

Extension of Fisher's least significant difference method to multi-armed group-sequential response-adaptive designs

Statistical Methods in Medical Research

2025, Vol. 34(9) 1780–1794

© The Author(s) 2025



Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/09622802251319896

journals.sagepub.com/home/smmWenyu Liu¹  and D Stephen Coad²

Abstract

Multi-armed multi-stage designs evaluate experimental treatments using a control arm at interim analyses. Incorporating response-adaptive randomisation in these designs allows early stopping, faster treatment selection and more patients to be assigned to the more promising treatments. Existing frequentist multi-armed multi-stage designs demonstrate that the family-wise error rate is strongly controlled, but they may be too conservative and lack power when the experimental treatments are very different therapies rather than doses of the same drug. Moreover, the designs use a fixed allocation ratio. In this article, Fisher's least significant difference method extended to group-sequential response-adaptive designs is investigated. It is shown mathematically that the information time continues after dropping inferior arms, and hence the error-spending approach can be used to control the family-wise error rate. Two optimal allocations were considered. One ensures efficient estimation of the treatment effects and the other maximises the power subject to a fixed total sample size. Operating characteristics of the group-sequential response-adaptive design for normal and censored survival outcomes based on simulation and redesigning the NeoSphere trial were compared with those of a fixed-sample design. Results show that the adaptive design attains efficient and ethical advantages, and that the family-wise error rate is well controlled.

Keywords

Censored survival outcome, doubly-adaptive biased coin design, efficient randomised-adaptive design, error-spending function, multiple treatment comparison, optimal allocation

1 Introduction

Group sequential analysis, where data are sequentially evaluated over time and which allows stopping the trial early for success or futility and dropping of inferior treatments at interim analyses, has been well accepted from a regulatory review perspective in terms of saving time and resources and being more ethical.¹ Incorporating response-adaptive randomisation in a group sequential design can be more ethical by assigning more patients to the more promising treatments utilising the cumulative patients' responses observed during the course of the trial. Some response-adaptive randomisation methods can target a pre-specified optimal allocation that is based on some optimality criterion. They are the so-called 'optimal' response-adaptive randomisation methods.

¹Translational Epidemiology Unit, Big Data Institute, Nuffield Department of Population Health, University of Oxford, UK

²School of Mathematical Sciences, Queen Mary, University of London, UK

Corresponding author:

Wenyu Liu, Translational Epidemiology Unit, Big Data Institute, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Oxford OX3 7LF, UK.

Email: wen74liu@gmail.com

Methods combining group sequential tests with response-adaptive randomisation have been proposed. Jennison and Turnbull² derived theory to support that the combined approach still maintains the overall error rates for two-armed normal trials with known variances. The authors proved that the joint distribution of the test statistics has a standard form similar to that for a group-sequential non-adaptive design, but with the additional feature that the information level can depend on previous test statistics. In addition, a reduction in the inferior treatment number can be achieved at a cost of a slight increase in the expected total sample size. Morgan and Coad³ compared several adaptive allocation rules in a group sequential setting for two-armed binary trials, including two urn-model type designs, the doubly-adaptive biased coin design (DBCD) and the sequential maximum likelihood estimation rule, which minimises the expected number of failures and is a special case of the DBCD. Among the designs that they investigated, the drop-the-loser rule⁴ is found to be the most efficient method for achieving the competing objectives of reducing the expected number of failures and the expected total sample size. Zhu and Hu⁵ studied the combined approach for two-armed clinical trials with normal and binary responses. By considering monitoring the adaptive design at a continuous information time, the authors proved that the sequence of test statistics converges to a Brownian motion in distribution and asymptotically satisfies the canonical joint distribution proposed by Jennison and Turnbull⁶ for standard group sequential designs. Extension of the methods to two-armed censored survival trials was discussed in Liu and Coad.⁷

For multi-armed clinical trials, in addition to the inflation of the type I error rate caused by sequential testing, one needs to ensure that the family-wise error rate induced by multiple treatment comparisons is preserved. One simple approach is the Bonferroni adjustment to control the nominal type I error rate for each pairwise test. The Bonferroni approach can be used in the group sequential setting and allows unequal allocation across treatment groups, and, if desired, different shapes of critical boundaries for different pairwise comparisons. However, the approach strongly controls the overall type I error rate, and it can be too conservative at the price of losing power. The multi-armed multi-stage (MAMS) design,^{8,9} which in essence is a pairwise comparison approach that simultaneously compares multiple arms with the control at each analysis, allows dropping of inferior treatments at interim analyses when considering efficacy and futility boundaries. It also strongly controls the family-wise error rate.^{10,11} Alternatively, one may choose to control the per-hypothesis error rate instead. Another approach to group sequential monitoring of multi-armed clinical trials is to use a global test. Jennison and Turnbull^{6,12} derived critical boundaries analogous to Pocock's and the O'Brien and Fleming boundaries. These were derived based on multi-armed normal trials with equal variances and equal treatment allocation. For the unequal variances case, Proschan et al.¹³ suggested obtaining the critical boundaries by simulation. Alternatively, by the significance level approach, the critical boundaries derived under the assumption of equal variances can be used to give an approximate test.⁶ Incorporating response-adaptive sampling in such a group sequential global test for multi-armed trials was investigated by Liu and Coad.¹⁴ Their simulation results showed that, using the significance level approach, the same critical boundaries can be used to give an approximate test and the use of the combined approach can preserve the advantages of both group sequential analysis and optimal response-adaptive randomisation.

Previous work in Liu and Coad^{7,14} has shown that the same critical boundaries as for a group-sequential non-adaptive design can be used approximately for both multi-armed and two-armed group-sequential response-adaptive clinical trials and yields good operating characteristics. The idea of this study is that work can be incorporated in an extension of the two-stage Fisher's least significant difference (LSD) method, which consists of (i) a global test that sequentially monitors the homogeneity of multi-armed treatment effects and (ii) unadjusted two-armed pairwise tests if the global null hypothesis is rejected. Moreover, it allows dropping inferior arms at interim analyses and using response-adaptive randomisation to assign more patients to the more promising treatments. Fisher's LSD method is more powerful than Scheffé's method and is considered to be one of the most powerful multiple comparison methods.¹⁵ It can serve as an alternative to monitoring pairwise comparisons using MAMS designs.

Proschan et al.¹³ considered the group sequential Fisher's LSD method in the cases of equal allocation and fixed unequal allocation determined prior to the commencement of the experiment. For unbalanced models, Hayter¹⁶ discussed how to adjust testing to control the family-wise error rate for the LSD method for trials with more than four treatments. No adjustment is needed for three-treatment comparisons. It is possible to increase the power for the remaining pairwise comparisons at a later stage when any pairwise null hypothesis has been rejected and the inferior treatment has been dropped. Follmann et al.¹⁷ proposed sequentially rejective procedures to alleviate the issue of power loss using critical boundaries for the subsequent tests based on the remaining number of treatments. Nevertheless, the authors pointed out that the sequentially rejective procedures do not necessarily strongly control the family-wise error rate. In practice, the choice between pairwise tests and a global test depends on the aim of a trial. Here, we will focus on Fisher's LSD method extended to the group sequential design with optimal response-adaptive randomisation. We will describe how the overall type I error rate is controlled in the group sequential Fisher's LSD method. More specifically, information time will be shown to be continuously increasing when the trial proceeds, even when inferior arms have been dropped at interim analyses. The combined approach in the Bayesian paradigm has been explored. With the advantage of flexibility in Bayesian designs,

one can also consider adding experimental arms to a platform clinical trial.¹⁸ However, unlike the frequentist approach, the Bayesian one does not focus on controlling the type I error rate, which is a common requirement by regulatory authorities. We focus on the frequentist approach here.

The form of the test for the analogue of Fisher's LSD method generalised to group-sequential response-adaptive designs is described in Section 2. It is also shown that the information time continues after dropping inferior arms for immediate and censored survival outcomes. Two response-adaptive randomisation functions are presented in Section 3. They are both a function of the current and the optimal allocation proportions. In particular, optimal allocation proportions for multi-armed trials and their update after some inferior arms have been dropped are discussed. Operating characteristics of the extension of Fisher's LSD method investigated via simulation of three-armed normal and censored survival trials and redesigning a four-armed binary trial are summarised in Section 4. Conclusions of the findings and some discussion of the trial design in practice are in Section 5. The definitions of the two randomisation functions and simulation results for some further scenarios are provided in the Appendix.

2 Form of test

An extension of Fisher's LSD method¹³ to group-sequential response-adaptive designs is investigated in this paper. First, a global test statistic is monitored sequentially to test for the homogeneity of treatment effects. If the global null hypothesis is rejected, unadjusted pairwise comparisons are conducted at this and subsequent looks if the trial proceeds. Inferior treatments can be dropped after pairwise comparisons. Note that each unadjusted pairwise comparison is essentially a two-sample test. One can conduct all pairwise comparisons. In a clinical trial, the primary interest is usually the comparison of each experimental treatment to the control. For simplicity, we illustrate the design for three-armed normal trials first, which can be applied to other types of responses and extended to more than three treatment arms, as described in later sections.

