For both methods, denote the total re-estimated sample size as $M_{total} = M_C + M_T$. The new sample sizes for the control and treatment groups within stage 2 are then defined as

$$m_C = M_C - n_{C,1} \quad , \quad m_T = M_T - n_{T,1}.$$

2.4.4 Illustrative Examples of SSR Approaches

We present examples illustrating how factors including the estimated mean Cohen's d at interim ($\bar{d}_1$), the timing of interim analysis ($\frac{n_1}{n}\times 100\%$), and the estimated sum of correlations ($2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1}$) influence the re-estimated sample size under the SSR-Power and SSR-CP formulas in (2.12) and (2.15). Consider a two-stage trial designed with $K = 6$ endpoints and a planned total of $N_{total} = 100$ participants. Let the allocation ratio $r = 1$, $\frac{N_{max,total}}{N_{total}} = 2$, one-sided $\alpha = 2.5\%$, $CP_{min} = 20\%$, and target power $= 80\%$.

Figure 2 shows the sample size increase percentage ($\frac{M_{total}}{N_{total}}\times 100\%$) in different zones for a trial using SSR-Power at the interim stage. Panels (a) and (b) both have an estimated sum of correlations of 9, but the timing of the interim analysis in panel (a) is 50%, earlier than 66.7% in panel (b). Panels (b) and (c) occur at the same interim timing of 66.7%, yet panel (c) includes a significantly larger estimate of the sum of correlations, at 19 compared to 9 in panel (b). The top-row subfigures show the conditional power at the interim. The promising zone (light yellow) is defined by the area between two red dotted lines, corresponding to conditional powers of 20% and 80%. Within this zone, parts where the sample size increases (i.e., sample size increase percentage is greater than 100%) are highlighted in green, while parts where the sample size remains unchanged are shown in black in the bottom-row subfigures. The dark red dot on the left of the bottom-row subfigures indicate the re-estimated sample size reaching the upper limit and the corresponding interim mean Cohen's d. By comparing panels (a) and (b), it is evident that later timing of the analysis results in a narrower promising zone, implying larger favorable (light blue) and unfavorable (gray) zones. This suggests that with more information available in the interim, there is less need to increase the sample size. In both panels (a) and (b), the re-estimated sample size never reaches the maximum (i.e., the 200% line). The actual maximum sample size increase percentages, marked by the pink points, are 132% for panel (a) and 118% for panel (b). However, after substantially increasing the estimated sum of correlations in panel (b) from 9 to 19 (as seen in panel (c)), the promising zone becomes much wider and $N_{max,total}$ is reached within this zone. This aligns with the expectation that higher variance necessitates a larger sample size to maintain the same level of precision.

Three scenarios of a trial applying SSR-CP at the interim analysis are depicted in the three panels of Figure 3. The zones based on conditional power here are identical to those in Figure 2. Patterns mirror those seen with SSR-Power: as the timing of the interim analysis progresses from 50% in panel (a) to 66.7% in panel (b), or when the estimated sum of correlations decreases from 19 in panel (c) to 9 in panel (b), there is a significant contraction in the promising zone. This reinforces the earlier point that later interim analyses or lower correlation estimates reduce the need for extensive sample size adjustments within the promising zone. Nevertheless, compared with SSR-Power, SSR-CP tends to inflate the sample size more, especially when the estimated sum of correlations is smaller, as in panels (a) and (b). Under SSR-CP, the re-estimated sample size reaches the maximum (200%) in panel (a) and nearly reaches it in panel (b), with an actual maximum of 190%.

2.4.5 Inverse Normal Combination Method

At the final analysis, the test statistic is obtained using the inverse normal combination method [31, 32], which combines independent global univariate test statistics from stages 1 and 2. The final test statistic is calculated as:

$$z^*_{final} = \sqrt{\frac{n_1}{n_1+n_2}}z^*_1 + \sqrt{\frac{n_2}{n_1+n_2}}z^*_2 \tag{2.17}$$

This approach allows for the integration of evidence from both stages of the trial, providing a robust measure of the

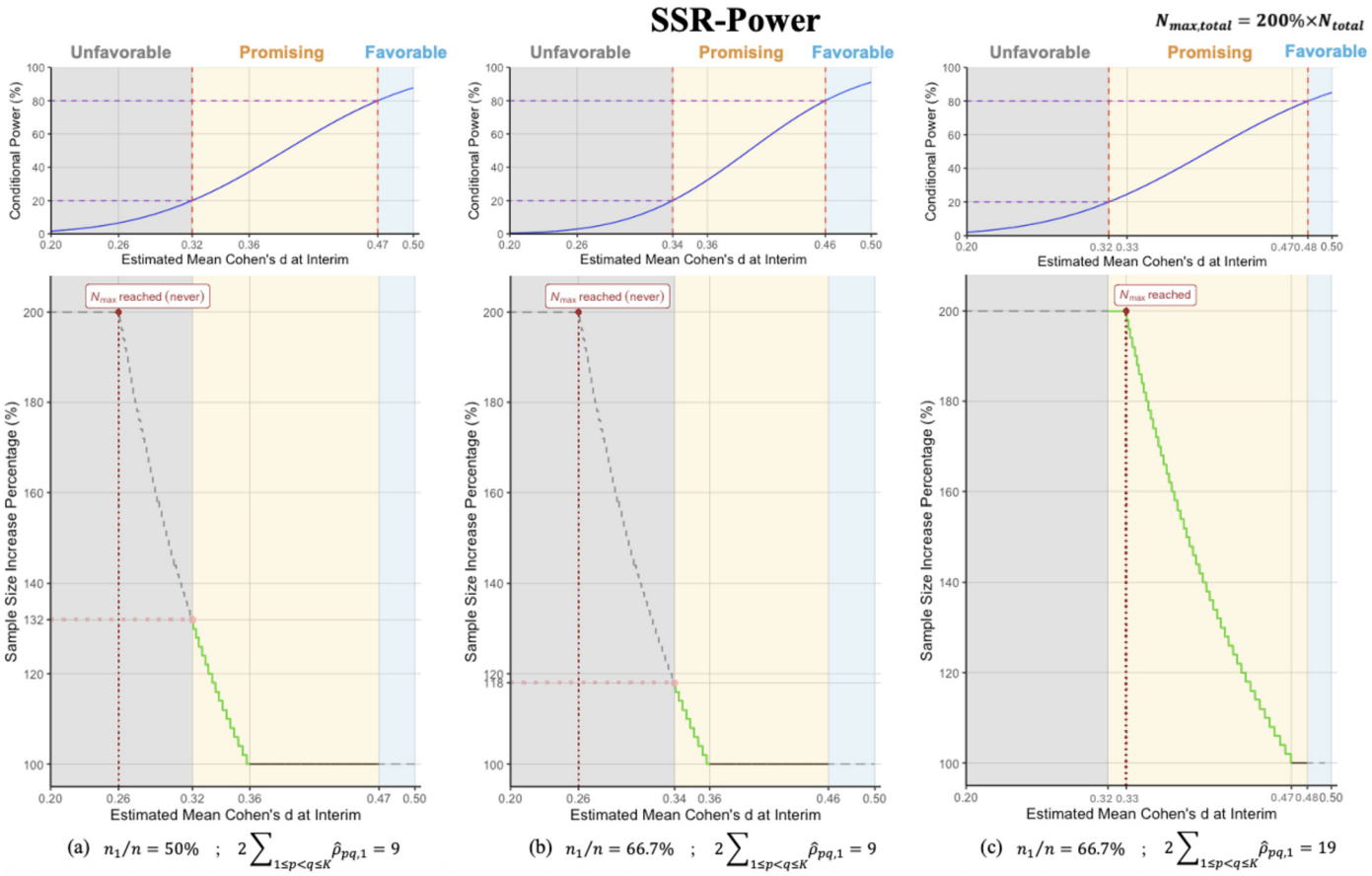

Figure 2: Illustrative example for SSR-Power: how the estimated mean Cohen's d at interim ($\bar{d}_1$), the timing of interim analysis ($\frac{n_1}{n} \times 100\%$), and the estimated sum of correlations ($2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1}$) influence the re-estimated sample size (2.12). Top row: conditional power at the interim and zoning (unfavorable - gray; promising - light yellow; favorable - light blue). Bottom row: sample size increase percentage, $\frac{M_{total}}{N_{total}} \times 100\%$ (green - increase; black - no change; gray dashed - theoretical continuation, not occurring in practice). Panels: a) Timing of interim analysis is 50%, estimated sum of correlations equals 9; b) Timing of interim analysis is 66.7%, estimated sum of correlations equals 9; c) Timing of interim analysis is 66.7%, estimated sum of correlations equals 19.

overall treatment effect. Meanwhile, it maintains the type I error by using the originally planned sample size weights regardless of the potential sample size adjustment for stage 2.

2.5 Procedures for Design and Analyses

Below we summarize procedures for implementing the developed two-stage sample size re-estimation design incorporating the totality of evidence for multiple endpoints. Suppose the interim analysis is conducted when $(100 \times \text{IF})\%$ of the planned participants have been recruited.

