Group-Sequential Response-Adaptive Designs for Comparing Several Treatments

Wenyu Liu

Submitted in partial fulfilment of the requirements of
the Degree of Doctor of Philosophy

School of Mathematical Sciences
Queen Mary, University of London

December 2016

Statement of originality

I, Wenyu Liu, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature:

Date:

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. D. S. Coad, for his continuous support, inspiration and guidance during my PhD study. In addition, I would like to thank my assessors, Professor P. D. Sasieni from the Wolfson Institute of Preventive Medicine and Dr. L. I. Pettit from the School of Mathematical Sciences, for their valuable comments on my work. Special thanks go to my family, especially my parents, for their endless spiritual and selfless financial support.

I would like to thank the Ministry of Education in Taiwan for funding my PhD. A paper based on Chapter 2 was selected for one of the Student Conference Awards at the 35th Annual Conference of the International Society for Clinical Biostatistics (ISCB) in Vienna, Austria, in August 2014. In addition, with support from the Queen Mary Postgraduate Research Fund, the work in Chapter 3 was presented at the 36th Annual Conference of the ISCB in Utrecht, The Netherlands, in August 2015. Part of the work in Chapter 4 was exhibited at the 42th Young Statisticians' Meeting in London in August 2016.

It is expected that each of the main chapters will be prepared as a journal paper. The work based on Chapter 2 is planned to be submitted to *Statistics in Medicine*. The target journal is the *Journal of Statistical Planning and Inference* for Chapter 3, and *Statistical Methods in Medical Research* for Chapter 4.

Abstract

Previous work on two-treatment comparisons has shown that the use of optimal response-adaptive randomisation with group sequential analysis can allocate more patients to the better-performing treatment while preserving the overall type I error rate. The sequence of test statistics for this adaptive design asymptotically satisfies the canonical joint distribution. The overall type I error rate can be controlled by utilising the error-spending approach. However, previous work focused on immediate responses. The application of the adaptive design to censored survival responses is investigated and different optimal response-adaptive randomised procedures compared. For a maximum duration trial, the information level at the final look is usually unpredictable. An approximate information time is defined.

Several treatments are often compared in a clinical trial nowadays. The adaptive design generalised to multi-arm clinical trials is studied. First, a global test is considered. The joint distribution of the sequence of test statistics no longer has the canonical distribution. However, the joint distribution can be derived, since the test statistic is a quadratic form of independent normal variables. Existing critical boundaries are based on normal responses and known variances with equal allocation and equal increments in information. Our results show that these boundaries can be used approximately for designs with other types of responses, unequal variances or unbalanced allocation.

If the global null hypothesis is rejected, then pairwise comparisons are conducted at the current and subsequent looks to investigate which treatment effects differ. This is an analogue of Fisher's least significant difference method that can control the family-wise error rate. The adaptive design can target any optimal allocation to achieve some optimality criterion, and allows dropping of inferior treatments at interim looks, which can be unequally spaced in information time. Optimal allocation proportions after dropping arms are described. The power is not adversely affected by unbalanced allocation.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Background

The use of a sequential test in clinical trials can require fewer patients than a fixed-sample design to achieve the same error probabilities (Jennison and Turnbull, 2000). It allows early stopping when either a significant treatment effect or futility is detected, and hence saves time and resources. Also, the use of sequential monitoring can ensure that the protocol is followed well and that the assumptions are not violated. Fully sequential designs that continuously evaluate treatment effects after each response observed are often unrealistic. Periodic group sequential designs, in which a number of interim analyses are conducted after groups of observations, are more practical. With these advantages, group sequential analysis has become standard practice for phase III clinical trials (Jennison and Turnbull, 2000). In addition, it is reported in recent studies that group sequential analysis is the most popular adaptive design used in medical studies (Dimairo et al., 2015; Hatfield et al., 2016).

The probability of falsely rejecting the null hypothesis increases when the number of group sequential tests is increased. Typically, the sample size and critical boundaries are determined to attain the pre-specified error probabilities. There are different discrete stopping boundaries, which require a pre-specified number of interim analyses and equally spaced information levels between looks. The boundaries of Pocock (1977) spend the same proportion of the type I error rate at each

look, whereas these of O'Brien and Fleming (1979) spend little of the type I error rate at early looks, and hence the critical value for the final look is close to that of a fixed-sample design.

However, it is often impracticable to pre-specify the number of interim analyses and they are sometimes conducted at unequally spaced information levels. For instance, when patient recruitment is slow, investigators may wish to postpone the time of conducting an analysis. It may even end up that the sample size at the planned end of the study is too small. Then the trial continues to recruit more patients and additional group sequential analyses are needed. To this end, Lan and DeMets (1983) proposed continuous critical boundaries derived from a user-selected error-spending function. An error-spending function spends the error rate with the information time, which is the ratio of the current information level to the expected information level at the final look. This approach allows the interim analyses to take place at any continuous information time, while preserving the nominal type I error probability.

In group sequential monitoring, trials can be terminated early. Hence, patients can be prevented from being exposed to inferior or unsafe treatments. In addition, response-adaptive randomisation, which skews the allocation proportion of the sample sizes towards the more promising treatments based on the responses observed, can further reduce the numbers of patients allocated to the inferior treatments compared to complete randomisation (Hu and Rosenberger, 2006). Nevertheless, a lower inferior treatment number usually leads to an increase in the expected total sample size (Jennison and Turnbull, 2001). Moreover, the use of response-adaptive randomisation may increase the variability in the allocation, which could adversely affect the power to detect the differences in treatment effects (Hu and Rosenberger, 2003). Therefore, response-adaptive randomisation methods usually seek to balance the competing goals of individual and collective

ethics, by allocating more patients to the better-performing treatment, minimising the total sample size and maximising the power.

There are different response-adaptive randomisation methods. One family is the urn-model type designs for dichotomised responses (Rosenberger and Lachin, 2002), including the randomised play-the-winner (RPW) design (Wei and Durham, 1978) and the drop-the-loser rule (Ivanova, 2003). The idea of this family of designs is that treatment assignment is determined by drawing a ball from an urn, and the composition of the balls is updated after each response. For example, for the RPW design, if a success is observed, an additional ball of the same type is added to the urn. Otherwise, an additional ball of the opposite type is added. Hence, the probability of drawing the type of ball corresponding to the better treatment increases. Another type of response-adaptive sampling rule is characterised by a loss function, which is a weighted sum of the sample sizes on each treatment, with a lower cost assigned to the better treatment (Jennison and Turnbull, 2000, 2001). Although both types of response-adaptive randomisation method can reduce the expected number of failures, the first type are not derived based on an optimality criterion.

For optimal response-adaptive randomised designs (Atkinson and Biswas, 2014; Antognini and Giovagnoli, 2015), a specific criterion is optimised based on the assumed response model to obtain an optimal allocation. For instance, a standard optimality criterion for binary responses is to minimise the expected number of failures. Different optimal response-adaptive randomisation procedures have been proposed, including the popular doubly-adaptive biased coin design (DBCD) (Eisele and Woodroofe, 1995) and the efficient randomised-adaptive design (ERADE) (Hu et al., 2009). An advantage of these designs is that they can target any specified optimal allocation based on some optimality criterion. Throughout the thesis, we will focus on optimal response-adaptive randomisation.

Jennison and Turnbull (2001) derived theory to support that the combined approach of group sequential analysis with response-adaptive randomisation 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. Their simulation results also show that the nominal error rates are attained for the combined approach. 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. In addition, Morgan (2003a) proposed two inferential methods for the treatment mean difference following such a group-sequential response-adaptive design. One considered approximate confidence intervals using a pivotal method, and the other constructed a bias-adjusted maximum likelihood estimator.

Morgan (2003b) investigated the combined approach for normal responses with unknown variances. As inaccurate estimates of the variances of the responses can influence the power considerably, she suggested using sample size re-estimation based on the new estimates of the variances updated by the observed responses. For two-armed binary trials, Morgan and Coad (2007) compared several adaptive allocation rules in a group sequential setting, including two urn-model type designs, the DBCD and sequential maximum likelihood estimation (SMLE) rule, which minimise the expected number of failures and is a special case of the DBCD. Among the designs they investigated, the drop-the-loser rule (Ivanova, 2003) 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 Hu (2010) studied the combined approach of group sequential analysis with optimal response-adaptive randomisation for two-armed clinical trials with

normal and binary responses. By considering monitoring the response-adaptive design at a continuous information time, Zhu and Hu (2010) proved that the sequence of test statistics converged to a Brownian motion in distribution and asymptotically satisfied the canonical joint distribution proposed by Jennison and Turnbull (2000) for standard group sequential designs. Continuous boundaries obtained by the error-spending approach (Lan and DeMets, 1983) can be used to control the nominal type I error rate. The simulation results in Zhu and Hu (2010) reveal that the use of the combined approach can preserve the advantages of both group sequential analysis and optimal response-adaptive randomisation. However, the authors focused on the popular DBCD for two-armed trials with immediate responses. The application of the combined approach to two-armed censored survival trials and to multi-armed experiments has not yet been studied. This will be investigated in the thesis, with various designs compared.

1.2. Literature review

1.2.1. Two-armed survival responses

Survival or time-to-event responses usually have a heavy upper tail. Therefore, the assumption of normality is not appropriate. One of the common statistical models for survival responses is the proportional hazards model (Cox and Oakes, 1984), which assumes that a unit increase in an explanatory variable will multiplicatively affect the hazard rate. The main focus is the hazard ratio, which can be estimated regardless of the unknown hazard function. However, the strong assumption of proportional hazards, which implies that the hazard rates for different treatments can never cross, may be unrealistic in practice.

Another commonly-seen method is the non-parametric logrank test. Group sequential monitoring of logrank tests has been discussed (Jennison and Turnbull,

2000). For survival responses, the information levels usually cannot be attained accurately, since they depend on the realised pattern of events and censoring. However, for the group sequential logrank test, the information level can be approximated by the observed number of events divided by four, under the assumption that the numbers at risk are similar in each treatment arm. Yet, equal allocation across treatment groups was considered. The sample size for each arm is chosen to attain the information level.

For optimal response-adaptive randomisation, the target allocation proportions are usually unbalanced and depend on the unknown parameters. Zhang and Rosenberger (2007) derived optimal allocations for two-armed censored survival trials with exponential and Weibull distributions. The optimal allocations are derived based on different optimality criteria, such as minimising the total sample size or the total expected hazard. Then the DBCD is applied to target the specified optimal allocation. The authors considered the priority queue data structure (Rosenberger and Seshaiyer, 1997), which assumes uniformly distributed staggered entry and right censoring. It is shown in theory that a delay in response has little effect on the asymptotic variance of the DBCD procedure. In addition, simulation results show that the use of the DBCD results in more patients being allocated to the more promising treatment without a loss of power. Nevertheless, a fixed-sample design was used.

