Group-sequential response-adaptive designs for multi-armed trials

Wenyu Liu[a] and D. Stephen Coad[b]

[a] *Institute of Cancer Research and Genomic Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, U.K.*

[b] *School of Mathematical Sciences, Queen Mary, University of London, Mile End Road, London E1 4NS, U.K.*

Abstract

Several experimental treatments are often compared with a common control in a clinical trial nowadays. A group sequential design incorporating response-adaptive randomisation can help to increase the probability of receiving a more promising treatment for patients in the trial and to detect a treatment effect early so as to benefit the whole population of interest. With such ethical advantages, the trial design has invoked investigation using the Bayesian approach. In the frequentist approach, the type I error rate of a multi-armed trial may involve two error elements, the inflated error rates caused by multiple treatment comparisons and sequential testing. In this study, a group sequential global test was considered. By monitoring the response-adaptive design at a continuous information time, calculation of the information time and two optimal response-adaptive sampling rules for multi-armed trials were described. Operating characteristics of the designs were investigated via simulation for censored exponential survival outcomes and using patient data sampled from a four-armed binary trial to demonstrate their practical applicability. Our results showed that, in general, the adaptive designs preserved ethical advantages in terms of reducing the average numbers of patients and

failures compared with a group-sequential non-adaptive randomised design, while not adversely affecting the power.

Keywords: binary outcome, censored survival outcome, global test, multiple treatment comparison, optimal allocation, power

1 Introduction

Comparing several treatments in a multi-armed trial is considered to be more efficient in terms of time, resource and sample size than conducting several conventional two-armed trials, each comparing an experimental arm with the control. Multiple treatment comparison can be conducted using either adjusted pairwise comparisons or global testing. For group sequential pairwise comparisons, one can simply use the Bonferroni adjustment to control the overall type I error rate (Follmann et al., 1994; Jennison and Turnbull, 2000). Specifically, for p pairwise tests at each look, α_k/p is the nominal type I error rate for each pairwise test, where α_k is the type I error rate spent by interim analysis k. The Bonferroni approach strongly controls the overall type I error rate. However, it can be too conservative at the price of losing power.

Other studies considered using multi-armed multi-stage (MAMS) designs to monitor multi-armed clinical trials (Magirr et al., 2012; Wason et al., 2014). MAMS designs simultaneously evaluate several regimens against a common control. Follmann et al. (1994) generalised the Pocock, O'Brien and Fleming, and Lan and DeMets boundaries to multi-armed trials considering pairwise comparisons, where Lan and DeMets (1983) is an alpha-spending approach and one can determine the rate at which alpha is used during the course of a two-treatment comparison. This approach preserves the nominal type I error probability without the requirement of pre-specifying the number of interim analyses. A disadvantage of the alpha-spending approach for multi-armed trials is that the expected sample size properties may not be desirable (Wason et al., 2016). With efficacy and futility boundaries, the designs allow dropping of inferior treatments at interim analyses. MAMS designs focus on

designs that strictly control the family-wise error rate in the strong sense, with the probability of falsely rejecting one or more null hypotheses being less than or equal to α (Bratton et al., 2016; Jaki et al., 2019). Given a pre-specified error rate, the number of patients needed per arm per stage and the critical boundaries are obtained by numerical computation.

In group sequential monitoring, when an inferior treatment is identified at an interim look, it is more ethical to adjust the allocation probabilities to assign more patients to the more promising arm(s). Incorporating response-adaptive randomisation in multi-armed designs has been discussed using the Bayesian approach (Wason and Trippa, 2014; Ventz et al., 2018; Ryan et al., 2020). In the frequentist approach, the joint distribution of the sequential Z test statistics for two-armed trials that combine group sequential analysis with response-adaptive randomisation has been shown to have a canonical form such that critical boundaries for standard group sequential designs can be utilised as an approximation (Jennison and Turnbull, 2000; Zhu and Hu, 2010; Liu and Coad, 2020). For multi-armed trials, Jennison and Turnbull (1991, 2000) derived exact critical boundaries for standard group sequential global tests analogous to Pocock's and the O'Brien and Fleming boundaries based on multi-armed normal trials with equal variances and equal treatment allocation. Whether or not the standard critical boundaries can still be used as an approximation to control the overall type I error rate in multi-armed group-sequential response-adaptive designs is of concern in this paper. The idea is that, if the standard group sequential critical boundaries for both two-treatment and multi-treatment comparisons can be applied to response-adaptive designs while preserving good operating characteristics, then the proposed global test statistics, followed by pairwise comparisons if the global null hypothesis is rejected, can be an alternative to monitoring pairwise comparisons for multi-armed multi-stage trials, which are more computationally demanding in obtaining the stopping boundaries. Different optimal response-adaptive sampling rules for binary and censored survival outcomes will be investigated.

Consider a motivating example of a four-armed binary trial. NeoSphere (Gianni et al., 2012) is a phase II randomised trial which compares the efficacy and safety of different combinations of treatments for women with breast cancer. Antibody trastuzumab with concomitant chemotherapy docetaxel is a conventional treatment for the cancer. The NeoSphere trial examined the activity of another anibody, pertuzumab, by assessing the effects of pertuzumab combined with either trastuzumab, docetaxel or both. The trial consisted of trastuzumab plus docetaxel (control), pertuzumab and trastuzumab plus docetaxel ($E1$), pertuzumab and trastuzumab ($E2$), and pertuzumab plus docetaxel ($E3$). There were 417 eligible women randomly assigned to the treatment groups with equal probabilities. The endpoint considered in the study was pathological complete response, which was dichotomised and serves as a surrogate for long-term efficacy. The complete response rate was 29% for the control, 45.8% for $E1$, 16.8% for $E2$ and 24% for $E3$. The study concluded that $E1$ had a significantly higher complete response rate compared to the conventional control group. Redesigning the clinical trial using response-adaptive randomisation will be investigated.

The structure of the remaining sections is as follows. In Section 2, a general form of the group sequential global test statistic comparing several treatments with a control and its sequential critical boundaries are described. In Section 3, two optimal allocations for multi-armed clinical trials are introduced. One ensures the most precise estimate of the parameter vector. The other maximises the power subject to a constraint on the total sample size or a function of the sample sizes. Then optimal response-adaptive randomisation procedures, which aim to target the pre-specified optimal allocations, are described. Results of the redesign of the motivating clinical trial and simulation for censored exponential survival outcomes are presented in Section 4, including the error probabilities, the expected number of patients, the expected number of failures and the average allocation proportion with its variability. Conclusions and further work are in Section 5.