2.1 Critical boundaries

Assume that the treatment responses for patients are normally distributed with mean μ and standard deviation σ . In what follows, let $E1$ and $E2$ denote the experimental treatments, and C refers to the control group. First, we wish to test the global null hypothesis $H_{G_0} : \mu_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \mu_G \neq \mathbf{0}$, where $\mu_G = (\mu_{E1} - \mu_C, \mu_{E2} - \mu_C)^T$ is a vector of treatment contrasts of the means. Let K be the number of group sequential analyses. The global test statistics S_k and the corresponding critical boundaries d_k at look k , $k = 1, \dots, K$, can be found in Jennison and Turnbull.^{6,12} For a group-sequential response-adaptive global test, the joint distribution of the sequence of test statistics $S_1, \dots, S_K$ does not have the standard canonical joint distribution. However, the significance level approach⁶ can be used to give an approximate test as long as the imbalance in the sample sizes is not too severe. More specifically, the critical boundaries originally derived based on equal variances and equal allocation can be applied to the observed responses to give an approximate test. If $S_k < d_k$, $k = 1, \dots, K - 1$, the trial proceeds with all of the treatments to the next interim analysis. If the trial reaches the end of the study and $S_K < d_K$, we accept H_{G_0} and claim that there is no difference between the experimental treatments and the control. If $S_k \geq d_k$, $k = 1, \dots, K$, we reject H_{G_0} and start pairwise comparisons.

In the pairwise comparisons, we wish to test the pairwise null hypotheses $H_0^{(j)} : \mu_{Ej} = \mu_C$ versus $H_a^{(j)} : \mu_{Ej} \neq \mu_C$ for $j = 1, 2$. These pairwise null hypotheses are tested repeatedly at each interim analysis. The two-sample test statistics Z_k and the sequential critical boundaries c_k based on the error-spending approach¹⁹ can be used to control the overall type I error rate, which does not require the number of looks to be pre-specified. Two error probabilities are considered: (i) the probability of rejecting at least one of the two null hypotheses and (ii) the probability of rejecting both null hypotheses. Let $Z_{jC,k}$, $j = 1, 2$, refer to the test statistic for comparing experimental treatment Ej with the control C at look k . Suppose that a higher value of the test statistic indicates that the corresponding experimental treatment has greater efficacy. Then there are several outcomes of the pairwise tests.

- If $Z_{1C,k} \geq c_k$ and $Z_{2C,k} \geq c_k$, we stop the trial and claim that both experimental treatments are superior to the control.
- If $Z_{1C,k} \leq -c_k$ and $Z_{2C,k} \leq -c_k$, we stop the trial and claim that both experimental treatments are inferior to the control.
- If $Z_{1C,k} \geq c_k$ and $Z_{2C,k} \leq -c_k$, we stop the trial and claim that $E1$ is superior and $E2$ is inferior to the control.
- If $Z_{1C,k} \leq -c_k$ and $Z_{2C,k} \geq c_k$, we stop the trial and claim that $E1$ is inferior and $E2$ is superior to the control.
- If $Z_{1C,k} \geq c_k$ and $-c_k < Z_{2C,k} < c_k$, we stop the trial and claim that $E1$ is superior to the control.
- If $-c_k < Z_{1C,k} < c_k$ and $Z_{2C,k} \geq c_k$, we stop the trial and claim that $E2$ is superior to the control.
- If $Z_{1C,k} \leq -c_k$ and $-c_k < Z_{2C,k} < c_k$, $k = 1, \dots, K - 1$, we drop $E1$ and continue with $E2$ and C to the next look. When $k = K$, we claim that $E1$ is inferior to the control, but that there is no difference between $E2$ and C .

- h. If $-c_k < Z_{1C,k} < c_k$ and $Z_{2C,k} \leq -c_k$, $k = 1, \dots, K-1$, we drop $E2$ and continue with $E1$ and C to the next look. When $k = K$, we claim that $E2$ is inferior to the control, but that there is no difference between $E1$ and C .
- i. If $-c_k < Z_{1C,k} < c_k$ and $-c_k < Z_{2C,k} < c_k$, $k = 1, \dots, K-1$, we continue with all of the treatments and conduct the pairwise tests again at the next look. When $k = K$, we claim that there is no difference between the treatments.

When the global null hypothesis H_{G_0} is rejected, the probability of encountering case i is small. However, it can occur, since different test statistics and critical boundaries are used for the global and pairwise tests. When the number of arms is J , there will be 3^{J-1} possible outcomes for the comparisons with the control.

2.2 Information time

2.2.1 Immediate responses

Information time for immediate responses is a ratio of the current sample size to the maximum sample size. For trials that allow dropping of inferior treatments, Follmann, Proschan and Geller¹⁷ showed that the information time at look k remains the same if some treatments are dropped at this look. More specifically, the information time continues and increases when the trial proceeds, with $t_0 = 0$ and $t_K = 1$. After dropping arms, it becomes the current number of subjects on the arms remaining in the trial divided by the total number of patients planned for these arms, that is,

$$t_k = \frac{n_k}{N} = \frac{\sum_{j=1}^J m_{j,k}}{\sum_{j=1}^J M_j} = \frac{\sum_{j \in C} m_{j,k}}{\sum_{j \in C} M_j} \in (0, 1], \quad k = 1, \dots, K \quad (1)$$

where n_k is the cumulative sample size at look k , N is the planned maximum sample size, $m_{j,k}$ is the cumulative number of patients on treatment j at look k , M_j is the total sample size on treatment j and C represents the current set of treatments remaining in the study, which is a subset of $\{1, \dots, J\}$.

For optimal response-adaptive randomisation, the total sample size on treatment j , M_j , is random. However, since M_j/N converges almost surely to the pre-specified target optimal allocation proportion for treatment j , ρ_j ,²⁰ M_j can be approximated by $\rho_j N$. The approximate information time becomes

$$t_k = \frac{\sum_{j \in C} m_{j,k}}{\sum_{j \in C} \hat{\rho}_j N} \in (0, 1], \quad k = 1, \dots, K \quad (2)$$

where $\hat{\rho}_j$ is the current estimate of ρ_j , which depends on the unknown parameters, and $\hat{\rho}_j N$ is an estimate of the planned total sample size on treatment j , M_j , with $\sum_{j \in C} \hat{\rho}_j \leq 1$, since $\sum_{j=1}^J \hat{\rho}_j = 1$. Details of the optimal allocation proportions $(\rho_1, \dots, \rho_J)$ for two or more arms are described in Section 3.1. Equation (1) shows that the information time continues after dropping arms. Applying (2) with the error-spending approach,¹⁹ the interim analysis can be planned at any continuous information time $t_k \in (0, 1]$.

2.2.2 Censored survival responses

For censored survival responses, the information time is proportional to the number of events. The information time can be approximated by the ratio of the expected number of events on the remaining arms at look k to the expected total number of events on the remaining arms, that is,

$$t_k = \frac{\hat{e}_k}{\hat{e}_K^{(k)}} = \frac{\sum_{j \in C} m_{j,k} \hat{e}_{j,k}}{\sum_{j \in C} M_j \hat{e}_{j,K}} \in (0, 1], \quad k = 1, \dots, K$$

where $\hat{e}_k$ is the estimated number of events at look k , $\hat{e}_K^{(k)}$ is the estimated number of events at the end of the trial based on the cumulative responses and $\hat{e}_{j,k}$ is the estimated probability of an event on arm j at look k . There are two candidates for the probability of an event. One is under the null hypothesis where the parameters are all equal and the other is under a specified alternative. Both can be used with the error-spending function to control the overall type I error rate.²¹

For simplicity, we consider the information time scale under H_{G_0} , and hence $\hat{\epsilon}_{j,k}$ and $\hat{\epsilon}_{j,K}$ can be replaced by $\hat{\epsilon}_k$ and $\hat{\epsilon}_K$, respectively. In addition, for optimal response-adaptive randomisation, M_j can be approximated by $\hat{\rho}_j N$. Then the approximate information time at interim analysis k can be written as

$$t_k = \frac{\sum_{j \in C} m_{j,k} \hat{\epsilon}_k}{\sum_{j \in C} \hat{\rho}_j N \hat{\epsilon}_K} \in (0, 1], \quad k = 1, \dots, K$$

Since the probability of an event at look k , ϵ_k , and the optimal allocation proportion for arm j , ρ_j , depend on the unknown parameters, the parameter estimates are used here. The accuracy of the parameter estimates increases in the later stages of the trial with a larger cumulative sample size. The use of an error-spending approach which allocates little type I error rate to the early group sequential tests seems to be more sensible.

