Sample Size Re-Estimation in Adaptive Enrichment Design

Ruitao Lin[1], Zhao Yang[2], Ying Yuan[1] and Guosheng Yin[1,2*],

[1]Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, Texas 77030, U.S.A

[2]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China

**email*: gyin@hku.hk

Abstract. Clinical trial participants are often heterogeneous, which is a fundamental problem in the rapidly developing field of precision medicine. Participants heterogeneity causes considerable difficulty in the current phase III trial designs. Adaptive enrichment designs provide a flexible and intuitive solution. At the interim analysis, we enrich the subgroup of trial participants who have a higher likelihood to benefit from the new treatment. However, it is critical to control the level of the test size and maintain adequate power after enrichment of certain subgroup of participants. We develop two adaptive enrichment strategies with sample size re-estimation and verify their feasibility and practicability through extensive simulations and sensitivity analyses. The simulation studies show that the proposed methods can control the overall type I error rate and exhibit competitive improvement in terms of statistical power and expected sample size. The proposed designs are exemplified with a real trial application.

KEY WORDS: Conditional power; enrichment strategies; patient heterogeneity; phase III trial designs; sample size re-estimation

1 Introduction

Clinical development of new therapies involves scientific, ethical and economic considerations, which is both time and resource consuming. Tremendous advances in our understanding of cancer biology and new developments in biotechnology provide numerical information and powerful tools to increase the efficiency of randomized clinical trials (RCTs). The scientific landscape has been redefined to achieve stratified and personalized medicine for patient subpopulations, which leads to a broad and exciting category of trial design, termed as the biomarker-guided adaptive enrichment design (BAED) [1]. Such enrichment designs, including the adaptive signature design [2], the adaptive threshold sample enrichment design [3] and the adaptive population enrichment design [4], have attracted an enormous amount of attention [5, 6, 7, 8, 9]. In the RCT methodology, the BAED prospectively uses biomarkers with strong credentials to enrich the subpopulation, in which the detection of a treatment effect is of a higher likelihood than it would be in the overall population. In such a design, enrollment would initially be open to all participants yet with the option to restrict the future enrollment in the mid-course of the trial to enrich the promising biomarker-specified subgroup only. It is widely believed that a suitably developed BAED can reduce the expected sample size (ESS) or study duration, improve power, enhance the probability of trial success, and reduce the cost and bias compared with the standard design without enrichment [6, 10, 11, 12, 13, 14].

Nevertheless, several limitations of adaptive enrichment designs have been noted. First, modification of trial participants during the mid-course may result in high uncertainty of the treatment effect and may also induce statistical bias. Second, the information borrowed between the biomarker-specified subgroups and the overall group might inappropriately guide the application of the investigated treatment. For example, if the treatment effect in one subgroup (e.g., the biomarker-positive group) is overly strong, the average treatment effect (ATE) in the overall group might also be statistically significant even if there is no treatment

effect in the other subgroup (e.g., the biomarker-negative group). In this case, declaring the effectiveness of the investigated treatment in the overall group is inappropriate. Third, the heterogeneity of trial participants may result in a dilution of the treatment effect, and thus cause an inappropriate and unethical application of the investigated treatment. Last, the initial sample size is computed for the overall group at the beginning of the trial. With the enrichment triggered at the interim analysis, it is not clear whether the remaining sample size can maintain adequate power for the enriched subgroup.

We study the two-stage, adaptive enrichment strategies with sample size re-estimation (SSR) to achieve adequate conditional power (CP), in which the enrichment and SSR procedures are implemented during the mid-course of the trial based on a single interim analysis. Futility and efficacy stopping boundaries can be pre-specified to guide the enrichment strategies and enrich the promising subgroup proceeding to the second stage. We re-estimate the sample size required in the second stage based on the interim data of the selected subgroup, with an aim to achieve adequate CP [15]. Finally, an efficacy test is employed to test the treatment effect in the enriched subgroup or the overall group. While existing works typically focus on participants' enrichment [3, 14, 16, 17, 18] or SSR [15, 19, 20, 21, 22], seldom are the two procedures combined [23]. We investigate the benefits and limitations of trial designs combining enrichment and SSR through a real trial example, extensive simulation studies and sensitivity analyses.

2 Methods

2.1 Trial setting

We consider a two-arm phase III trial to compare an experimental treatment with a standard treatment based on a continuous outcome. The trial participants are heterogeneous and can be distinguished by some baseline characteristics or predictive biomarkers. For simplicity, we assume there are two subgroups, i.e., subgroup 1 (biomarker positive) and subgroup 2

(biomarker negative). Suppose the proportion of subgroup 1 is ρ. The problem of interest is to test the treatment effect in the adaptively enriched subgroup and the overall group, given that subgroups are well defined prior to the start of the trial. One-sided hypothesis testing procedures are used with the null and alternative hypotheses specified as

$$\begin{aligned} H_{00} : \theta_0 = 0 \quad &\text{versus} \quad H_{10} : \theta_0 > 0 \\ H_{01} : \theta_1 = 0 \quad &\text{versus} \quad H_{11} : \theta_1 > 0 \\ H_{02} : \theta_2 = 0 \quad &\text{versus} \quad H_{12} : \theta_2 > 0 \end{aligned}$$

where $\theta_0 = \rho\theta_1 + (1-\rho)\theta_2$ is the treatment effect for the overall group with θ_1 and θ_2 respectively being the subgroup-specific treatment effects for subgroups 1 and 2. Note that H_{00} versus H_{10} evaluates whether there is a treatment effect in the overall group, while H_{01} versus H_{11} and H_{02} versus H_{12} test the treatment efficacy in subgroups 1 and 2, respectively.

Traditionally, the ATE is often estimated, which implicitly assumes homogeneous treatment effects (i.e., $\theta_0 = \theta_1 = \theta_2$) across two subgroups. However, the treatment effects in subgroups may vary considerably from the ATE, leading to heterogeneous treatment effects (i.e., $\theta_0 \neq \theta_1 \neq \theta_2$). In this case, adaptive enrichment designs are recommended to account for heterogeneity and enrich the subgroup in which the treatment effect can be more readily demonstrated. To determine whether the treatment works for patients in the enriched subgroup or the overall group, we describe the general framework of the adaptive enrichment strategy with SSR.

The proposed methods can accommodate various types of endpoints, including continuous, binary, and survival outcomes, as long as a test statistic can be well defined, e.g., the log-rank test statistic for the survival endpoints. For illustrative purpose, we focus on the continuous case. Specifically, in a two-arm trial with a continuous outcome, let Y_i denote the outcome for the ith subject in the experimental group for $i = 1, \ldots, N_Y$, and let X_i denote the outcome for the ith subject in the standard group, for $i = 1, \ldots, N_X$. The outcomes in the two groups are assumed to be independent and normally distributed with an equal and known variance σ^2, i.e., $Y_i \sim \mathrm{N}(\mu_{Y_g}, \sigma^2)$ and $X_i \sim \mathrm{N}(\mu_{X_g}, \sigma^2)$, where μ_{Y_g} denotes the

true mean of subgroup g for $g = 1, 2$ under the experimental treatment, and μ_{X_g} denotes the true mean of subgroup g under the standard treatment. Denote $\mu_{Y_0} = \rho\mu_{Y_1} + (1-\rho)\mu_{Y_2}$ and $\mu_{X_0} = \rho\mu_{X_1} + (1-\rho)\mu_{X_2}$ and, as a result, $\theta_0 = \mu_{Y_0} - \mu_{X_0} = \rho\theta_1 + (1-\rho)\theta_2$ is the overall treatment effect, with $\theta_1 = \mu_{Y_1} - \mu_{X_1}$ and $\theta_2 = \mu_{Y_2} - \mu_{X_2}$ being the subgroup-specific treatment effects for subgroups 1 and 2, respectively.