2 Form of test

2.1 Information time

Let N be the maximum number of patients for a trial with J arms and let K be the number of group sequential analyses. For multi-armed trials with immediate responses, the information time at look k, t_k, is proportional to the number of patients obtained so far.

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

where $m_{j,k}$ is the cumulative number of patients for treatment j, $j = 1, ..., J$, at look k, $m_{j,K} = M_j$, and $n_k = \sum_{j=1}^{J} m_{j,k}$ is the cumulative sample size at look k, $n_K = N$.

For survival responses, the information time is proportional to the number of events. As described in Kim et al. (1995), and Liu and Coad (2020), it can be expressed as

$$t_k = \frac{\sum_{j=1}^{J} m_{j,k}\hat{\epsilon}_k}{\sum_{j=1}^{J} M_j\hat{\epsilon}_K} = \frac{n_k\hat{\epsilon}_k}{N\hat{\epsilon}_K} \in (0, 1], \quad k = 1, ..., K,$$

where ϵ_k is the probability of an event, which can be estimated using empirical or model-based methods.

2.2 Group sequential global test statistic

Consider testing a vector of J treatment contrasts. The global null hypothesis is $H_{G_0} : \boldsymbol{\theta}_G = 0$ versus $H_{G_a} : \boldsymbol{\theta}_G \neq 0$, where $\boldsymbol{\theta}_G = (\theta_1 - \theta_J,\ \theta_2 - \theta_J, ...,\ \theta_{J-1} - \theta_J)^T$ considering arm J as the control arm. The test statistic at group sequential test k is

$$S_k = \hat{\boldsymbol{\theta}}_{G_k}^T\ \hat{\Sigma}_k^{-1}\ \hat{\boldsymbol{\theta}}_{G_k}, \quad k = 1, ..., K,$$

where $\hat{\boldsymbol{\theta}}_{G_k}$ is the maximum likelihood estimator of $\boldsymbol{\theta}_G$ based on the responses obtained so far and $\hat{\Sigma}_k^{-1}$ is the corresponding estimated covariance matrix.

For binary outcomes, $\boldsymbol{\theta}_G$ is the vector of treatment contrasts of the probabilities of success $p_j,\ j = 1, ..., J$, and

$$\hat{\Sigma}_k = \begin{pmatrix} \frac{\hat{p}_{1,k}\hat{q}_{1,k}}{m_{1,k}} & 0 & ... & 0 \\ 0 & \frac{\hat{p}_{2,k}\hat{q}_{2,k}}{m_{2,k}} & & 0 \\ & & \ddots & \\ 0 & 0 & ... & \frac{\hat{p}_{J-1,k}\hat{q}_{J-1,k}}{m_{J-1,k}} \end{pmatrix} + \frac{\hat{p}_{J,k}\hat{q}_{J,k}}{m_{J,k}}\mathbf{1}\mathbf{1}^T,$$

where $\hat{p}_{j,k}$ is the maximum likelihood estimate of the probability of success for treatment j at look k and $\hat{q}_{j,k} = 1 - \hat{p}_{j,k}$. Here, $\mathbf{1} = (1, ..., 1)^T$ is the vector with $J - 1$ ones.

For censored exponential survival outcomes, $\boldsymbol{\theta}_G$ is the vector of treatment contrasts of the survival means $\theta_j,\ j = 1, ..., J$. The maximum likelihood estimate of the mean survival time for treatment j evaluated at look k is $\hat{\theta}_{j,k} = \sum_{i=1}^{m_{j,k}} y_{i,j,k}/r_{j,k}$ and $\text{var}(\hat{\theta}_{j,k}) = \theta_j^2/E(r_{j,k})$, where $y_{i,j,k}$ is the survival time for patient $i,\ i = 1, ..., m_{j,k}$, on treatment j at interim look k and $r_{j,k}$ is the cumulative number of events on treatment j at look k, then we obtain

$$\hat{\Sigma}_k = \begin{pmatrix} \frac{\hat{\theta}_{1,k}^2}{r_{1,k}} & 0 & ... & 0 \\ 0 & \frac{\hat{\theta}_{2,k}^2}{r_{2,k}} & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \frac{\hat{\theta}_{J-1,k}^2}{r_{J-1,k}} \end{pmatrix} + \frac{\hat{\theta}_{J,k}^2}{r_{J,k}}\mathbf{1}\mathbf{1}^T.$$

The matrix $\hat{\Sigma}_k$ is nonsingular and its inverse exists. Under H_{G_0}, the marginal distribution of S_k is asymptotically χ^2_{J-1}, since S_k is a quadratic form of asymptotically normal variables. Under H_{G_a}, the distribution is asymptotically noncentral chi-squared with noncentrality parameter

$$\eta_k = \boldsymbol{\theta}_G^T \, \Sigma_k^{-1} \, \boldsymbol{\theta}_G = \sum_{j=1}^{J-1} \frac{m_{j,k}}{p_j q_j}(p_j - p_J)^2 - \frac{1}{\sum_{j=1}^{J} \frac{m_{j,k}}{p_j q_j}} \left\{ \sum_{j=1}^{J-1} \frac{m_{j,k}}{p_j q_j}(p_j - p_J) \right\}^2$$

for binary outcomes and

$$\eta_k = \sum_{j=1}^{J-1} \frac{r_{j,k}}{\theta_j^2}(\theta_j - \theta_J)^2 - \frac{1}{\sum_{j=1}^{J} \frac{r_{j,k}}{\theta_j^2}} \left\{ \sum_{j=1}^{J-1} \frac{r_{j,k}}{\theta_j^2}(\theta_j - \theta_J) \right\}^2$$

for censored exponential survival outcomes. The function η_k is concave and $\partial \eta_k / \partial r_{j,k} \geq 0$ in the second case. In other words, when the cumulative number of events on any treatment arm, $r_{j,k}$, $j = 1, ..., J$, increases, the noncentrality parameter η_k is increased. Note that the forms are similar to those considered in fixed-sample designs (Tymofyeyev et al., 2007; Sverdlov et al., 2011).