1.2.2. Multi-armed clinical trials

Several treatments are often compared in a clinical trial nowadays (Follmann et al., 1994). For multiple comparisons, one needs to ensure that the family-wise error rate is preserved, since the more pairwise comparisons that are made, the higher the probability that a null hypothesis will be rejected. For all pairwise comparisons, $p = J(J-1)/2$ tests are conducted, where J is the number of treatments.

For comparisons with a common control, there are $p = J - 1$ tests. To this end, many approaches based on fixed-sample designs have been proposed. One simple approach is the Bonferroni adjustment, which uses α/p as the nominal type I error rate for each pairwise test. In group sequential monitoring, pairwise comparisons are repeatedly carried out at each look. One also needs to ensure that the overall type I error rate is controlled. The Bonferroni approach can be extended to group sequential designs by replacing α with α_k, the type I error rate spent by interim analysis k.

In addition, Follmann et al. (1994) obtained exact critical boundaries for group sequential pairwise tests for tests on means and survival distributions using logrank tests, under the assumption of equal variances or censoring distributions for each arm. The authors pointed out that the difference between the critical values for the exact methods and those using the Bonferroni adjustment is very modest. Moreover, the Bonferroni approach is more flexible, which allows unequal allocation across treatment groups, and, if desired, different shapes of critical boundaries for different pairwise comparisons. Both the Bonferroni and exact methods strongly control the overall type I error rate. However, they can be too conservative at the price of losing power.

Another approach to group sequential monitoring of multi-armed clinical trials is to use a global test. Jennison and Turnbull (1991, 2000) derived critical boundaries analogous to Pocock's and the O'Brien and Fleming boundaries. These are derived based on multi-armed normal trials with equal variances and equal treatment allocation. For the unequal variances case, Proschan et al. (1994) 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 as an approximation (Jennison and Turnbull, 2000). Nevertheless, these studies do not consider the incorporation of response-adaptive sampling

rules in the group sequential analysis. The optimal response-adaptive randomisation procedures have been generalised to multi-armed trials using a fixed-sample design, including the DBCD (Hu and Zhang, 2004) and the ERADE (Zhang, 2016). Whether the combined approach for multi-armed trials still preserves the error rates while targeting some optimality criterion is of concern in this thesis. However, 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.

1.2.3. Multi-armed clinical trials with dropping of arms

In group sequential monitoring, when an inferior treatment is identified at an interim look, it is unethical to continue assigning patients to that arm. Fisher's least significance difference (LSD) method is considered to be one of the most powerful multiple comparison methods (Christensen, 2002). A group sequential Fisher's LSD method was proposed by Proschan et al. (1994). 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 the pairwise comparisons. However, the study of Proschan et al. (1994) only considered the cases of equal allocation and fixed unequal allocation determined prior to the commencement of the experiment. Implementation of optimal response-adaptive randomisation in a fixed-sample Fisher's LSD design has been studied, including Tymofyeyev et al. (2007) for binary responses and Sverdlov et al. (2011) for censored survival responses. These studies considered the popular DBCD. One objective of the thesis is to extend this work to group sequential designs, with different optimal response-adaptive sampling rules compared.

Other studies considered using multi-armed multi-stage (MAMS) designs to monitor multi-armed clinical trials (Magirr et al., 2012; Wason et al., 2016). MAMS designs simultaneously evaluate several regimens against a common control. With efficacy and futility boundaries, the designs allow dropping of inferior treatments at interim analyses. MAMS designs are shown to strongly control the type I error rate, which means that the probability of falsely rejecting one or more null hypotheses is less than or equal to α (Bratton et al., 2016). However, response-adaptive randomisation has not been incorporated in MAMS designs. The number of patients needed per arm per stage and the critical boundaries are obtained by numerical computation. In this thesis, Fisher's LSD method generalised to group-sequential response-adaptive designs will be the focus.

1.3. Structure of chapters

In Chapter 2, the study of Zhu and Hu (2010) that combines group sequential analysis with the DBCD for two-armed trials with immediate responses is described and compared to that using the ERADE by simulation in Section 2.2. An extension of the combined approach to censored survival responses is provided in Section 2.3. The issue of right censoring in group sequential analysis is taken into account in the model. For instance, if a subject has not responded at the time when an interim test is conducted, the true survival time is greater than the observed survival time. Right censoring caused by lost to follow-up due to death from an unrelated cause or emigration is also considered. For maximum duration trials, an approximation to the information time for censored survival responses is obtained, which depends on the assumed models for patient entry, survival time and censoring. The error-spending approach can be applied to control the overall type I error rate. In addition, a placebo-controlled binary trial is redesigned using

the combined approach.

In Chapters 3 and 4, extensions of the combined approach to several treatment comparisons are investigated. Chapter 3 considers a global test. The global test statistics for different types of responses are described in Section 3.1. Critical boundaries derived by Jennison and Turnbull (1991, 2000) are applied as an approximation for the designs. For optimal response-adaptive randomisation, the DBCD and the ERADE generalised to multi-treatment trials are given in Section 3.2. Two optimal allocations for multi-armed trials are considered. One ensures the most efficient treatment effect estimates and the other maximises the power of tests of homogeneity while fixing the total sample size. Properties of the group-sequential response-adaptive design for multi-treatment trials are investigated by simulation in Section 3.3, where both equal and unequal increments in information time are considered. In addition, results for a fixed-sample design with one analysis conducted at the end of the trial are provided alongside for comparison.

Chapter 4 explores an analogue of Fisher's LSD method generalised to group-sequential response-adaptive designs. The global and pairwise tests are discussed in Section 4.1. Any treatment that is inferior to the control can be discarded after the pairwise comparisons. The information time for trials that allow dropping arms is described in Section 4.2. Optimal allocation proportions after dropping treatments are described in Section 4.3. Properties of the analogue of Fisher's LSD method generalised to group-sequential optimal response-adaptive designs are investigated by simulation in Section 4.4. In addition, simulation results of redesigning a four-armed clinical trial are summarised.

Conclusions are drawn in Chapter 5, together with a discussion of the limitations and possible extensions to the research. Programs for the simulation studies are written in the statistical software R. In addition, for censored survival responses,

the optimal allocations depend on the probability of an event, which is calculated in Appendix A based on the assumed models for patient arrival, survival and censoring time described in Section 2.3.2. For the global tests on means, Appendix B gives the derivation of the noncentrality parameter for the group-sequential chi-squared test statistics.

2. Group-sequential response-adaptive designs for two-armed trials

2.1. Immediate responses

2.1.1. Form of test

Information time

Suppose that N is the planned maximum number of patients for a trial with K group sequential analyses. For immediate responses including normal and binary endpoints, the information level is proportional to the number of subjects recruited (Jennison and Turnbull, 2000). As the trial proceeds, we obtain more responses and gain more information. The information time at group sequential test k is

$$t_k = \frac{\mathcal{I}_k}{\mathcal{I}_K} = \frac{\sum_{j=1}^{2} m_{j,k}}{\sum_{j=1}^{2} M_j} = \frac{n_k}{N}, \quad k = 1, ..., K, \tag{2.1}$$

where $\mathcal{I}_k$ denotes the information level at group sequential test k, $m_{j,k}$ is the cumulative number of patients on treatment j, $j = 1, 2$, at look k, $m_{j,K} = M_j$, and $n_k = \sum_{j=1}^{2} m_{j,k}$ is the cumulative total sample size at look k, $n_K = N$.

Interim analyses can be conducted at any continuous information time $t_k \in (0, 1]$. For example, suppose that we wish to conduct the first interim test when about

one quarter of the maximum sample size has been recruited. Then the information time at the first look is set as $t_1 = 0.25$ and about $n_1 = \lceil t_1 N \rceil$ patients are needed. Here, $\lceil x \rceil$ is the smallest integer greater than or equal to x. If the trial continues to the end of the study without early termination, then $t_K = 1$ and the maximum number of patients N is reached.

Test statistics

Normal responses

Assume that the response for patient $i,\ i = 1, ..., m_{j,k}$, on treatment $j,\ j = 1, 2$, is normal, $Y_{i,j} \sim N(\mu_j, \sigma_j^2)$. Then the sample mean for arm j at look k is

$$\hat{\mu}_{j,k} = \frac{\sum_{i=1}^{m_{j,k}} Y_{i,j}}{m_{j,k}}.$$

Suppose that the parameter of interest is the difference in treatment effects, $\phi = \mu_1 - \mu_2$. We wish to test the null hypothesis $H_0 : \phi = 0$ versus the alternative $H_a : \phi \neq 0$. The test statistic at interim analysis $k,\ k = 1, ..., K$, is expressed as

$$Z_k = \frac{\hat{\mu}_{1,k} - \hat{\mu}_{2,k}}{\sqrt{\frac{\hat{\sigma}_{1,k}^2}{m_{1,k}} + \frac{\hat{\sigma}_{2,k}^2}{m_{2,k}}}} \sim N(0, 1) \tag{2.2}$$

approximately under H_0, where

$$\hat{\sigma}_{j,k}^2 = \frac{1}{m_{j,k} - 1} \sum_{i=1}^{m_{j,k}} (Y_{i,j} - \hat{\mu}_{j,k})^2$$

is the sample variance for arm j at look k.

For optimal response-adaptive randomisation, the cumulative number of patients on treatment j at look k, $m_{j,k}$, is random. However, the allocation proportions converge almost surely to the pre-specified optimal allocation proportions derived

based on some optimality criterion (Hu and Rosenberger, 2003). More precisely, let ρ_j be the target optimal allocation proportion for treatment j. Previous work has shown that M_j/N converges to ρ_j almost surely for a fixed-sample design. Hence, M_j can be approximated by $\rho_j N$.

For group-sequential optimal response-adaptive designs, where repeated analyses are conducted using the cumulative responses, we have $m_{j,k} = \rho_j n_k$ approximately. Since ρ_j is usually a function of the unknown parameters, let $\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \hat{\rho}_2)$ be the optimal allocation evaluated based on the responses obtained so far. Then (2.2) is approximately

$$Z_k = \frac{\hat{\mu}_{1,k} - \hat{\mu}_{2,k}}{\sqrt{\frac{\hat{\sigma}^2_{1,k}}{\lceil \hat{\rho}_1 n_k \rceil} + \frac{\hat{\sigma}^2_{2,k}}{\lceil \hat{\rho}_2 n_k \rceil}}}.$$

Some commonly-used optimal allocations will be introduced in Section 2.1.2. Here, the required cumulative sample size at look k, n_k, to achieve the specified information level can be obtained from (2.1).