2.2 General framework

In a two-stage procedure, the first stage enrolls patients with a pre-defined biomarker indicator characterizing patients' heterogeneity (i.e., subgroup 1 or subgroup 2). Patients are equally randomized to the experimental and standard treatment arms. An interim analysis is conducted after a total of N_1 patients are enrolled. We then construct the subgroup-specific test statistics for subgroups 1 and 2,

$$t_1 = \frac{\bar{\mu}_{Y_1} - \bar{\mu}_{X_1}}{\sqrt{4\sigma^2/(\rho N_1)}} \quad \text{and} \quad t_2 = \frac{\bar{\mu}_{Y_2} - \bar{\mu}_{X_2}}{\sqrt{4\sigma^2/\{(1-\rho)N_1\}}},$$

where $\bar{\mu}_{Y_g}$ and $\bar{\mu}_{X_g}$ denote the sample means respectively for the treatment and standard arms within subgroup g, $g = 1, 2$. If the prevalence rate ρ is unknown *a priori*, it can be estimated by the empirical proportion of subgroup 1 using the data observed in the first stage. Similarly, the test statistic t_0 of the overall group at stage 1 can be written as

$$t_0 = \frac{\bar{\mu}_{Y_0} - \bar{\mu}_{X_0}}{\sqrt{4\sigma^2/N_1}} = \sqrt{\rho}t_1 + \sqrt{1-\rho}t_2$$

which is a weighted average of the stage 1 subgroup-specific statistics t_1 and t_2, with $\bar{\mu}_{Y_0}$ and $\bar{\mu}_{X_0}$ being the sample means of the overall group in the experimental and standard arms, respectively.

The three interim test statistics are used to determine the enriched subgroups and early termination. Let $g^\dagger$ denote the selected subgroup after the interim analysis:

$$g^\dagger = \begin{cases} 0, & \text{if no subgroup is enriched;} \\ 1, & \text{if subgroup 1 is enriched;} \\ 2, & \text{if subgroup 2 is enriched,} \end{cases}$$

and let $t^{(1)} \equiv t_{g^\dagger}$ denote the test statistic for the selected subgroup $g^\dagger$ at stage 1. For the adaptively enriched subgroup or the overall group, we propose to re-estimate the sample size N_2 required in the second stage and the critical value c_α used in the final analysis, with an aim to reach the conditional power $(1-\beta_2)$ while preserving the overall type I error rate at a pre-specified level of α.

At the end of the second stage, we conduct the final analysis to test the treatment efficacy in the enriched subgroup or the overall group. Particularly, the test statistic at stage 2 (using stage 2 data only) is given by

$$t^{(2)} = \frac{\hat{\mu}_{Y_{g^\dagger}} - \hat{\mu}_{X_{g^\dagger}}}{\sqrt{4\sigma^2/N_2}},$$

where $\hat{\mu}_{Y_{g^\dagger}}$ and $\hat{\mu}_{X_{g^\dagger}}$ denote the sample means respectively for the treatment and control arms within the selected subgroup $g^\dagger$, based on the observed data in the second stage. The final test statistic based on the complete data combined from stages 1 and 2 can be written as

$$T = \sqrt{w}t^{(1)} + \sqrt{1-w}t^{(2)},$$

where $w = N_1^\dagger/N$ is the information fraction at the interim analysis, with $N_1^\dagger$ being the sample size for the selected subgroup $g^\dagger$ at stage 1 and $N = N_1^\dagger + N_2$ being the total sample size for the selected subgroup. In particular, if enrichment is triggered at the interim analysis, $N_1^\dagger = \rho N_1$ or $(1-\rho)N_1$, depending on whether subgroup 1 or 2 is selected to stage 2, and otherwise, $N_1^\dagger = N_1$ with no enrichment.

Let $\delta_{g^\dagger}$ be the true standardized treatment effect for subgroup $g^\dagger$, i.e., $\delta_{g^\dagger} = (\mu_{Y^\dagger} - \mu_{X^\dagger})/\sigma$. Under $\delta_{g^\dagger}$, the stage 2 statistic $t^{(2)}$ follows a normal distribution,

$$t^{(2)} \mid \delta_{g^\dagger} \sim \mathrm{N}\left(\frac{\delta_{g^\dagger}}{\sqrt{4/N_2}}, 1\right).$$

Let $CP_{\delta_{g^\dagger}}(c_\alpha, N_2|t^{(1)})$ be the conditional probability that T exceeds c_α given $t^{(1)} = t_{g^\dagger}$ under the true effect size $\delta_{g^\dagger}$, which is known as the conditional power (CP),

$$CP_{\delta_{g^\dagger}}(c_\alpha, N_2|t^{(1)}) \equiv \Pr\{T \geqslant c_\alpha | t^{(1)}, \delta_{g^\dagger}\} = 1 - \Phi\left\{\frac{c_\alpha - t^{(1)}\sqrt{w}}{\sqrt{1-w}} - \frac{\sqrt{N_2}\delta_{g^\dagger}}{2}\right\} \tag{2.1}$$

In such a case, N_2 can be re-estimated by setting $CP_{\delta_{g^\dagger}}(c_\alpha, N_2|t^{(1)}) = 1-\beta_2$.

The overall type I error rate is $\int_{-\infty}^{\infty} CP_0(c_\alpha, N_2|t^{(1)})f(t^{(1)})\mathrm{d}t^{(1)}$, where $f(\cdot)$ is the density function of the test statistic $t^{(1)}$. To control the overall type I error rate, we specify a conditional error function $A(\cdot)$ [15], which is an increasing function over $[0,1]$ and satisfies

$$\int_{-\infty}^{\infty} A(t^{(1)})f(t^{(1)})\mathrm{d}t^{(1)} = \alpha. \tag{2.2}$$

By equating $CP_0(c_\alpha, N_2|t^{(1)}) = A(t^{(1)})$ and solving the value for c_α, the type I error rate can be preserved. In this case,

$$c_\alpha = t^{(1)}\sqrt{w} + z_{A(t^{(1)})}\sqrt{1-w} = \frac{\sqrt{N_1^\dagger}t^{(1)} + \sqrt{N_2}z_{A(t^{(1)})}}{\sqrt{N_1^\dagger + N_2}},$$

where z_A denotes the $100(1-A)$th percentile of the standard normal distribution. Under the proposed enrichment strategies, $f(t^{(1)})$ does not hold a standard functional form and depends on the subgroup-specific statistics t_1 and t_2. Therefore, the computation of the overall type I error rate consists of several double integrals.

Furthermore, replacing $\delta_{g^\dagger}$ by its maximum likelihood estimate, $\hat{\delta}_{g^\dagger} = t_{g^\dagger}\sqrt{4/N_1^\dagger}$, we can compute the sample size N_2 to achieve the CP of $1-\beta_2$,

$$N_2 = N_1^\dagger \left\{ \frac{z_{A(t^{(1)})} + z_{\beta_2}}{t^{(1)}} \right\}^2, \tag{2.3}$$

$$c_\alpha = \frac{(t^{(1)})^2 + z_{A(t^{(1)})}\left\{z_{A(t^{(1)})} + z_{\beta_2}\right\}}{\sqrt{(t^{(1)})^2 + \left\{z_{A(t^{(1)})} + z_{\beta_2}\right\}^2}}, \tag{2.4}$$

where $N_1^\dagger$ and $t^{(1)} = t_{g^\dagger}$ are the sample size and test statistic for the selected subgroup $g^\dagger$ at stage 1, respectively. Note that choosing the value of N_2 without correcting the final critical value may inflate the overall type I error rate in the standard SSR [15]. Such an inflation of the type I error rate also prevails for enrichment designs with SSR.