2.3 Stopping boundaries

For standard group-sequential non-adaptive designs, Jennison and Turnbull (1991) showed that the sequence of test statistics is Markov. More specifically, the probability distribution of S_{k+1} depends only on S_k and not on $\{S_1, ..., S_{k-1}\}$. The joint distribution of $\{S_1, ..., S_{k+1}\}$ can be constructed recursively by multiplying the conditional distributions of S_{k+1} given S_k for $k \geq 1$. Then the critical boundaries d_k can be obtained recursively based on the multivariate joint distribution. The boundaries analogous to Pocock's and the O'Brien and Fleming boundaries for sequential t, χ^2 and F tests can be found in Jennison and Turnbull (1991). Based on the significance level approach, the critical boundaries can be used to give an approximate test for adaptive randomised trials, as long as the imbalance in the sample sizes is not too severe (Jennison and Turnbull, 1991, 2000). The stopping rules are as follows. For $k = 1, ..., K-1$, we reject H_{G_0} if $S_k \geq d_k$; otherwise, we continue to the next look. For $k = K$, we reject H_{G_0} if $S_k \geq d_k$ and accept H_{G_0} otherwise. Pairwise comparisons can be conducted after reject-

ing the global null hypothesis to investigate which experimental treatments have different efficacies to the control. Conventional critical boundaries for two-armed designs can be used.

3 Optimal response-adaptive randomisation

Two optimal allocations for multi-armed trials are considered. One ensures the most precise estimate of the parameter vector and the other maximises the power subject to a fixed total sample size. The two optimal allocations reduce to Neyman allocation when $J = 2$. However, for $J \geq 3$, they have different characteristics. The optimal allocations were derived based on a fixed-sample design. However, they can be used in group sequential designs, since the target optimal allocation would not be affected by the number of interim tests.

3.1 Optimal allocations

3.1.1 D_A-optimal allocation

Suppose that the parameter of interest is $A^T\boldsymbol{\beta} = (\beta_1 - \beta_J, ..., \beta_{J-1} - \beta_J)^T$, where A^T is a $(J-1) \times J$ matrix. Then the D_A-optimal allocation (Wong and Zhu, 2008) ensures the most precise estimate of $A^T\boldsymbol{\beta}$ by minimising the determinant of $\text{cov}(A^T\hat{\boldsymbol{\beta}}) = A^T M^{-1}(\xi)A$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$, over all possible randomisation designs ξ, or, equivalently, minimising $\Phi(\xi) = \log|A^T M^{-1}(\xi)A|$. The D_A-optimal allocation yields the smallest confidence ellipsoid for $A^T\boldsymbol{\beta}$. In practice, one can use the general equivalence theorem (Kiefer and Wolfowitz, 1960) to obtain the D_A-optimal allocation by solving the system of equations

$$d_A(x_j, \xi^*) = J - 1, \quad j = 1, ..., J,$$

where $d_A(x_j, \xi^*)$ is the directional derivative of the criterion $\Phi(\xi^*)$, x_j is the indicator for treatment j, ξ^* is the optimum value of ξ and $J-1$ is the rank of A.

For multi-armed binary trials, the D_A-optimal allocation, where ρ_j is the allocation proportion for treatment j and $0 \leq \rho_j \leq 1/(J-1)$, can be derived by solving

$$d_A(x_j, \xi^*) = \frac{1}{\rho_j} - \frac{1/(p_j q_j)}{\sum_{l=1}^{J} 1/(p_l q_l)\rho_l} = J - 1, \quad j = 1, ..., J, \tag{1}$$

where $1/(p_j q_j)$ is inversely proportional to the variance for treatment j.

Similarly, for exponentially distributed survival outcomes, the D_A-optimal allocation can be obtained by solving the equations

$$d_A(x_j, \xi^*) = \frac{1}{\rho_j} - \frac{\epsilon_j/\theta_j^2}{\sum_{l=1}^{J} (\epsilon_l/\theta_l^2)\rho_l} = J - 1, \quad j = 1, ..., J, \tag{2}$$

where θ_j is the mean survival time for treatment j, ϵ_j is the probability of an event on arm j and ϵ_j/θ_j^2 is inversely proportional to the variance for arm j. The D_A-optimal design consistently allocates more patients to the treatments that have larger variances for the responses. For exponentially distributed survival responses, the variance of the responses on arm j is $\theta_j^2/E(r_j)$. In this case, the variance increases when the mean survival time for treatment j is increased. Therefore, for exponential survival responses, the D_A-optimal allocation is always ethical. However, for normal and binary responses, the most efficient design may assign more patients to the inferior treatments.

3.1.2 Optimal allocation based on nonlinear programming

To find a design that maximises the power of a test, one can consider maximising the noncentrality parameter η, since the power increases as η is increased. Tymofyeyev et al. (2007) investigated the optimality rule which maximises the noncentrality parameter such that

$$\sum_{j=1}^{J} v_j M_j \leq C \quad \text{and} \quad \frac{M_j}{N} \geq B \quad \text{for } j = 1, ..., J,$$

where $(v_1, ..., v_J)$ is a vector of some positive weights, M_j is the sample size for treatment j, $\sum_{j=1}^{J} M_j = N$ and the lower bound $B \in [0, 1/J]$. When $(v_1, ..., v_J) = (1, ..., 1)$ and $B = 0$, the solution maximises the power subject to the constraint that the total sample size does not exceed a fixed value, which is an analogue of Neyman allocation. When $(v_1, ..., v_J) = (q_1, ..., q_J)$, where q_j is the failure probability for treatment j, the derived optimal allocation minimises the expected number of failures for a fixed power, which is an analogue of the optimal allocation derived by Rosenberger et al. (2001) generalised to $J \geq 3$ treatments. A general solution $(\rho_1, ..., \rho_J)$ for any vector of weights does not exist, and numerical methods are required. The solution in the case of $(v_1, \ ..., v_J) = (1, ..., 1)$ is given below.

Let $p_1 = ... = p_s > p_{s+1} \geq ... \geq p_{J-g} > p_{J-g+1} = ... = p_J$ for some positive integers s and g. When $B \in [0, \tilde{B}]$, $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_J, 1/J)$, the solution $(\rho_1, ..., \rho_J)$ is obtained by

$$\begin{aligned} &\rho_1 = ... = \rho_s = \frac{1}{s}\left(QB + \frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_J q_J}}\right), \\ &\rho_{s+1} = ... = \rho_{J-g} = B, \\ &\rho_{J-g+1} = ... = \rho_J = \frac{1}{g}\{1 - B(K - s - g) - s\rho_1\}, \end{aligned} \tag{3}$$

where

$$\begin{aligned} \tilde{B}_1 &= \frac{1}{s - Q}\frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_J q_J}}, \\ \tilde{B}_J &= \frac{1}{J + Q - s}\frac{\sqrt{p_J q_J}}{\sqrt{p_1 q_1} + \sqrt{p_J q_J}} \end{aligned}$$

and

$$\begin{aligned} Q = &\frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_J q_J}}\left\{\sum_{j=s+1}^{J-g}\frac{p_J q_J}{p_j q_j} - (J - s - g)\right\} \\ &- \frac{\sqrt{p_1 q_1 p_J q_J}}{p_1 - p_J}\sum_{j=s+1}^{J-g}\frac{p_j - p_J}{p_j q_j}. \end{aligned}$$

When $B > \tilde{B}$, the optimal allocation proportions are fixed. That is, they are functions of B but not the parameters of interest. In this paper, the

situation $B \in [0, \tilde{B}]$ is considered. Generalisation of such optimal allocation to censored exponential survival responses can be found in Sverdlov et al. (2011).