3 Optimal response-adaptive randomisation

3.1 Optimal allocation proportions for multi-armed trials

We consider two target optimal allocation proportions for multi-armed clinical trials. D_A -optimal allocation was derived based on heteroscedastic models where the variances for the responses across treatments are unequal and an unbalanced allocation can be more efficient. Wong and Zhu²² first derived D_A -optimal allocation for multi-armed normal trials using a fixed sample and Sverdlov et al.²³ further generalised the D_A -optimal allocation to censored survival responses. In brief, the D_A -optimal allocation is obtained by minimising the determinant of the covariance matrix of the vector of treatment contrasts estimator. That is, the smallest confidence ellipsoid for the vector of treatment contrasts is found so that the D_A -optimal allocation ensures the most efficient estimates of the treatment effects. The other optimal allocation for multi-armed trials that we consider is based on non-linear programming (NP), which maximises the power subject to the constraint that the total sample size does not exceed a fixed value. NP allocation involves a user-specified lower bound for the allocation proportions, B . This allocation is an analogue of Neyman allocation and has a closed form for the solution.^{23,24} NP can be extended to find solutions that maximise the power subject to a fixed weighted sample size or minimise the total expected hazard subject to a constraint on the power, which is an analogue of the optimal allocation derived by Rosenberger et al.²⁵ generalised to $J \geq 3$ treatments. There is no general solution to such extensions and numerical methods are required to obtain the solution. Here, we consider NP that maximises the power subject to a fixed total sample size.

Note that both of the above optimal allocation proportions ($\rho_1, \dots, \rho_J$) for multi-armed clinical trials were originally derived based on a fixed-sample design. They can be applied in group sequential designs, since the target optimal allocation would not be affected by the number of interim tests. In this paper, we extend the optimal response-adaptive randomisation to the case where inferior treatments are dropped at interim analyses. In this case, the allocation proportions for the remaining arms increase.¹⁷ The new optimal allocation proportion for treatment j , ρ'_j , after some arms have been dropped becomes

$$\rho'_j = \frac{M_j}{\sum_{l \in C} M_l} = \frac{\hat{\rho}_j N}{\sum_{l \in C} \hat{\rho}_l N} = \frac{\hat{\rho}_j}{\sum_{l \in C} \hat{\rho}_l}, \quad \sum_{j \in C} \rho'_j = 1 \quad (3)$$

where C is the current set of arms, M_j is the total sample size for treatment j , which can be approximated by $\hat{\rho}_j N$ for optimal response-adaptive designs, and $\hat{\rho}_j$ is the estimated optimal allocation proportion for treatment j without dropping of inferior treatments based on the responses available.

3.2 Response-adaptive randomisation procedures

Two optimal response-adaptive randomisation procedures, the doubly-adaptive biased coin design (DBCD)²⁶ and the efficient randomised-adaptive design (ERADE),²⁷ aimed at targeting the pre-specified optimal allocation for multi-armed trials are considered. Both the DBCD and the ERADE are functions of the current and target treatment allocation proportions, and the treatment allocation for the two response-adaptive randomisation procedures depends on the cumulative data. The target optimal allocation is also a function of the unknown parameters of the treatment effect sizes. In practice, parameter estimates based on the current responses are used. A learning phase using methods such as permuted-block randomisation is usually required to obtain initial parameter estimates. Then the DBCD and the ERADE for multi-armed trials can be implemented. The applications of the two optimal response-adaptive randomisation procedures with group sequential analysis in two-armed and multi-armed trials have been elaborated upon in Zhu and Hu⁵ and Liu and Coad.^{7,14} We briefly describe the two procedures in the Appendix.

Table 1. Simulated type I error rates for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = \mu_{E2} = \mu_C = 1$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 300$.

Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.046	0.002	299.2	(8.7)	0.334	(0.026)	0.333	(0.026)	0.333	(0.026)
DBCD _{D_A}	0.046	0.004	299.4	(6.8)	0.453	(0.016)	0.355	(0.019)	0.192	(0.019)
ERADE _{D_A}	0.044	0.003	299.3	(7.9)	0.451	(0.012)	0.355	(0.016)	0.194	(0.017)

4 Simulation studies

4.1 Three-armed normal trials

Consider comparing $J = 3$ treatments, including the control, at $K = 3$ group sequential analyses using the analogue of Fisher's LSD method. First, for the test of homogeneity, the global null hypothesis $H_{G_0} : \mu_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_0} : \mu_G \neq \mathbf{0}$, where $\mu_G = (\mu_{E1} - \mu_C, \mu_{E2} - \mu_C)^T$, is considered. The nominal type I error rate $\alpha = 0.05$ was set. The O'Brien and Fleming boundaries (18.36, 9.18, 6.12) for the sequential global chi-squared test statistics at information times ($t_1 = 0.33, t_2 = 0.67, t_3 = 1$) were used to give an approximate test. If H_{G_0} is rejected, then the pairwise null hypotheses $H_0^{(j)} : \mu_{Ej} = \mu_C$ versus the alternative hypotheses $H_a^{(j)} : \mu_{Ej} \neq \mu_C, j = 1, 2$, are tested subsequently. The O'Brien and Fleming boundaries (3.731, 2.504, 1.994) were used to give an approximate test for each pairwise Z test. These critical boundaries were derived based on normal responses with equal variances.^{6,28} Previous studies have shown that these critical boundaries can be applied to unequal variances or unequal treatment allocation as long as the imbalance in the sample sizes is not too severe.^{6,7,14} By using the error-spending approach, critical boundaries for group sequential tests taking place at unequally-spaced information times can also be obtained.

The two group-sequential response-adaptive designs DBCD and ERADE are compared with the group-sequential non-adaptive complete randomisation (CR) design in terms of the error probabilities, the expected number of patients (ENP), the expected number of failures (ENF) and the average allocation proportions with standard deviations. The following two error probabilities are considered: the probability of rejecting at least one of the two pairwise null hypotheses, that is, $\tilde{\alpha}^I$ under the null hypotheses and power^I under the alternative hypotheses; and the probability of rejecting both pairwise null hypotheses, that is, $\tilde{\alpha}^{II}$ under the null hypotheses and power^{II} under the alternative hypotheses. For the optimal response-adaptive designs, permuted-block randomisation was used for the first 10% of the N patients to obtain initial parameter estimates. Then the DBCD and ERADE functions with tuning parameters $\gamma = \gamma' = 2$ were used to compute the adaptive allocation probability. For normal responses, the D_A -optimal allocation was used as the optimal allocation $\rho = (\rho_{E1}, \rho_{E2}, \rho_C)$. Treatments inferior to the control were allowed to be dropped after the pairwise tests. After dropping treatments, (3) is used to obtain the optimal allocation proportions for the remaining arms. The results are based on 10,000 replicates. The Monte Carlo simulation error is around 0.002 for $\tilde{\alpha}^I$; ranges from 0.0004 to 0.0006 for $\tilde{\alpha}^{II}$; ranges from 0.002 to 0.005 for power^I; and ranges from 0.003 to 0.005 for power^{II}.

The true underlying effect of each treatment influences the decisions of treatment dropping and trial termination. Here, we consider three potential scenarios: (i) the control is inferior to the two experimental treatments, (ii) one experimental treatment is superior and one is inferior to the control and (iii) both experimental therapies are inferior to the control. The order of $E1$ and $E2$ does not affect the results.

As shown in Table 1, the simulated type I error rate $\tilde{\alpha}^I$ is close to 0.05 for all of the designs. Generally, $\tilde{\alpha}^I$ is within one standard error of 0.05. Under H_{G_0} , the chance of rejecting both pairwise null hypotheses, $\tilde{\alpha}^{II}$, is very small. Most trials continued to the end of the study without early termination and dropping of inferior arms. The ENP is about the same as the maximum number of patients, N . In addition, for the DBCD and the ERADE, the D_A -optimal allocation without dropping of arms is (0.454, 0.356, 0.191), which assigns more patients to the treatment with a larger variance in the responses. The desired D_A -optimal allocation proportions are well targeted. Also, in this case, the standard deviations of the allocation proportions are lower for the response-adaptive designs than for CR.

Table 2 considers the case where the control is inferior to the two experimental therapies. In this case, the ENP is about 264 for the group sequential CR design and 260 for the group-sequential response-adaptive designs, since the trials are allowed to stop early for superiority at interim analyses. Compared to CR, the response-adaptive designs can increase the power while using fewer patients. For example, the ERADE increases the power by nearly 4% while using five fewer patients on average compared to CR. Again, the target D_A -optimal allocation without dropping of inferior treatments is (0.454, 0.356, 0.191). Both adaptive designs target the D_A -optimal allocation proportions well, with the ERADE consistently having lower standard deviations.