Based on this general framework, the selection of the conditional error function $A(\cdot)$ and the determination of the density function $f(t^{(1)})$ are critical to different enrichment strategies. We present the technical details of each strategy in the following sections.

2.3 Enrichment with early stopping for both futility and efficacy based on SSR (EFE-SSR)

As shown in Figure 1, the enrichment and early stopping strategies in EFE-SSR are solely guided by the pre-specified futility and efficacy stopping boundaries ($0 \leqslant l < u$) for the test statistics. At the interim analysis, we examine the treatment effect in both subgroups 1 and 2 based on the subgroup-specific test statistics t_1 and t_2 simultaneously. Let g^* index the subgroup that has a larger subgroup-specific test statistic, i.e., $g^* = \arg\max_g\{t_g; g = 1, 2\}$, and similarly let $g' = \arg\min_g\{t_g; g = 1, 2\}$ denote the subgroup that has a smaller subgroup-specific statistic. The trial can be carried out as follows:

(1) If $t_{g'} \geqslant u$, terminate the trial and declare treatment efficacy in the overall group.

(2) If $t_{g^*} \geqslant u$ and $t_{g'} \leqslant l$, terminate the trial and declare treatment efficacy in subgroup g^* and treatment futility in subgroup g'.

(3) If $t_{g^*} \geqslant u$ and $l < t_{g'} < u$, terminate the trial and declare treatment efficacy in subgroup g^* and inconclusive treatment effect in subgroup g'.

(4) If $l < t_{g^*} < u$ and $t_{g'} \leqslant l$, terminate the enrollment of patients in subgroup g' for futility, and enrich subgroup g^* in the second stage; that is, the selected subgroup $g^\dagger = g^*$.

(5) If $t_{g^*} \leqslant l$, terminate the trial and declare treatment futility in the overall group.

(6) If $l < t_{g'} \leqslant t_{g^*} < u$, the trial proceeds to the second stage without enrichment; that is, the selected group is the overall group with $g^\dagger = 0$.

At the end of the second stage, the final test is conducted to evaluate the treatment efficacy in the enriched subgroup $g^\dagger = g^*$ or the overall group $g^\dagger = 0$.

Following Proschan and Hunsberger [15], the *circular conditional error function* is employed to protect the overall type I error rate at a pre-specified α level, namely,

$$A_{\text{cir}}(t^{(1)}; l, u) = \begin{cases} 0, & \text{if} \quad t^{(1)} \leqslant l, \\ 1 - \Phi\big(\sqrt{u^2 - (t^{(1)})^2}\big), & \text{if} \quad l < t^{(1)} < u \\ 1, & \text{if} \quad t^{(1)} \geqslant u \end{cases} \tag{2.5}$$

where $t^{(1)}$ is the enriched subgroup-specific test statistic, and l and u are the futility and efficacy stopping boundaries at the interim analysis respectively.

In general, we can calculate the value of u given $l = z_{\alpha^*}$ based on the prespecified enrichment rules and equations (2.2) and (2.5), such that the overall type I error rate can be preserved at the nominal level α, where z_{α^*} is the $100(1-\alpha^*)$th percentile of the standard normal distribution and $\alpha < \alpha^* \leqslant 0.5$. In this case, α^* can be treated as the futility cutoff of the p-value after enrolling the first N_1 patients, and its value can be easily specified by the investigators. In EFE-SSR, if the p-value of subgroup g^* with the largest test statistic in stage 1 exceeds α^*, the trial will be terminated early for futility.

Compared to the standard SSR without enrichment, the calculation of the overall type I error rate in EFE-SSR, i.e., $\int_{-\infty}^{\infty} A_{\text{cir}}(t^{(1)}) f(t^{(1)}) \mathrm{d}t^{(1)}$, becomes more complex, because $f(t^{(1)})$ is a mixture distribution function due to the adaptive enrichment strategies. Generally, the overall type I error rate can be split into three components,

$$\begin{aligned} \alpha &= \int_l \int_{-\infty}^{l} A_{\text{cir}}(t_1)\phi(t_1)\phi(t_2)\mathrm{d}t_2\mathrm{d}t_1 + \int_l \int_{-\infty}^{l} A_{\text{cir}}(t_2)\phi(t_1)\phi(t_2)\mathrm{d}t_1\mathrm{d}t_2 \\ &\quad + \int_l \int_l A_{\text{cir}}(t_0)\phi(t_1)\phi(t_2)\mathrm{d}t_1\mathrm{d}t_2 \\ &= 2 \times \int_{z_{\alpha^*}}^{u} \left\{ \left(1 - \Phi\left(\sqrt{u^2 - t_1^2}\right)\right) \phi(t_1) + 1 - \Phi(u) \right\} (1 - \alpha^*)\, \mathrm{d}t_1 \\ &\quad + \int_{z_{\alpha^*}} \int_{z_{\alpha^*}} \left\{ \left(1 - \Phi\left(\sqrt{u^2 - t_0^2}\right)\right) \mathrm{I}(z_{\alpha^*} \leqslant t_0 < u) + \mathrm{I}(t_0 \geqslant u) \right\} \phi(t_1)\phi(t_2)\mathrm{d}t_1\mathrm{d}t_2, \end{aligned}$$

where $\mathrm{I}(\cdot)$ is the indicator function. As a result, the efficacy stopping boundary u can be solved numerically based on the above equation. For example, when $\alpha = 0.05$ and $l = z_{\alpha^*} = 0.84$ with $\alpha^* = 0.2$, the value of u is 2.234.

Once the futility and efficacy stopping boundaries l and u are obtained, we construct the adaptive enrichment strategies based on t_1 and t_2. For the enriched subgroup or the overall group proceeding to the second stage, we first calculate the value of $A_{\text{cir}}(t^{(1)})$ using equation (2.5) and then determine N_2 and c_α using (2.3) and (2.4), respectively. To implement EFE-SSR, the choice of the prespecified design parameters N_1 and l can be determined by extensive simulation studies under various scenarios, so that the resulting design is equipped with desirable operating characteristics.

On a side note, it is also possible to adapt EFE-SSR to a strategy that only considers early stopping for futility. Due to its unique construction, the efficacy stopping boundary u in (2.5) cannot be infinite, and thus the circular conditional error function cannot be applied to the situation when early stopping for efficacy is not allowed. For such a special strategy, we instead propose to use the linear conditional error function $A_{\text{lin}}(\cdot)$ to control the overall type I error; see the supplementary material for more technical details. As shown by Proschan and Hunsberger [15], the linear and circular error functions have similar performances through careful calibration. Although the linear error function is also suitable for EFE-SSR, it involves two parameters that should be determined in advance. To reduce the burden of parameter specification, we use the circular function $A_{\text{cir}}(\cdot)$ for the enrichment strategies if early terminations for both futility and efficacy are included at the interim analysis.

2.4 Enrichment with early stopping for both futility and efficacy based on the ϵ rule and SSR (EFE-ϵ-SSR)

Compared to EFE-SSR, EFE-ϵ-SSR includes both futility and efficacy stopping boundaries ($l_\epsilon \geqslant 0$ and $u_\epsilon \leqslant 0$) as well as an additional parameter $\epsilon \geqslant 0$, which is the indifference margin, as shown in Figure 2. Such an indifference margin should be prespecified and it defines the cutoff of the difference between the subgroup-specific test statistics. In other words, if the distance in the test statistic between two subgroups is smaller than ϵ, then the treatment

effects are deemed indistinguishable between subgroups, corresponding to the homogeneous case; otherwise, it corresponds to the heterogeneous case and it is more desirable to enrich the subgroup with a larger treatment effect. At the interim analysis, the trial based on EFE-ϵ-SSR proceeds as follows:

(1) If $t_{g^*} \leqslant l_\epsilon$, then terminate the trial and declare treatment futility in the overall group.