3.2 Optimal response-adaptive randomisation procedures

Optimal response-adaptive randomisation procedures are used to target the pre-specified optimal allocation. Two such procedures for multi-armed trials are introduced below. Since the optimal allocations are functions of unknown parameters, a learning phase using permuted-block randomisation for the first 10% of the N patients is applied throughout the simulations in this paper. This burn-in learning phase sample size appears to be large enough for the initial parameter estimates to be reasonably reliable. In practice, the burn-in sample size should be considered on a case-by-case basis for different design parameters. Permuted-block randomisation balances the sample sizes across the treatment groups. When initial parameter estimates are obtained, the following randomisation procedures can be implemented.

3.2.1 Doubly-adaptive biased coin design (DBCD)

Suppose that $m_j^{(i)}$ is the cumulative sample size on treatment j after i patients, $i = 1, ..., N$. Let $m_j^{(i)}/i$ and $\hat{\rho}_j^{(i)}$ be the current and target allocation proportions for treatment j, $j = 1, ..., J$, 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

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

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

ness. The DBCD is the most deterministic when $\gamma \to \infty$, whereas the procedure is the most random when $\gamma = 0$. The value $\gamma = 2$ is commonly used for a reasonable trade-off between variability and power. When $m_j^{(i)}/i > \hat{\rho}_j^{(i)}$, the probability that the next patient will be assigned to arm j is decreased and vice versa. At an extreme case such as $m_j^{(i)}/i = 1$, it is impossible that the next patient will be assigned to arm j. The allocation probability g_j is updated after each response observed.

3.2.2 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

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

Similar to the DBCD, the ERADE allocation probability depends on the current and the optimal allocation proportions. Here, the tuning parameter γ' can be any positive number. Through personal communication with L.-X. Zhang, a value $2 \leq \gamma' \leq 4$ is suggested to achieve a high power while allowing a reasonable degree of randomness. The ERADE has been shown to attain the Cramér-Rao lower bound for the variance of the allocation proportions (Zhang, 2016). That is, the use of the ERADE guarantees the least variability in the allocation among all response-adaptive randomisation methods.

The optimal response-adaptive randomisation procedures require the optimal allocation proportions $\rho_1, ..., \rho_J$ to be continuous and twice continuously differentiable. The D_A-optimal allocation satisfies these conditions. However, the closed-form solution for the optimal allocation based on nonlinear

programming is discontinuous when the parameters are all equal. Smoothing techniques are required.

4 Simulation results

4.1 Redesign of a four-armed binary trial

We now revisit the Neosphere trial. Let $p_C = 0.29$, $p_{E1} = 0.458$, $p_{E2} = 0.168$ and $p_{E3} = 0.24$. The global null hypothesis $H_{G_0} : \boldsymbol{p}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{p}_G \neq \mathbf{0}$ with $\boldsymbol{p}_G = (p_{E1} - p_C,\ p_{E2} - p_C,\ p_{E3} - p_C)^T$ is tested. The nominal type I error rate is set to 0.05 and $K = 3$ group sequential tests are planned at equally-spaced information times. The O'Brien and Fleming critical boundaries (23.76, 11.88, 7.92) for $J = 4$ treatments are used as an approximation. Results for the group sequential complete randomisation (CR) design and the fixed-sample CR and response-adaptive designs are provided alongside for comparison. For the fixed-sample designs, the critical boundary is 7.81. The adaptive designs are investigated by simulation with 5,000 replicates in terms of the error probabilities, the expected number of patients (ENP), the expected number of failures (ENF), the allocation proportions and the corresponding variability (s.d.).

The first 40 patients are randomly assigned using permuted-block randomisation with ratio 1:1:1:1 to obtain initial parameter estimates. Then the optimal response-adaptive randomisation procedures are performed. The tuning parameters $\gamma = \gamma^{'} = 2$ are set for the DBCD and the ERADE functions. For the D_A-optimal allocation, the target allocation proportions for the four-treatment trial can be obtained by solving the system of equations in (1), where $J = 4$. For the nonlinear programming (NP) allocation, the user-specified lower bound for the allocation proportions B is set to be 0.20 to satisfy $B \in [0, \tilde{B}]$, where $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_4, 0.25)$. The closed-form solution for the NP optimal allocation requires the order of the parameters to be $p_1 > p_2 \geq p_3 > p_4$. Let $p_1 = p_{E1},\ p_2 = p_C,\ p_3 = p_{E3}$ and $p_4 = p_{E2}$. Then, from (3), the NP optimal allocation which maximises the power subject to

the total sample size not exceeding a fixed value is

$$\rho_1 = QB + \frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_4 q_4}},$$
$$\rho_2 = \rho_3 = B,$$
$$\rho_4 = 1 - 2B - \rho_1,$$

where

$$\tilde{B}_1 = \frac{1}{1-Q} \frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_4 q_4}},$$
$$\tilde{B}_4 = \frac{1}{3+Q} \frac{\sqrt{p_4 q_4}}{\sqrt{p_1 q_1} + \sqrt{p_4 q_4}}$$

and

$$Q = \frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_4 q_4}} \left(\sum_{j=2}^{3} \frac{p_4 q_4}{p_j q_j} - 2 \right)$$
$$- \frac{\sqrt{p_1 q_1 p_4 q_4}}{p_1 - p_4} \sum_{j=2}^{3} \frac{p_j - p_4}{p_j q_j}.$$

When $B = 0$, the solution maximises the power but reduces to Neyman allocation for $J = 2$, where patients are assigned to the best and the worst treatments only. When $B = 0.25$, the solution becomes equal allocation.

Table 1: Simulated type I error rate for redesigning the NeoSphere trial using complete randomisation and response-adaptive randomisation, $p_C = 0.29,\ p_{E1} = 0.29,\ p_{E2} = 0.29,\ p_{E3} = 0.29$ and $N = 417$.