In practice, the actual observed number of patients is used for $m_{j,k}$ in (2.2). Unequal increments in the observed information levels may occur. Critical boundaries that control the overall type I error rate can be obtained by using the error-spending approach (Lan and DeMets, 1983), which spends the type I error rate as a function of the information time. More details will be given later.

Binary responses

For binary endpoints, the responses are dichotomous rather than continuous. For instance, the responses may indicate whether or not a patient responds to a given treatment. The response for patient i on treatment j, $Y_{i,j}$, is $\text{Bin}(1, p_j)$. Here, the parameter p_j is the probability of success on treatment j, and $q_j = 1 - p_j$ is the failure rate. For arm j at look k, we have the estimates

$$\hat{p}_{j,k} = \frac{\sum_{i=1}^{m_{j,k}} Y_{i,j}}{m_{j,k}} \qquad \text{and} \qquad \hat{q}_{j,k} = 1 - \hat{p}_{j,k}.$$

Consider testing $H_0 : \phi = 0$ versus $H_a : \phi \neq 0$, where $\phi = p_1 - p_2$, to compare the two proportions of success. The test statistic at look k can be written as

$$Z_k = \frac{\hat{p}_{1,k} - \hat{p}_{2,k}}{\sqrt{\frac{\hat{p}_{1,k}\hat{q}_{1,k}}{m_{1,k}} + \frac{\hat{p}_{2,k}\hat{q}_{2,k}}{m_{2,k}}}} = \frac{\hat{p}_{1,k} - \hat{p}_{2,k}}{\sqrt{\frac{\hat{p}_{1,k}\hat{q}_{1,k}}{\lceil \hat{\rho}_1 n_k \rceil} + \frac{\hat{p}_{2,k}\hat{q}_{2,k}}{\lceil \hat{\rho}_1 n_k \rceil}}}. \tag{2.3}$$

Under H_0, the marginal distribution of Z_k in (2.3) is asymptotically standard normal for large sample sizes using the Central Limit Theorem.

Joint distribution of the sequence of test statistics

A common form of the joint distribution of $\{Z_1, ..., Z_K\}$ has been derived by Jennison and Turnbull (2000), which is called the canonical joint distribution and stated as follows. Given information levels $\{\mathcal{I}_1, ..., \mathcal{I}_K\}$, (i) $\{Z_1, ..., Z_K\}$ is multivariate normal, (ii) $E(Z_k) = \phi\sqrt{\mathcal{I}_k}$ and (iii) $\text{cov}(Z_{k_1},\ Z_{k_2}) = \sqrt{\mathcal{I}_{k_1}/\mathcal{I}_{k_2}},\ 1 \leq k_1 \leq k_2 \leq K$. Let $\hat{\phi}_k$ be the parameter estimate at look k, so that $\mathcal{I}_k = \{\text{var}(\hat{\phi}_k)\}^{-1}$. This form applies exactly for normal responses with known variances and approximately for other types of endpoints, including right-censored survival responses.

For a two-sample normal test using an adaptive sampling rule, which seeks to balance the competing goals of lowering the number of patients allocated to the inferior treatment and reducing the expected total sample size, Jennison and Turnbull (2001) showed that the standard form of the canonical joint distribution still holds, provided that the group sizes are computed to satisfy the given information levels.

Consider two-armed trials using optimal response-adaptive randomisation, which

can target any specified optimal allocation, and allow an interim analysis to be taken at a continuous information time. Zhu and Hu (2010) proved that the sequence of test statistics converges to Brownian motion in distribution and that the joint distribution asymptotically satisfies the canonical joint distribution. More specifically, they showed that (i) $\{Z_1, ..., Z_K\}$ is multivariate normal, (ii) $E(Z_k) = \mu\sqrt{Nt_k}$ and (iii) $\text{cov}(Z_{k_1},\ Z_{k_2}) = \sqrt{t_{k_1}/t_{k_2}},\ 0 \leq t_{k_1} \leq t_{k_2} \leq 1$, where, for normal responses, $\mu = (\mu_1 - \mu_2)/\sqrt{\sigma_1^2/\rho_1 + \sigma_2^2/\rho_2}$, N is the maximum sample size and t_k is the information time at look k. As the canonical joint distribution still holds asymptotically for group-sequential response-adaptive designs, the required error probabilities can also be achieved using the same approach as for a non-adaptive randomised design.

The error-spending approach

The error-spending approach can be used to maintain the error rate at the specified value for any observed information sequence $\{\mathcal{I}_1, ..., \mathcal{I}_K\}$, provided that $\mathcal{I}_k$ is conditionally independent of previous parameter estimates $\{\hat{\phi}_1, ..., \hat{\phi}_{k-1}\}$ given $\{\mathcal{I}_1, ..., \mathcal{I}_{k-1}\}$. Otherwise, the standard form of the canonical joint distribution will fail to hold. There are two types of error-spending functions, the α-spending function and the β-spending function which are used to control the type I and type II error rates, respectively. Throughout this thesis, we consider the former error-spending function.

An α-spending function, $\alpha(t_k)$, represents how much of the cumulative type I error rate is to be spent at information time $t_k = \mathcal{I}_k/\mathcal{I}_K$. It is a continuous and monotonically non-decreasing function with $\alpha(0) = 0$ and $\alpha(1) = \alpha$. Proschan et al. (2006) discussed three such functions. One approximates the O'Brien and Fleming boundaries (O'Brien and Fleming, 1979), one approximates Pocock's boundaries (Pocock, 1977) and the third one is the linear α-spending function. We have

$$\begin{aligned}\alpha_{O-F}(t_k) &= 2\{1-\Phi(z_{\alpha/2}/\sqrt{t_k})\},\\ \alpha_P(t_k) &= \alpha\log\{1+(e-1)t_k\},\\ \alpha_L(t_k) &= \alpha t_k,\end{aligned} \tag{2.4}$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$ and Φ denotes the standard normal distribution function.

The O'Brien and Fleming boundaries spend little type I error probability during the early stages of a trial, and, if the last look is reached, the type I error rate will be close to that of a fixed-sample design. In contrast, Pocock's boundaries are more likely to reject H_0 earlier, and hence have a large critical value at the final stage to attain the overall type I error rate. The linear boundaries are, in general, between the two. The critical boundaries to be used need to be pre-specified in the study protocol.

Based on the joint distribution of the sequence of test statistics, the critical boundaries $\{c_1, ..., c_K\}$ can be calculated recursively using the equation

$$P_{\phi=0}(|Z_1| < c_1, ..., |Z_{k-1}| < c_{k-1}, |Z_k| \geq c_k) = \alpha(t_k) - \alpha(t_{k-1}).$$

For $k = 1$, the type I error probability to be spent is $P_{\phi=0}(|Z_1| \geq c_1) = \alpha(t_1)$. The critical boundary c_1 can be easily obtained by inverting the standard normal distribution function. If the trial progresses, the probability of crossing the stopping boundary at the second look is $P_{\phi=0}(|Z_1| < c_1, |Z_2| \geq c_2) = \alpha(t_2) - \alpha(t_1)$, where Z_1 and Z_2 follow a bivariate normal distribution. By integrating out Z_1 from the joint distribution of Z_1 and Z_2, c_2 can be obtained. Similarly, the critical boundaries for the kth interim analysis, $c_k,\ k > 2$, are computed through integration of a multivariate normal distribution. This method does not require the number of group sequential analyses, K, to be pre-specified.

Stopping rules

Under group-sequential monitoring, a decision for early termination can be made if there is sufficient evidence of a treatment effect or futility at interim analyses. Otherwise, the trial proceeds to recruit more participants for more information until it reaches the end of the trial. For a two-sided group-sequential monitoring trial, the stopping rules are given below.

• For $k = 1, ..., K-1$, stop the trial and reject the null hypothesis if $|Z_k| \geq c_k$; otherwise, continue to the next interim analysis.
• For $k = K$, reject H_0 if $|Z_k| \geq c_k$, and accept H_0 otherwise.

2.1.2. Optimal response-adaptive randomisation

Optimal allocations

Optimal response-adaptive randomised designs aim to target the pre-specified optimal allocation derived based on some optimality criterion. Some common optimal allocations for testing treatment differences with immediate responses are introduced below. These optimal allocations were derived using a fixed-sample design with one analysis conducted at the end of the trial, which is the case $K = 1$ in Section 2.1.1. The pre-specified optimal allocation should remain unchanged regardless of the number of interim analyses.

Neyman allocation

Neyman allocation minimises the total sample size under a variance constraint. For normal responses, we wish to minimise $N = M_1 + M_2$ with respect to M_1, subject to a fixed variance, $\sigma_1^2/M_1 + \sigma_2^2/(N - M_1) = C$, where C is a constant. The solution obtained by Jennison and Turnbull (2000) using a loss function specialised to the sum of the sample sizes on each treatment is

$$\rho_1 = \frac{\sigma_1}{\sigma_1 + \sigma_2} \quad \text{and} \quad \rho_2 = 1 - \rho_1. \tag{2.5}$$

For binary endpoints, under a variance constraint, $p_1q_1/M_1 + p_2q_2/(N - M_1) = C$, the solution is

$$\rho_1 = \frac{\sqrt{p_1q_1}}{\sqrt{p_1q_1} + \sqrt{p_2q_2}} \quad \text{and} \quad \rho_2 = 1 - \rho_1.$$

Neyman allocation has the merit of efficiency. It maximises the power of a two-sample Z test for a fixed sample size. However, the solution is not always ethical. The most efficient allocation may assign more patients to the inferior treatment.

Optimal allocation

For binary responses, an ethical optimal criterion is to minimise the total expected number of failures, $q_1M_1 + q_2M_2$, under a variance constraint. Rosenberger et al. (2001) obtained the solution as

$$\rho_1 = \frac{\sqrt{p_1}}{\sqrt{p_1} + \sqrt{p_2}} \quad \text{and} \quad \rho_2 = 1 - \rho_1. \tag{2.6}$$

There are other optimal allocations derived based on different optimality criteria. Here, we focus on these two widely-used optimal allocations, Neyman allocation for normal responses and optimal allocation for binary responses.

The optimal allocations (2.5) and (2.6) depend on the unknown parameters. In practice, the current parameter estimates are used, which are updated after each response observed.