Initial Design

1. Calculate the initially planned sample size (N_{total}) with (2.11), where the number of endpoints (K), pairwise correlation between endpoints (ρ_{pq}), target type I error (α), and target power ($1-\beta$) should all be predetermined.
2. Obtain the planned total, stage 1 and stage 2 sample sizes for the control group (N_C, $n_{C,1}$ and $n_{C,2}$) and treatment group (N_T, $n_{T,1}$ and $n_{T,2}$), respectively.

$$\begin{aligned} N_C = n = \lceil N_{total}/(1+r) \rceil \quad &, \quad N_T = \lceil rN_C \rceil \\ n_{C,1} = n_1 = \lceil \text{IF} \times N_C \rfloor \quad &, \quad n_{T,1} = \lceil rn_1 \rceil \\ n_{C,2} = n_2 = N_C - n_1 \quad &, \quad n_{T,2} = N_T - n_{T,1}. \end{aligned}$$

Interim Stage

1. Perform one-sided global hypothesis testing for H_0: $\bar{\theta} = 0$; H_1: $\bar{\theta} > 0$ with interim data, and obtain the test statistics z_1^*.
   - Reject H_0 and stop the trial if $z_1^* > Z_{\alpha_1}$;
   - Fail to reject H_0 and continue the trial if $z_1^* \le Z_{\alpha_1}$.
2. If the trial continues, calculate the conditional power by (2.10). If $CP_{min} < CP(\bar{d}_1) < 1-\beta$, re-estimate the

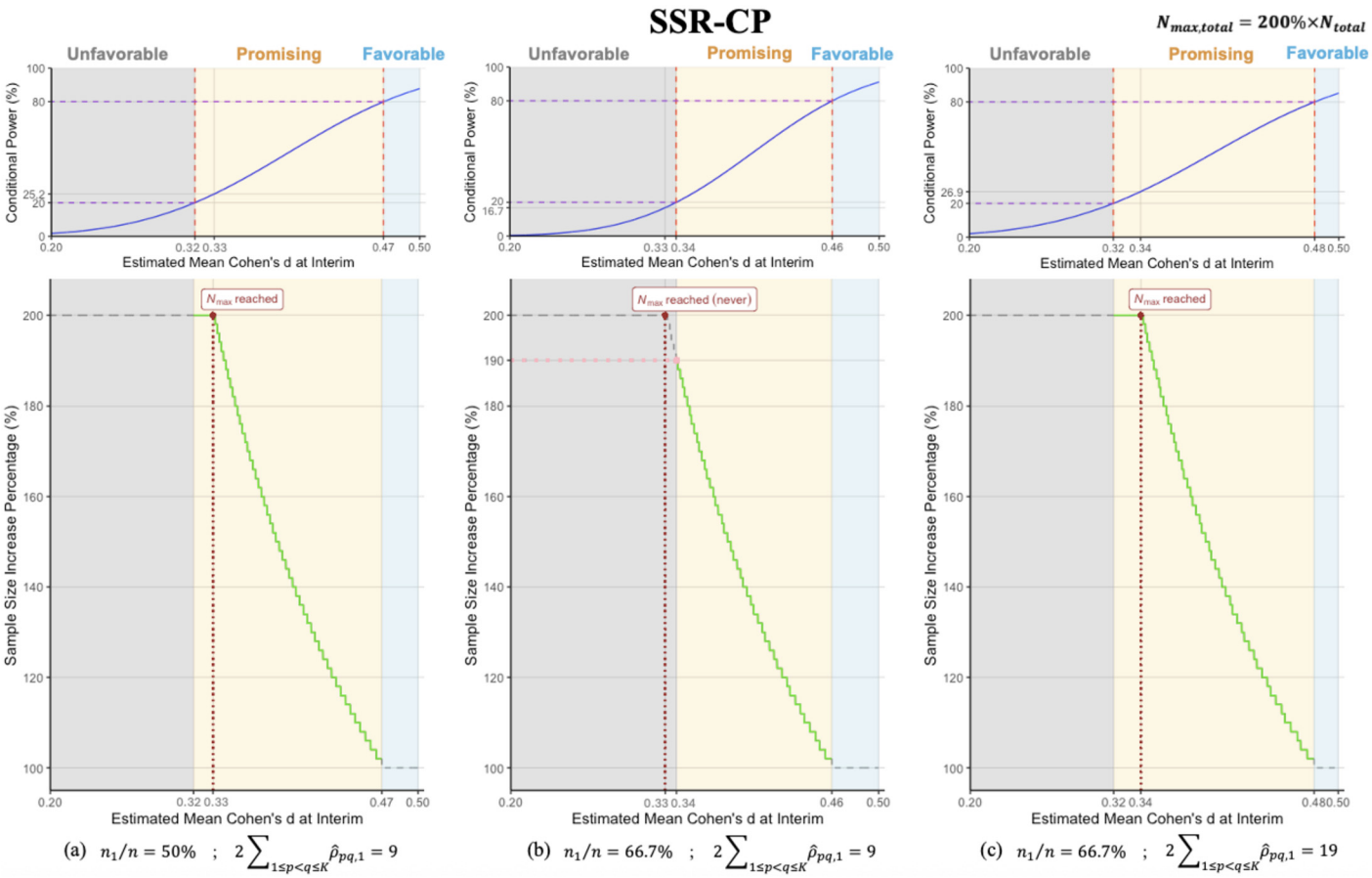

Figure 3: Illustrative example for SSR-CP: how the estimated mean Cohen's d at interim ($\bar{d}_1$), the timing of interim analysis ($\frac{n_1}{n} \times 100\%$), and the estimated sum of correlations ($2\sum_{1\le p<q\le K}\hat{\rho}_{pq,1}$) influence the re-estimated sample size (2.15). Top row: conditional power at the interim and zoning (unfavorable - gray; promising - light yellow; favorable - light blue). Bottom row: sample size increase percentage, $\frac{M_{total}}{N_{total}} \times 100\%$ (green - increase; black - no change; gray dashed - theoretical continuation, not occurring in practice). Panels: a) Timing of interim analysis is 50%, estimated sum of correlations equals 9; b) Timing of interim analysis is 66.7%, estimated sum of correlations equals 9; c) Timing of interim analysis is 66.7%, estimated sum of correlations equals 19.

sample size by SSR-Power or SSR-CP for stage 2 as m_C and m_T.

Final Analysis Stage

1. Perform one-sided global hypothesis testing for H_0: $\bar{\theta} = 0$; H_1: $\bar{\theta} > 0$ with stage 2 data only, and obtain the test statistics z_2^*.
2. Calculate the final stage test statistics z_{final}^* by using the inverse normal combination method that combines independent global univariate test statistics from stages 1 and 2 by (2.17).
3. Final rejection rule:
   - Reject H_0 if $z_{final}^* > Z_{\alpha_2}$;
   - Otherwise, fail to reject H_0.

3. SIMULATION STUDY

To investigate the impact of the proposed adaptive design on type I error control and power in various scenarios, we performed an extensive series of simulations with a two-stage setting and a target power of 80%. The trial considered a total of six clinically important endpoints, generated from a multivariate normal distribution with arm-specific means $\boldsymbol{\mu}_T$, $\boldsymbol{\mu}_C$ and a common covariance matrix $\boldsymbol{\Sigma}$. The control mean was set to $\boldsymbol{\mu}_C = \boldsymbol{\mu} = (4, 6, 4, 5, 7, 6)^T$, and the treatment mean was defined by the overall treatment effect and the standard deviation of each endpoint as $\boldsymbol{\mu}_{T,k} = \boldsymbol{\mu}_{C,k} + \bar{\theta}\sigma_k$, $\forall k = 1, \ldots, 6$. For $\boldsymbol{\Sigma}$, two sets of values were considered: $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. $\boldsymbol{\Sigma}_0$ assumes equal variances and equal correlations across the six endpoints, having the diagonal values of 1 and the off-diagonal values of 0.3. That is, standard deviation $\sigma_k = 1$, $\forall k = 1, \ldots, 6$ and correlation coefficient $\rho_{pq} = 0.3$, $\forall 1 \le p < q \le 6$. In contrast, $\boldsymbol{\Sigma}_1$ as-

sumes varying variances and correlations across endpoints, with $\boldsymbol{\sigma} = (0.5, 0.5, 1, 1, 2, 2)^T$ and

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.1 & 0.3 & 0.7 & 0.1 & 0.3 \\ 0.1 & 1 & 0.7 & 0.1 & 0.3 & 0.7 \\ 0.3 & 0.7 & 1 & 0.1 & 0.3 & 0.7 \\ 0.7 & 0.1 & 0.1 & 1 & 0.1 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.1 & 1 & 0.7 \\ 0.3 & 0.7 & 0.7 & 0.3 & 0.7 & 1 \end{bmatrix}.$$

Thus,

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.25 & 0.025 & 0.15 & 0.35 & 0.1 & 0.3 \\ 0.025 & 0.25 & 0.35 & 0.05 & 0.3 & 0.7 \\ 0.15 & 0.35 & 1 & 0.1 & 0.6 & 1.4 \\ 0.35 & 0.05 & 0.1 & 1 & 0.2 & 0.6 \\ 0.1 & 0.3 & 0.6 & 0.2 & 4 & 2.8 \\ 0.3 & 0.7 & 1.4 & 0.6 & 2.8 & 4 \end{bmatrix}.$$

The simulation can be reproduced using the `SSRTE_simstudy()` function from our `R` package `SSRTE`, which is available on GitHub. Installation instructions are provided in Section 5.

3.1 Type I Error Control

Type I error was evaluated under the null hypothesis H_0: $\bar{\theta} = 0$, i.e., there is no overall treatment effect ($\boldsymbol{\mu}_T = \boldsymbol{\mu}_C$).

We evaluated 60 scenarios derived from a combination of four key factors. First, we considered five different settings for the initially planned two-stage total sample size with a consistent allocation ratio of 1, denoted $N_{total} = 60, 120, 180, 240, 2000$. Second, we evaluated three different sample size adaptation approaches: i) no SSR, ii) SSR-Power, and iii) SSR-CP. Regarding the analysis of the totality of evidence, we compared two methods: the exact small sample OLS test ("OLS") and the permutation test ("Permutation"). Lastly, two different timings of interim analysis (50% and 66.7%) were investigated. We conducted 1 million simulations for the scenarios using the OLS method and 10,000 simulations for the scenarios using the permutation method, with 2,000 permutations in each simulation. The target significance level for one-sided hypothesis testing was set at $\alpha = 2.5\%$.