					$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)								
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.050	415.7	(13.9)	295.1	(9.9)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD_{D_A}	0.061	414.7	(20.9)	294.4	(14.8)	0.250	(0.012)	0.250	(0.012)	0.250	(0.012)	0.250	(0.011)
ERADE_{D_A}	0.058	414.9	(17.5)	294.6	(12.4)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)
DBCD_{NP}	0.069	413.6	(25.1)	293.6	(17.8)	0.244	(0.108)	0.244	(0.116)	0.243	(0.115)	0.268	(0.105)
ERADE_{NP}	0.057	414.3	(21.4)	294.2	(15.2)	0.243	(0.108)	0.246	(0.115)	0.245	(0.114)	0.267	(0.105)
					Fixed-sample design								
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.054	417	(0)	296.1	(0)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD_{D_A}	0.056	417	(0)	296.1	(0)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)
ERADE_{D_A}	0.058	417	(0)	296.1	(0)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)
DBCD_{NP}	0.055	417	(0)	296.1	(0)	0.242	(0.108)	0.245	(0.117)	0.246	(0.116)	0.266	(0.104)
ERADE_{NP}	0.055	417	(0)	296.1	(0)	0.241	(0.107)	0.246	(0.115)	0.245	(0.115)	0.268	(0.105)

Under the null hypothesis, the type I error rate in Table 1 for the fixed-sample

designs is well controlled in general. However, for the combined approach using the DBCD, $\tilde{\alpha}$ is inflated. For the ERADE designs, $\tilde{\alpha}$ lies within three standard errors of 0.05. This may be due to the fact that critical boundaries derived based on normal responses and equal variances with equal allocation are used as an approximation here. Under the null hypothesis where the parameters are all equal, the probability of early termination is small. The differences in the ENP and the ENF for the group sequential and fixed-sample designs are small. In addition, under H_{G_0}, the optimal allocation proportions are close to equal allocation, with the D_A-optimal allocation consistently having the least variation in the allocation proportions.

Since there are significant differences in the treatment effects, a high-powered test was obtained for all designs: see Table 2. This agreed with Gianni et al. (2012) that patients who received pertuzumab and trastuzumab plus docetaxel ($E1$) had a significantly improved pathological complete response rate compared to those who received the control, where a two-sided Mantel-Haenszel test was used.

The total number of failures in the NeoSphere trial was 296. A similar figure for the expected number of failures (ENF) was found for the fixed-sample CR design. If fixed-sample response-adaptive designs were used, about two fewer failures on average would be avoided using the D_A-optimal allocation and around 22 fewer could be achieved using the NP allocation. In addition, if group-sequential response-adaptive designs were used, a further reduction in the ENF could be obtained. Since the expected number of patients (ENP) for the group sequential designs was substantially lower than for the fixed-sample designs, the ENF was also decreased. If the group-sequential response-adaptive design with D_A-optimal allocation was applied, around 86 failures could be avoided. If the NP allocation was used, about 109 fewer failures could be achieved.

The ENF$'$ for the group sequential designs is calculated based on $N = 417$ patients to compare with the fixed-sample designs. When trials stop at an

interim analysis, the rest of the patients are assigned to the better-performing treatment and the expected number of failures for the rest of the patients is $(1-p_1)E(N_{rest})$, where p_1 is the probability of success for the best-performing treatment and N_{rest} denotes the number of remaining patients. In practice, trials stop when a decision is made. The ENF$^{'}$ is also consistently lower than the ENF for the fixed-sample designs, since the rest of the patients are assigned to the most promising treatment if trials stop early. For instance, for the NP allocation, the ENF$^{'}$ for the group sequential designs is about 15 less than the ENF for the fixed-sample designs. The other designs achieve around 20 fewer failures.

4.2 Three-armed censored survival trials

Consider testing $H_{G_0} : \boldsymbol{\theta}_G = \mathbf{0}$ versus $H_{G_a} : \boldsymbol{\theta}_G \neq \mathbf{0}$ with $\boldsymbol{\theta}_G = (\theta_{E1} - \theta_C,\ \theta_{E2} - \theta_C)^T$, where θ_j refers to the mean survival time for treatment j. This testing problem has been investigated by Sverdlov et al. (2011) using a fixed-sample design with the DBCD, and their simulation settings were based on a head and neck cancer experiment (Fountzilas et al., 2004). Here, similar simulation settings are considered. The duration of the trial D is 96 months. Independent exponentially distributed survival times and uniformly distributed arrival and censoring times are assumed. The D_A-optimal allocation and the NP allocation are used as the target allocations for the optimal response-adaptive designs. For the NP allocation, the user-specified lower bound for the allocation proportions B is set to be 0.20 to satisfy $B \in [0, \tilde{B}]$, where $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_3, 1/3)$. The nominal type I error rate is set to 0.05. There are $K = 3$ group sequential analyses planned at equally- and unequally-spaced information times. The O'Brien and Fleming boundaries (18.36, 9.18, 6.12) are used as an approximation. Again, $\gamma = \gamma^{'} = 2$ are set for the DBCD and the ERADE functions. Results for the group sequential CR design and fixed-sample CR and response-adaptive designs are also provided for comparison. For the fixed-sample designs, the critical value is 5.99. The maximum number of patients, N, is computed to achieve around 80% power for the group sequential CR design. The simulation results are based

on 5,000 replicates.

From Table 3, we find that the critical boundaries derived based on standard group sequential designs can be used as an approximation here. The type I error rate for all of the designs is less than 0.01 from 0.05. The differences in the ENP and the ENF among the designs are small under H_{G_0}.
Under the alternative hypothesis, from Table 4, the response-adaptive designs using the NP allocation can achieve a higher power and reduce the ENP and the ENF compared with the other designs. For instance, the use of the NP allocation can increase the power by around 4% compared to the D_A-optimal allocation. Meanwhile, about seven fewer patients on average are used and nine events are prevented. However, compared with the NP allocation, the D_A-optimal rule has more accuracy and precision in targeting the optimal allocation proportions.

Tables 5 and 6 compare the designs under group sequential monitoring with equal and unequal increments in information time. The fixed-sample designs are provided alongside for comparison. The maximum number of patients, N, is computed by simulation to attain around 80% power for the group sequential CR design. Compared with the settings in Tables 3 and 4, the mean survival time for each treatment is increased. When the mean survival time is 24 months as in Table 3, the probability of an event is 0.62, whereas, when the mean survival time is 45 months, as in Table 5, the probability of an event is 0.45. The probability of a censored response is increased for longer mean survival times.