Optimal response-adaptive randomisation procedures

Response-adaptive randomisation assigns patients according to previous treatment allocations and responses. Permuted-block randomisation can be used early on to obtain initial parameter estimates. This method balances the sample sizes across the treatment groups. For a two-treatment comparison, block sizes of

$\{2, 4, 6, 8, ...\}$ can be chosen. The sample sizes for the treatment groups are equal within each block. For three-treatment comparisons, block sizes of $\{3, 6, 9, ...\}$ can be used. A more unpredictable allocation can be achieved by randomly selecting the block size. After obtaining initial parameter estimates, the two optimal response-adaptive randomisation procedures below can be implemented.

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 optimal allocation proportions for treatment j, $j = 1, 2$, evaluated based on the responses available. Eisele and Woodroofe (1995) proposed a response-adaptive allocation probability, which is a function of the current and optimal allocation proportions. The probability that the $(i+1)$th patient will be assigned to treatment 1 is

$$g_1 = \begin{cases} \dfrac{\hat{\rho}_1^{(i)} \left\{ \dfrac{\hat{\rho}_1^{(i)}}{m_1^{(i)}/i} \right\}^{\gamma}}{\hat{\rho}_1^{(i)} \left\{ \dfrac{\hat{\rho}_1^{(i)}}{m_1^{(i)}/i} \right\}^{\gamma} + \hat{\rho}_2^{(i)} \left\{ \dfrac{\hat{\rho}_2^{(i)}}{m_2^{(i)}/i} \right\}^{\gamma}}, & \text{if } \ 0 < m_j^{(i)}/i < 1, \\ 1 - m_1^{(i)}/i, & \text{if } \ m_1^{(i)}/i = 0, 1, \end{cases} \tag{2.7}$$

where $0 \leq \gamma \leq \infty$ is a constant that determines the degree of randomness of the allocation procedure. The procedure is the most random when $\gamma = 0$. In this case, $g_1 = \hat{\rho}_1^{(i)}$ if $0 < m_j^{(i)}/i < 1$, which corresponds to the sequential maximum likelihood estimation procedure (Melfi and Page, 1998). The randomisation procedure is the most deterministic when γ approaches infinity. The design is then the same as Thompson's (1933) procedure, where $g_1 = 1$ if $m_1^{(i)}/i < \hat{\rho}_1^{(i)}$ and $g_1 = 0$ if $m_1^{(i)}/i \geq \hat{\rho}_1^{(i)}$. Many studies use $\gamma = 2$, which can achieve a high power while allowing a reasonable degree of randomness.

Efficient randomised-adaptive design (ERADE)

Similar to the DBCD function, the allocation probability function for the ERADE

depends on the current allocation proportion $m_j^{(i)}/i$ and the estimated target allocation proportion $\hat{\rho}_j^{(i)},\ j = 1, 2$. However, the ERADE function (Hu et al., 2009) is discontinuous. The probability that the next patient will be assigned to treatment 1 is

$$g_1 = \begin{cases} \gamma' \hat{\rho}_1^{(i)}, & \text{if } m_1^{(i)}/i > \hat{\rho}_1^{(i)}, \\ \hat{\rho}_1^{(i)}, & \text{if } m_1^{(i)}/i = \hat{\rho}_1^{(i)}, \\ 1 - \gamma' \{1 - \hat{\rho}_1^{(i)}\}, & \text{if } m_1^{(i)}/i < \hat{\rho}_1^{(i)}, \end{cases} \tag{2.8}$$

where $0 \leq \gamma' < 1$ is a constant that controls the degree of randomisation. The ERADE allocation procedure becomes more deterministic when γ' approaches zero. A value of γ' between 0.4 and 0.7 is recommended (Hu et al., 2009).

The allocation probability for treatment 1, g_1, using (2.7) or (2.8) is updated sequentially after each response observed.

Asymptotic properties of optimal response-adaptive randomisation

Response-adaptive randomisation procedures are often compared in terms of optimality, variability and power. Previous work has shown that the limiting allocation proportions using optimal response-adaptive randomisation converge to the target optimal allocation proportions almost surely (Hu and Rosenberger, 2003; Zhang and Rosenberger, 2006). Optimality can be achieved by both the DBCD and the ERADE with reasonably small variability in the allocation proportions. In particular, the ERADE generally has a lower variability than the DBCD. Hu et al. (2006) derived an asymptotic Cramér-Rao lower bound for the variance of the allocation proportions. The DBCD has been shown to attain the lower bound only when $\gamma \to \infty$ (Zhang and Rosenberger, 2006). However, the ERADE has been proved to always attain the lower bound (Hu et al., 2009).

Some previous work found that the use of optimal response-adaptive randomisa-

tion increases the variance of the allocation proportions, which results in an adverse effect on the power of tests (Melfi and Page, 1998; Hu and Rosenberger, 2003). However, the results obtained by Zhu and Hu (2010) and Tymofyeyev et al. (2007) show an increase in power. It is argued that the use of optimal response-adaptive randomisation can lower the expected number of treatment failures without a loss of power, or even lead to a higher power compared to complete randomisation.

In terms of the convergence rate, Zhang (2016) indicated that the ERADE may not converge as fast as the DBCD in some situations. This may be because the ERADE function is discontinuous for two-armed clinical trials, which can be less stable as the allocation probabilities jump from one value to another.

2.2. Simulation studies for immediate responses

In this section, the simulation setting is similar to that of Zhu and Hu (2010). However, more randomisation procedures are compared and different information sequences considered.

2.2.1. Normal responses

Suppose that the maximum number of patients is $N = 500$. Permuted-block randomisation is used for the first 10% of the maximum sample size to obtain initial parameter estimates. Then the optimal response-adaptive randomisation procedures, the DBCD and the ERADE, are applied using $\gamma = 2$ and $\gamma^{'} = 0.5$, respectively. Neyman allocation is used as the target optimal allocation. For comparison with a non-adaptive randomised design, the results for complete randomisation (CR) are included.

For group sequential analysis, there are $K = 3$ sequential tests planned at the

unequally spaced information times (0.2, 0.5, 1) and (0.5, 0.8, 1). The former conducts early interim looks, whereas the latter has interim analyses after reaching half of the maximum information level. The overall type I error rate $\alpha = 0.05$ is set, and the O'Brien and Fleming critical boundaries derived by the error-spending approach are used. The critical boundaries for the tests taken at the two information sequences are (4.877, 2.963, 1.969) and (2.963, 2.266, 2.028), respectively. Results of fixed-sample designs based on N patients are provided alongside for comparison, which are a special case of the group sequential design with $K = 1$ and $t_K = 1$. The critical value in this case is 1.960.

The designs are compared in terms of the error probabilities, the expected number of patients (ENP), the average allocation proportion for treatment 1 and the corresponding variability. Rather than being adjusted to get a particular information, the maximum number of patients N is fixed for all of the designs. The results are based on 5,000 replicates.

Table 2.1.: Simulated type I error rate for two-armed normal trials with Neyman allocation in group sequential and fixed-sample designs, $\mu_1 = \mu_2 = 1,\ \sigma_1 = 1,\ \sigma_2 = 2,\ N = 500$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)				
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.048	499.0	(15.8)	0.500	(0.021)
DBCD	0.046	499.5	(11.2)	0.334	(0.019)
ERADE	0.040	499.4	(12.2)	0.334	(0.015)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)				
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.053	497.2	(19.6)	0.500	(0.021)
DBCD	0.057	496.6	(21.8)	0.334	(0.019)
ERADE	0.044	497.5	(18.9)	0.334	(0.015)
	Fixed-sample design				
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.048	500	(0)	0.501	(0.021)
DBCD	0.048	500	(0)	0.334	(0.019)
ERADE	0.051	500	(0)	0.334	(0.015)

As can be seen from Table 2.1, all of the designs can well attain the pre-specified value of the type I error rate, 0.05, with usually a discrepancy of less than three standard errors. More specifically, $\tilde{\alpha}$ lies within (0.041,0.059), except for the ERADE when $(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1), where $\tilde{\alpha}$ just falls outside of this range. Under the null hypothesis, the ENP is similar for all of the designs, since the chance of early termination is small. For the optimal response-adaptive randomised designs, the target Neyman allocation proportion for treatment 1 is $\rho_1 = 0.333$ from (2.5). Both the DBCD and the ERADE target ρ_1 well, with the ERADE having a lower standard deviation for the allocation proportion.

Table 2.2.: Simulated power for two-armed normal trials with Neyman allocation in group sequential and fixed-sample designs, $\mu_1 = 1.4,\ \mu_2 = 1,\ \sigma_1 = 1,\ \sigma_2 = 2,\ N = 500$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)				
Procedure	Power	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.796	458.1	(93.4)	0.500	(0.023)
DBCD	0.847	450.2	(99.9)	0.335	(0.021)
ERADE	0.838	450.9	(99.4)	0.334	(0.016)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)				
Procedure	Power	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.797	416.2	(86.1)	0.500	(0.023)
DBCD	0.837	404.5	(88.4)	0.334	(0.022)
ERADE	0.826	407.4	(88.3)	0.334	(0.017)
	Fixed-sample design				
Procedure	Power	ENP	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.805	500	(0)	0.501	(0.021)
DBCD	0.856	500	(0)	0.334	(0.019)
ERADE	0.855	500	(0)	0.334	(0.015)

Under the alternative hypothesis, from Table 2.2, a higher power is achieved by the optimal response-adaptive randomised designs while using a lower ENP. For instance, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), the DBCD has around a 4% higher power than CR, and reduces the ENP by about 12. For the group sequential designs which allow early stopping, the ENP is significantly decreased compared to the fixed-sample designs. However, the corresponding standard deviation of the ENP

is large, since there are only three possible values that the ENP can take. More specifically, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), the number of patients can be 250, 400 or 500 when the trial stops at information time $t_1,\ t_2$ or t_3, respectively. The frequencies of the first two values are 807 and 2,174 for CR, 986 and 2,312 for the DBCD, and 948 and 2,259 for the ERADE. The target Neyman allocation proportion is well achieved for both optimal response-adaptive randomised designs, with the ERADE consistently having a lower variability in the allocation proportion.

2.2.2. Binary responses

For binary responses, the optimal allocation derived by Rosenberger et al. (2001) is used as the target optimal allocation. In addition, the expected number of failures (ENF) is computed at the time when a decision is made. If a trial stops early, the rest of the patients are assigned to the more promising treatment and the total expected number of failures (ENF$'$) is obtained. Here, ENF$'$ is provided to compare with the ENF for the fixed-sample designs based on the maximum sample size N. In practice, trials stop when a decision is made. The other simulation settings are the same as in Section 2.2.1.