Stage 1 and stage 2 data were generated separately. Stage 1 data for the treatment and control groups were based on the initial sample size planned: $\mathbf{Y_{T,1}} \in \mathbb{R}^{rn \times 6}$ and $\mathbf{Y_{C,1}} \in \mathbb{R}^{n \times 6} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Stage 2 data were based on the new sample size, which was only generated if the trial did not stop at the interim analysis: $\mathbf{Y_{T,2}} \in \mathbb{R}^{rm \times 6}$ and $\mathbf{Y_{C,2}} \in \mathbb{R}^{m \times 6} \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

3.2 Power Assessment

Unlike type I error, the power was evaluated under the alternative hypothesis H_1: $\bar{\theta} > 0$. Therefore, the scenarios considered for power assessment differed slightly from those for type I error control. Specifically, we evaluated 24 scenarios influenced by four factors. Three of these factors were the same as in the type I error simulations: three sample size adaptation approaches, two global tests for totality of evidence, and two timings of interim analysis. However, instead of fixed sample sizes, we fixed the target power at 80% and introduced two settings for the *initially planned sample size* (IPSS, i.e., N_{total}), labeled "Original" and "Underestimated." Based on the sample size formula (2.11) and the true sums of correlations (9 under $\boldsymbol{\Sigma}_0$ and 11 under $\boldsymbol{\Sigma}_1$), the underlying true mean Cohen's d ($\bar{\theta}$) required to achieve a N_{total} of 100 and 200 is as follows:

- For $N_{\text{total}} = 100$: $\bar{\theta} = 0.362$ under $\boldsymbol{\Sigma}_0$; $\bar{\theta} = 0.385$ under $\boldsymbol{\Sigma}_1$.
- For $N_{\text{total}} = 200$: $\bar{\theta} = 0.256$ under $\boldsymbol{\Sigma}_0$; $\bar{\theta} = 0.272$ under $\boldsymbol{\Sigma}_1$.

In the Original setting, we simulated data assuming a true mean Cohen's d of 0.362 under $\boldsymbol{\Sigma}_0$ or 0.385 under $\boldsymbol{\Sigma}_1$ and planned for 100 participants in total. In contrast, the Underestimated scenario simulated data with a true mean Cohen's d of 0.256 under $\boldsymbol{\Sigma}_0$ or 0.272 under $\boldsymbol{\Sigma}_1$, yet still planned for 100 participants, effectively underestimating the required sample size of 200. This scenario thus reflects an overly optimistic assumption about the treatment effect size, where it is presumed to be 0.362 or 0.385 instead of the actual 0.256 or 0.272.

In addition to reporting empirical power, we also reported the expected sample size (ESS) and the maximum sample size (MSS) observed in the simulated trials. The ESS was calculated as the average of the actual sample sizes in all simulated trials. For trials that stopped at the interim stage, the actual sample size enrolled up to that point was used. The MSS was the maximum sample size observed in all simulations.

4. SIMULATION RESULTS

Because the simulation results under $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ are highly similar, we present only the results under $\boldsymbol{\Sigma}_0$ in the main text and include the results under $\boldsymbol{\Sigma}_1$ in Appendix B.

4.1 Type I Error Control

The results of the empirical type I error associated with each scenario under $\boldsymbol{\Sigma}_0$ are shown in Figure 4. From left to right, the first two panels display results when 50% of participants have been recruited in the interim, while the two right panels correspond to a recruitment level of 66.7% in the interim. Moreover, the first and third panels depict results from the exact small sample OLS test, while the second and fourth panels present results from the permutation test. The x-axis represents the initial sample sizes planned (N_{total}), and the different sample size adaptation scenarios are discriminated by lines with different shapes and colors.

For each interim analysis timing, for the exact small sample OLS test, the empirical type I error remains close to the

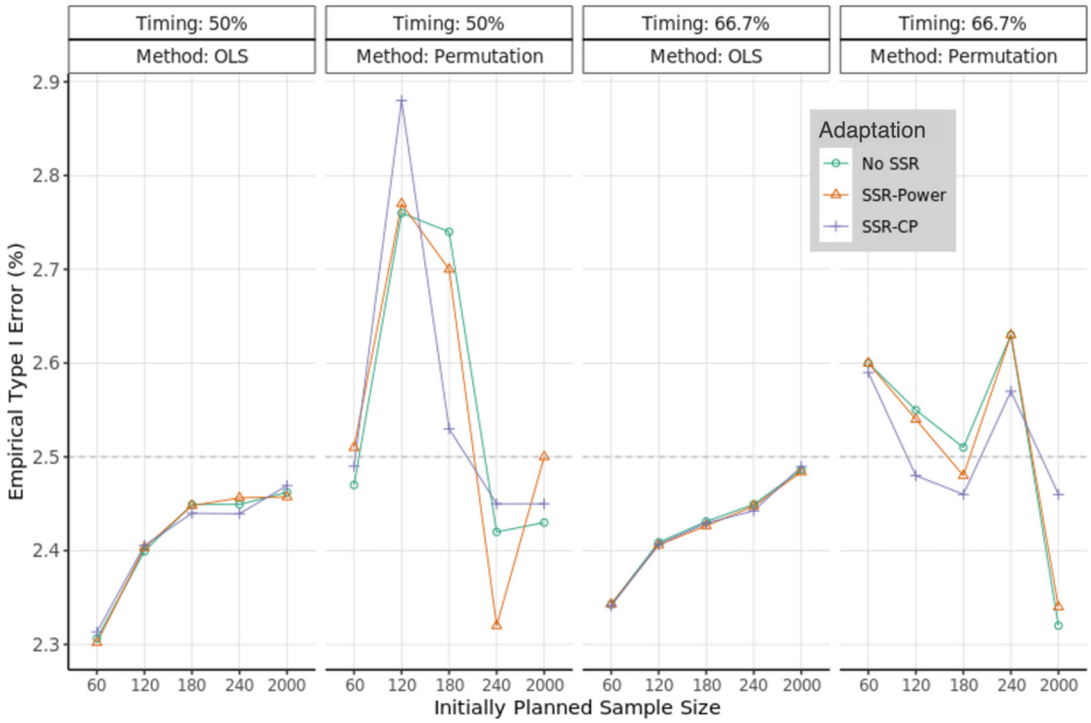

Figure 4: Empirical type I error associated with each scenario (under $\mathbf{\Sigma}_0$), with a one-sided significance level of $\alpha = 2.5\%$.

nominal level of 2.5% (marked by the dashed line) and is strongly controlled across different sample sizes and adaptation settings. There is an increasing trend from smaller population to larger population. That is, for smaller sample size the type I error control is more conservative. Although type I error tends to increase with sample size, it is still well controlled within the 2.5% line. This reflects the design of this method, which aims to provide strict type I error control for a small population. A rationale for this upward trend is that the degrees of freedom of the exact t-test depend on the sample sizes (see (2.5)), where a larger sample size leads to higher degrees of freedom and a higher p-value of the test statistic, making it easier to falsely reject the null hypothesis. Moreover, the type I error rates of all three adaptations for the OLS test are nearly identical, particularly when the interim analysis is conducted at a later stage. This further confirms that our SSR design results in minimal inflation of the type I error.

In contrast, scenarios with the permutation test exhibit greater variability in type I error rates, with the highest empirical type I error almost reaching 2.9%. However, the error rates fluctuating around 2.5% are reasonably controlled.

Comparing the two different timings of interim analysis, there is significantly less variation in adaptations when the analysis occurred later, specifically at 66.7%. Both the OLS test and the permutation test demonstrate type I errors closer to the target of 2.5%. This alignment is more pronounced with the permutation test, while the OLS test shows only marginal closeness to the target across all initially planned sample sizes, with the exception of when the sample size is set at 180 or 240.

4.2 Power Assessment

Table 1 details the empirical power, ESS and MSS for different interim timings, IPSS types, and adaptation methods for both OLS and permutation tests, regardless of the interim results. Typically, power reaches the target of 80% without the need for sample size re-estimation (SSR) when the original planned sample size is used. However, the power drops to around 50% when the initially planned sample size is underestimated, regardless of the adaptation method used.

Notably, the permutation test consistently shows slightly higher power compared to the exact small sample OLS test. Meanwhile, the OLS test tends to exhibit larger ESS values. For most cases, the MSS for both tests is twice the initially planned sample sizes, with the exception of the SSR-Power adaptation with both the permutation and OLS tests, particularly when the interim analysis time is 66.7%.

The gains in power from SSR are modest in the Table 1, likely due to the fact that only about 24.8% of scenarios actually requiring a sample size re-estimation, that is, falls in promising zone (see Table 2). This phenomenon is also noted by Mehta and Pocock [26]. Despite this, SSR-CP consistently achieves higher power across all settings compared to SSR-Power, necessitating larger ESS and MSS.

A more detailed and fair assessment is presented in Table 3, which focuses on zone-specific empirical power based

Table 1. Empirical power for each adaptation, regardless of the interim results (under $\boldsymbol{\Sigma}_0$), with a target power of 80%.

Timing of Interim	IPSS Type	Adaptation	OLS Test			Permutation Test		
			Power	ESS	MSS	Power	ESS	MSS
50%	Original	No SSR	78.24	94	100	78.72	92	100
		SSR-Power	78.95	96	200	79.28	94	192
		SSR-CP	81.07	110	200	81.42	107	200
	Underestimated	No SSR	48.83	98	100	49.86	97	100
		SSR-Power	49.60	100	200	50.53	99	178
		SSR-CP	52.72	111	200	53.33	110	200
66.70%	Original	No SSR	77.86	88	100	77.90	87	100
		SSR-Power	78.26	89	170	78.33	88	158
		SSR-CP	79.55	99	200	79.45	97	200
	Underestimated	No SSR	48.41	95	100	49.15	94	100
		SSR-Power	48.77	95	180	49.44	95	170
		SSR-CP	50.58	103	200	50.90	102	200

Note: "IPSS Type" represents initially planned sample size settings, which has two types: "Original" and "Underestimated." "ESS" and "MSS" represent the expected sample size and maximum sample size, respectively. When there is no SSR and the sample size is fixed, the MSS is always the same as N_{total}. The significance level α is 2.5%.