Tables 5 and 6 demonstrate similar behaviour to Tables 3 and 4. In brief, the adaptive designs with the NP optimal allocation can achieve a higher power and reduce the ENP and the ENF, whereas those using the D_A-optimal allocation have lower variation in the allocation proportions compared with the NP optimal allocation. In addition, the critical boundaries can be used as an approximation to preserve the type I error rate for multi-armed censored survival trials with equal and unequal increments in information time.

Table 2: Simulated power for redesigning the NeoSphere trial using complete randomisation and response-adaptive randomisation, $p_C = 0.29,\ p_{E1} = 0.458,\ p_{E2} = 0.168,\ p_{E3} = 0.24$ and $N = 417$.

						$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)									
Procedure	power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.987	304.5	(70.3)	216.5	(50.1)	277.9	(12.2)	0.249	(0.024)	0.250	(0.024)	0.251	(0.024)	0.250	(0.024)
DBCD$_{D_A}$	0.991	298.8	(72.5)	210.9	(51.5)	275.3	(12.4)	0.256	(0.015)	0.267	(0.013)	0.229	(0.021)	0.248	(0.017)
ERADE$_{D_A}$	0.987	297.3	(72.6)	209.8	(51.6)	275.1	(12.4)	0.256	(0.014)	0.267	(0.011)	0.229	(0.020)	0.248	(0.015)
DBCD$_{NP}$	0.994	284.0	(71.5)	187.3	(47.2)	259.5	(8.9)	0.198	(0.037)	0.465	(0.038)	0.152	(0.031)	0.185	(0.032)
ERADE$_{NP}$	0.993	282.0	(72.4)	186.4	(48.0)	259.7	(9.2)	0.199	(0.038)	0.460	(0.040)	0.154	(0.029)	0.187	(0.031)
						Fixed-sample design									
Procedure	power	ENP	(s.d.)	ENF	(s.d.)	-	-	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.989	417	(0)	296.5	(2.1)	-	-	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
DBCD$_{D_A}$	0.990	417	(0)	294.3	(1.3)	-	-	0.256	(0.012)	0.267	(0.011)	0.230	(0.015)	0.248	(0.012)
ERADE$_{D_A}$	0.989	417	(0)	294.3	(1.0)	-	-	0.256	(0.009)	0.266	(0.008)	0.230	(0.013)	0.248	(0.010)
DBCD$_{NP}$	0.994	417	(0)	274.4	(2.8)	-	-	0.198	(0.030)	0.470	(0.033)	0.145	(0.029)	0.187	(0.030)
ERADE$_{NP}$	0.995	417	(0)	274.7	(2.6)	-	-	0.198	(0.028)	0.468	(0.030)	0.147	(0.027)	0.188	(0.027)

The target D_A-optimal and NP allocations are (0.256, 0.266, 0.230, 0.248) and (0.200, 0.479, 0.121, 0.200), respectively.

Table 3: Simulated type I error rate for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = \theta_{E2} = \theta_C = 24$ and $N = 312$.

Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.041	311.4	(7.0)	193.9	(4.4)	0.333	(0.025)	0.333	(0.025)	0.333	(0.025)
DBCD$_{D_A}$	0.058	310.3	(12.3)	193.2	(7.7)	0.333	(0.034)	0.333	(0.034)	0.333	(0.034)
ERADE$_{D_A}$	0.057	310.7	(10.7)	193.5	(6.6)	0.333	(0.031)	0.334	(0.031)	0.333	(0.030)
DBCD$_{NP}$	0.058	310.5	(10.8)	193.4	(6.7)	0.336	(0.091)	0.332	(0.090)	0.332	(0.090)
ERADE$_{NP}$	0.052	310.9	(10.0)	193.6	(6.2)	0.333	(0.088)	0.334	(0.089)	0.332	(0.089)

Table 4: Simulated power for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = 34,\ \theta_{E2} = 24,\ \theta_C = 20$ and $N = 312$.

Procedure	power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.733	294.7	(32.9)	178.8	(20.1)	0.334	(0.026)	0.333	(0.026)	0.333	(0.026)
DBCD$_{D_A}$	0.785	284.9	(38.9)	170.0	(23.8)	0.409	(0.029)	0.324	(0.039)	0.268	(0.041)
ERADE$_{D_A}$	0.788	284.8	(38.8)	170.0	(23.7)	0.407	(0.026)	0.325	(0.035)	0.269	(0.037)
DBCD$_{NP}$	0.827	277.2	(40.5)	161.8	(25.0)	0.533	(0.092)	0.229	(0.062)	0.238	(0.050)
ERADE$_{NP}$	0.824	278.0	(40.4)	162.5	(24.9)	0.526	(0.091)	0.234	(0.065)	0.240	(0.048)

The target D_A-optimal and NP allocations are (0.406, 0.323, 0.271) and (0.544, 0.200, 0.256), respectively.

Table 5: Simulated type I error rate for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = \theta_{E2} = \theta_C = 45$ and $N = 600$.

				($t_1,\ t_2,\ t_3$)=(0.33, 0.67, 1)							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.040	599.2	(10.5)	269.6	(4.7)	0.334	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD$_{D_A}$	0.048	598.4	(14.6)	269.2	(6.6)	0.334	(0.029)	0.333	(0.030)	0.333	(0.029)
ERADE$_{D_A}$	0.050	598.5	(14.5)	269.3	(6.5)	0.333	(0.026)	0.333	(0.026)	0.333	(0.026)
DBCD$_{NP}$	0.045	598.4	(14.7)	269.2	(6.6)	0.331	(0.084)	0.334	(0.084)	0.335	(0.085)
ERADE$_{NP}$	0.048	598.5	(13.8)	269.3	(6.2)	0.332	(0.081)	0.336	(0.081)	0.332	(0.081)
				($t_1,\ t_2,\ t_3$)=(0.5, 0.8, 1)							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.039	599.5	(6.2)	269.7	(2.8)	0.334	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD$_{D_A}$	0.049	598.8	(11.0)	269.4	(5.0)	0.333	(0.029)	0.333	(0.030)	0.333	(0.029)
ERADE$_{D_A}$	0.051	599.0	(9.4)	269.5	(4.2)	0.333	(0.027)	0.334	(0.027)	0.333	(0.026)
DBCD$_{NP}$	0.048	598.9	(9.9)	269.4	(4.4)	0.336	(0.085)	0.333	(0.084)	0.332	(0.082)
ERADE$_{NP}$	0.044	599.1	(8.3)	269.6	(3.7)	0.335	(0.083)	0.331	(0.081)	0.334	(0.083)
				Fixed-sample design							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.041	600	(0)	269.9	(0.0)	0.334	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD$_{D_A}$	0.051	600	(0)	269.9	(0.0)	0.334	(0.029)	0.333	(0.029)	0.333	(0.029)
ERADE$_{D_A}$	0.050	600	(0)	269.9	(0.0)	0.334	(0.026)	0.333	(0.027)	0.333	(0.027)
DBCD$_{NP}$	0.048	600	(0)	269.9	(1.8)	0.332	(0.088)	0.333	(0.087)	0.334	(0.088)
ERADE$_{NP}$	0.042	600	(0)	269.9	(0.0)	0.332	(0.086)	0.334	(0.085)	0.334	(0.087)

Table 6: Simulated power for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = 59,\ \theta_{E2} = 45,\ \theta_C = 37$ and $N = 600$.