Under the null hypothesis, from Table 2.3, the type I error rate for all of the designs is less than three standard errors away from 0.05, except for CR with $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), where a slightly conservative type I error rate, $\tilde{\alpha} = 0.040$, is obtained. The differences in the ENP and the ENF between the group sequential and the fixed-sample designs are small under H_0. Both optimal response-adaptive designs target well the optimal allocation proportion for arm 1 obtained from (2.6). Here, the target allocation proportion for treatment 1 is $\rho_1 = 0.5$ under the null hypothesis. The ERADE consistently has a lower standard deviation for $\tilde{\rho}_1$ among all designs.

Under the alternative hypothesis, from Table 2.4, the expected numbers of pa-

Table 2.3.: Simulated type I error rate for two-armed binary trials with optimal allocation in group sequential and fixed-sample designs, $p_1 = p_2 = 0.5$, $N = 500$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.046	499.2	(14.8)	249.6	(13.6)	0.500	(0.021)
DBCD	0.048	499.2	(14.1)	249.9	(13.4)	0.500	(0.016)
ERADE	0.048	499.1	(15.4)	249.7	(13.7)	0.500	(0.012)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.040	497.4	(18.7)	248.8	(14.6)	0.500	(0.021)
DBCD	0.050	496.8	(21.0)	248.3	(15.4)	0.501	(0.016)
ERADE	0.050	497.0	(20.5)	248.6	(15.1)	0.500	(0.012)
	Fixed-sample design						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.053	500	(0)	250.1	(11.0)	0.500	(0.022)
DBCD	0.053	500	(0)	250.1	(11.0)	0.500	(0.016)
ERADE	0.059	500	(0)	250.1	(11.0)	0.500	(0.011)

tients and failures can be significantly reduced using group sequential analysis compared to the fixed-sample designs, despite the small treatment difference. For example, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), around 87 fewer patients on average are used, and there is a reduction of about 37 in the ENF compared to the fixed-sample designs. The power and the ENP for the response-adaptive designs are similar to CR. However, there are about two fewer failures on average achieved by the response-adaptive designs compared to CR. In addition, the optimal response-adaptive randomisation procedures assign more patients to the better-performing treatment. About 53% of the patients are allocated to treatment 2 with the higher probability of success 62.5%.

Simulation results for other scenarios where the difference in the probabilities of success is larger show similar conclusions. For binary responses, when the treatment difference increases, the sample size decreases considerably. Only a small number of failures is saved by using the response-adaptive designs instead of CR. Also, the power and the ENP are comparable for CR, and the DBCD and the ERADE.

Table 2.4.: Simulated power for two-armed binary trials with optimal allocation in group sequential and fixed-sample designs, $p_1 = 0.5,\ p_2 = 0.625,\ N = 500$.

		$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.810	454.8	(96.4)	199.1	(43.4)	216.0	(12.1)	0.500	(0.023)
DBCD	0.809	454.6	(96.7)	197.3	(43.4)	214.3	(12.1)	0.470	(0.017)
ERADE	0.810	455.4	(96.0)	197.7	(43.1)	214.5	(11.9)	0.470	(0.013)
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.799	413.8	(87.9)	181.1	(39.7)	213.5	(11.5)	0.500	(0.024)
DBCD	0.797	413.1	(88.4)	179.4	(40.1)	212.0	(11.5)	0.471	(0.017)
ERADE	0.796	413.8	(87.0)	179.6	(39.3)	211.9	(11.3)	0.470	(0.012)
		Fixed-sample design							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	-	-	$\tilde{\rho}_1$	(s.d.)
CR	0.814	500	(0)	218.9	(11.2)	-	-	0.500	(0.022)
DBCD	0.813	500	(0)	217.3	(11.2)	-	-	0.472	(0.015)
ERADE	0.814	500	(0)	217.2	(11.1)	-	-	0.472	(0.010)

2.2.3. Redesigning a placebo-controlled clinical trial

A multi-centre, placebo-controlled trial that investigated the efficacy of zidovudine in reducing the risk of maternal-infant HIV transmission was studied by Connor et al. (1994). A total number of 477 HIV-infected women were randomly assigned to the experimental treatment or the placebo group with equal probabilities. Of the 239 pregnant women receiving the experimental treatment, 8.3% of the infants were HIV-infected, whereas 25.5% of the 238 from the placebo group were diagnosed as HIV positive.

A redesign of the placebo-controlled HIV trial using a group-sequential response-adaptive design was investigated by Zhu and Hu (2010). The success probabilities for the experimental treatment and the control group were assumed to be $p_1 = 0.917$ and $p_2 = 0.745$, respectively. However, the authors focused on the DBCD. Here, a comparison of different designs is made. Three group sequential tests are planned at information times (0.2, 0.5, 1). The critical boundaries derived

based on the linear spending function are used. The other simulation settings are the same as in Section 2.2.2.

Table 2.5.: Simulated type I error rate for redesigning a two-armed binary trial with optimal allocation, $p_1 = p_2 = 0.745,\ N = 477$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.057	468.4	(50.8)	119.3	(16.0)	0.500	(0.022)
DBCD	0.055	469.1	(48.5)	119.5	(15.5)	0.500	(0.014)
ERADE	0.054	469.9	(45.9)	119.8	(15.0)	0.500	(0.008)

Table 2.5 shows that, under the null hypothesis, the critical boundaries can control the overall type I error rate for all of the designs. The ENP, the ENF and $\tilde{\rho}_1$ are similar for the response-adaptive and non-adaptive designs.

Table 2.6.: Simulated power for redesigning a two-armed binary trial with optimal allocation, $p_1 = 0.917,\ p_2 = 0.745,\ N = 477$.

Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.9992	209.8	(111.0)	35.4	(18.6)	57.6	(10.1)	0.499	(0.031)
DBCD	0.9992	212.0	(111.9)	34.9	(18.3)	56.9	(9.7)	0.526	(0.020)
ERADE	0.9996	211.0	(107.7)	34.6	(17.6)	56.7	(9.4)	0.529	(0.011)

Under the alternative hypothesis, from Table 2.6, a high power is achieved for all of the designs. In Zhu and Hu (2010), the powers are 0.999 and 0.997 for CR and the DBCD, respectively. This means that the difference in treatment effects can be easily detected. The optimal allocation (Rosenberger et al., 2001) aims to minimise the total expected number of failures. However, here just marginal decreases in the ENF and the ENF$'$ are shown for the DBCD and the ERADE compared to CR. Although the response-adaptive designs achieve a lower number of failures while using a slightly larger number of patients, there is not much gain compared to the CR design. Hence, in this case, use of the group-sequential CR design is preferred.

2.3. Extension to censored survival responses

2.3.1. Information time

Maximum information trials consider a pre-determined target information level. The information level for survival responses is proportional to the number of events. A trial continues until the maximum information level is achieved. However, for survival responses, the outcomes such as death are often unpredictable, and depend on the recruitment rate, censoring and so on. A trial may not achieve the required information level at the end of the study, or the information level may be attained soon after the trial begins. In these cases, a maximum duration trial, where the maximum length of the trial is pre-determined, may be more feasible in practice.

For maximum duration trials, the number of events at the final look is not known until the trial reaches the end of the study. Hence, at each interim analysis, a predicted value for the final information level evaluated at look k, $\hat{\mathcal{I}}_K^{(k)},\ k = 1, ..., K$, is needed. Then the information time for survival responses at group sequential test k can be expressed as

$$t_k = \frac{\mathcal{I}_k}{\hat{\mathcal{I}}_K^{(k)}} = \frac{e_k}{\hat{e}_K^{(k)}}, \tag{2.9}$$

where e_k is the observed number of events at look k and $\hat{e}_K^{(k)}$ is the expected total number of events evaluated at that look (Jennison and Turnbull, 2000).

Kim et al. (1995) considered

$$t_k = \begin{cases} \dfrac{e_k}{\hat{e}_K^{(k)}}, & \text{if } k < K \text{ and } e_k \leq \hat{e}_K^{(k)}, \\ 1, & \text{otherwise.} \end{cases} \tag{2.10}$$

They explained that the total expected number of events can be estimated based on the assumed survival model. However, there are two candidates for the esti-

mate of e_K. One is under the null hypothesis of no treatment difference and the other is based on the specified alternative hypothesis. Kim et al. (1995) showed that the overall type I error rate can be preserved under both hypotheses for a logrank test. The power depends on the actual information level obtained.

For parametric tests, (2.9) can be approximated by

$$t_k = \frac{\hat{e}_k}{\hat{e}_K^{(k)}} = \frac{\sum_{j=1}^{2} m_{j,k}\hat{\epsilon}_{j,k}}{\sum_{j=1}^{2} m_{j,K}\hat{\epsilon}_{j,K}}, \quad k = 1, ..., K,$$

where $m_{j,k}$ is the cumulative sample size for treatment j at look k and $\hat{\epsilon}_{j,k}$ is the probability of an event for treatment j evaluated at look k, which depends on the assumed models. For uniformly distributed arrival and censoring times, and exponentially distributed survival time with mean θ_j, the probability of an event is

$$\epsilon_{j,k} = 1 - \frac{\theta_j}{D}\left\{1 + \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\} - \frac{\theta_j}{Dt_k}\left(1 - \frac{2\theta_j}{D}\right)\left\{1 - \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\}, \quad (2.11)$$

where D is the maximum duration of the trial. Details of the derivation are provided in Appendix A. The probability of an event increases as the length of the trial is increased. Also, for group sequential designs, $\epsilon_{j,k}$ is larger at later looks than at early ones. More specifically, as D and/or t_k increase, the probability of an event is increased. Since the $\epsilon_{j,k}$ are functions of unknown parameters, for the first look when $k = 1$, initial parameter estimates from a previous study or obtained in the learning phase using permuted-block randomisation can be used. Then the parameter estimates based on the cumulative responses can be used for $k \geq 2$.

As mentioned above, there are two candidates for the estimated number of events, which result in two information time scales. The type I error rate can be guaranteed by using either information time scale (Kim et al., 1995). For simplicity, we

consider the estimated number of events, $\hat{e}_k$, under the assumption that the null hypothesis is true ($\theta_1 = \theta_2$). Then the subscript j for $\epsilon_{j,k}$ denoting treatment can be suppressed. The approximate information time at look k becomes

$$t_k = \frac{\hat{e}_k}{\hat{e}_K^{(k)}} = \frac{\sum_{j=1}^{2} m_{j,k}\hat{\epsilon}_k}{\sum_{j=1}^{2} m_{j,K}\hat{\epsilon}_K} = \frac{n_k\hat{\epsilon}_k}{N\hat{\epsilon}_K} \in (0,1], \quad k = 1, ..., K. \tag{2.12}$$

In (2.9) and (2.10), $t_k = 1$ can occur if $k \leq K$. However, in (2.12), $t_k = 1$ occurs only when $k = K$. That is, for a maximum duration trial, the type I error rate will be spent only when the trial reaches the maximum length of the study. The error-spending approach can be used to control the overall type I error rate. Now suppose that we wish to conduct the first interim analysis when about one third of the expected total number of events is obtained. Then the first interim analysis is planned at $t_1 = 1/3$, and, from (2.12), $n_1 = \lceil t_1 N\hat{\epsilon}_K/\hat{\epsilon}_1 \rceil$ is the approximate number of patients needed at the first look.