Table 2. Empirical probabilities of interim results for adaptations with SSR (under $\boldsymbol{\Sigma}_0$).

Timing of Interim	IPSS Type	Interim Result	Percentage (%)	
			OLS	Permutation
50%	Original	Cont. - Favorable	17.21	13.38
		Cont. - Promising	31.45	30.84
		Cont. - Unfavorable	39.53	39.90
		Reject, Stop	11.81	15.88
	Underestimated	Cont. - Favorable	8.80	7.39
		Cont. - Promising	24.78	24.57
		Cont. - Unfavorable	62.27	62.23
		Reject, Stop	4.14	5.81
66.7%	Original	Cont. - Favorable	0.03	0.09
		Cont. - Promising	21.31	19.44
		Cont. - Unfavorable	42.51	42.93
		Reject, Stop	36.15	37.54
	Underestimated	Cont. - Favorable	0.01	0.04
		Cont. - Promising	15.80	14.23
		Cont. - Unfavorable	68.34	68.70
		Reject, Stop	15.84	17.03

Note: In the "Interim Results" column, "Cont." stands for "Continue," indicating that the trial will not be rejected at interim, with subsequent zone decisions based on the interim conditional power. These percentages are the same for both SSR-Power and SSR-CP adaptations.

Table 3. Empirical power for adaptations with SSR, conditional on interim results (under $\boldsymbol{\Sigma}_0$), with a target power of 80%.

Timing of Interim	IPSS Type	Interim Result & Adaptation	OLS Test			Permutation Test		
			Power	ESS	MSS	Power	ESS	MSS
50%	Original	Cont. - Favorable	96.57	100	100	96.94	100	100
		Cont. - Promising & SSR-Power	90.67	106	200	90.73	106	192
		Cont. - Promising & SSR-CP	97.40	149	200	97.67	149	200
		Cont. - Unfavorable	55.67	100	100	56.27	100	100
		Reject, Stop	100.00	50	50	100.00	50	50
	Underestimated	Cont. - Favorable	89.21	100	100	89.17	100	100
		Cont. - Promising & SSR-Power	75.98	107	200	77.61	107	178
		Cont. - Promising & SSR-CP	88.57	153	200	89.01	153	200
		Cont. - Unfavorable	30.15	100	100	30.63	100	100
		Reject, Stop	100.00	50	50	100.00	50	50
66.7%	Original	Cont. - Favorable	95.11	100	100	88.89	100	100
		Cont. - Promising & SSR-Power	92.72	105	170	93.00	105	158
		Cont. - Promising & SSR-CP	98.76	151	200	98.77	152	200
		Cont. - Unfavorable	52.52	100	100	52.71	100	100
		Reject, Stop	100.00	66	66	100.00	66	66
	Underestimated	Cont. - Favorable	89.15	100	100	75.00	100	100
		Cont. - Promising & SSR-Power	82.32	105	180	82.64	106	170
		Cont. - Promising & SSR-CP	93.73	153	200	92.90	155	200
		Cont. - Unfavorable	29.13	100	100	30.01	100	100
		Reject, Stop	100.00	66	66	100.00	66	66

Note: The zone-specific power for interim results as "Cont. - Favorable" and "Cont. - Unfavorable" are the same for both SSR adaptation types; while the power for any interim result being "Reject, Stop" is consistently 100%.

on interim results. When the interim analysis occurs after 50% of the participants have been recruited and the initial sample size is underestimated, the power increases significantly. For OLS and permutation tests without SSR, the power is approximately 48.8% and 49.9% (row 4 of Table 1), respectively, but increases to 76.0% and 77.6% with SSR-Power (row 7 of Table 3), and even further to 88.6% and 89.0% with SSR-CP (row 8 of Table 3). A similar pattern is observed at the timing of the interim analysis 66.7%, where the power without SSR (OLS: 48.4%, Permutation: 49.2%, see Table 1) increases to 82.3% and 82.6% for SSR-Power and to 93.7% and 92.9% for SSR-CP (see Table 3). It is evident that SSR-Power generally aligns closely with the target power of 80%, while SSR-CP often exceeds this target, leading to significantly higher ESS values in all scenarios.

5. APPLICATION: DESIGN AND ANALYSIS ILLUSTRATION WITH R PACKAGE SSRTE

We present an illustrative example demonstrating how users can implement the proposed design using our user-friendly `R` package `SSRTE`, available on https://github.com/Slan1997/SSRTE. The package can be easily installed and loaded with the following `R` commands:

```
devtools::install_github("Slan1997/SSRTE")
library(SSRTE)
```

Suppose we aim to design a two-stage clinical trial with six continuous endpoints, an allocation ratio of $r = 1$, a one-sided type I error rate of $\alpha = 2.5\%$, a target power of $1 - \beta = 80\%$, and an interim analysis planned at 50% information fraction.

At the initial design stage, based on limited prior information, we assume that the best estimate of the mean Cohen's d ($\bar{\theta}$) is 0.4, and the pairwise correlation among endpoints is $\rho_{pq} = 0.5$ for all $1 \leq p < q \leq 6$. The initially planned total and stage 1 sample sizes can be calculated using the function `get_initial_tot_ss()` as follows:

```
init_ss <- get_initial_tot_ss(
  beta0     = 0.20,
  alpha0    = 0.025,
  n_endpts  = 6,
  r         = 1,
  timing0   = 0.50,
  theta_k   = 0.40,
```

```
  rho       = diag(1 - 0.5, n_endpts) +
              matrix(0.5, n_endpts, n_endpts))
init_ss
# $N_tot0            [1] 114.4628
# $N_ct              [1] 58
# $N_trt             [1] 58
# $n_ct              [1] 29
# $n_trt             [1] 29
# $N_interim_actual  [1] 58
# $N_total_actual    [1] 116
# $timing_actual     [1] 0.5
```

The planned total sample size is 116, with 58 participants in the interim stage, equally divided between treatment and control groups.

Next, stage 1 data is collected. For illustration purposes, a simulated data set is used as the true data, with a true control arm mean vector $\boldsymbol{\mu}_C = (4, 6, 4, 5, 7, 6)^T$, a true mean Cohen's d of 0.3, shared standard deviations across arms $\boldsymbol{\sigma} = (1.5, 0.5, 1.6, 1, 0.6, 1.2)^T$ and the true correlation matrix

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & 0.1 & 0.3 & 0.5 & 0.1 & 0.3 \\ 0.1 & 1 & 0.5 & 0.1 & 0.3 & 0.5 \\ 0.3 & 0.5 & 1 & 0.1 & 0.3 & 0.5 \\ 0.5 & 0.1 & 0.1 & 1 & 0.1 & 0.3 \\ 0.1 & 0.3 & 0.3 & 0.1 & 1 & 0.5 \\ 0.3 & 0.5 & 0.5 & 0.3 & 0.5 & 1 \end{bmatrix}.$$

From a reproducible research perspective, using a simulated dataset that can be publicly shared is preferable to relying on real data that cannot be made available.

The control and treatment data for stage 1 are generated using the following code:

```
sim1 <- simulate_example_one_stage_data(
  n_trt  = 29,
  n_ct   = 29,
  n_endpts = 6,
  exp_mean_cohen_d = 0.3,
  mu_ct = c(4, 6, 4, 5, 7, 6),
  sd  = c(1.5, 0.5, 1.6, 1, 0.6, 1.2),
  rho = matrix(c(
    1.0, 0.1, 0.3, 0.5, 0.1, 0.3,
    0.1, 1.0, 0.5, 0.1, 0.3, 0.5,
    0.3, 0.5, 1.0, 0.1, 0.3, 0.5,
    0.5, 0.1, 0.1, 1.0, 0.1, 0.3,
    0.1, 0.3, 0.3, 0.1, 1.0, 0.5,
    0.3, 0.5, 0.5, 0.3, 0.5, 1.0
  ), nrow = 6, byrow = TRUE),
  seed = 42)
str(sim1)
# List of 2
#  $ y_trt: num [1:29, 1:6] 6.687 7.454 2.167 ...
#  $ y_ct : num [1:29, 1:6] 5.13 4.16 4.29 ...
```

The interim analysis is then performed using the function `interim_analysis()`. Suppose the exact OLS test and SSR-CP are used, we can create an interim analysis report by following commands (if the permutation test is used, a random seed is needed to ensure reproducibility):

```
# set.seed(1) # seed needed for "Permutation"
ia <- interim_analysis(
  y_trt1 = sim1$y_trt,
  y_ct1  = sim1$y_ct,
  N_trt = 58,
  N_ct  = 58,
  N_ct_max = 120,
  timing_actual = 0.5,
  alpha0 = 0.025,
  beta0 = 0.2,
  alloc_rate = 1,
  n_endpts = 6,
  SSR_type = "SSR-CP",
  global_test_type = "exact OLS")
print(ia)
```

```
── Interim analysis ──────────────────────────────────
Decision: Fail to reject H0 at interim; continue the trial.
Global test: exact OLS
SSR type: SSR-CP
── Results ───────────────────────────────────────────
Interim boundary (C1): 2.963
Test statistic (Z): 2.070
p-value: 0.019
── Estimates ─────────────────────────────────────────
Mean Cohen's d: 0.348
A_obs (SE of bar_t_1): 0.611
Sum of pairwise correlations: 7.427
── SSR ───────────────────────────────────────────────
Estimated conditional power: 0.413
Promising zone? yes
Estimated control N (before capping): 90
Max allowed control N: 120
── Re-estimated Sample sizes ─────────────────────────
Stage 1 | Control: 29, Treatment: 29
Stage 2 | Control: 61, Treatment: 61
Final | Control: 90, Treatment: 90
```

Figure 5: Interim analysis report from `SSRTE` package.

The output of `print(ia)` is presented in Figure 5, which indicates that the null hypothesis is not rejected at interim ($z_1^* = 2.07 < Z_{\alpha_1} = 2.96$). The conditional power is estimated as 41.3%, which falls within the promising zone $(0.2, 0.8)$, leading to a re-estimation of the sample size: 61 participants per arm for stage 2 (122 total) and 90 per arm for the overall final (180 total).