(2) Otherwise, compare the difference between the subgroup-specific test statistics, $\Delta_t = t_{g^*} - t_{g'}$, with ϵ.

(2a) If $\Delta_t \geqslant \epsilon$, perform an efficacy test in subgroup g^*,

- If $t_{g^*} \geqslant u_\epsilon$, then terminate the trial and declare treatment efficacy in subgroup g^*.
- Otherwise, enrich subgroup g^* in the second stage; that is, $g^\dagger = g^*$.

(2b) Otherwise, i.e., $\Delta_t < \epsilon$, perform an efficacy test in the overall group.

- If $t_0 > u_\epsilon$, then terminate the trial and declare treatment efficacy in the overall group.
- Otherwise, the trial proceeds to the second stage without enrichment; that is, the selected subgroup is the overall group with $g^\dagger = 0$.

(3) Based on the observed data, we calculate the critical value used at the final analysis and re-estimate the sample size N_2 to reach adequate CP while controlling the overall type I error rate.

Similar to EFE-SSR, the circular conditional error function $A_{\text{cir}}(t^{(1)}; l_\epsilon, u_\epsilon)$ with futility and efficacy stopping boundaries (l_ϵ, u_ϵ) is employed to preserve the overall type I error rate

at a pre-specified α level; that is,

$$\begin{aligned}\alpha \;=\; & \int_{-\infty}^{\infty}\int_{\max\{t_1+\epsilon,l_\epsilon\}}^{\infty} A_{\text{cir}}(t_2)\phi(t_1)\phi(t_2)\mathrm{d}t_2\mathrm{d}t_1 + \int_{-\infty}^{\infty}\int_{\max\{t_2+\epsilon,l_\epsilon\}}^{\infty} A_{\text{cir}}(t_1)\phi(t_1)\phi(t_2)\mathrm{d}t_1\mathrm{d}t_2 \\ & + \int_{-\infty}^{\infty}\int_{\max\{t_1,l_\epsilon\}}^{t_1+\epsilon} A_{\text{cir}}(t_0)\phi(t_1)\phi(t_2)\mathrm{d}t_2\mathrm{d}t_1 + \int_{-\infty}^{\infty}\int_{\max\{t_2,l_\epsilon\}}^{t_2+\epsilon} A_{\text{cir}}(t_0)\phi(t_1)\phi(t_2)\mathrm{d}t_1\mathrm{d}t_2.\end{aligned}$$

Given the futility stopping boundary l_ϵ and the prespecified type I error rate α, the efficacy stopping boundary u_ϵ can be solved from the above equation numerically. For example, when $\alpha = 0.05$, $\epsilon = 0.2$, and $l_\epsilon = z_{\alpha^*} = 0.84$ with $\alpha^* = 0.2$, the value of u_ϵ is 2.189. For the enriched subgroup or the overall group proceeding to the second stage, the stage 2 sample size N_2 and the final cutoff c_α can be determined using (2.3) and (2.4), respectively. In addition, as shown in the supplementary material, the EFE-ϵ-SSR strategy can also be adapted, based on the linear conditional error function, to the situation when only futility stopping is included.

Compared with EFE-SSR, the inclusion of ϵ in EFE-ϵ-SSR adds another layer of flexibility in controlling whether the overall group will be kept in the second stage. In general, the smaller the value of ϵ, the higher the likelihood that only one subgroup will be enriched in stage 2, thus leading to a smaller sample size on average. Therefore, when the experimental treatment is efficacious in only one subgroup, the EFE-ϵ-SSR rule is more appealing than EFE-SSR as the former generally leads to a higher probability of enriching the correct subgroup. On the other hand, when the experimental treatment works for both subgroups, the EFE-SSR rule usually has a higher probability to keep the overall group in stage 2. The simulation study reported in Section 4 also confirms such a statement; for example, see Table 3.

3 Real Trial Example

To illustrate the proposed methods, we apply them to the NeoSphere trial [24], which was a randomized multicenter, open-label trial comparing the neoadjuvant pertuzumab and

trastuzumab (experimental treatment) with the combination of pertuzumab and docetaxel (standard treatment) in women with locally advanced, inflammatory, or early HER2-positive breast cancer. The primary endpoint was the pathological complete response (pCR). Trial participants were heterogeneous and classified by the hormone receptor expression. The trial was designed to explore the efficacy and safety of the treatment. At the end of the trial, 18 of 107 patients (data missing for one patient) receiving the treatment had pCR, including 3 out of 51 patients with ER-positive, PR-positive or both (subgroup 1), and 15 out of 55 patients with ER-negative and PR-negative (subgroup 2), while 23 out of 96 patients receiving the standard treatment had pCR, including 8 out of 46 in subgroup 1 and 15 out of 50 in subgroup 2. See Gianni et al. [24] for more details. Based on the observed data, it is reasonable to assume the prevalence ratio $\rho = 0.5$. To test the treatment efficacy of the experimental treatment in patients with respect to the standard treatment, we consider a study extension based on the observed trial data and construct three null hypotheses:

H_{00}: The experimental and standard treatments are equally effective in the overall group.

H_{01}: The experimental and standard treatments are equally effective in subgroup 1.

H_{02}: The experimental and standard treatments are equally effective in subgroup 2.

To cast this example in the context of a normal mean with a known variance, we apply the variance-stabilizing arcsin transformation for the empirical proportions (see the supplementary material of statistical techniques for more details). As a result, the interim test statistics (empirical effect sizes) of subgroup 1, subgroup 2 and the overall group computed using the NeoSphere trial data are 1.819 (0.370), 0.306 (0.060) and 1.273 (0.179), respectively.

We set the overall type I error rate to be 0.05 and the indifference margin ϵ to be 0.5. Suppose the futility stopping boundary at the interim analysis is 1.036, which corresponds to a p-value of 0.15. In other words, if the p-value at stage 1 for subgroup g^* is greater than 0.15, then the trial would be terminated for futility. For comparison, we also include

the standard SSR procedure [15] without enriching any subgroup in stage 2. Based on the circular conditional error function, the interim efficacy stopping boundaries in SSR, EFE-SSR, and EFE-ϵ-SSR are determined as 1.821, 2.194 and 2.212, respectively. Since SSR does not include an enrichment strategy, the overall group proceeds to the second stage because the overall group test statistic lies inside the continuation region of $(1.036, 1.821)$. In contrast, with the enrichment strategies imposed under the proposed designs, only subgroup 1 is enriched in the second stage as the test statistic in subgroup 1 (i.e., 1.819) is larger than the futility stopping boundary of 1.036, while that in subgroup 2 (i.e., 0.306) is smaller than 1.036 as well as $t_1 - \epsilon$.

As shown in Table S1 of the supplementary material, the additional sample size N_2 and the critical value can be estimated for each method. Figure 3 panel (a) presents the changes of CP and the p-value required at the final analysis with respect to a range of values N_2. It shows that for all methods, the higher the CP, the larger the required sample size in stage 2. Although EFE-SSR and EFE-ϵ-SSR have larger efficacy stopping boundaries as well as requiring larger critical values at the final analysis, the required sample sizes at stage 2 using these two methods are much smaller than that of SSR. This is a consequence of enrichment as only subgroup 1, with a stronger benefit from the treatment and a larger test statistic, is selected to stage 2 under EFE-SSR and EFE-ϵ-SSR. On the other side, the SSR method does not distinguish the treatment effects between subgroups and selects the overall group to stage 2. Thus, the overall treatment effect is diluted by the small effect of subgroup 2, which in turn leads to the increase in sample size.