Here, $\hat{\epsilon}_k$, $k = 1, ..., K$, is an estimate based on the sample. The accuracy of the parameter estimates increases when the sample size is increased. At early looks with small sample sizes, the approximate information level in (2.12) can be inaccurate. Any deviations between the observed and target information levels may affect the type I error rate. However, the use of the O'Brien and Fleming test can alleviate the issue, since little type I error rate is spent during the early stages. In addition, related work of Proschan et al. (1992) and Jennison and Turnbull (2000) investigated the effect of applying critical boundaries derived based on equally spaced information levels to the actual observed unequal information sequence. It was found that the maximum increase in the type I error rate seems to be robust under various scenarios of departure from equal group sizes.

2.3.2. Model assumptions

The proposed group-sequential response-adaptive design for survival responses allows staggered entry and takes into account the issue of right-censoring. Let D be the length of a maximum duration trial. The information times $t_0 = 0$ and $t_K = 1$ refer to the commencement and the end of the trial, respectively. Suppose that group sequential tests take place at information time $t_k \in (0, 1],\ k = 1, ..., K$. Then the calendar time at which the kth interim analysis occurs can be expressed as Dt_k. Assume that patient arrival time is uniformly distributed. The arrival time for patient i who arrived before or at the kth look is $A_i \sim U(0,\ Dt_k)$. Assume that the survival time for patient i on treatment j, $S_{i,j}$, follows an exponential distribution with mean θ_j, where $\theta_j > 0$. Then the density function of $S_{i,j}$ is

$$f(s_{i,j}; \theta_j) = \frac{1}{\theta_j} \exp\left(-\frac{s_{i,j}}{\theta_j}\right) \quad \text{for } s_{i,j} > 0.$$

The survival function, which is the probability that the time of an event will be later than $s_{i,j}$, is

$$P(S_{i,j} > s_{i,j}) = \exp\left(-\frac{s_{i,j}}{\theta_j}\right),$$

and the hazard rate for treatment j is θ_j^{-1}. In addition, the censoring time for patient i, C_i, is assumed to be uniformly distributed from zero to D. Here, the treatment groups are assumed to have the same arrival and censoring time distributions. Patients' arrival, survival and censoring times are assumed to be independent of each other.

Under the above model assumptions, the observed survival response for patient $i,\ i = 1, ..., m_{j,k}$, on treatment $j,\ j = 1, 2$, at group sequential test $k,\ k = 1, ..., K$, can be expressed as $Y_{i,j,k} = \min(S_{i,j},\ C_i,\ Dt_k - A_i)$. Here, the duration of the trial, D, and the arrival time of patient i, A_i, start from the beginning of the study, while the survival time $S_{i,j}$ and the censoring time C_i commence from the arrival of that patient. For example, suppose that the number of group sequential tests

is $K = 3$. At the first interim analysis, we have $Y_{i,j,1} = \min(S_{i,j},\ C_i,\ Dt_1 - A_i)$, where $A_i \sim U(0,\ Dt_1)$.

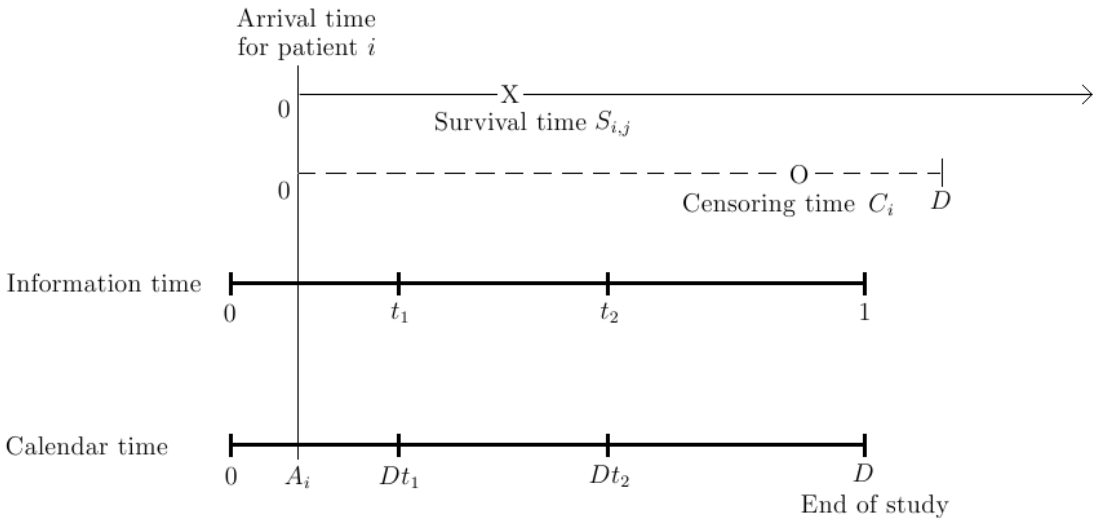

Figure 2.1.: An example of a patient's arrival time, survival time and censoring time

If $Y_{i,j,1} = S_{i,j}$, then the patient's response is an event. If $Y_{i,j,1} = C_i$, then the patient's response is right-censored due to loss to follow up. If $Y_{i,j,1} = Dt_1 - A_i$, then the outcome is right-censored because the patient has not yet responded at the first look. The patient's response will then be followed up at later looks. Let $E_{i,j} = \min(S_{i,j}, C_i)$. If the response occurs between the first and the second looks, then $Dt_1 < A_i + E_{i,j} \leq Dt_2$ and we have $Y_{i,j,2} = E_{i,j}$. An example is shown in Figure 2.1, where $E_{i,j} = S_{i,j}$ and $Y_{i,j,2} = S_{i,j}$. If $Dt_2 < A_i + E_{i,j} \leq Dt_3$, then we have $Y_{i,j,2} = Dt_2 - A_i$ and $Y_{i,j,3} = E_{i,j}$. If the patient has not responded by the end of the trial, then $A_i + E_{i,j} > Dt_3$, $Y_{i,j,2} = Dt_2 - A_i$ and $Y_{i,j,3} = Dt_3 - A_i$. Similarly, at the second interim test, we have $Y_{i,j,2} = \min(S_{i,j},\ C_i,\ Dt_2 - A_i)$, where $A_i \sim U(0,\ Dt_2)$. If $Y_{i,j,2} = Dt_2 - A_i$, then we follow up the response at the next look. If $Dt_2 < A_i + E_{i,j} \leq Dt_3$, then $Y_{i,j,3} = E_{i,j}$. Otherwise, $Y_{i,j,3} = Dt_3 - A_i$. For the final look, however, no response will be followed up.

Likelihood function, point estimate and variance

Suppose that two independent random samples $\{y_{i,j,k}, \delta_{i,j,k}, i = 1, \ldots, m_{j,k}\}$ for treatment $j,\ j = 1, 2$, are obtained. Here, $\delta_{i,j,k} = 1$ if the response of patient i on arm j at look k is an event and $\delta_{i,j,k} = 0$ if the response is censored. Under the above model, the likelihood function of θ_j for treatment j based on the responses obtained so far can be expressed as

$$L(\theta_j) = \prod_{i=1}^{m_{j,k}} \left\{ \frac{1}{\theta_j} \exp\left(-\frac{y_{i,j,k}}{\theta_j}\right) \right\}^{\delta_{i,j,k}} \exp\left(-\frac{y_{i,j,k}}{\theta_j}\right)^{1-\delta_{i,j,k}}, \tag{2.13}$$

where the first part of the product in (2.13) refers to the survival times and the second part represents the censoring times. The log-likelihood function is

$$l(\theta_j) = \log L(\theta_j) = -r_{j,k} \log(\theta_j) - \frac{\sum_{i=1}^{m_{j,k}} y_{i,j,k}}{\theta_j},$$

where $r_{j,k} = \sum_{i=1}^{m_{j,k}} \delta_{i,j,k}$ is the cumulative number of events at look k. Thus, we have

$$\frac{dl(\theta_j)}{d\theta_j} = -r_{j,k}\theta_j^{-1} + \left(\sum_{i=1}^{m_{j,k}} y_{i,j,k} \right) \theta_j^{-2} = 0$$

for a maximum. Hence, we obtain the maximum likelihood estimate of the mean survival time θ_j to be

$$\hat{\theta}_{j,k} = \frac{\sum_{i=1}^{m_{j,k}} y_{i,j,k}}{r_{j,k}},$$

which is the sum of the observed survival times divided by the number of events obtained so far.

The Fisher information is

$$\begin{aligned} I(\theta_j) = -E\left\{ \frac{d^2 l(\theta_j)}{d\theta_j^2} \right\} &= -E\left\{ r_{j,k}\theta_j^{-2} - 2\left(\sum_{i=1}^{m_{j,k}} Y_{i,j,k} \right) \theta_j^{-3} \right\} \\ &= -E(r_{j,k})\theta_j^{-2} + 2E\left(\sum_{i=1}^{m_{j,k}} Y_{i,j,k} \right) \theta_j^{-3}. \end{aligned}$$

Here, if there were no censored data, $\sum_{i=1}^{m_{j,k}} Y_{i,j,k} \sim \Gamma(m_{j,k}, \theta_j^{-1})$. However, if

there is censoring, then $\sum_{i=1}^{m_{j,k}} Y_{i,j,k} \sim \Gamma(r_{j,k}, \theta_j^{-1})$ and $E(\sum_{i=1}^{m_{j,k}} Y_{i,j,k})$ can be approximated by $\theta_j E(r_{j,k})$ (Cox and Oakes, 1984). So the Fisher information is approximately

$$I(\theta_j) = -E(r_{j,k})\theta_j^{-2} + 2\theta_j E(r_{j,k})\theta_j^{-3} = E(r_{j,k})\theta_j^{-2}.$$

Consequently, we have

$$\text{var}(\hat{\theta}_{j,k}) = I(\theta_j)^{-1} = \frac{\theta_j^2}{E(r_{j,k})}$$

as the approximate variance of $\hat{\theta}_{j,k}$ for treatment j evaluated based on the cumulative responses.