A new set of stage 2 data is simulated to represent the true data collected in this stage:

```
sim2 <- simulate_example_one_stage_data(
  n_trt = 61, n_ct = 61, ..., seed = 58)
```

```
str(sim2)
# List of 2
# $ y_trt: num [1:61, 1:6] 3.49 4.63 2.74 ...
# $ y_ct : num [1:61, 1:6] 4.28 4.4 3.15 ...
```

The same underlying distributional parameters as those used in `sim1` are applied here and are therefore omitted.

Finally, the final analysis is conducted using the function `final_analysis()`, which takes the interim analysis results and stage 2 data as inputs:

```
fa <- final_analysis(
  interim    = ia,
  y_trt2     = sim2$y_trt,
  y_ct2      = sim2$y_ct,
  alloc_rate = 1)
print(fa)
```

```
── Final analysis ──────────────────────────────
Decision: Reject H0 at final analysis.
── Combined result ─────────────────────────────
Final Z: 3.477
Final p-value: 0.000
Final boundary (C2): 1.969
── Interim component ───────────────────────────
Interim Z (Z1): 2.070
Interim boundary (C1): 2.963
Timing: 0.500
── Stage 2 component ───────────────────────────
Stage 2 Z (Z2): 2.848
Stage 2 p-value: 0.002
Stage 2 n: control = 61, treatment = 61
── Stage 2 estimates ───────────────────────────
Mean Cohen's d (stage 2): 0.359
A_obs2 (SE of bar_t_2): 0.670
Sum rho (stage 2): 10.159
```

Figure 6: Final analysis report from `SSRTE` package.

The final analysis report (Figure 6) shows that $z^*_{final} = 3.45 > Z_{\alpha_2} = 1.97$, and therefore the null hypothesis is rejected.

6. DISCUSSION

We have proposed a sample size re-estimation design based on the promising zone framework that incorporates two distinct totality of evidence methods to perform a global test of mean effect size across multiple endpoints. Specifically, we begin by considering an exact OLS test tailored for small sample sizes, as the asymptotic theory in the approximate z-test becomes unreliable in rare populations where sample sizes are typically very small [3]. Additionally, we adopt a nonparametric test due to its robustness and independence from assumptions about the underlying distribution of the test statistics. For the SSR approach, we explore two distinct methods: one is based on power (SSR-Power) and the other on conditional power (SSR-CP). To facilitate these approaches, we have derived the initial sample size formula as well as a conditional power formula capable of handling global tests for multiple endpoints.

To assess the impact of our approach on type I error control and power, we conducted a simulation study under various scenarios, including different timings of interim analyses and initial sample size settings, within a two-stage SSR design framework involving six continuous endpoints.

The results of our simulation study demonstrate that the proposed design, which incorporates both the totality of evidence tests and SSR methods, effectively controls type I error rates in trials of rare diseases with multiple continuous endpoints. In particular, the OLS test offers stricter type I error control in smaller sample settings, while the permutation test experiences a large variability in all scenarios due to its random resampling nature. Regarding power assessment, in general, the SSR design maintains adequate power even when the initially planned sample size is underestimated or the treatment effect size is overly optimistic. Specifically, the permutation test achieves slightly higher power and requires fewer samples in most scenarios. Consequently, we recommend the permutation test for situations where strict type I error control is not required, and when users prefer a less assumption-dependent approach with a moderately timed interim analysis. Furthermore, SSR-CP consistently achieves higher power than SSR-Power but tends to require larger sample sizes and more frequently reaches the maximum allowable sample size. Lastly, conducting the interim analysis at a later stage consistently helps reduce the variability in the type I error of the permutation test and enhances the trial's power for both tests.

Our study has limitations. We consider situations where there is no clear hierarchical structure among endpoints. For prioritized endpoints, more specialized methods may be considered [42, 43]. Moreover, our current simulation assumes that all endpoints are continuous, whereas some trials can include other data types. For example, when dealing with binary endpoints, using Cohen's d as the global effect size measure may not be appropriate, since the assumption of common variance is violated by nature. In Chapter 4.3.8 of Zhang et al. (2023) [3], the authors proposed using z-scores to handle mixed types of endpoints and conducted simulations comparing different tests, such as the z-score OLS, the permutation test and hybrid OLS. However, their z-scores for all continuous, binary and ordinal endpoints are obtained by "subtracting the pooled groups' mean and dividing the pooled groups' standard deviation." Thus, these z-scores also assume a common variance, which suffers from the same limitation as Cohen's d when applied to binary endpoints. More research is needed to develop methods that can appropriately accommodate different types of endpoints.

Furthermore, to test the totality of evidence we use the unweighted average of Cohen's d across endpoints, which

implicitly assumes that all endpoints have equal clinical importance. When endpoints have heterogeneous clinical relevance to the treatment, a weighted average of Cohen's d can be used instead to reflect their relative importance. Note that the use of Cohen's d in this work does not replace endpoint-specific estimands or model-based analyses. Rather, it is used as a standardized parameter to facilitate a global assessment of treatment effect across multiple endpoints measured on different scales. This approach is intended for settings in which multiple clinically relevant endpoints exist, a single primary endpoint may be difficult to prespecify, and a global assessment is scientifically justified. Endpoint-specific analyses and clinically interpretable estimands remain essential and complementary components of the overall analysis strategy.

In addition, we only implement efficacy boundaries in our sequential design. However, it is possible to add futility boundaries if there is sufficient insight on the minimum clinically important difference. In such cases, during interim analysis, the critical values for efficacy and futility, denoted as $Z_{e,1}$ and $Z_{f,1}$, respectively, could be used to refine the rejection rule at the interim stage as follows:

- Reject H_0 and stop the trial if $\bar{z}_1^* > Z_{e,1}$;
- Fail to reject H_0 and stop the trial if $\bar{z}_1^* < Z_{f,1}$;
- Fail to reject H_0 and continue the trial if $Z_{f,1} \leq \bar{z}_1^* \leq Z_{e,1}$.

Finally, the current design may lead to an excessive boost in power when initial estimates are accurate, as it allows for retaining or increasing the sample size but not decreasing it. This could potentially result in an inefficient use of resources for some studies. Therefore, decisions on whether or not to increase the sample size must be made carefully and tailored to the specific requirements of each study. The study with adaptive sample size is more complex in conduct; therefore, it is essential to perform extensive simulations to assess the operating characteristics. In addition, for trial integrity considerations, information access and the process of performing SSR should be well planned in advance.

APPENDIX A. DERIVATIONS

A.1 Covariance & Correlation

For each pair of endpoints p and q where $1 \leq p < q \leq K$, their correlation can be derived as follows:

$$\begin{aligned}
&Cov(\bar{Y}_{T,p,\ell} - \bar{Y}_{C,p,\ell}, \bar{Y}_{T,q,\ell} - \bar{Y}_{C,q,\ell}) \\
&= Cov(\bar{Y}_{T,p,\ell}, \bar{Y}_{T,q,\ell}) + Cov(\bar{Y}_{C,p,\ell}, \bar{Y}_{C,q,\ell}) \\
&= Cov(\frac{1}{rn}\sum_{i=1}^{rn} Y_{T,p,\ell,i}, \frac{1}{rn}\sum_{i=1}^{rn} Y_{T,q,\ell,i}) \\
&\quad + Cov(\frac{1}{n}\sum_{j=1}^{n} Y_{C,p,\ell,j}, \frac{1}{n}\sum_{j=1}^{n} Y_{C,q,\ell,j}) \\
&= \frac{1}{(rn)^2}\sum_{i=1}^{rn} Cov(Y_{T,p,\ell,i}, Y_{T,q,\ell,i}) \\
&\quad + \frac{1}{n^2}\sum_{j=1}^{n} Cov(Y_{C,p,\ell,j}, Y_{C,q,\ell,j}) \\
&= \frac{1}{(rn)^2} rn\rho_{pq}\sigma_p\sigma_q + \frac{1}{n^2} n\rho_{pq}\sigma_p\sigma_q \\
&= (\frac{1}{rn} + \frac{1}{n})\rho_{pq}\sigma_p\sigma_q. \qquad (A.1)
\end{aligned}$$

For z-scores from two endpoints p and q, $\forall 1 \leq p < q \leq K$, the correlation between them is given by

$$\begin{aligned}
Corr(z_{p,\ell}, z_{q,\ell}) &= Cov(z_{p,\ell}, z_{q,\ell})/(1 \cdot 1) \\
&= Cov\left(\frac{\bar{Y}_{T,p,\ell} - \bar{Y}_{C,p,\ell}}{\sqrt{\frac{1}{rn} + \frac{1}{n}}\sigma_p}, \frac{\bar{Y}_{T,q,\ell} - \bar{Y}_{C,q,\ell}}{\sqrt{\frac{1}{rn} + \frac{1}{n}}\sigma_q}\right) \\
&\overset{\text{by (A.1)}}{=\!=} \frac{1}{(\frac{1}{rn} + \frac{1}{n})\sigma_p\sigma_q}(\frac{1}{rn} + \frac{1}{n})\rho_{pq}\sigma_p\sigma_q \\
&= \rho_{pq}. \qquad (A.2)
\end{aligned}$$

A.2 Sample Size Formula by Power

By the definition of power, we can get the expression of sample size:

$$\begin{aligned}
P\left\{\frac{\bar{z}_1}{\sigma_{\bar{z}}} > Z_\alpha \middle| \frac{\mu_{\bar{z}}^*}{\sigma_{\bar{z}}}\right\} &= 1 - \beta \\
P\left\{\frac{\bar{z}_1 - \mu_{\bar{z}}^*}{\sigma_{\bar{z}}} > Z_\alpha - \frac{\mu_{\bar{z}}^*}{\sigma_{\bar{z}}}\right\} &= 1 - \beta \\
\frac{\mu_{\bar{z}}^*}{\sigma_{\bar{z}}} - Z_\alpha &= Z_\beta \\
\frac{\frac{\sqrt{N_{total}}}{\sqrt{(\frac{1}{r}+1)(r+1)}}\bar{\theta}^*}{\sqrt{\frac{1}{K^2}(K + 2\sum_{1\leq p<q\leq K}\rho_{pq})}} &= Z_\alpha + Z_\beta
\end{aligned}$$