We observe that EFE-SSR and EFE-ϵ-SSR produce similar values of N_2 and c_α, which is due to the value of the indifference margin $\epsilon = 0.5$ that facilitates similar enrichment decisions between the two designs. Given the same futility stopping boundary, the efficacy stopping boundary of EFE-ϵ-SSR ($u_\epsilon = 2.194$) is relatively smaller than that of EFE-SSR ($u = 2.212$) and, as a result, EFE-ϵ-SSR spends more type I error at the interim than EFE-

SSR. This implies that less type I error would be spent at the final analysis for EFE-ϵ-SSR, thus leading to a smaller sample size of EFE-ϵ-SSR. Such a phenomenon can also be observed in Figure 3 (b): with different values of the stage 1 test statistic, EFE-ϵ-SSR (with $\epsilon = 0.5$) uniformly yield a smaller value of N_2 than EFE-SSR.

In fact, the additional design parameter ϵ in EFE-ϵ-SSR offers extra flexibility in controlling the enrichment stringency: the smaller the value of ϵ, the more restrictive of the enrichment at the interim. In an extreme situation with $\epsilon = 0$, only one subgroup having the maximum test statistic can be selected to stage 2 based on EFE-ϵ-SSR. On the other side, with a large value of ϵ, the likelihood of no enrichment increases. For illustration purpose, we additionally experiment $\epsilon = 1.6$ for EFE-ϵ-SSR. As the difference in the test statistic between two subgroups is $1.819 - 0.306 = 1.513$, which is smaller than $\epsilon = 1.6$, there is not enrichment at the interim and patients from both subgroups will be enrolled in stage 2. As a result, the required sample size in EFE-ϵ-SSR increases to 868, because of the diluted overall treatment effect as well as the increased $N_1^\dagger$. Another interesting finding from Figure 3 (b) is that for EFE-ϵ-SSR, the conditional error curves are close regardless of the changes in ϵ. This indicates that as long as the enriched subgroups are the same (for example, when the treatment is truly efficacious in one subgroup but futile in the other subgroup), different values of ϵ may lead to very similar sample sizes at the second stage.

4 Numerical Study

4.1 Simulation configurations

We compare the performances of EFE-SSR and EFE-ϵ-SSR with the standard design without enrichment (S\E), the adaptive enrichment design (AED) without early stopping and SSR (see the supplementary material for more details), and the standard SSR without enrichment through extensive simulations. Simulation results of the proposed methods without early stopping for efficacy can be found in the supplementary material. Suppose that a balanced

two-arm phase III trial with heterogeneous trial participants is conducted. The primary endpoint is the percentage of change in tumor sizes assessed by diagnostic imaging. One-sided hypothesis tests are used. Assume the proportion of subgroup 1 is $\rho = 0.50$, and the variance is known with $\sigma^2 = 1$. In the S\E design, we consider the treatment effect for the overall group is $\theta_0 = 0.20$, and thus 620 patients are required to achieve 80% power at the 0.05 significance level. We set the value of the threshold $\epsilon = 0.2$ and the futility stopping boundary to be 0.842 (corresponding to a p-value of 0.20) for all methods. The efficacy stopping boundaries are then determined to be 1.852, 2.234, and 2.189 in SSR, EFE-SSR, and EFE-ϵ-SSR, respectively. The sample size required in the second stage and the critical value used at the final analysis are updated given the data observed at the interim analysis. A total of 5,000 simulated studies are carried out for each method under each scenario.

4.2 Performance evaluation

To quantify the performances of the proposed designs, several metrics are utilized: (1) Overall type I error rate (which is defined as the probability of rejecting the null hypothesis when the treatment is inactive), which should be controlled at a nominal level, i.e., $\alpha = 0.05$; (2) power (which is defined as the probability of rejecting at least one null hypothesis), for which the larger the better; and (3) expected sample size (ESS), which is defined as the average sample size over all simulated trials, and a smaller value of ESS should be considered more desirable; and (4) the percentage of enrichment (which is defined as the proportion of the enriched subgroup proceeding to the second stage among all simulated trials) for subgroups 1 and 2.

4.3 Sensitivity analysis

In practice, the performances of the proposed methods are typically influenced by four factors: the proportion ρ of subgroup 1, the initial sample size N_1 at the first stage, the futility stopping boundary, and the threshold ϵ. In general, the larger the value of ρ the larger

effective sample size, leading to higher statistical power if subgroup 1 is more sensitive to the experimental treatment. The smaller the value of N_1, the higher the uncertainty of the estimation at the interim analysis, resulting in lower statistical power. The smaller the value of the futility stopping boundary, the smaller the likelihood of futility stopping, resulting in a larger sample size as well as higher statistical power. The smaller the value of the threshold ϵ, the higher the chance of enrichment. Hence, we consider various values of ρ, N_1, the futility stopping boundary, and ϵ to evaluate the robustness and efficiency of the proposed methods.

4.4 Results

As shown in Table 1, the overall type I error rates are well controlled for AED, SSR, EFE-SSR, and EFE-ϵ-SSR under various settings of the null hypothesis. Under the null hypothesis, SSR yields the smallest ESS because it has the narrowest continuation region and thus leads to the highest percentage of early termination. Comparing EFE-ϵ-SSR with EFE-SSR, the former approach on average has a smaller ESS than the latter under the null hypothesis. This is because compared to EFE-SSR, EFE-ϵ-SSR with $\epsilon = 0.20$ generally yields a higher likelihood to enrich only one subgroup in the second stage.

Table 2 summarizes the overall power and ESS under various alternative hypotheses, and Table 3 presents percentages of enrichment and the probabilities of rejecting H_{00}, H_{01}, and H_{02} for the designs under comparison. Table 2 shows that the SSR design performs on average better than S\E in terms of both ESS and power. Compared to S\E, AED gains 10.0% in power due to enrichment. In most scenarios of Table 2, EFE-SSR and EFE-ϵ-SSR yield comparable results, and both outperform SSR in terms of power and ESS, with almost a 7.0% gain in power and more than 10.0% reduction in ESS. In the case where only one subgroup can benefit from the treatment, such as scenarios 1, 4, and 7, the advantages of EFE-SSR and EFE-ϵ-SSR over SSR are obvious due to the accurate identification of the

beneficial subgroup at the interim. Table 2 also reveals that when the treatment effect in one subgroup is promising and that in the other subgroup is futile or marginal, such as scenarios 1 and 2, EFE-ϵ-SSR has a larger chance to identify the correct subgroup due to its more restrictive enrichment criterion (as shown in Table 3), and thus leads to a smaller ESS than EFE-SSR without deterioration in power. Such a reduction in ESS with EFE-ϵ-SSR is more prominent with a smaller futility stopping boundary (as shown in Figure **??** of the supplementary material particularly when the futility stopping boundary is smaller than 0.7). Furthermore, according to Table 3, EFE-SSR has a higher chance to reject the null hypothesis for the overall group H_{00} compared with other competitive methods, which may also lead to a higher chance of making erroneous conclusions in scenarios 1, 2, 4, 5 and 7, where heterogeneous treatment effect is present across subgroups.

Supplementary Figure **??** presents the power and ESS under a range of values of ρ using all designs under consideration. The power of AED, SSR, EFE-SSR, and EFE-ϵ-SSR increases substantially when the proportion ρ becomes large if subgroup 1 is more sensitive to the experimental treatment and vice versa. Moreover, the proposed EFE-SSR and EFE-ϵ-SSR outperform SSR in terms of both power and ESS, especially when the proportion ρ is small, i.e., $\rho < 0.8$.

Supplementary Figure **??** presents the power and ESS under a range of values of N_1. The power and ESS of SSR, EFE-SSR, and EFE-ϵ-SSR increase as the value of N_1 becomes large. In particular, the larger value of N_1, the higher power and ESS required for the trial. The performance of the AED is less sensitive to the value of N_1 as the ESS is fixed. From our experience, the value of N_1 is recommended to be one-half of the initial sample size based on the standard design without enrichment.