2.3.3. Test statistic

Consider the parameter of interest as the difference in the mean survival times for the two treatments, $\phi = \theta_1 - \theta_2$. The null hypothesis is $H_0 : \phi = 0$ versus the alternative hypothesis $H_a : \phi \neq 0$. Suppose that $\hat{\phi}_k = \hat{\theta}_{1,k} - \hat{\theta}_{2,k}$ is the maximum likelihood estimate of ϕ at look k. Under the assumption that the responses on the two arms are independent, we obtain

$$\text{var}(\hat{\phi}_k) = \text{var}(\hat{\theta}_{1,k} - \hat{\theta}_{2,k}) = I(\theta_1)^{-1} + I(\theta_2)^{-1} = \theta_1^2/E(r_{1,k}) + \theta_2^2/E(r_{2,k}).$$

Based on the assumed model, $E(r_{j,k})$ can be approximated by $m_{j,k}\epsilon_{j,k}$, where $m_{j,k}$ is the sample size on treatment j at look k and $\epsilon_{j,k}$ is the probability of having an event. In practice, the observed number of events on arm j at look k, $r_{j,k}$, is used. Then the test statistic at look k can be expressed as

$$Z_k = \frac{\hat{\theta}_{1,k} - \hat{\theta}_{2,k}}{\sqrt{\frac{\hat{\theta}_{1,k}^2}{r_{1,k}} + \frac{\hat{\theta}_{2,k}^2}{r_{2,k}}}}, \qquad k = 1, ..., K. \tag{2.14}$$

The test statistic (2.14) is approximately normal for large sample sizes. Note that it cannot be calculated until there is at least one event on both treatment arms, so that $r_{j,k} > 0,\ j = 1, 2$.

2.3.4. Optimal response-adaptive randomisation

Although survival responses are usually not immediately available after the treatments are assigned, the optimal response-adaptive randomisation procedures described in Section 2.1.2 can still be applied, provided that some responses have been obtained. In fact, Zhang and Rosenberger (2007) showed that a moderate delay in censored survival responses has only a modest effect on the asymptotic properties of the DBCD. Here, some optimal allocations for censored survival responses discussed in Zhang and Rosenberger (2007) are introduced.

Neyman allocation

For the test statistic considered, Neyman allocation is found by minimising the total sample size $M_1 + (N - M_1)$ with respect to M_1 under a variance constraint, $\theta_1^2/E(r_1) + \theta_2^2/E(r_2) = C$, where the expected number of events on arm j, $E(r_j) = M_j\epsilon_j$. The solution is

$$\rho_1 = \frac{\theta_1\sqrt{\epsilon_2}}{\theta_1\sqrt{\epsilon_2} + \theta_2\sqrt{\epsilon_1}} \quad \text{and} \quad \rho_2 = 1 - \rho_1.$$

Optimal allocation

For survival responses, optimal allocation aims to minimise the total expected hazard, $M_1\theta_1^{-1} + (N - M_1)\theta_2^{-1}$, with respect to M_1 under the above variance constraint. The solution is now

$$\rho_1 = \frac{\sqrt{\theta_1^3\epsilon_2}}{\sqrt{\theta_1^3\epsilon_2} + \sqrt{\theta_2^3\epsilon_1}} \quad \text{and} \quad \rho_2 = 1 - \rho_1. \tag{2.15}$$

Here, the optimal allocation proportions are functions of the unknown parameters.

In addition, the probability of an event on treatment j, ϵ_j, which depends on the assumed models, is also a function of the unknown parameters, as shown in (2.11). In practice, the current parameter estimates based on the responses available are used. Then the estimated target optimal allocation proportion for treatment 1 can be used in the DBCD and ERADE functions in Section 2.1.2 to obtain the allocation probability for the next patient.

2.3.5. Simulation study

Consider the model assumptions described in Section 2.3.2. We wish to compare the two survival means using test statistic (2.14). Similar parameter settings are used as in Rosenberger and Seshaiyer (1997) and Zhang and Rosenberger (2007). The duration of the trial is $D = 1.5936$. The maximum sample size $N = 800$ is chosen to achieve around 80% power. Here, the optimal allocation that minimises the total expected hazard for censored survival responses is applied. The other simulation settings are the same as in Section 2.2.2.

As can be seen in Table 2.7, the type I error rate for all of the designs is within 0.01 deviation from the pre-specified value, 0.05. A slightly conservative $\tilde{\alpha}$ is obtained for CR, which is within three standard errors of 0.05. For the response-adaptive designs, $\tilde{\alpha}$ usually lies within 1-3 standard errors of 0.05. For $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), the value of $\tilde{\alpha}$ for the ERADE is just outside of the range of three standard errors, (0.041, 0.059). The results reveal that, in general, the analogous O'Brien and Fleming boundaries derived by the error-spending approach based on normal responses can be applied as an approximate result to censored survival responses.

Under the null hypothesis, the target optimal allocation yields equal allocation, and hence the expected number of patients (ENP) and the expected number of failures (ENF) are similar across designs. However, the use of the response-adaptive designs seems to increase the variability in the allocation proportion compared to

Table 2.7.: Simulated type I error rate for two-armed censored survival trials with optimal allocation in group sequential and fixed-sample designs, $\theta_1 = \theta_2 = 1,\ N = 800$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.043	799.9	(4.8)	298.0	(1.8)	0.500	(0.017)
DBCD	0.052	798.9	(15.8)	297.6	(5.9)	0.501	(0.059)
ERADE	0.051	798.8	(16.9)	297.6	(6.3)	0.500	(0.052)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.041	797.7	(15.9)	297.1	(5.9)	0.500	(0.017)
DBCD	0.057	796.2	(24.2)	296.6	(9.0)	0.501	(0.062)
ERADE	0.060	796.0	(24.1)	296.5	(9.0)	0.501	(0.054)
	Fixed-sample designs						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.042	800	(0)	298.0	(0)	0.500	(0.017)
DBCD	0.055	800	(0)	298.0	(0)	0.501	(0.059)
ERADE	0.052	800	(0)	298.0	(0)	0.500	(0.053)

CR. The standard deviations of $\tilde{\rho}_1$ for the DBCD and the ERADE are just over three times higher than for CR.

Under the alternative hypothesis, from Table 2.8, the use of optimal response-adaptive randomisation significantly reduces the ENF for both group sequential and fixed-sample designs compared to CR. For the group sequential designs, we also have a lower ENP for the response-adaptive randomised designs than for CR. Moreover, the power is not adversely affected and can be higher in the response-adaptive randomised designs. For example, for $(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1), the power for the ERADE is increased by around 1% compared to CR, while using 17 fewer patients and reducing the number of failures by about 10.

Here, the optimal allocation proportion for treatment 1, ρ_1, calculated from (2.15) is 0.652. The average allocation proportion, $\tilde{\rho}_1$, for the DBCD and the ERADE is within 0.01 deviation from ρ_1. In addition to the high accuracy in targeting the optimal allocation proportion, a reasonably small standard deviation for $\tilde{\rho}_1$ is achieved, with the ERADE consistently having a lower variability than the DBCD.

Table 2.8.: Simulated power for two-armed censored survival trials with optimal allocation in group sequential and fixed-sample design, $\theta_1 = 1.4,\ \theta_2 = 1,\ N = 800$.

		$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.815	767.7	(81.7)	255.1	(27.2)	264.5	(3.5)	0.500	(0.017)
DBCD	0.823	749.7	(97.5)	240.0	(33.6)	254.7	(5.9)	0.656	(0.070)
ERADE	0.826	750.4	(96.9)	240.4	(33.1)	254.9	(5.5)	0.652	(0.061)
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	ENF$'$	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.808	730.5	(64.3)	242.7	(21.4)	263.0	(2.8)	0.500	(0.018)
DBCD	0.805	717.5	(78.9)	229.7	(28.4)	253.8	(5.8)	0.657	(0.070)
ERADE	0.821	715.9	(79.2)	229.1	(28.3)	253.7	(5.5)	0.658	(0.064)
		Fixed-sample designs							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	-	-	$\tilde{\rho}_1$	(s.d.)
CR	0.769	800	(0)	265.8	(1.1)	-	-	0.500	(0.017)
DBCD	0.779	800	(0)	256.0	(3.8)	-	-	0.653	(0.060)
ERADE	0.775	800	(0)	256.1	(3.5)	-	-	0.651	(0.054)

Compared to the fixed-sample designs, the group sequential designs clearly have a lower ENP due to early stopping under H_a, and hence a reduced ENF is obtained. As mentioned previously, the quantity ENF$'$ for the group sequential designs is computed to compare with ENF for the fixed-sample designs based on the maximum sample size N. If a decision for early termination is made, the rest of the patients are assigned to the better-performing treatment arm. Here, a slightly lower ENF$'$ than the ENF is obtained. For instance, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), there are about three fewer failures on average for the group-sequential response-adaptive designs than for the fixed-sample response-adaptive designs.

2.4. Conclusions

The combined approach of group sequential analysis with optimal response-adaptive randomisation has been studied and shown to be more ethical in terms of reducing the average sample size and the expected number of failures for binary responses. A generalisation of the combined approach to censored survival responses using

an estimate of the information time based on the model assumed is studied in this chapter. Then critical boundaries derived using the error-spending approach can be applied to the designs. The simulation results reveal that incorporating adaptive sampling rules in group sequential designs also preserves the error rates while assigning more patients to the more promising treatment.

Both optimal response-adaptive randomisation procedures target the pre-specified desired allocation well with reasonably small variability. Among the response-adaptive designs, the ERADE consistently has a lower standard deviation for the allocation proportion compared to the DBCD.

3. Group-sequential response-adaptive designs for multi-armed trials without dropping of inferior arm(s)

3.1. Form of test

3.1.1. Information time

Let N be the maximum number of patients for a trial with K group sequential analyses and let J be the number of arms. For normal and binary multi-armed trials without allowing dropping of inferior treatments, the information time at look k is

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

where $\mathcal{I}_k$ denotes the information level at group sequential test k, $m_{j,k}$ is the cumulative number of observations 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, $n_K = N$. Here, the formula for the information time has the same form as (2.1) for two-armed trials.

For censored survival responses, the information level is proportional to the number of events. Consider the approximate information level described in Section

2.3.1. Then we have

$$t_k = \frac{\mathcal{I}_k}{\mathcal{I}_K} = \frac{\hat{e}_k}{\hat{e}_K} = \frac{\sum_{j=1}^{J} m_{j,k}\hat{\epsilon}_{j,k}}{\sum_{j=1}^{J} M_j\hat{\epsilon}_{j,K}},$$

where $\hat{e}_k$ is the estimated number of events at look k, which depends on the model assumptions. As mentioned in Section 2.3.1, there are two candidates for the estimate of e_k. One is under the null hypothesis and the other is under a specified alternative hypothesis. The overall type I error rate can be controlled in either case (Kim et al., 1995). For simplicity, the estimate of e_k under the null hypothesis, where the parameters are all equal, is considered. Then the approximate information time becomes

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

Here, the subscript j for $\hat{\epsilon}_{j,k}$ denoting the treatment is suppressed, since the parameters are all equal.