Table 2. Simulated powers for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 2$, $\mu_{E2} = 1.5$, $\mu_C = 1$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 300$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.782	0.230	264.5	(48.9)	0.333	(0.027)	0.333	(0.028)	0.333	(0.028)
DBCD _{D_A}	0.808	0.289	260.6	(50.9)	0.452	(0.016)	0.356	(0.020)	0.192	(0.020)
ERADE _{D_A}	0.809	0.285	260.1	(50.9)	0.451	(0.013)	0.355	(0.017)	0.194	(0.019)

Table 3. Simulated powers for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 1$, $\mu_{E2} = 2$, $\mu_C = 1.5$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 300$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.617	0.083	280.0	(39.5)	0.331	(0.028)	0.334	(0.027)	0.335	(0.027)
DBCD _{D_A}	0.642	0.089	280.3	(38.4)	0.446	(0.028)	0.358	(0.024)	0.197	(0.020)
ERADE _{D_A}	0.650	0.083	280.4	(38.4)	0.445	(0.026)	0.357	(0.022)	0.198	(0.019)

Table 4. Simulated powers for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 1.5$, $\mu_{E2} = 1$, $\mu_C = 2$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 300$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.984	0.222	265.1	(26.7)	0.363	(0.035)	0.274	(0.049)	0.363	(0.035)
DBCD _{D_A}	0.971	0.255	270.5	(26.0)	0.489	(0.039)	0.292	(0.053)	0.218	(0.025)
ERADE _{D_A}	0.970	0.249	270.4	(26.1)	0.488	(0.039)	0.293	(0.053)	0.219	(0.024)

Table 3 considers the case where one experimental treatment is superior and one is inferior to the control. In this case, trials may stop early with the claim that a treatment is superior or drop the inferior treatment at an interim analysis. Here, the ENP is about 280 for the group sequential designs. Compared to the previous table, power^I reduces to about 61% for CR and 65% for the response-adaptive designs, since the contrasts of the means between the experimental treatments and the control are smaller in this case. For the DBCD and the ERADE, the target D_A -optimal allocation is (0.454, 0.356, 0.191) if no treatment is dropped. If $E1$ is dropped, the target allocation becomes (0, ρ'_{E2} , ρ'_C), where $\rho'_{E2} = \rho_{E2}/(\rho_{E2} + \rho_C) = 0.356/(0.356 + 0.191) = 0.651$ and $\rho'_C = 1 - \rho'_{E2} = 0.349$. Hence, the average allocation proportion for $E1$, $\tilde{\rho}_{E1}$, is a little lower than 0.454. However, since the contrast between $E1$ and the control is small, the chance of dropping $E1$ may not be very high. Therefore, the difference between $\tilde{\rho}_{E1}$ and ρ_{E1} is small.

Table 4 considers the case where both experimental therapies are inferior to the control. The target D_A -optimal allocation is (0.454, 0.356, 0.191) if no inferior treatment is dropped. If $E1$ is dropped, the target D_A -optimal allocation becomes (0, 0.651, 0.349). If $E2$ is dropped, it becomes (0.704, 0, 0.296). Compared to the previous tables, the standard deviations for the allocation proportions increase for all of the designs, especially the adaptive ones, since the chance of dropping arms is increased in this case. With smaller standard deviations σ_{E2} and σ_C , the difference between $E2$ and the control can be statistically significant. However, the response-adaptive designs achieve a slightly lower power than CR. This may be due to the slightly greater variability in the allocation proportions for the DBCD and the ERADE compared to CR. Nevertheless, power^I is still quite high for all of the designs.

4.2 Three-armed censored survival trials

Consider testing the contrasts of the mean survival times between two experimental treatments and the control, where independent exponentially distributed survival times and uniformly distributed arrival and censoring times are assumed. This testing problem was investigated by Sverdlov, Tymofyeyev and Wong²³ using a fixed-sample design with the DBCD, and their simulation settings were based on the head and neck cancer experiment.²⁹ We wish to test the global null hypothesis $H_{G_0} : \theta_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \theta_G \neq \mathbf{0}$, where $\theta_G = (\theta_{E1} - \theta_C, \theta_{E2} - \theta_C)^T$. If H_{G_0} is rejected, then the pairwise null hypotheses $H_0^{(j)} : \theta_{Ej} = \theta_C$ versus the alternative hypotheses $H_a^{(j)} : \theta_{Ej} \neq \theta_C$, $j = 1, 2$, are tested. For NP allocation, the lower bound for the allocation proportions is set to be 0.2. The nominal type I error rate, the approximate critical boundaries and the other simulation settings are the same as in Section 4.1. The Monte Carlo simulation error is around 0.002 for $\tilde{\alpha}^I$; ranges from 0.0006 to 0.0010 for $\tilde{\alpha}^{II}$; ranges from 0.004 to 0.005 for power^I; ranges from 0.001 to 0.005 for power^{II}.

Table 5. Simulated type I error rates for three-armed censored survival trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\theta_{E1} = \theta_{E2} = \theta_C = 24$, $N = 312$.

Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.033	0.006	311.8	(3.4)	203.7	(8.0)	0.334	(0.025)	0.333	(0.025)	0.333	(0.025)
DBCD $_{D_A}$	0.053	0.010	311.4	(6.8)	203.6	(8.2)	0.334	(0.034)	0.333	(0.035)	0.333	(0.034)
ERADE $_{D_A}$	0.049	0.008	311.3	(6.9)	203.2	(9.5)	0.334	(0.031)	0.333	(0.031)	0.333	(0.031)
DBCD $_{NP}$	0.040	0.004	311.4	(6.6)	203.2	(9.3)	0.333	(0.091)	0.334	(0.091)	0.333	(0.089)
ERADE $_{NP}$	0.042	0.004	311.4	(6.5)	203.3	(9.2)	0.334	(0.088)	0.333	(0.087)	0.333	(0.087)

Table 6. Simulated powers for three-armed censored survival trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\theta_{E1} = 34$, $\theta_{E2} = 24$, $\theta_C = 20$, $N = 312$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.721	0.090	295.5	(32.4)	186.3	(25.6)	0.334	(0.026)	0.333	(0.026)	0.333	(0.026)
DBCD $_{D_A}$	0.770	0.106	285.9	(37.7)	176.3	(30.0)	0.408	(0.029)	0.324	(0.037)	0.267	(0.040)
ERADE $_{D_A}$	0.769	0.105	286.3	(37.5)	176.6	(29.6)	0.406	(0.026)	0.324	(0.034)	0.269	(0.037)
DBCD $_{NP}$	0.800	0.047	280.4	(39.7)	168.6	(31.7)	0.532	(0.091)	0.229	(0.062)	0.239	(0.050)
ERADE $_{NP}$	0.791	0.048	280.9	(39.6)	169.2	(31.6)	0.526	(0.089)	0.233	(0.062)	0.241	(0.048)

Table 7. Simulated powers for three-armed censored survival trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\theta_{E1} = 20$, $\theta_{E2} = 34$, $\theta_C = 24$, $N = 312$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.540	0.023	306.4	(18.6)	197.0	(16.1)	0.332	(0.027)	0.334	(0.026)	0.334	(0.026)
DBCD $_{D_A}$	0.624	0.041	300.6	(26.2)	189.7	(21.5)	0.267	(0.043)	0.408	(0.029)	0.324	(0.038)
ERADE $_{D_A}$	0.621	0.041	300.9	(25.6)	190.0	(21.4)	0.269	(0.041)	0.406	(0.027)	0.324	(0.036)
DBCD $_{NP}$	0.602	0.012	296.4	(30.0)	182.7	(25.4)	0.241	(0.050)	0.529	(0.090)	0.229	(0.061)
ERADE $_{NP}$	0.598	0.014	296.2	(29.9)	182.8	(25.3)	0.242	(0.049)	0.526	(0.089)	0.232	(0.062)

Under the null hypothesis, from Table 5, the simulated type I error rate $\tilde{\alpha}^I$ is close to the nominal value for the response-adaptive designs, with less than one standard error deviation from 0.05 for the DBCD and the ERADE targeting the D_A -optimal allocation, and within three standard errors for the designs targeting the NP allocation. However, a conservative $\tilde{\alpha}^I$ is obtained for CR. Under H_{G_0} , the ENP and the ENF are similar for all of the designs. In addition, the average allocation proportions for all of the designs are close to equal allocation, with the response-adaptive ones targeting the NP allocation having larger standard deviations for the allocation proportions compared to the other designs.

Under the alternative hypothesis, in Table 6, the target D_A -optimal allocation without dropping of treatments is (0.406, 0.323, 0.271) and the optimal allocation based on nonlinear programming is (0.544, 0.2, 0.256). In this case, the NP allocation assigns more patients to the best treatment and fewer to the worst treatment compared to the D_A -optimal allocation. The DBCD and the ERADE targeting the NP allocation yield about eight fewer failures compared with the response-adaptive designs targeting the D_A -optimal allocation. In addition, the NP designs increase the power by around 3% and require about six fewer patients on average than the response-adaptive designs with D_A -optimal allocation.