$\Longleftrightarrow$

$$N_{total} = \frac{1}{K^2}(K + 2\sum_{1\leq p<q\leq K}\rho_{pq})\left(\frac{Z_\alpha + Z_\beta}{\frac{\bar{\theta}^*}{\sqrt{(\frac{1}{r}+1)(1+r)}}}\right)^2. \quad (A.3)$$

A.3 Conditional Power

For $\ell = 1, 2$, $\bar{z}_\ell = \frac{1}{K}\sum_{k=1}^{K}\frac{(\bar{Y}_{T,k,\ell} - \bar{Y}_{C,k,\ell})\sqrt{n_\ell}}{\sqrt{\frac{1}{r}+1}\sigma_k}$. We have $\bar{z}_\ell \sim N\left(\frac{\bar{\theta}\sqrt{n_\ell}}{\sqrt{\frac{1}{r}+1}}, \sigma_{\bar{z}}^2\right)$, where $\sigma_{\bar{z}}^2 = \frac{1}{K^2}(K + 2\sum_{1\leq p<q\leq K}\rho_{pq})$. Since stages are independent, $Cov(\bar{z}_1, \bar{z}_2) = 0$.

For the regular test statistics at final stage without SSR (denoted as $\bar{z}_f$; total sample size for control group $n = n_1 + n_2$), we can show the following:

$$\bar{z}_f = \frac{1}{K}\sum_{k=1}^{K}\frac{(\bar{Y}_{T,k,f} - \bar{Y}_{C,k,f})\sqrt{n}}{\sqrt{\frac{1}{r}+1}\sigma_k}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\frac{\left[\frac{n_1\bar{Y}_{T,k,1}+n_2\bar{Y}_{T,k,2}}{n} - \frac{n_1\bar{Y}_{C,k,1}+n_2\bar{Y}_{C,k,2}}{n}\right]\sqrt{n}}{\sqrt{\frac{1}{r}+1}\sigma_k}$$
$$= \frac{n_1}{n}\frac{\sqrt{n}}{\sqrt{n_1}}\bar{z}_1 + \frac{n_2}{n}\frac{\sqrt{n}}{\sqrt{n_2}}\bar{z}_2$$
$$= \sqrt{\frac{n_1}{n}}\bar{z}_1 + \sqrt{\frac{n_2}{n}}\bar{z}_2.$$

Moreover,

$$E[\bar{z}_f] = \sqrt{\frac{n_1}{n}}\frac{\bar{\theta}\sqrt{n_1}}{\sqrt{\frac{1}{r}+1}} + \sqrt{\frac{n_2}{n}}\frac{\bar{\theta}\sqrt{n_2}}{\sqrt{\frac{1}{r}+1}} = \frac{\bar{\theta}\sqrt{n}}{\sqrt{\frac{1}{r}+1}};$$
$$Var(\bar{z}_f) = (\frac{n_1}{n}+\frac{n_2}{n})\sigma_{\bar{z}}^2 = \sigma_{\bar{z}}^2.$$

That is, $\bar{z}_f \sim N\left(\frac{\bar{\theta}\sqrt{n}}{\sqrt{\frac{1}{r}+1}}, \sigma_{\bar{z}}^2\right)$. Further,

$$Corr(\bar{z}_1, \bar{z}_f) = \frac{Cov(\bar{z}_1, \sqrt{\frac{n_1}{n}}\bar{z}_1 + \sqrt{\frac{n_2}{n}}\bar{z}_2)}{\sigma_{\bar{z}}\sigma_{\bar{z}}}$$
$$= \frac{\sqrt{\frac{n_1}{n}}\sigma_{\bar{z}}^2}{\sigma_{\bar{z}}^2} = \sqrt{\frac{n_1}{n}}.$$

From above, we know the conditional variable $\bar{z}_f|\bar{z}_1$ also follows a normal distribution, with mean and variance as follows:

$$E[\bar{z}_f|\bar{z}_1] = \frac{\bar{\theta}\sqrt{n}}{\sqrt{\frac{1}{r}+1}} + \sqrt{\frac{n_1}{n}}(\bar{z}_1 - \frac{\bar{\theta}\sqrt{n_1}}{\sqrt{\frac{1}{r}+1}});$$
$$Var(\bar{z}_f|\bar{z}_1) = \sigma_{\bar{z}}^2(1-\frac{n_1}{n}) = \frac{n_2}{n}\sigma_{\bar{z}}^2.$$

Thus, at interim stage, the mean Cohen's d $\bar{\theta}$ is estimated by the observed mean Cohen's d $\bar{d}_1$. Then, the observed mean z-score is actually the t-statistic $\bar{t}_1 = \frac{\bar{d}_1\sqrt{n_1}}{\sqrt{\frac{1}{r}+1}}$. Therefore,

$$E[\bar{z}_f|\bar{z}_1 = \bar{t}_1] = E[\bar{z}_f|\bar{\theta} = \bar{d}_1]$$
$$= \frac{\bar{d}_1\sqrt{n}}{\sqrt{\frac{1}{r}+1}} + \sqrt{\frac{n_1}{n}}(\bar{t}_1 - \frac{\bar{d}_1\sqrt{n_1}}{\sqrt{\frac{1}{r}+1}})$$
$$= \frac{\bar{d}_1\sqrt{n}}{\sqrt{\frac{1}{r}+1}}$$

That is, $\bar{z}_f|\bar{z}_1 = \bar{t}_1 \sim N(\frac{\bar{d}_1\sqrt{n}}{\sqrt{\frac{1}{r}+1}}, \frac{n_2}{n}\sigma_{\bar{z}}^2)$.

The conditional probability CP at stage 1 can be derived as

$$CP(\bar{d}_1) = Pr\left\{\bar{z}_f \geq Z_{\alpha_2}|\bar{z}_1 = \bar{t}_1\right\}$$
$$= 1 - \Phi\left(\frac{Z_{\alpha_2} - \frac{\bar{d}_1\sqrt{n}}{\sqrt{\frac{1}{r}+1}}}{\sqrt{\frac{n_2}{n}}\sigma_{\bar{z}}}\right)$$

This can be further simplified as,

$$CP(\bar{d}_1) = 1 - \Phi\left(\frac{\sqrt{n}Z_{\alpha_2} - \frac{n}{\sqrt{n_1}}\bar{t}_1}{\sqrt{n_2}\sigma_{\bar{z}}}\right)$$
$$= 1 - \Phi\left(\frac{\sqrt{n}Z_{\alpha_2} - \frac{n_1+n_2}{\sqrt{n_1}}\bar{t}_1}{\sqrt{n_2}\sigma_{\bar{z}}}\right)$$
$$= 1 - \Phi\left(\frac{\sqrt{n}Z_{\alpha_2}}{\sqrt{n_2}\sigma_{\bar{z}}} - \frac{\sqrt{n_1}\bar{t}_1}{\sqrt{n_2}\sigma_{\bar{z}}} - \frac{\sqrt{n_2}\bar{t}_1}{\sqrt{n_1}\sigma_{\bar{z}}}\right)$$
$$= 1 - \Phi\left(\frac{1}{\sigma_{\bar{z}}}\left(\frac{\sqrt{n}Z_{\alpha_2} - \sqrt{n_1}\bar{t}_1}{\sqrt{n_2}} - \frac{\sqrt{n_2}\bar{t}_1}{\sqrt{n_1}}\right)\right)$$
$$= \Phi\left(\frac{1}{\sigma_{\bar{z}}}\left(\frac{\sqrt{n_1}\bar{t}_1 - \sqrt{n}Z_{\alpha_2}}{\sqrt{n_2}} + \frac{\sqrt{n_2}\bar{t}_1}{\sqrt{n_1}}\right)\right)$$

Since $\sigma_{\bar{z}}$ is unknown, we have to estimate it with

$$\hat{\sigma}_{\bar{z}} = \sqrt{\frac{1}{K^2}(K + 2\sum_{1\leq p<q\leq K}\hat{\rho}_{pq,1})}.$$

In summary, the final form of the approximated conditional power at interim should be

$$CP(\bar{d}_1) = \Phi\left(\frac{\frac{\sqrt{n_1}\bar{t}_1 - \sqrt{n}Z_{\alpha_2}}{\sqrt{n_2}} + \frac{\sqrt{n_2}\bar{t}_1}{\sqrt{n_1}}}{\sqrt{\frac{1}{K^2}(K + 2\sum_{1\leq p<q\leq K}\hat{\rho}_{pq,1})}}\right) \quad (A.4)$$

APPENDIX B. ADDITIONAL SIMULATION RESULTS

The simulation results under $\mathbf{\Sigma}_1$ are provided below.

DECLARATION OF POTENTIAL CONFLICTS OF INTEREST

At the time the research was conducted, Yong Lin, Philip He, and Di Shu were employees of Daiichi Sankyo, Inc. Please note that the views and opinions expressed in this paper are those of the authors and are not intended to reflect the views and/or opinions of their employer(s).

ACKNOWLEDGMENTS

We thank the guest editors and reviewers for their constructive comments and suggestions, which have greatly improved the paper.