Supplementary Figure **??** presents the power and ESS under a range of futility stopping boundaries. Under the null hypothesis, all the considered methods can preserve the overall type I error rate under different values of the futility stopping boundary. The AED does not

allow any early termination, hence its power and ESS are invariant to the futility stopping boundary. Under the alternative hypothesis, the power and ESS of SSR, EFE-SSR, and EFE-ϵ-SSR increase as the futility stopping boundary decreases. This is because with a smaller futility stopping boundary, more trials would proceed to stage 2, leading to increase in both sample size and power. However, when the treatment is truly effective in at least one subgroup (as shown in the last panel of Figure **??**), the performances of EFE-SSR and EFE-ϵ-SSR are insensitive to the futility stopping boundary, because they can identify the correct subgroup with high probability. Based on the sensitivity analysis, we recommend choosing the futility stopping boundary from $[0.7, 1.1]$ to guarantee high overall power and low ESS.

Furthermore, we consider three different values of ϵ for EFE-ϵ-SSR: $\epsilon = 0, 0.5$, and 0.8. In particular, with $\epsilon = 0$, the EFE-ϵ-SSR strategy only enriches one subgroup at the interim analysis. As shown in the Supplementary Tables **??** and **??**, the power of EFE-ϵ-SSR is relatively invariant to values of ϵ. On the other side, the ESS generally decreases as the value of ϵ decreasesm because a small value of ϵ facilitates a more stringent enrichment strategy. The ESS of EFE-ϵ-SSR under $\epsilon = 0.8$ is much larger than that under $\epsilon \leqslant 0.5$ without a significant gain in power. This is because the larger the value of ϵ, the less probability the enrichment would be triggered, and thus EFE-ϵ-SSR shrinks towards the standard SSR without enrichment. On the other side, when $\epsilon = 0$, only one subgroup can be enrolled. As shown in Supplementary Table **??**, if the experimental treatment works in both subgroups, setting $\epsilon = 0$ may yield undesirable performance with a high subgroup-specific type II error rate in the unselected subgroup. As a result, a relatively small but positive value is recommended for the threshold ϵ, i.e., $0.1 \leqslant \epsilon \leqslant 0.5$.

5 Software

To facilitate the use of the proposed designs, we have developed an `R` package "esDesign" that allows users to calculate the futility and/or efficacy stopping boundaries, calibrate the value of the threshold, estimate the sample size required at the second stage, compute the critical value used at the final analysis and conduct simulation studies. The software is freely available on `CRAN` (https://cran.r-project.org/web/packages/esDesign/index.html).

6 Discussion

Adaptive enrichment design with sample size re-estimation inherits the advantages of both strategies of enriching the subgroup for which the treatment effect appears to be strong and re-estimating the required sample size in the second stage to ensure adequate conditional power. We have verified the feasibility and practicability of the combination of enrichment and SSR through a real trial example and extensive simulation studies. Compared with the traditional adaptive designs, such as SSR or AED, not only can the adaptive enrichment strategies with SSR increase the power substantially, but they also reduce the ESS, while preserving the overall type I error rate at a nominal level. It may be due to the fact that the trials declaring early stopping for futility by SSR are transformed to enrich a proportion of trial participants, who have a higher likelihood to benefit from the investigated treatment. This is very attractive as it enhances the probability of trial success, facilitates the advancement of drug development, and provides easy-to-implement approaches. The proposed methods may be extended to trials with survival endpoints, for which the log-rank test is often used for hypothesis testing. It is also warranted to explore the cases with multiple endpoints or a primary endpoint and a secondary endpoint.

Acknowledgements

We thank the two referees, Associate Editor, and Editor for their many constructive and insightful comments that have led to significant improvements in the article. Lin's research was supported in part by grants P30 CA016672 and P50 CA221703 from the National Cancer Institute (NCI), Yuan's research was supported in part by grants P50 CA098258 and P30 CA016672 from the NCI, and Yin's research was supported in part by a grant No. 17308420 from the Research Grants Council of Hong Kong.

References

[1] Antoniou, M., Jorgensen, A. L., Kolamunnage-Dona, R. (2016). Biomarker-guided adaptive trial designs in phase II and phase III: A methodological review. *Plos One* **11**, e0149803.

[2] Freidlin, B. and Simon, R. (2005). Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research* **21**, 7872–7878.

[3] Liu, A., Liu, C., Li, Q., Yu, K. F., and Yuan, V. W. (2010). A threshold sample-enrichment approach in a clinical trial with heterogeneous subpopulations. *Clinical Trials* **7**, 537–545.

[4] Wang, S. J., James Hung, H. M., and O'Neill, R. T. (2009). Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal* **51**, 358–374.

[5] Freidlin, B. and Korn, E. L. (2014). Biomarker enrichment strategies: matching trial design to biomarker credentials. *Nature Reviews Clinical Oncology* **11**, 81–90.

[6] Matsui, S. and Crowley, J. (2018). Biomarker-Stratified Phase III Clinical Trials: enhancement with a subgroup-focused sequential design. *Clinical Cancer Research* **24**, 994–1001.

[7] Tajik, P., Zwinderman, A. H., Mol, B. W., and Bossuyt, P. M. (2013). Trial designs for personalizing cancer care: a systematic review and classification. *Clinical Cancer Research* **19**, 4578–4588.

[8] Freidlin, B., Korn, E. L., and Gray, R. (2014). Marker sequential test (MaST) design.*Clinical Trials* **11**, 19–27.

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

[10] Beckman, R. A., Clark, J., and Chen, C. (2011). Integrating predictive biomarkers and classifiers into oncology clinical development programmes. *Nature reviews Drug discovery* **10**, 735–748.

[11] Kelloff, G. J. and Sigman, C. C. (2012). Cancer biomarkers: selecting the right drug for the right patient. *Nature Reviews Drug Discovery* **11**, 201–214.

[12] Mandrekar, S. J. and Sargent, D. J. (2009). Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *Journal of Clinical Oncology* **27**, 4027–4034.

[13] Freidlin, B., McShane, L. M., and Korn, E. L. (2010). Randomized clinical trials with biomarkers: design issues. *Journal of the National Cancer Institute* **102**, 152–160.

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

[15] Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics* **51**, 1315–1324.

[16] Magnusson, B. P. and Turnbull, B. W. (2013). Group sequential enrichment design incorporating subgroup selection. *Statistics in Medicine* **32**, 2695–2714.

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

[18] Jenkins, M., Stone, A., and 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.

[19] Bauer, P. and Kohne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics* **50**, 1029–1041.

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

[21] Jennison, C. and Turnbull, B. W. (2003). Midcourse sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* **22**, 971–993.

[22] Gao, P., Ware, J. H., and Mehta, C. (2008). Sample size re-estimation for adaptive sequential design in clinical trials. *Journal of Biopharmaceutical Statistics* **18**, 1184–1196.

[23] Mehta, C., Gao, P., Bhatt, D. L., Harrington, R. A., Skerjanec, S., and Ware, J. H. (2009). Optimizing trial design: sequential, adaptive, and enrichment strategies. *Circulation* **119**, 597–605.

[24] Gianni, L., Pienkowski, T., Im, Y. H., Roman, L., Tseng, L. M., Liu, M. C., et al. (2012). Efficacy and safety of neoadjuvant pertuzumab and trastuzumab in women with

locally advanced, inflammatory, or early HER2-positive breast cancer (NeoSphere): a randomised multicentre, open-label, phase 2 trial. *The Lancet Oncology* **13**, 25–32.

Table 1: Comparison of the overall type I error rate α (%), expected sample size (ESS) and the probabilities of rejecting null hypotheses H_{00}, H_{01} and H_{02} (%) using the standard sample size re-estimation (SSR) procedure without enrichment, adaptive enrichment design (AED) without early stopping, and the proposed enrichment strategies with SSR (EFE-SSR and EFE-ϵ-SSR) under the null scenario with $(\theta_1, \theta_2) = (0, 0)$ and $\epsilon = 0.2$.