For Table 7, where one experimental treatment is superior and one is inferior to the control, the target D_A -optimal allocation is (0.271, 0.406, 0.323) and the NP allocation is (0.256, 0.544, 0.2) without dropping of treatments. If $E1$ is dropped, the new target allocation becomes (0, 0.557, 0.443) for the D_A -optimal allocation and (0, 0.731, 0.269) for the NP allocation. The DBCD and the ERADE targeting the NP allocation consistently have higher standard deviations for the average allocation proportions compared to the other designs. However, the designs using the NP sampling rule achieve around seven fewer failures compared to those using the D_A -optimal allocation and about fifteen fewer failures compared to CR, since more patients are assigned to the best treatment, $E2$ in this case. In addition, the designs targeting the NP allocation reduce the ENP by about four compared to the ones using the D_A -optimal allocation. Although the power of the tests is also reduced, this may be due to the simulated type I error rates being smaller for the designs targeting the NP allocation than those targeting the D_A -optimal allocation.

For Table 8, where the experimental treatments are inferior to the control, the target D_A -optimal allocation is (0.271, 0.323, 0.406) and the NP optimal allocation is (0.256, 0.2, 0.544). If $E1$ is dropped, the new target allocation becomes (0,

Table 8. Simulated powers for three-armed censored survival trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\theta_{E1} = 20$, $\theta_{E2} = 24$, $\theta_C = 34$, $N = 312$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	ENF	(s.d.)	$\tilde{p}_{E1}$	(s.d.)	$\tilde{p}_{E2}$	(s.d.)	$\tilde{p}_C$	(s.d.)
CR	0.716	0.386	303.9	(19.0)	193.8	(15.4)	0.324	(0.034)	0.337	(0.028)	0.338	(0.028)
DBCD _{D_A}	0.772	0.471	299.0	(25.2)	187.0	(20.6)	0.259	(0.053)	0.327	(0.043)	0.413	(0.034)
ERADE _{D_A}	0.789	0.475	299.4	(25.0)	187.3	(20.5)	0.261	(0.051)	0.327	(0.040)	0.411	(0.032)
DBCD _{NP}	0.811	0.476	293.0	(29.1)	179.4	(24.2)	0.229	(0.053)	0.229	(0.058)	0.542	(0.090)
ERADE _{NP}	0.812	0.481	292.7	(29.5)	179.3	(24.6)	0.232	(0.054)	0.233	(0.059)	0.536	(0.092)

0.443, 0.557) for the D_A -optimal allocation and (0, 0.269, 0.731) for the NP allocation. If $E2$ is dropped, the new target allocation becomes (0.400, 0, 0.600) for the D_A -optimal allocation and (0.320, 0, 0.680) for the NP allocation. Compared to CR, the average allocation proportions for the DBCD and the ERADE assign more patients to the best treatment (C) and fewer to the least efficacious treatment ($E1$), and hence the ENF is reduced. More specifically, about fifteen fewer failures are achieved for the response-adaptive designs targeting the NP allocation and about seven fewer failures for the designs using the D_A -optimal allocation. Among the response-adaptive designs, the DBCD and the ERADE targeting the NP allocation attain a higher power while using about six fewer patients on average than those based on the D_A -optimal allocation.

4.3 Redesigning a four-armed binary trial

The NeoSphere trial³⁰ is redesigned using the adaptive designs with dropping of inferior treatments at interim analyses. The probabilities of success for the treatments are $p_C = 0.29$, $p_{E1} = 0.458$, $p_{E2} = 0.168$ and $p_{E3} = 0.24$. First, a chi-squared test statistic is monitored for tests of homogeneity. The global null hypothesis is $H_{G_0} : \mathbf{p}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \mathbf{p}_G \neq \mathbf{0}$ with $\mathbf{p}_G = (p_{E1} - p_C, p_{E2} - p_C, p_{E3} - p_C)^T$. Three group sequential analyses are planned at equally-spaced information times. The nominal type I error rate 0.05 was set. The O'Brien and Fleming critical boundaries derived based on normal responses with equal variances are used as an approximation. For the four-treatment comparison chi-squared statistics, the sequence of critical boundaries at the three equally-spaced information times is (23.76, 11.88, 7.92). If the global null hypothesis is rejected, then pairwise Z tests comparing each experimental treatment to the control are carried out at the current and subsequent looks. The pairwise null hypotheses $H_0^{(j)} : p_{Ej} = p_C$ versus the alternative hypotheses $H_a^{(j)} : p_{Ej} \neq p_C$, $j = 1, 2, 3$, are tested. Let $Z_{jC,k}$ refer to the pairwise test statistic for comparing treatment Ej with the control C at look k . The sequence of critical boundaries for each pairwise test is (3.731, 2.504, 1.994).

Suppose that a higher value of the test statistic indicates that the corresponding experimental treatment has greater efficacy. For the adaptive designs, early termination is allowed for treatment efficacy or futility. The possible cases when trials continue to the next look are shown below.

- (i) If $Z_{1C,k} \leq -c_k$, $|Z_{2C,k}| < c_k$ and $|Z_{3C,k}| < c_k$, $k = 1, \dots, K-1$, then $E1$ is dropped, and $E2$, $E3$ and the control are continued to the next look.
- (ii) If $|Z_{1C,k}| < c_k$, $Z_{2C,k} \leq -c_k$ and $|Z_{3C,k}| < c_k$, $k = 1, \dots, K-1$, then $E2$ is dropped, and $E1$, $E3$ and the control are continued to the next look.
- (iii) If $|Z_{1C,k}| < c_k$, $|Z_{2C,k}| < c_k$ and $Z_{3C,k} \leq -c_k$, $k = 1, \dots, K-1$, then $E3$ is dropped, and $E1$, $E2$ and the control are continued to the next look.
- (iv) If $Z_{1C,k} \leq -c_k$, $Z_{2C,k} \leq -c_k$ and $|Z_{3C,k}| < c_k$, $k = 1, \dots, K-1$, then $E1$ and $E2$ are dropped, and $E3$ and the control are continued to the next look.
- (v) If $Z_{1C,k} \leq -c_k$, $|Z_{2C,k}| < c_k$ and $Z_{3C,k} \leq -c_k$, $k = 1, \dots, K-1$, then $E1$ and $E3$ are dropped, and $E2$ and the control are continued to the next look.
- (vi) If $|Z_{1C,k}| < c_k$, $Z_{2C,k} \leq -c_k$ and $Z_{3C,k} \leq -c_k$, $k = 1, \dots, K-1$, then $E2$ and $E3$ are dropped, and $E1$ and the control are continued to the next look.
- (vii) If $|Z_{1C,k}| < c_k$, $|Z_{2C,k}| < c_k$ and $|Z_{3C,k}| < c_k$, $k = 1, \dots, K-1$, then all treatments are continued to the next look.

In addition, fixed-sample designs are provided alongside for comparison, where the critical boundaries are 7.81 and 1.96 for the global and pairwise tests, respectively. For optimal response-adaptive randomisation, the D_A -optimal allocation and the optimal allocation based on nonlinear programming (NP) were used. For the NP allocation, the lower bound for the allocation proportions B is set to be 0.2. The Monte Carlo simulation error is around 0.002 for $\tilde{\alpha}^I$; ranges from 0.0004 to 0.0005 for $\tilde{\alpha}^{II}$; around 0.002 for power^I; and ranges from 0.0005 to 0.0008 for power^{II}.