Accepted 14 February 2026

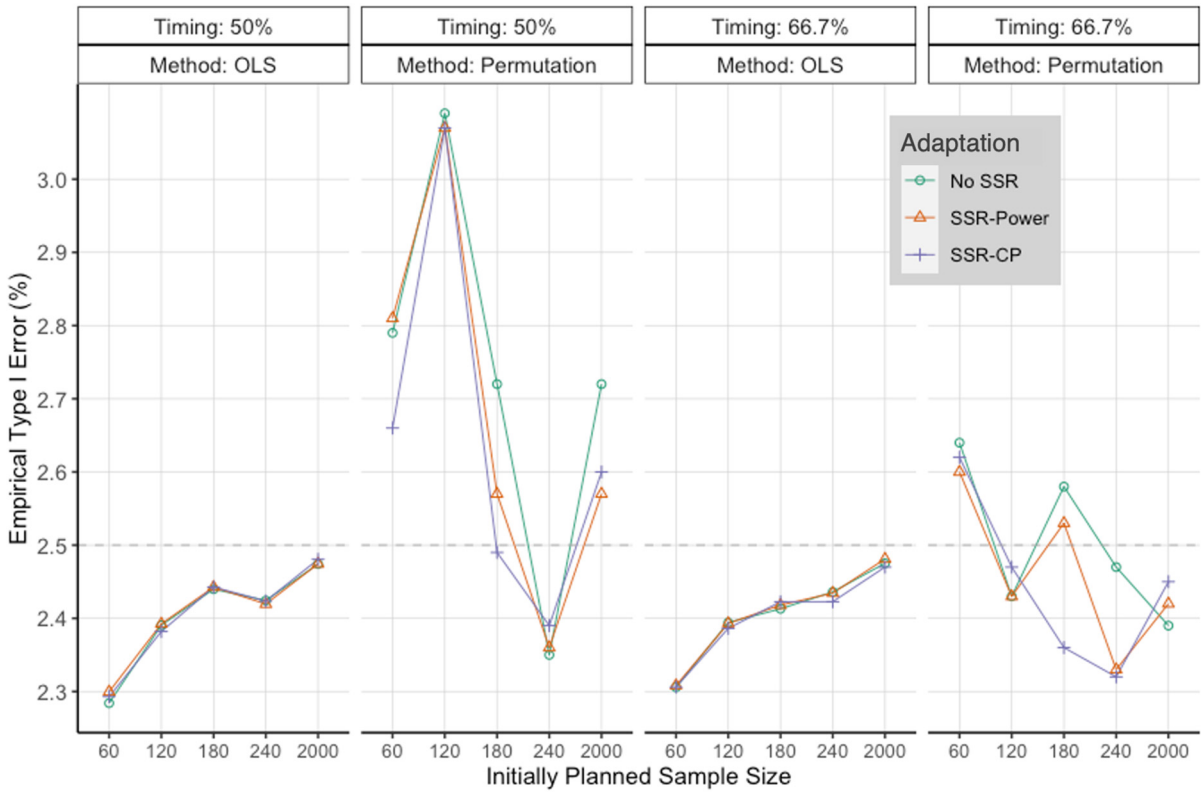

Figure B1: Empirical type I error associated with each scenario (under $\mathbf{\Sigma}_1$), with a one-sided significance level of $\alpha = 2.5\%$.

Table B1. Empirical power for each adaptation, regardless of the interim results (under $\mathbf{\Sigma}_1$), with a target power of 80%.

Timing of Interim	IPSS Type	Adaptation	OLS Test			Permutation Test		
			Power	ESS	MSS	Power	ESS	MSS
50%	Original	No SSR	78.19	94	100	78.29	92	100
		SSR-Power	79.57	98	200	79.52	95	200
		SSR-CP	81.58	110	200	81.53	108	200
	Underestimated	No SSR	48.63	98	100	49.58	97	100
		SSR-Power	50.22	102	200	51.31	101	200
		SSR-CP	53.39	113	200	54.33	112	200
66.70%	Original	No SSR	77.80	88	100	78.17	87	100
		SSR-Power	78.73	90	198	79.10	89	188
		SSR-CP	80.17	100	200	80.57	99	200
	Underestimated	No SSR	48.32	95	100	48.89	94	100
		SSR-Power	49.33	97	192	49.78	96	184
		SSR-CP	51.45	105	200	51.78	105	200

Note: "IPSS Type" represents initially planned sample size settings, which has two types: "Original" and "Underestimated." "ESS" and "MSS" represent the expected sample size and maximum sample size, respectively. When there is no SSR and the sample size is fixed, the MSS is always the same as N_{total}. The significance level α is 2.5%.

Table B2. Empirical probabilities of interim results for adaptations with SSR (under $\mathbf{\Sigma}_1$).

Timing of Interim	IPSS Type	Interim Result	Percentage (%)	
			OLS	Permutation
50%	Original	Cont. - Favorable	21.74	17.49
		Cont. - Promising	31.82	31.23
		Cont. - Unfavorable	34.76	35.21
		Reject, Stop	11.69	16.07
	Underestimated	Cont. - Favorable	11.78	10.55
		Cont. - Promising	26.83	26.51
		Cont. - Unfavorable	57.33	57.15
		Reject, Stop	4.07	5.79
66.7%	Original	Cont. - Favorable	0.48	0.59
		Cont. - Promising	26.65	24.32
		Cont. - Unfavorable	36.84	37.13
		Reject, Stop	36.03	37.96
	Underestimated	Cont. - Favorable	0.25	0.29
		Cont. - Promising	21.21	20.26
		Cont. - Unfavorable	62.82	62.75
		Reject, Stop	15.71	16.70

Note: In the "Interim Results" column, "Cont." stands for "Continue," indicating that the trial will not be rejected at interim, with subsequent zone decisions based on the interim conditional power. These percentages are the same for both SSR-Power and SSR-CP adaptations.

Table B3. Empirical power for adaptations with SSR, conditional on interim results (under $\mathbf{\Sigma}_1$), with a target power of 80%.

Timing of Interim	IPSS Type	Interim Result & Adaptation	OLS Test			Permutation Test		
			Power	ESS	MSS	Power	ESS	MSS
50%	Original	Cont. - Favorable	96.03	100	100	96.05	100	100
		Cont. - Promising & SSR-Power	90.45	112	200	90.81	111	200
		Cont. - Promising & SSR-CP	96.78	151	200	97.25	150	200
		Cont. - Unfavorable	52.44	100	100	51.95	100	100
		Reject, Stop	100.00	50	50	100.00	50	50
	Underestimated	Cont. - Favorable	87.73	100	100	88.25	100	100
		Cont. - Promising & SSR-Power	74.88	114	200	76.31	114	200
		Cont. - Promising & SSR-CP	86.73	155	200	87.70	155	200
		Cont. - Unfavorable	27.44	100	100	27.96	100	100
		Reject, Stop	100.00	50	50	100.00	50	50
66.7%	Original	Cont. - Favorable	95.24	100	100	96.61	100	100
		Cont. - Promising & SSR-Power	92.68	109	198	92.23	110	188
		Cont. - Promising & SSR-CP	98.11	147	200	98.27	150	200
		Cont. - Unfavorable	47.61	100	100	48.86	100	100
		Reject, Stop	100.00	66	66	100.00	66	66
	Underestimated	Cont. - Favorable	88.64	100	100	75.86	100	100
		Cont. - Promising & SSR-Power	81.80	111	192	82.08	111	184
		Cont. - Promising & SSR-CP	91.80	151	200	91.95	152	200
		Cont. - Unfavorable	25.54	100	100	25.86	100	100
		Reject, Stop	100.00	66	66	100.00	66	66

Note: The zone-specific power for interim results as "Cont. - Favorable" and "Cont. - Unfavorable" are the same for both SSR adaptation types; while the power for any interim result being "Reject, Stop" is consistently 100%.

REFERENCES

[1] Congress, U. S. Orphan drug act. https://www.fda.gov/industry/designating-orphan-product-drugs-and-biological-products/orphan-drug-act. In *Public Law 97-414, 97th Congress* (1983). 4 January 1983.

[2] Griggs, R. C., Batshaw, M., Dunkle, M., Gopal-Srivastava, R., Kaye, E., Krischer, J., Nguyen, T., Paulus, K., Merkel, P. A. et al. Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular genetics and metabolism* **96**(1) 20–26 (2009).

[3] Zhang, L., Liu, L., Yi, B., Miao, X., Nair, N., Liu, X., Jazic, I. and Laird, G. Clinical trial design and analysis considerations for rare diseases. In *Drug Development for Rare Diseases* 36–92. Chapman and Hall/CRC, (2023).

[4] Xu, H., Liu, Y. and Beckman, R. A. Adaptive endpoints selection with application in rare disease. *Statistics in Biopharmaceutical Research* **16**(1) 55–63 (2024).

[5] Finkel, R. S., Mercuri, E., Darras, B. T., Connolly, A. M., Kuntz, N. L., Kirschner, J., Chiriboga, C. A., Saito, K., Servais, L., Tizzano, E. et al. Nusinersen versus sham control in infantile-onset spinal muscular atrophy. *New England Journal of Medicine* **377**(18) 1723–1732 (2017).

[6] Cox, G. F. The art and science of choosing efficacy endpoints for rare disease clinical trials. *American Journal of Medical Genetics Part A* **176**(4) 759–772 (2018).

[7] Food, U. S. and Administration, D. Multiple endpoints in clinical trials: Guidance for industry (2022). Accessed: 2024-08-29. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials

[8] Huque, M. F. and Sankoh, A. J. A reviewer's perspective on multiple endpoint issues in clinical trials. *Journal of Biopharmaceutical Statistics* **7**(4) 545–564 (1997).

[9] Caputo, A., Racine, A., Paule, I., Tariot, P. N., Langbaum, J. B., Coello, N., Riviere, M. -E., Ryan, J. M., Lopez Lopez, C., Graf, A. et al. Rationale for the selection of dual primary endpoints in prevention studies of cognitively unimpaired individuals at genetic risk for developing symptoms of alzheimer's disease. *Alzheimer's Research & Therapy* **15**(1) 45 (2023).