		SSR		AED					EFE-SSR					EFE-ϵ-SSR				
N_1	ρ	α	ESS	α	ESS	H_{00}	H_{01}	H_{02}	α	ESS	H_{00}	H_{01}	H_{02}	α	ESS	H_{00}	H_{01}	H_{02}
156	0.3	5.4	257	4.7	620	0.3	2.1	2.3	5.4	288	1.4	1.9	2.1	4.9	283	0.3	2.2	2.4
156	0.5	5.4	257	4.6	620	0.2	2.3	2.1	4.9	285	1.3	1.9	1.7	4.9	282	0.3	2.7	1.9
156	0.7	5.4	257	4.7	620	0.2	2.3	2.1	5.0	282	1.5	1.7	1.8	4.9	285	0.2	2.3	2.4
310	0.3	5.2	504	4.6	620	0.2	2.1	2.3	4.8	558	1.3	2.0	1.6	4.5	555	0.1	2.6	1.8
310	0.5	5.2	504	4.4	620	0.3	2.2	2.0	4.6	576	1.4	1.4	1.8	4.8	561	0.2	2.3	2.3
310	0.7	5.2	504	4.4	620	0.2	2.2	2.0	5.5	568	1.5	2.1	1.9	4.6	568	0.2	2.4	2.0

Note: N_1 is the total sample size in stage 1, and ρ is the true prevalence rate of subgroup 1. The nominal type I error rate of 5% and conditional power 80% are specified for the two-stage adaptive designs.

Table 2: Comparison of power (%) and expected sample size (ESS) using the standard design without enrichment (S\E), adaptive enrichment design (AED) without early stopping, sample size re-estimation (SSR) design, and the proposed enrichment strategies with SSR (EFE-SSR and EFE-ϵ-SSR) under alternative scenarios with $\epsilon = 0.2$.

			S\E		AED		SSR		EFE-SSR		EFE-ϵ-SSR	
Scenario	(θ_1, θ_2)	θ_0	Power	ESS	Power	ESS	Power	ESS	Power	ESS	Power	ESS
1	(0.2, 0.0)	0.10	33.5	620	52.6	620	37.7	672	50.5	619	50.5	608
2	(0.2, 0.1)	0.15	57.5	620	57.0	620	60.4	671	58.8	632	57.5	609
3	(0.2, 0.2)	0.20	79.5	620	73.3	620	78.3	620	75.2	573	71.0	561
4	(0.3, 0.0)	0.15	57.5	620	81.7	620	60.4	671	78.1	548	78.4	553
5	(0.3, 0.1)	0.20	79.5	620	83.1	620	78.3	620	81.1	531	79.6	523
6	(0.3, 0.2)	0.25	93.1	620	88.8	620	89.5	526	88.0	491	86.5	489
7	(0.4, 0.0)	0.20	79.5	620	93.9	620	78.3	620	92.3	456	92.7	440
8	(0.4, 0.2)	0.30	98.1	620	96.8	620	96.0	445	95.6	406	94.4	408

Note: θ_1, θ_2 and θ_0 are the treatment effects of subgroups 1, 2, and the overall group, respectively. The sample size in stage 1 is $N_1 = 310$, and the prevalence rate of subgroup 1 is $\rho = 0.5$. The nominal type I error rate of 5% and conditional power 80% are specified for the two-stage adaptive designs.

Table 3: Comparison of the percentage of enrichment in subgroup 1 (ES_1) or 2 (ES_2) and the probabilities of rejecting null hypotheses H_{00}, H_{01} and H_{02} (%) using the adaptive enrichment design (AED) without early stopping and the proposed enrichment strategies with SSR (EFE-SSR and EFE-ϵ-SSR) under alternative scenarios with $\epsilon = 0.2$.

			AED					EFE-SSR					EFE-ϵ-SSR				
Scenario	(θ_1, θ_2)	θ_0	ES_1	ES_2	H_{00}	H_{01}	H_{02}	ES_1	ES_2	H_{00}	H_{01}	H_{02}	ES_1	ES_2	H_{00}	H_{01}	H_{02}
1	(0.2, 0.0)	0.10	76.6	16.1	1.8	49.6	1.2	75.4	12.5	8.4	41.1	1.1	80.2	16.2	1.7	47.1	1.7
2	(0.2, 0.1)	0.15	60.5	29.8	4.7	41.6	10.8	53.6	22.6	22.2	30.5	6.2	62.2	29.8	5.2	41.7	10.6
3	(0.2, 0.2)	0.20	45.4	43.3	7.9	33.6	31.8	32.2	36.7	39.3	17.1	18.8	46.5	45.8	7.5	32.1	31.4
4	(0.3, 0.0)	0.15	88.1	7.4	2.1	81.7	0.7	82.6	5.9	14.0	63.7	0.5	86.6	9.1	2.4	74.4	1.6
5	(0.3, 0.1)	0.20	76.8	15.9	4.9	71.7	6.5	64.3	12.5	32.3	45.8	3.0	73.2	20.7	4.8	67.2	7.7
6	(0.3, 0.2)	0.25	61.0	28.9	8.8	57.6	22.3	44.6	20.7	52.5	27.8	7.7	58.6	32.7	8.7	55.3	22.5
7	(0.4, 0.0)	0.20	93.9	3.0	2.2	93.1	0.4	83.6	2.9	17.8	74.4	0.1	93.1	4.8	1.6	90.4	0.7
8	(0.4, 0.2)	0.30	76.8	15.9	7.1	76.5	13.2	52.5	10.0	60.6	32.5	2.5	71.2	22.2	6.7	74.5	13.2

Note: θ_1, θ_2 and θ_0 are the treatment effects of subgroups 1, 2, and the overall group, respectively. The sample size in stage 1 is $N_1 = 310$, and the prevalence rate of subgroup 1 is $\rho = 0.5$. The nominal type I error rate of 5% and conditional power 80% are specified for the two-stage adaptive designs.