Table 9. Simulated type I error rates for redesigning NeoSphere trial using complete randomisation and response-adaptive randomisation with dropping of inferior treatments, $p_C = 0.29$, $p_{E1} = 0.29$, $p_{E2} = 0.29$, $p_{E3} = 0.29$, $N = 417$. $(t_1, t_2, t_3) = (0.33, 0.67, 1)$

Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{p}_C$	(s.d.)	$\tilde{p}_{E1}$	(s.d.)	$\tilde{p}_{E2}$	(s.d.)	$\tilde{p}_{E3}$	(s.d.)
CR	0.042	0.002	416.4	(8.0)	295.7	(5.7)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD _{D_A}	0.044	0.002	416.1	(12.0)	295.4	(8.5)	0.250	(0.012)	0.250	(0.013)	0.250	(0.013)	0.250	(0.012)
ERADE _{D_A}	0.047	0.003	416.0	(11.9)	295.4	(8.5)	0.250	(0.010)	0.250	(0.010)	0.250	(0.010)	0.250	(0.010)
DBCD _{NP}	0.042	0.002	416.2	(9.3)	295.5	(6.6)	0.261	(0.072)	0.239	(0.083)	0.260	(0.071)	0.240	(0.061)
ERADE _{NP}	0.045	0.002	416.0	(11.8)	295.4	(8.4)	0.259	(0.074)	0.241	(0.084)	0.260	(0.074)	0.240	(0.065)

Fixed-sample design

 $(t_1, t_2, t_3) = (0.33, 0.67, 1)$

CR	0.041	0.003	417	(0)	296.1	(0)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD _{D_A}	0.044	0.003	417	(0)	296.1	(0)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)
ERADE _{D_A}	0.045	0.003	417	(0)	296.1	(0)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)
DBCD _{NP}	0.041	0.002	417	(0)	296.1	(0)	0.249	(0.071)	0.250	(0.072)	0.252	(0.073)	0.250	(0.072)
ERADE _{NP}	0.040	0.002	417	(0)	296.1	(0)	0.250	(0.073)	0.250	(0.073)	0.252	(0.074)	0.249	(0.073)

As can be seen in Table 9, under the null hypothesis, the type I error rate $\tilde{\alpha}^I$ is within three standard errors of 0.05 for both the group sequential and the fixed-sample designs. The adaptive designs that combine group sequential analysis with optimal response-adaptive randomisation procedures, which allow dropping of inferior treatments during the course of the trial, can well preserve the overall type I error rate. Under the null hypothesis, the D_A -optimal allocation becomes equal allocation. For the NP allocation, patients are sequentially assigned according to the order of the current parameter estimates. The variability in the allocation proportions is much higher in this case than for the other designs.

Under the alternative hypothesis, in Table 10, a similar power is obtained for all of the designs. However, use of the group-sequential response-adaptive designs can reduce the ENP and the ENF compared to the group sequential CR design. More specifically, about four fewer patients on average and five fewer failures are achieved for the adaptive designs with the D_A -optimal allocation, and around eight fewer patients on average and 24 fewer failures can be achieved for those with the NP allocation.

Both optimal response-adaptive designs, the DBCD and the ERADE, can target the optimal allocations well. The use of the ERADE with the D_A -optimal allocation consistently yields the lowest standard deviations for the allocation proportions compared to the other designs. Similar conclusions can be drawn for the fixed-sample designs. Compared to these, the group sequential designs can require 52-61 fewer patients on average and prevent around 40 failures while attaining similar error probabilities.

5 Discussion

The above simulation results show that the group-sequential response-adaptive designs with dropping of inferior treatments can well control the overall type I error rate. More precisely, the probability of falsely rejecting one or more pairwise null hypotheses when the parameters are all equal is less than or equal to 5%. In addition, the combined approach can achieve a higher or similar power while using fewer patients compared to the group sequential CR design. Furthermore, fewer failures are obtained for the adaptive designs. It is concluded that the combined approach can be more ethical in terms of reducing the total sample size and the total number of failures in a trial, since early stopping for efficacy and futility and dropping of inferior arms at interim looks are allowed. Both optimal response-adaptive designs can target the specified optimal allocations well, with the ERADE consistently having a lower variability in the allocation proportions than the DBCD. Comparing the two optimal allocations derived based on different optimality criteria, in general, the adaptive designs with the D_A -optimal allocation have a lower variance for the allocation proportions, whereas the NP allocation can achieve a good power while minimising the average number of patients. In addition to the two optimal allocations considered in this paper, one can choose other treatment allocations based on different optimality criteria in response-adaptive randomisation.

Our previous study of the redesigning of the four-armed NeoSphere trial using the combined approach but without dropping of treatments in Liu and Coad¹⁴ showed that the type I error rate was slightly inflated. This was because the critical boundaries under the assumption of equal variances for all treatments were used as an approximation. However, heterogeneity increases when the treatments have unequal variances, which could result in a higher probability of rejecting

Table 10. Simulated powers for redesigning neoSphere trial using complete randomisation and response-adaptive randomisation with dropping of inferior treatments, $p_C = 0.29$, $p_{E1} = 0.458$, $p_{E2} = 0.168$, $p_{E3} = 0.24$, $N = 417$. $(t_1, t_2, t_3) = (0.33, 0.67, 1)$

Procedure	power ^I	power ^{II}	ENP	(s.d.)	ENF	(s.d.)	$\bar{p}_C$	(s.d.)	$\bar{p}_{E1}$	(s.d.)	$\bar{p}_{E2}$	(s.d.)	$\bar{p}_{E3}$	(s.d.)
CR	0.940	0.007	364.0	(63.7)	257.9	(45.3)	0.255	(0.025)	0.255	(0.024)	0.238	(0.034)	0.252	(0.026)
DBCD _{DA}	0.946	0.007	360.9	(66.2)	253.8	(46.7)	0.261	(0.020)	0.272	(0.018)	0.218	(0.037)	0.249	(0.023)
ERADE _{DA}	0.945	0.006	361.9	(65.8)	254.6	(46.4)	0.261	(0.019)	0.272	(0.016)	0.218	(0.036)	0.250	(0.021)
DBCD _{NP}	0.944	0.003	359.8	(68.0)	243.3	(46.0)	0.207	(0.022)	0.400	(0.023)	0.192	(0.023)	0.201	(0.015)
ERADE _{NP}	0.949	0.003	359.9	(68.1)	243.4	(46.1)	0.208	(0.023)	0.399	(0.022)	0.192	(0.022)	0.201	(0.015)

Fixed-sample design

 $(t_1, t_2, t_3) = (0.33, 0.67, 1)$

CR	0.942	0	417	(0)	296.5	(2.1)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD _{DA}	0.948	0	417	(0)	294.3	(1.2)	0.256	(0.011)	0.267	(0.011)	0.230	(0.015)	0.248	(0.012)
ERADE _{DA}	0.948	0	417	(0)	294.3	(1.0)	0.256	(0.009)	0.266	(0.008)	0.230	(0.013)	0.248	(0.010)
DBCD _{NP}	0.956	0	417	(0)	282.7	(1.5)	0.204	(0.018)	0.396	(0.018)	0.200	(0.009)	0.200	(0.010)
ERADE _{NP}	0.955	0	417	(0)	282.7	(1.2)	0.205	(0.019)	0.395	(0.017)	0.200	(0.007)	0.201	(0.010)

• The target D_A -optimal allocation ($p_C, p_{E1}, p_{E2}, p_{E3}$) is (0.256, 0.266, 0.230, 0.248) and the NP optimal allocation is (0.2, 0.4, 0.2, 0.2) if no arm has been dropped.

• If E2 is dropped, the optimal allocation becomes (0.332, 0.345, 0, 0.322) for the D_A -optimal allocation and (0.228, 0.545, 0, 0.228) for the NP allocation.

• If E3 is dropped, the optimal allocation becomes (0.340, 0.354, 0.306, 0) for the D_A -optimal allocation and (0.250, 0.599, 0.151, 0) for the NP allocation.

• If both E2 and E3 are dropped, the optimal allocation becomes (0.490, 0.510, 0, 0) for the D_A -optimal allocation and (0.295, 0.705, 0, 0) for the NP allocation.

the global null hypothesis. Nevertheless, if the global null hypothesis is rejected, pairwise comparisons are conducted subsequently in the Fisher's LSD method presented in this paper. The critical boundaries for the pairwise Z tests can be applied for different variances and unequal numbers of patients on the treatment arms, since the sequence of test statistics still asymptotically has the canonical joint distribution.⁶ Therefore, although a false rejection of the global null hypothesis could be made in the first stage where the critical boundaries derived under the assumption of equal variances were used, this would just lead to the commencement of pairwise comparisons, and the error probabilities for Fisher's LSD method are based on the subsequent pairwise tests.

Our simulation studies on normal, binary and censored survival responses show that the extension of Fisher's LSD method to group-sequential response-adaptive designs can be applied to different responses while preserving good operating characteristics. In practice, considerations on whether or not to incorporate response-adaptive randomisation (RAR) in a group sequential clinical trial and to strongly or weakly control the family-wise error rate include, but are not limited to, (a) the number of treatments being compared and the heterogeneity of the treatment effects, (b) the length of time to obtain the intermediate/surrogate outcome measurements, and (c) the phase and aim of the clinical trial. RAR design is worth applying when more than two treatments are compared with heterogeneous treatment effects, and rapidly accumulating outcomes are available. Although adaptive designs can be applied across different phases of a clinical trial, from an early-phase dose escalation study to a late-phase confirmatory trial, the latter usually requires strong control of the overall type I error rate. Conversely, a relaxed type I error rate may be used for the former, since the main objective is usually to quickly identify promising treatments with tolerable toxicities and some sign of efficacy. Whether or not to control the family-wise error rate has been widely discussed.³¹ The main argument for not adjusting for multiplicity is that, if the same comparisons were made in separate trials, control of the family-wise error rate would not be required. In the second stage of our proposed group-sequential response-adaptive Fisher's LSD method, unadjusted two-sample pairwise comparison is considered for each hypothesis test comparing any experimental treatments remaining in the trial with the control.