					$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)								
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.807	556.3	(62.1)	247.0	(27.7)	263.6	(4.3)	0.334	(0.019)	0.333	(0.019)	0.333	(0.019)
DBCD$_{D_A}$	0.829	545.4	(66.2)	237.7	(29.9)	258.5	(5.0)	0.401	(0.027)	0.331	(0.033)	0.269	(0.035)
ERADE$_{D_A}$	0.835	547.0	(65.9)	238.4	(29.8)	258.6	(5.0)	0.400	(0.023)	0.331	(0.030)	0.269	(0.033)
DBCD$_{NP}$	0.853	536.8	(67.6)	229.6	(31.1)	253.6	(7.0)	0.506	(0.093)	0.239	(0.066)	0.254	(0.048)
ERADE$_{NP}$	0.844	537.9	(66.9)	230.4	(30.8)	254.0	(7.0)	0.499	(0.095)	0.243	(0.070)	0.257	(0.048)
					$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)								
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.791	566.4	(40.5)	251.4	(18.1)	264.2	(3.0)	0.334	(0.019)	0.333	(0.019)	0.333	(0.019)
DBCD$_{D_A}$	0.822	558.2	(45.1)	243.2	(20.9)	259.1	(4.0)	0.401	(0.026)	0.331	(0.032)	0.268	(0.034)
ERADE$_{D_A}$	0.818	558.6	(45.1)	243.5	(20.8)	259.2	(3.9)	0.399	(0.023)	0.331	(0.029)	0.270	(0.031)
DBCD$_{NP}$	0.846	552.2	(46.1)	236.1	(22.2)	254.3	(6.3)	0.507	(0.093)	0.236	(0.064)	0.257	(0.048)
ERADE$_{NP}$	0.842	552.2	(45.9)	236.5	(22.3)	254.6	(6.5)	0.500	(0.096)	0.240	(0.068)	0.259	(0.048)
					Fixed-sample design								
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	-	-	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.763	600	(0)	266.3	(1.2)	-	-	0.334	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD$_{D_A}$	0.804	600	(0)	261.5	(1.7)	-	-	0.399	(0.024)	0.331	(0.029)	0.270	(0.030)
ERADE$_{D_A}$	0.797	600	(0)	261.5	(1.6)	-	-	0.398	(0.022)	0.330	(0.027)	0.272	(0.028)
DBCD$_{NP}$	0.835	600	(0)	256.4	(5.4)	-	-	0.510	(0.092)	0.230	(0.062)	0.261	(0.051)
ERADE$_{NP}$	0.831	600	(0)	256.6	(5.3)	-	-	0.506	(0.091)	0.231	(0.061)	0.263	(0.050)

The target D_A-optimal and NP allocations are (0.400, 0.330, 0.270) and (0.519, 0.200, 0.281), respectively.

5 Discussion

5.1 Conclusions

Previous work on two-armed clinical trials has shown that combining group sequential analysis with response-adaptive randomisation preserves the advantages of both techniques while controlling the error rates. This paper investigates the combined approach in multi-armed clinical trials. Based on the results obtained, the critical boundaries derived based on standard group sequential designs can be used as an approximation for the adaptive design for different types of outcomes. Moreover, response-adaptive randomisation can be more efficient than complete randomisation in multi-arm trials.

Compared with the group sequential CR design, the adaptive designs can increase the power of the tests of homogeneity while decreasing the average numbers of patients and failures. 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 higher power while minimising the average number of patients.

We acknowledge that some aspects of the NeoSphere trial in Section 4.1 have been simplified for illustrative purposes. In practice, considerations on whether or not to use the combined approach include but are not limited to (a) the number of treatments being compared, (b) the target sample size/number of events, (c) the length of time to obtain the intermediate/surrogate outcome measurements and (d) the availability of a response-adaptive randomisation algorithm, especially if several centres are involved. It is acknowledged that the use of such an adaptive design is more computationally demanding than conventional randomisation algorithms such as minimisation, and is less appealing for rare diseases or in two-armed trials.

5.2 Further work

The global test focuses on a test of homogeneity. The critical boundaries used are based on the joint distribution of the test statistics assuming that sampling for all treatments continues to the end of the trial. Dropping of inferior treatments violates this underlying assumption. Further work on the multi-armed adaptive designs that allow dropping inferior arms at interim analyses will be presented separately.

This paper focused on treatment contrasts as the parameters of interest. A rate ratio or hazard ratio are also common outcomes. Biswas et al. (2011) investigated the optimal allocations for different types of outcomes. However, the multi-armed adaptive designs, which combine group sequential tests with adaptive randomisation techniques, presented in this paper have not been investigated using outcomes of a rate/hazard ratio, yet may be of interest for future study.

In the Bayesian paradigm, MAMS designs with adaptive randomisation have been explored. The posterior estimation, which is updated by the observed cumulative outcomes with a pre-specified prior, is used to allocate patients adaptively. The posterior distribution is also computed to see if it meets the pre-specified Bayesian decision rule to stop the trial early at an interim analysis. With the advantage of flexibility in Bayesian designs, Ventz et al. (2018) further considered adding experimental arms to a platform clinical trial. 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. To extend the frequentist design in this paper to allow arms to be added while controlling the type I error rate, one will need to show that the information time formula described in Section 2.1 continues to hold after adding arms, which may be worth exploring. The operating characteristics of such a design can be compared with those of a Bayesian approach.