[10] Hochberg, Y. A sharper bonferroni procedure for multiple tests of significance. *Biometrika* **75**(4) 800–802 (1988). https://doi.org/10.1093/biomet/75.4.800. MR0995126

[11] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*. 65–70 (1979). MR0538597

[12] Dmitrienko, A., Bretz, F., Westfall, P. H., Troendle, J., Wiens, B. L., Tamhane, A. C. and Hsu, J. C. Multiple testing methodology. In *Multiple testing problems in pharmaceutical statistics* 53–116. Chapman and Hall/CRC, (2009).

[13] Ristl, R., Urach, S., Rosenkranz, G. and Posch, M. Methods for the analysis of multiple endpoints in small populations: a review. *Journal of biopharmaceutical statistics* **29**(1) 1–29 (2019).

[14] McMenamin, M., Berglind, A. and MS Wason, J. Improving the analysis of composite endpoints in rare disease trials. *Orphanet journal of rare diseases* **13**. 1–9 (2018).

[15] Hamasaki, T., Evans, S. R. and Asakura, K. Design, data monitoring, and analysis of clinical trials with co-primary endpoints: a review. *Journal of biopharmaceutical statistics* **28**(1) 28–51 (2018).

[16] Verbeeck, J., Dirani, M., Bauer, J. W., Hilgers, R. -D., Molenberghs, G. and Nabbout, R. Composite endpoints, including patient reported outcomes, in rare diseases. *Orphanet Journal of Rare Diseases* **18**(1) 262 (2023).

[17] O'Brien, P. C. Procedures for comparing samples with multiple endpoints. *Biometrics*. 1079–1087 (1984). https://doi.org/10.2307/2531158. MR0786180

[18] Logan, B. R. and Tamhane, A. C. On o'brien's ols and gls tests for multiple endpoints. *Lecture Notes-Monograph Series*. 76–88 (2004). https://doi.org/10.1214/lnms/1196285627. MR2118593

[19] Dallow, N. S., Leonov, S. L. and Roger, J. H. Practical usage of o'brien's ols and gls statistics in clinical trials. *Pharmaceutical statistics* **7**(1) 53–68 (2008).

[20] Sun, H., Davison, B. A., Cotter, G., Pencina, M. J. and Koch, G. G. Evaluating treatment efficacy by multiple end points in phase ii acute heart failure clinical trials: analyzing data using a global method. *Circulation: Heart Failure* **5**(6) 742–749 (2012).

[21] Li, D., McDonald, C. M., Elfring, G. L., Souza, M., McIntosh Dae Hyun Kim, J. and Wei, L. -J. Assessment of treatment effect with multiple outcomes in 2 clinical trials of patients with duchenne muscular dystrophy. *JAMA network open* **3**(2) e1921306–e1921306 (2020).

[22] Lancaster, H. O. The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics* **3**(1) 20–33 (1961). https://doi.org/10.1111/j.1467-842x.1961.tb00058.x. MR0130742

[23] Dai, H., Leeder, J. S. and Cui, Y. A modified generalized fisher method for combining probabilities from dependent tests. *Frontiers in genetics* **5** (2014).

[24] Food, U. S. and Administration, D. Adaptive design clinical trials for 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 (2019). Accessed: 2024-08-29.

[25] Food, U. S. and Administration, D. Rare diseases: Considerations for development of drugs and biological products. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/rare-diseases-considerations-development-drugs-and-biological-products (2023). Accessed: 2024-08-29.

[26] Mehta, C. R. and Pocock, S. J. Adaptive increase in sample size when interim results are promising: a practical guide with examples. *Statistics in medicine* **30**(28) 3267–3284 (2011). https://doi.org/10.1002/sim.4102. MR2861612

[27] Chen, Y. J., DeMets, D. L. and Lan, K. G. Increasing the sample size when the unblinded interim result is promising. *Statistics in medicine* **23**(7) 1023–1038 (2004).

[28] Ping Gao, Ware, J. H. and Mehta, C. Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* **18**(6) 1184–1196 (2008). https://doi.org/10.1080/10543400802369053. MR2522185

[29] Edwards, J. M., Walters, S. J., Kunz, C. and Julious, S. A. A systematic review of the "promising zone". *design. Trials* **21**. 1–10 (2020).

[30] Wang, J. and Ren, Q. A note on the promising zone approach in adaptive trial design. *Statistics in Biopharmaceutical Research* **14**(1) 132–137 (2022).

[31] Cui, L., Hung, H. J. and Wang, S. -J. Modification of sample size in group sequential clinical trials. *Biometrics* **55**(3) 853–857 (1999).

[32] Lehmacher, W. and Wassmer, G. Adaptive sample size calculations in group sequential trials. *Biometrics* **55**(4) 1286–1290 (1999).

[33] Cohen, J. *Statistical power analysis for the behavioral sciences*. Routledge, New York (1988).

[34] Catherine, O., Fritz, Morris, P. E. and Richler, J. J. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* **141**(1) 2 (2012).

[35] Lan, K. G. and DeMets, D. L. Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3) 659–663 (1983). https://doi.org/10.2307/2336502. MR0725380

[36] Keaven, A. *gsDesign: Group Sequential Design* (2024). R package version 3.6.2

[37] Kim, K. and Demets, D. L. Design and analysis of group sequential tests based on the type i error spending rate function. *Biometrika* **74**(1) 149–154 (1987). https://doi.org/10.1093/biomet/74.1.149. MR0885927

[38] Liddy, M., Chen, Ibrahim, J. G. and Chu, H. Flexible stopping boundaries when changing primary endpoints after unblinded interim analyses. *Journal of biopharmaceutical statistics* **24**(4) 817–833 (2014). https://doi.org/10.1080/10543406.2014.901341.

MR3210433
[39] Meurer, W. J. and Tolles, J. Interim analyses during group sequential clinical trials. *JAMA* **326**(15) 1524–1525 (2021).
[40] Pocock, S. J. Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**(2) 191–199 (1977).
[41] O'Brien, P. C. and Fleming, T. R. A multiple testing procedure for clinical trials. *Biometrics*. 549–556 (1979).
[42] Mohamed Amine Bayar, Le Teuff, G. and Koenig, F. Marie-Cecile Le Deley, and Stefan Michiels. Group sequential adaptive designs in series of time-to-event randomised trials in rare diseases: A simulation study. *Statistical Methods in Medical Research* **29**(6) 1483–1498 (2020). https://doi.org/10.1177/0962280219862313. MR4106952
[43] Felker, G. M. and Maisel, A. S. A global rank end point for clinical trials in acute heart failure. *Circulation: Heart Failure* **3**(5) 643–646 (2010).
[44] Kaizer, A. M., Belli, H. M., Ma, Z., Nicklawsky, A. G., Roberts, S. C., Wild, J., Wogu, A. F., Xiao, M. and Sabo, R. T. Recent innovations in adaptive trial designs: a review of design opportunities in translational research. *Journal of Clinical and Translational Science* **7**(1), e125 (2023).
[45] Bauer, P., Bretz, F., Dragalin, V., König, F. and Wassmer, G. Twenty-five years of confirmatory adaptive designs: opportunities and pitfalls. *Statistics in Medicine* **35**(3) 325–347 (2016). https://doi.org/10.1002/sim.6472. MR3455501
[46] Asakura, K., Hamasaki, T. and Evans, S. R. Interim evaluation of efficacy or futility in group-sequential trials with multiple co-primary endpoints. *Biometrical Journal* **59**(4) 703–731 (2017). https://doi.org/10.1002/bimj.201600026. MR3672692
[47] Arup, K., Sinha, Moye, L. III, Piller, L. B., Yamal, J. -M., Barcenas, C. H., Lin, J. and Davis, B. R. Adaptive group-sequential design with population enrichment in phase 3 randomized controlled trials with two binary co-primary endpoints. *Statistics in medicine* **38**(21) 3985–3996 (2019). https://doi.org/10.1002/sim.8216. MR3999259
[48] Hung, H. J., Wang, S. -J. and O'Neill, R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of biopharmaceutical statistics* **17**(6) 1201–1210 (2007). https://doi.org/10.1080/10543400701645405. MR2414571
[49] Arup, K., Sinha, Moye, L. III, Piller, L. B., Yamal, J. -M., Barcenas, C. H., Song, J. and Davis, B. R. Simultaneous population enrichment and endpoint selection in phase 3 randomized controlled trials: An adaptive group sequential design with two binary alternative primary endpoints. *Communications in Statistics-Theory and Methods* **53**(10) 3728–3741 (2024). https://doi.org/10.1080/03610926.2022.2163180. MR4728341
[50] Sugitani, T., Bretz, F. and Maurer, W. A simple and flexible graphical approach for adaptive group-sequential clinical trials. *Journal of biopharmaceutical statistics* **26**(2) 202–216 (2016).
[51] Posch, M. and Bauer, P. Adaptive two stage designs and the conditional error function. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **41**(6) 689–696 (1999). https://doi.org/10.1002/bimj.200510236. MR2247055
[52] Proschan, M. A. and Hunsberger, S. A. Designed extension of studies based on conditional power. *Biometrics*. 1315–1324 (1995).
[53] Müller, H. -H. and Schäfer, H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* **57**(3) 886–891 (2001). https://doi.org/10.1111/j.0006-341X.2001.00886.x. MR1859823
[54] Buyse, M. Generalized pairwise comparisons of prioritized outcomes in the two-sample problem. *Statistics in medicine* **29**(30) 3245–3257 (2010). https://doi.org/10.1002/sim.3923. MR2758717

Lan Shi. Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN.
E-mail address: lan.shi@vanderbilt.edu

Yong Lin. Biostatistics & Data Management, Daiichi Sankyo, Inc., Basking Ridge, NJ.
E-mail address: yong.lin@daiichisankyo.com

Philip He. Biostatistics & Data Management, Daiichi Sankyo, Inc., Basking Ridge, NJ.
E-mail address: philip.he@daiichisankyo.com

Di Shu. Biostatistics & Data Management, Daiichi Sankyo, Inc., Basking Ridge, NJ.
E-mail address: di.shu@daiichisankyo.com