Trial design is usually considered on a case-by-case basis and carefully planned before the start of the trial. The proposed combined approach possesses the efficient and ethical advantages of adaptive designs. For instance, the interim analyses can take place at any continuous information time. Inferior treatments can be dropped at an interim analysis while the information time continues. Adding new arms to an ongoing clinical trial has practical advantages, but is even more complicated and requires a rigorous plan. Such a development is beyond the scope of this paper. In brief, previous work has described how the family-wise type I error rate, and any-pair and all-pairs powers can be calculated when adding new

arms to a platform trial using an extension of Dunnett's method.³² The authors suggested the following strategies. If the new research arm is deemed to be very different to the arms already in the trial, then no strong control of the family-wise type I error rate is required. One can just control the pairwise type I error rate and power each research arm separately. On the other hand, if the research arms are similar, such as different dosages of a drug, then one needs to adjust the error rate based on the total number of pairwise comparisons, the allocation ratio for the new or ongoing comparisons and the overlapping shared control information time of the new research arm versus the other ones, that is, the correlation between the test statistics for the pairwise comparisons.

Acknowledgements

This work was carried out whilst the first author was in receipt of scholarship from the Ministry of Education in Taiwan for her Ph.D. study at Queen Mary, University of London. The authors also wish to thank the Editor and two referees for their comments, which have led to an improved paper.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Wenyu Liu  <https://orcid.org/0000-0002-6682-2053>

References

- Lewis FI. An introduction to group sequential methods: planning and multi-aspect optimization. arXiv 2023. <http://arxiv.org/abs/2303.01040>.
- Jennison C and Turnbull BW. Group sequential tests with outcome-dependent treatment assignment. *Seq Anal* 2001; **20**: 209–234.
- Morgan CC and Coad DS. A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Stat Med* 2007; **26**: 1937–1954.
- Ivanova A. A play-the-winner-type urn design with reduced variability. *Metrika* 2003; **58**: 1–13.
- Zhu H and Hu F. Sequential monitoring of response-adaptive randomized clinical trials. *Ann Stat* 2010; **38**: 2218–2241.
- Jennison C and Turnbull BW. *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman and Hall/CRC Press, 2000.
- Liu W and Coad DS. Group-sequential response-adaptive designs for censored survival outcomes. *J Stat Plan Inference* 2020; **205**: 293–305.
- Magirr D, Jaki T and Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**: 494–501.
- Wason J, Stallard N, Bowden J, et al. A multi-stage drop-the-losers design for multi-arm clinical trials. *Stat Methods Med Res* 2017; **26**: 508–524.
- Bratton DJ, Parmar MKB, Phillips PPJ, et al. Type I error rates of multi-arm multi-stage clinical trials: strong control and impact of intermediate outcomes. *Trials* 2016; **17**: 309.
- Jaki T, Pallmann P and Magirr D. The R package MAMS for designing multi-arm multi-stage clinical trials. *J Stat Softw* 2019; **88**: 1–25.
- Jennison C and Turnbull BW. Exact calculations for sequential t, X² and F tests. *Biometrika* 1991; **78**: 133–141.
- Proschan MA, Follmann DA and Geller NL. Monitoring multi-armed trials. *Stat Med* 1994; **13**: 1441–1452.
- Liu W and Coad DS. Group-sequential response-adaptive designs for multi-armed trials. *Seq Anal* 2023; **42**: 112–128.
- Christensen R. *Plane answers to complex questions: the theory of linear models, 3rd edn*. New York, NY: Springer, 2013.
- Hayter AJ. The maximum familywise error rate of Fisher's least significant difference test. *J Am Stat Assoc* 1986; **81**: 1000–1004.
- Follmann DA, Proschan MA and Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**: 325–336.
- Ventz S, Cellamare M, Parmigiani G, et al. Adding experimental arms to platform clinical trials: randomization procedures and interim analyses. *Biostatistics* 2018; **19**: 199–215.
- Lan KKG and DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**: 659–663.
- Hu F and Zhang L-X. Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann Stat* 2004; **32**: 268–301.
- Kim K, Boucher H and Tsiatis AA. Design and analysis of group sequential logrank tests in maximum duration versus information trials. *Biometrics* 1995; **51**: 988–1000.
- Wong WK and Zhu W. Optimum treatment allocation rules under a variance heterogeneity model. *Stat Med* 2008; **27**: 4581–4595.

23. Sverdlov O, Tymofyeyev Y and Wong WK. Optimal response-adaptive randomized designs for multi-armed survival trials. *Stat Med* 2011; **30**: 2890–2910.
24. Tymofyeyev Y, Rosenberger WF and Hu F. Implementing optimal allocation in sequential binary response experiments. *J Am Stat Assoc* 2007; **102**: 224–234.
25. Rosenberger WF, Stallard N, Ivanova A, et al. Optimal adaptive designs for binary response trials. *Biometrics* 2001; **57**: 909–913.
26. Eisele JR and Woodroffe MB. Central limit theorems for doubly adaptive biased coin designs. *Ann Stat* 1995; **23**: 234–254.
27. Zhang L-X. Response-adaptive randomization: an overview of designs and asymptotic theory. arXiv 2014. <http://arxiv.org/abs/1412.1553>.
28. Proschan MA, Lan KKG and Wittes JT. *Statistical monitoring of clinical trials: a unified approach*. New York, NY: Springer, 2010.
29. Fountzilas G, Ciuleanu E, Dafni U, et al. Concomitant radiochemotherapy vs radiotherapy alone in patients with head and neck cancer: a Hellenic Cooperative Oncology Group Phase III Study. *Med Oncol* 2004; **21**: 95–107.
30. Gianni L, Pienkowski T, Im Y-H, et al. 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. *Lancet Oncol* 2012; **13**: 25–32.
31. Wason JMS, Stecher L and Mander AP. Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* 2014; **15**: 364.
32. Choodari-Oskooei B, Bratton DJ, Gannon MR, et al. Adding new experimental arms to randomised clinical trials: impact on error rates. *Clin Trials* 2020; **17**: 273–284.

Appendix

AI. Response-adaptive randomisation methods

Doubly-adaptive biased coin design (DBCD)

Suppose that $m_j^{(i)}$ is the cumulative sample size on treatment j after i patients, $i = 1, \dots, N$. Let $m_j^{(i)}/i$ and $\hat{\rho}_j^{(i)}$ be the current and target allocation proportions for treatment j , $j = 1, \dots, J$, based on the treatment assignments and responses obtained so far. Then the probability that the next patient will be assigned to treatment j is given by

$$g_j = \begin{cases} \frac{\hat{\rho}_j^{(i)} \left\{ \frac{\hat{\rho}_j^{(i)}}{m_j^{(i)}/i} \right\}^\gamma}{\sum_{l=1}^J \hat{\rho}_l^{(i)} \left\{ \frac{\hat{\rho}_l^{(i)}}{m_l^{(i)}/i} \right\}^\gamma} & \text{if } 0 < m_j^{(i)}/i < 1, \\ 1 - m_j^{(i)}/i & \text{if } m_j^{(i)}/i = 0, 1 \end{cases}$$

where $\gamma \in [0, \infty)$ is a tuning parameter that controls the degree of randomness.

Efficient randomised-adaptive design (ERADE)

Let

$$\psi(x) = 1 + \sqrt{(x^{2\gamma'} - 1) \vee 0}$$

be a weight function. Here, $a \vee b = \max(a, b)$. The probability that the next patient will be assigned to treatment j is given by

$$g_j = \begin{cases} \frac{\hat{\rho}_j^{(i)} \psi\left(\frac{\hat{\rho}_j^{(i)}}{m_j^{(i)}/i}\right)}{\sum_{l=1}^J \hat{\rho}_l^{(i)} \psi\left(\frac{\hat{\rho}_l^{(i)}}{m_l^{(i)}/i}\right)} & \text{if } 0 < m_j^{(i)}/i < 1, \\ 1 - m_j^{(i)}/i & \text{if } m_j^{(i)}/i = 0, 1 \end{cases}$$

where the tuning parameter γ' which determines the randomness in the allocation can be any positive number.

Table A.1. Simulated type I error rates for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = \mu_{E2} = \mu_C = 1$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 90$.

Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.045	0.003	89.9	(2.0)	0.333	(0.048)	0.334	(0.048)	0.333	(0.048)
DBCD _{D_A}	0.052	0.003	89.8	(2.6)	0.450	(0.029)	0.355	(0.038)	0.194	(0.039)
ERADE _{D_A}	0.051	0.003	89.7	(3.0)	0.447	(0.027)	0.354	(0.035)	0.199	(0.036)

Table A.2. Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 3$, $\mu_{E2} = 2$, $\mu_C = 1$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 90$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.833	0.278	78.7	(14.8)	0.333	(0.051)	0.334	(0.051)	0.333	(0.051)
DBCD _{D_A}	0.865	0.329	76.0	(15.8)	0.451	(0.031)	0.360	(0.040)	0.190	(0.040)
ERADE _{D_A}	0.874	0.331	76.0	(15.8)	0.448	(0.026)	0.358	(0.036)	0.194	(0.038)

Table A.3. Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 1$, $\mu_{E2} = 3$, $\mu_C = 2$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 90$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.645	0.115	85.9	(10.3)	0.326	(0.055)	0.338	(0.050)	0.336	(0.050)
DBCD _{D_A}	0.675	0.117	85.5	(10.8)	0.430	(0.048)	0.356	(0.040)	0.214	(0.033)
ERADE _{D_A}	0.679	0.117	85.5	(10.8)	0.429	(0.046)	0.355	(0.038)	0.216	(0.032)

Table A.4. Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $\mu_{E1} = 2$, $\mu_{E2} = 1$, $\mu_C = 3$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$, $\sigma_C = 1$, $N = 90$.

Procedure	power ^I	power ^{II}	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.989	0.269	88.0	(7.6)	0.372	(0.055)	0.255	(0.068)	0.374	(0.057)
DBCD _{D_A}	0.976	0.288	88.1	(7.5)	0.482	(0.056)	0.251	(0.077)	0.268	(0.043)
ERADE _{D_A}	0.977	0.280	88.1	(7.5)	0.482	(0.054)	0.251	(0.075)	0.268	(0.041)

A2. More simulation scenarios

A2.1. Three-armed normal trials

Similar to Section 4.1, here some scenarios with a smaller maximum sample size $N = 90$ are considered. Under the null hypothesis, $\mu_{E1} = \mu_{E2} = \mu_C = 1$, $\sigma_{E1} = 4$, $\sigma_{E2} = 2$ and $\sigma_C = 1$. There are three scenarios for the alternative hypothesis: (i) both experimental treatments are better than the control, (ii) one experimental treatment is better and the other is worse than the control, and (iii) both experimental treatments are worse than the control. The target D_A -optimal allocations are the same as before.

Under the null hypothesis, the overall type I error rate in Table A1 is still well-controlled, even though the maximum sample size is reduced. However, the variability of the treatment allocation has increased. Compared to the results in Section 4.1, the powers under the alternative hypotheses in Tables A2 to A4 are slightly higher here, probably due to the greater contrasts between the means for the experimental arms and the control, together with the increased variability in the treatment allocation.

A2.2. Four-armed binary trials

An example of a scenario for four-armed binary trials that is prone to weak control of the type I error rate is now investigated. Here, one of the experimental arms is substantially inferior to the control, but the remaining two arms are very similar to the control. Assume that the probabilities of success for the control and experimental treatments are $p_C = 0.4$, $p_{E1} = 0.1$, $p_{E2} = 0.41$ and $p_{E3} = 0.42$. Given the nominal type I error rate of 0.05, to achieve around 85% power^I, that is, the probability of rejecting at least one of the pairwise null hypotheses, a maximum sample size $N = 144$ is planned. Group sequential analyses at information times $(t_1, t_2, t_3) = (0.33, 0.67, 1)$ that allow dropping inferior arms were considered. Type

I error rates and powers obtained from the simulation results in such a scenario comparing the use of CR and the DBCD and ERADE with the two optimal allocations based on 10,000 replicates are shown in Table A5.

Clearly, the response-adaptive randomisation methods with D_A -optimal allocation had a slightly inflated type I error rate with around 6% higher power. On the other hand, the response-adaptive randomisation methods with NP allocation when the lower bound for the allocation proportions B is set to be 0.2 had the lowest type I error rate. However, the standard deviation of the allocation proportion $\hat{p}$ for each treatment is the largest for NP allocation. Note that the NP allocation rules are different depending on the values of B and the parameters p , or their estimates $\hat{p}$. The target NP allocation* shown was calculated based on $p_C = 0.4$, $p_{E1} = 0.1$, $p_{E2} = 0.41$ and $p_{E3} = 0.42$. On the other hand, the actual allocation $\hat{p}$ for each treatment in the simulation was calculated based on $\hat{p}$, which might be using a different allocation rule depending on whether or not $B \in (0, \tilde{B})$, where $\tilde{B}$ is a function of the parameter estimates: see the allocation rules described in Tymofyeyev et al.²⁴ The NP allocation has the greatest variation in the allocation proportions, since the actual allocation $\hat{p}$ for each treatment jumped around different values based on different allocation rules.

Table A.5. Simulated type I error rates when $p_C = 0.4$, $p_{E1} = 0.4$, $p_{E2} = 0.4$, $p_{E3} = 0.4$ and powers when $p_C = 0.4$, $p_{E1} = 0.1$, $p_{E2} = 0.41$, $p_{E3} = 0.42$ for four-armed binary trials with $N = 144$ and dropping of inferior treatments.

$p_C = 0.4, p_{E1} = 0.4, p_{E2} = 0.4, p_{E3} = 0.4$														
Procedure	$\hat{\alpha}^I$	$\hat{\alpha}^{II}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{p}_C$	(s.d.)	$\tilde{p}_{E1}$	(s.d.)	$\tilde{p}_{E2}$	(s.d.)	$\tilde{p}_{E3}$	(s.d.)
CR	0.047	0.003	143.6	(4.3)	86.2	(2.6)	0.250	(0.034)	0.250	(0.034)	0.250	(0.034)	0.250	(0.034)
DBCD $_{D_A}$	0.053	0.005	143.5	(4.7)	86.1	(2.8)	0.250	(0.018)	0.250	(0.018)	0.250	(0.018)	0.250	(0.018)
ERADE $_{D_A}$	0.055	0.005	143.4	(5.7)	86.0	(3.4)	0.250	(0.016)	0.250	(0.017)	0.250	(0.016)	0.250	(0.017)
DBCD $_{NP}$	0.039	0.002	143.8	(3.3)	86.3	(2.0)	0.251	(0.074)	0.249	(0.071)	0.251	(0.072)	0.250	(0.072)
ERADE $_{NP}$	0.040	0.002	143.7	(3.7)	86.2	(2.2)	0.249	(0.073)	0.250	(0.073)	0.251	(0.073)	0.250	(0.073)
$p_C = 0.4, p_{E1} = 0.1, p_{E2} = 0.41, p_{E3} = 0.42$														
CR	0.854	0.004	138.0	(9.3)	91.0	(7.4)	0.260	(0.038)	0.221	(0.046)	0.260	(0.037)	0.259	(0.037)
DBCD $_{D_A}$	0.868	0.004	139.4	(8.1)	89.9	(6.3)	0.274	(0.024)	0.177	(0.048)	0.274	(0.025)	0.275	(0.024)
ERADE $_{D_A}$	0.860	0.004	139.5	(8.3)	90.1	(6.4)	0.273	(0.023)	0.179	(0.047)	0.273	(0.024)	0.274	(0.023)
DBCD $_{NP}$	0.807	0.002	135.9	(12.1)	92.7	(9.1)	0.237	(0.056)	0.296	(0.076)	0.246	(0.061)	0.221	(0.040)
ERADE $_{NP}$	0.808	0.003	135.3	(12.4)	92.3	(9.3)	0.237	(0.056)	0.295	(0.075)	0.246	(0.060)	0.222	(0.041)

• If no arm has been dropped, the target D_A -optimal allocation $(p_C, p_{E1}, p_{E2}, p_{E3})$ is (0.266, 0.200, 0.267, 0.267) and the NP optimal allocation* is (0.4, 0.2, 0.2, 0.2).

• If $E1$ is dropped, the optimal allocation becomes (0.333, 0, 0.333, 0.334) for the D_A -optimal allocation and (0.5, 0, 0.25, 0.25) for the NP allocation*.

• If $E2$ is dropped, the optimal allocation becomes (0.363, 0.272, 0, 0.364) for the D_A -optimal allocation and (0.5, 0.25, 0, 0.25) for the NP allocation*.

• If $E3$ is dropped, the optimal allocation becomes (0.364, 0.272, 0.364, 0) for the D_A -optimal allocation and (0.5, 0.25, 0.25, 0) for the NP allocation*.

• If both $E1$ and $E2$ are dropped, the optimal allocation becomes (0.499, 0, 0, 0.501) for the D_A -optimal allocation and (0.67, 0, 0, 0.33) for the NP allocation*.

• If both $E1$ and $E3$ are dropped, the optimal allocation becomes (0.5, 0, 0.5, 0) for the D_A -optimal allocation and (0.67, 0, 0.33, 0) for the NP allocation*.

• If both $E2$ and $E3$ are dropped, the optimal allocation becomes (0.572, 0.428, 0, 0) for the D_A -optimal allocation and (0.67, 0.33, 0, 0) for the NP allocation*.