Seamless phase II/III designs allow dose or treatment selection at one or more interim analyses. A single arm is selected and the comparative efficacy with the control evaluated (Stallard and Todd, 2011; Thall, 2008). This approach was further extended to allow any number of treatments to continue at each stage as long as the number is pre-specified and not data-dependent. Jennison and Turnbull (2006) showed that seamless phase II/III group sequential designs control the family-wise error rate in the strong sense. It is reasonable for dose selection but might be too conservative in some circumstances, especially when the treatments being compared are very different. One may explore generalising seamless phase II/III designs to the response-adaptive setting. A short-term outcome should be used at interim analyses.

Exponential survival times are considered in this paper, which require the strong assumption of a constant hazard. The D_A-optimal design for multi-armed trials assuming Weibull survival times has been investigated (Sverdlov et al., 2014). Their approach required interim data to be analysed at pre-specified points in the trial. The proposed global testing in the present paper can be extended to the Weibull survival case. However, unlike the simple exponential case, there will be no closed form for the probability of an event on arm j, ϵ_j, in (2). Nevertheless, an alternative is to consider the empirical estimate of ϵ_j using the actual observed events. This approach allows an interim analysis to be conducted at any time, since the critical boundaries for pairwise comparisons following the completion of global testing were derived based on an alpha-spending function.

Acknowledgements

This work was carried out whilst the first author was in receipt of funding from the Ministry of Education in Taiwan. The authors also wish to thank three referees for their comments, which have led to an improved paper.

References

Biswas, A., Mandal, S., and Bhattacharya, R. (2011). Multi-Treatment Optimal Response-Adaptive Designs for Phase III Clinical Trials. *Journal of the Korean Statistical Society*, 40(1):33–44.

Bratton, D. J., Parmar, M. K. B., Phillips, P. P. J., and Choodari-Oskooei, B. (2016). Type I Error Rates of Multi-Arm Multi-Stage Clinical Trials: Strong Control and Impact of Intermediate Outcomes. *Trials*, 17(1):309.

Follmann, D. A., Proschan, M. A., and Geller, N. L. (1994). Monitoring Pairwise Comparisons in Multi-Armed Clinical Trials. *Biometrics*, 50(2):325–36.

Fountzilas, G., Ciuleanu, E., Dafni, U., and Plataniotis, G. (2004). Concomitant Radiochemotherapy vs Radiotherapy Alone in Patients with Head and Neck Cancer. *Medical Oncology*, 21(2):95–107.

Gianni, L., Pienkowski, T., Im, Y.-H., and Roman, L. (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. *Lancet Oncology*, 13(1):25–32.

Jaki, T., Pallmann, P., and Magirr, D. (2019). The R Rackage MAMS for Designing Multi-Arm Multi-Stage Clinical Trials. *Journal of Statistical Software*, 88(4):1–25.

Jennison, C. and Turnbull, B. W. (1991). Exact Calculations for Sequential t, χ^2 and F Tests. *Biometrika*, 78(1):133–141.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Application to Clinical Trials.* London: Chapman and Hall/CRC.

Jennison, C. and Turnbull, B. W. (2006). Confirmatory Seamless Phase II/III Clinical Trials with Hypotheses Selection at Interim: Opportunities and Limitations. *Biometrical Journal.*, 48(4):650–655.

Kiefer, J. and Wolfowitz, J. (1960). The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics*, 12:363–6.

Kim, K., Boucher, H., and Tsiatis, A. A. (1995). Design and Analysis of Group Sequential Logrank Tests in Maximum Duration Versus Information Trials. *Biometrics*, 51(3):988–1000.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete Sequential Boundaries for Clinical Trials. *Biometrika*, 70(3):659–63.

Liu, W. and Coad, D. S. (2020). Group-Sequential Response-Adaptive Designs for Censored Survival Outcomes. *Journal of Statistical Planning and Inference*, 205:293–305.

Magirr, D., Jaki, T., and Whitehead, J. (2012). A Generalized Dunnett Test for Multi-Arm Multi-Stage Clinical Studies with Treatment Selection. *Biometrika*, 99(2):494–501.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001). Optimal Adaptive Designs for Binary Response Trials. *Biometrics*, 57(3):909–13.

Ryan, E. G., Lamb, S. E., Williamson, E., and Gates, S. (2020). Bayesian Adaptive Designs for Multi-Arm Trials: An Orthopaedic Case Study. *Trials*, 21(1):83.

Stallard, N. and Todd, S. (2011). Seamless Phase II/III Designs. *Statistical Methods in Medical Research*, 20(6):623–34.

Sverdlov, O., Ryeznik, Y., and Wong, W. K. (2014). Efficient and Ethical Response-Adaptive Randomization Designs for Multi-Arm Clinical Trials with Weibull Time-to-Event Outcomes. *Journal of Biopharmaceutical Statistics*, 24(4):732–54.

Sverdlov, O., Tymofyeyev, Y., and Wong, W. K. (2011). Optimal Response-Adaptive Randomized Designs for Multi-Armed Survival Trials. *Statistics in Medicine*, 30(24):2890–910.

Thall, P. F. (2008). A Review of Phase 2–3 Clinical Trial Designs. *Lifetime Data Analysis*, 14(1):37–53.

Tymofyeyev, Y., Rosenberger, W. F., and Hu, F. (2007). Implementing Optimal Allocation in Sequential Binary Response Experiments. *Journal of the American Statistical Association*, 102(477):224–34.

Ventz, S., Cellamare, M., Parmigiani, G., and Trippa, L. (2018). Adding Experimental Arms to Platform Clinical Trials: Randomization Procedures and Interim Analyses. *Biostatistics*, 19(2):199–215.

Wason, J., Magirr, D., Law, M., and Jaki, T. (2016). Some Recommendations for Multi-Arm Multi-Stage Trials. *Statistical Methods in Medical Research*, 25(2):716–27.

Wason, J., Stallard, N., Bowden, J., and Jennison, C. (2014). A Multi-Stage Drop-the-Losers Design for Multi-Arm Clinical Trials. *Statistical Methods in Medical Research*, 26(1):508–24.

Wason, J. and Trippa, L. (2014). A Comparison of Bayesian Adaptive Randomization and Multi-Stage Designs for Multi-Arm Clinical Trials. *Statistics in Medicine*, 33(13):2206–21.

Wong, W. K. and Zhu, W. (2008). Optimum Treatment Allocation Rules under a Variance Heterogeneity Model. *Statistics in Medicine*, 27(22):4581–95.

Zhang, L.-X. (2016). Response-Adaptive Randomization: An Overview of Designs and Asymptotic Theory. Unpublished manuscript.

Zhu, H. and Hu, F. (2010). Sequential Monitoring of Response-Adaptive Randomized Clinical Trials. *The Annals of Statistics*, 38(4):2218–41.