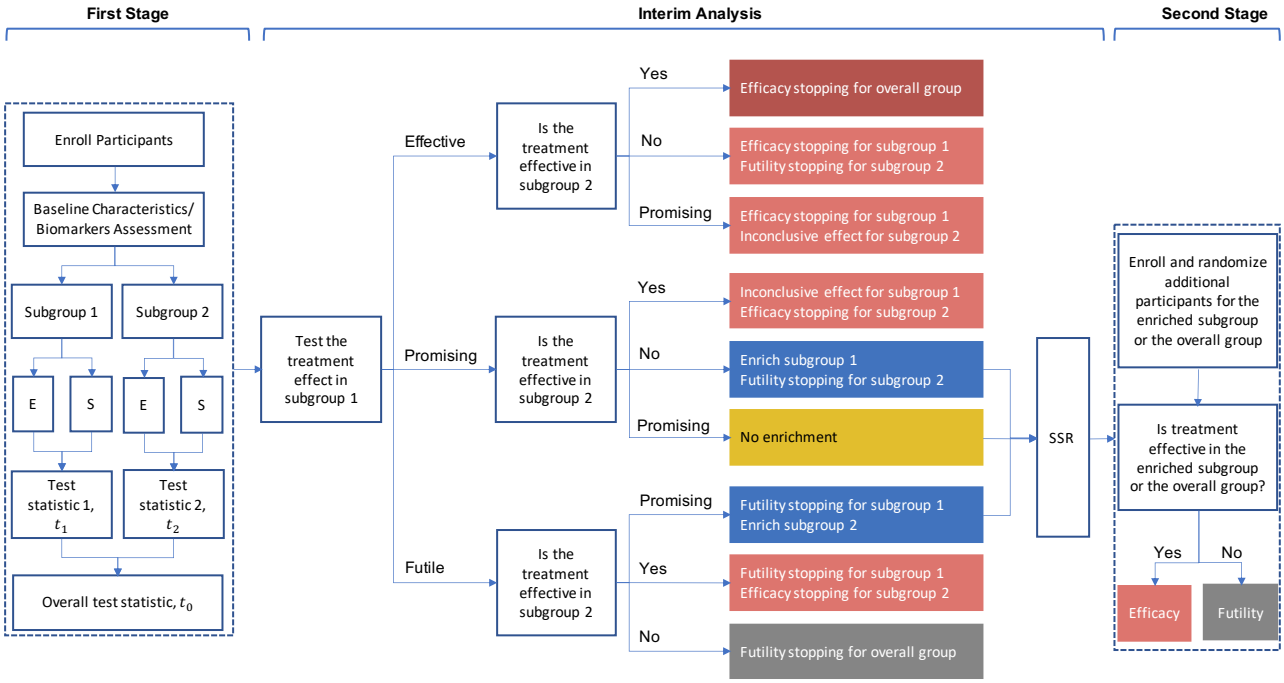

Figure 1: Schema of the enrichment strategy with early stopping for both futility and efficacy based on sample size re-estimation (EFE-SSR), with "E" representing the experimental treatment and "S" representing the standard treatment.

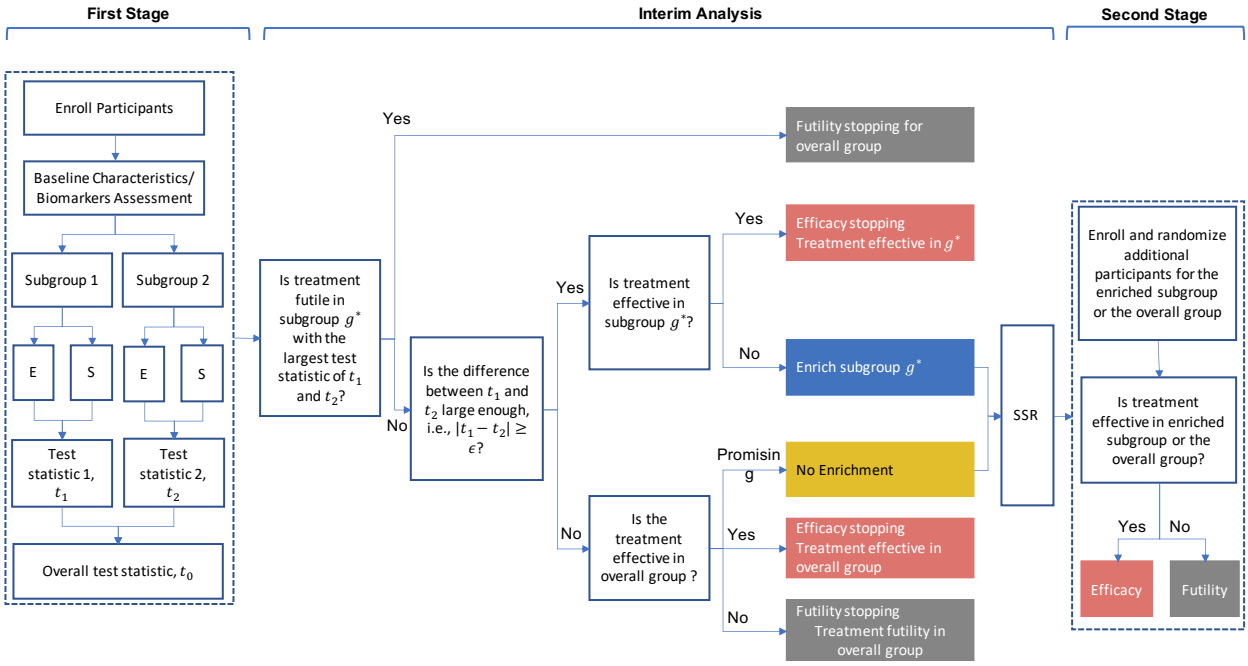

Figure 2: Schema of the enrichment strategy with early stopping for both futility and efficacy based on the ϵ rule and sample size re-estimation (EFE-ϵ-SSR), with "E" representing the experimental treatment and "S" representing the standard treatment.

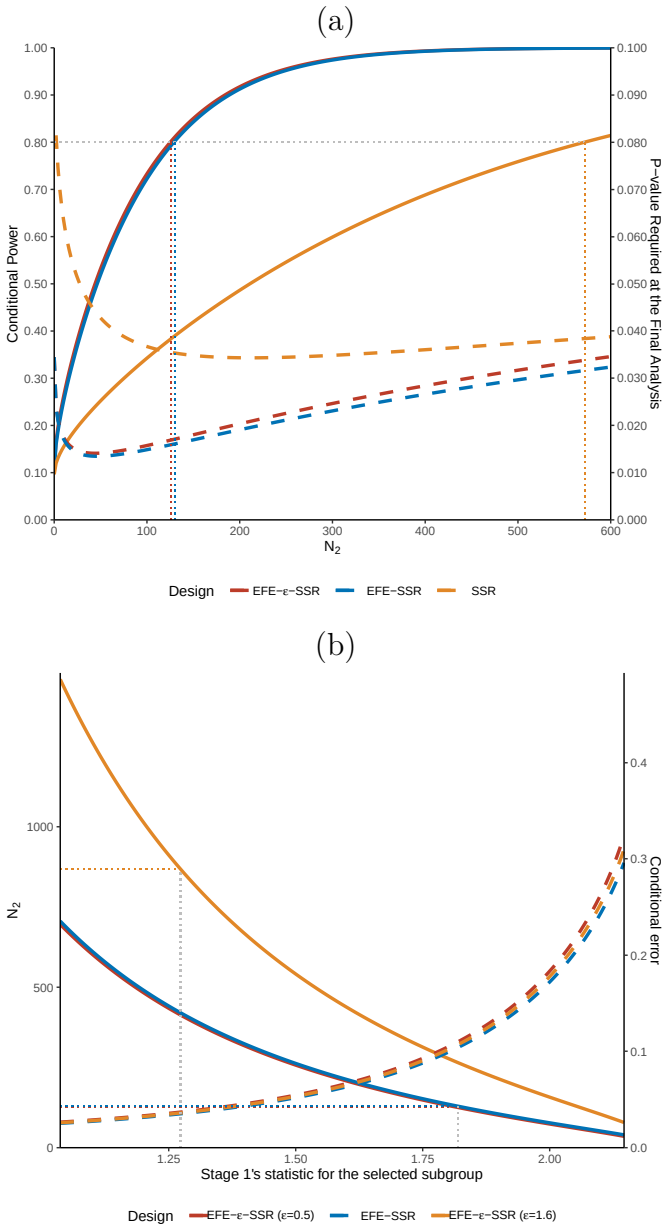

Figure 3: (a): Changes of conditional power (solid lines) and p-value (dashed lines) required at the second stage with respect to N_2 in the hypothetical example using SSR, EFE-SSR, and EFE-ϵ-SSR; the dotted lines correspond to the conditional power of 80%. (b): Required sample size (solid lines) and conditional error (dashed lines) at the second stage with respect to the stage 1 statistic for the selected subgroup in the hypothetical example using EFE-SSR and EFE-ϵ-SSR. Two different values of ϵ ($\epsilon = 0.5$ and $\epsilon = 1.6$) are considered for EFE-ϵ-SSR. The vertical dotted lines in panel (b) exhibit the test statistics for subgroup 1 (EFE-SSR and EFE-ϵ-EER with $\epsilon = 0.5$) and the overall group (EFE-ϵ-EER with $\epsilon = 1.6$). The horizontal lines denote the respective sample sizes at the second stage.