3.1.2. Global test statistics

Normal responses with a common variance

Consider testing the homogeneity of J normal means with a common variance σ^2. We wish to test the global null hypothesis $H_{G_0} : \mu_1 = ... = \mu_J$ versus the alternative hypothesis $H_{G_a} : \neg H_{G_0}$. This testing problem has been studied by Jennison and Turnbull (1991, 2000). For equal allocation, the global test statistic at look k is

$$S_k = \frac{m_k}{\sigma^2}\sum_{j=1}^{J}(\bar{Y}_{j,k} - \bar{Y}_{.k})^2, \quad k = 1, ..., K,$$

where m_k is the cumulative sample size on each arm at look k, $\bar{Y}_{j,k}$ is the sample mean of the cumulative responses on treatment j at look k and $\bar{Y}_{.k} = \sum_{j=1}^{J}\bar{Y}_{j,k}/J$ is the overall mean based on the responses obtained so far. The test statistic S_k is

essentially the standardised between-arms sum-of-squares statistic. The marginal distribution of S_k is exactly χ^2 with $J-1$ degrees of freedom under H_{G_0}. Under H_{G_a}, the distribution is noncentral chi-squared with noncentrality parameter

$$\eta_k = \frac{m_k}{\sigma^2}\sum_{j=1}^{J}(\mu_j - \bar{\mu})^2, \qquad k = 1, ..., K,$$

where $\bar{\mu} = (\mu_1 + ... + \mu_J)/J$. The noncentrality parameter η_k increases as the cumulative sample size m_k is increased. In other words, the noncentrality parameter and the power of tests increase in later looks with a larger cumulative sample size.

For unequal treatment allocation, the global test statistic at interim test k becomes

$$S_k = \frac{1}{\sigma^2}\sum_{j=1}^{J} m_{j,k}(\bar{Y}_{j,k} - \bar{Y}_{.k})^2, \qquad k = 1, ..., K, \tag{3.1}$$

where the overall mean evaluated at look k is now

$$\bar{Y}_{.k} = \frac{\sum_{j=1}^{J} m_{j,k}\bar{Y}_{j,k}}{\sum_{j=1}^{J} m_{j,k}} = \frac{\sum_{j=1}^{J} m_{j,k}\bar{Y}_{j,k}}{n_k}.$$

For optimal response-adaptive randomisation, the sample size for treatment j at look k, $m_{j,k}$, is random . However, it can be approximated, since the allocation proportions converge almost surely to the pre-specified optimal allocation proportions (Hu and Zhang, 2004). We have $m_{j,k} = \rho_j \sum_{j=1}^{J} m_{j,k} = \rho_j n_k$ approximately. The marginal distribution of S_k in (3.1) is asymptotically χ^2 with $J-1$ degrees of freedom under H_{G_0}, provided that the unbalanced allocation is not too severe. Under H_{G_a}, it is asymptotically noncentral chi-squared with noncentrality parameter

$$\eta_k = \frac{1}{\sigma^2}\sum_{j=1}^{J} m_{j,k}(\mu_j - \bar{\mu})^2, \qquad k = 1, ..., K,$$

where

$$\bar{\mu} = \frac{\sum_{j=1}^{J} m_{j,k}\mu_j}{\sum_{j=1}^{J} m_{j,k}} = \frac{\sum_{j=1}^{J} m_{j,k}\mu_j}{n_k}.$$

Normal responses with a pooled variance estimate

In practice, σ^2 is usually unknown. A pooled sample variance can be used as an estimate of σ^2. If the sample sizes are equal for all treatments, the following F statistic is monitored:

$$F_k = \frac{m_k}{(J-1)\ S_k^2} \sum_{j=1}^{J} (\bar{Y}_{j,k} - \bar{Y}_{.k})^2, \qquad k = 1, ..., K,$$

where S_k^2 is the pooled sample variance evaluated at look k given by

$$S_k^2 = \frac{\sum_{j=1}^{J} S_{j,k}^2}{J} \quad \text{and} \quad S_{j,k}^2 = \frac{1}{m_k - 1} \sum_{i=1}^{m_k} (Y_{i,j} - \bar{Y}_{j,k})^2.$$

Here, $F_k \sim F_{v_1,v_2}$ with $v_1 = J - 1$ and $v_2 = J(m_k - 1)$ under H_{G_0} (Jennison and Turnbull, 1991, 2000).

For optimal response-adaptive randomisation where treatment allocation is usually unequal, the test statistic is

$$F_k = \frac{1}{(J-1)\ S_k^2} \sum_{j=1}^{J} m_{j,k} (\bar{Y}_{j,k} - \bar{Y}_{.k})^2, \qquad k = 1, ..., K,$$

with

$$S_k^2 = \frac{\sum_{j=1}^{J} (m_{j,k} - 1) S_{j,k}^2}{\sum_{j=1}^{J} m_{j,k} - J} \quad \text{and} \quad S_{j,k}^2 = \frac{1}{m_{j,k} - 1} \sum_{i=1}^{m_{j,k}} (Y_{i,j} - \bar{Y}_{j,k})^2.$$

As long as the unbalanced treatment allocation is not too severe, the test statistic is asymptotically F distributed with degrees of freedom $v_1 = J - 1$ and $v_2 = J(m_{.k} - 1)$ under H_{G_0}, where $m_{.k} = \sum_{j=1}^{J} m_{j,k}/J$ can be used.

Normal responses with unequal variances

For equal allocation, Proschan et al. (1994) used the following chi-squared statistic for testing homogeneity:

$$S_k = (Z_{1,k}, ..., Z_{J-1,k})\hat{\Sigma}_k^{-1}(Z_{1,k}, ..., Z_{J-1,k})^T, \qquad k = 1, ..., K,$$

where $Z_{j,k},\ j = 1, ..., J-1$, is the standardised normal statistic for comparing treatment j with the average of the other arms. For example,

$$Z_{1,k} = \{\bar{Y}_{1,k} - (\bar{Y}_{2,k} + ... + \bar{Y}_{J,k})/(J-1)\}/\sqrt{\hat{V_{1,k}}},$$

with $\hat{V}_{1,k} = \{\hat{\sigma}^2_{1,k} + (\hat{\sigma}^2_{2,k} + ... + \hat{\sigma}^2_{J,k})/(J-1)^2\}/m_k$. Also, $\hat{\Sigma}_k$ is the $(J-1)\times(J-1)$ estimated covariance matrix of $(Z_{1,k}, ..., Z_{J-1,k})$ at look k. Here, arm J is omitted to prevent the issue of singularity. One may choose to leave out any treatment j and the same value of the test statistic will be obtained.

Now suppose that we wish to test $H_{G_0} : \boldsymbol{\mu}_G = \mathbf{0}$ versus $H_{G_a} : \boldsymbol{\mu}_G \neq \mathbf{0}$, where $\boldsymbol{\mu}_G = (\mu_1 - \mu_J,\ \mu_2 - \mu_J, ...,\ \mu_{J-1} - \mu_J)^T$ is a vector of treatment contrasts taking arm J as the reference group. For unequal treatment allocation, the global test statistic is

$$S_k = \hat{\boldsymbol{\mu}}^T_{G_k}\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\mu}}_{G_k}, \qquad k = 1, ..., K, \tag{3.2}$$

where $\hat{\boldsymbol{\mu}}_{G_k} = (\bar{Y}_{1,k} - \bar{Y}_{J,k},\ \bar{Y}_{2,k} - \bar{Y}_{J,k}, ...,\ \bar{Y}_{J-1,k} - \bar{Y}_{J,k})^T$ is the maximum likelihood estimator of $\boldsymbol{\mu}_G$ evaluated at interim test k and

$$\hat{\Sigma}_k = \begin{pmatrix} \frac{\hat{\sigma}^2_{1,k}}{m_{1,k}} & 0 & ... & 0 \\ 0 & \frac{\hat{\sigma}^2_{2,k}}{m_{2,k}} & & 0 \\ & & \ddots & \\ 0 & 0 & ... & \frac{\hat{\sigma}^2_{J-1,k}}{m_{J-1,k}} \end{pmatrix} + \frac{\hat{\sigma}^2_{J,k}}{m_{J,k}}\mathbf{1}\mathbf{1}^T. \tag{3.3}$$

Here, $\mathbf{1} = (1, ..., 1)^T$ is the vector with $J-1$ ones. In addition, $\hat{\Sigma}_k$ is nonsingular and its inverse exists.

Since $\hat{\boldsymbol{\mu}}_{G_k} \sim N_{J-1}(\boldsymbol{\mu}_G, \Sigma_k)$, we have $\boldsymbol{V} = \hat{\boldsymbol{\mu}}_{G_k} - \boldsymbol{\mu}_G \sim N_{J-1}(\mathbf{0}, \Sigma_k)$. Thus, we can write

$$\begin{aligned} E(S_k) &= E(\hat{\boldsymbol{\mu}}_{G_k}^T \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\mu}}_{G_k}) \simeq E\{(\boldsymbol{V} + \boldsymbol{\mu}_G)^T \Sigma_k^{-1} (\boldsymbol{V} + \boldsymbol{\mu}_G)\} \\ &= E(\boldsymbol{V}^T \Sigma_k^{-1} \boldsymbol{V}) + 2\boldsymbol{\mu}_G^T \Sigma_k^{-1} E(\boldsymbol{V}) + \boldsymbol{\mu}_G^T \Sigma_k^{-1} \boldsymbol{\mu}_G \\ &= E(\boldsymbol{V}^T \Sigma_k^{-1} \boldsymbol{V}) + \boldsymbol{\mu}_G^T \Sigma_k^{-1} \boldsymbol{\mu}_G. \end{aligned}$$

Let $\boldsymbol{Z} = \Sigma_k^{-1/2} \boldsymbol{V} \sim N_{J-1}(\mathbf{0}, \mathbf{I}_{J-1})$. Then we have

$$\begin{aligned} E(S_k) &\simeq E(\boldsymbol{Z}^T \boldsymbol{Z}) + \boldsymbol{\mu}_G^T \Sigma_k^{-1} \boldsymbol{\mu}_G \\ &= J - 1 + \boldsymbol{\mu}_G^T \Sigma_k^{-1} \boldsymbol{\mu}_G. \end{aligned}$$

Under H_{G_0}, the marginal distribution of S_k is asymptotically χ^2 with $J-1$ degrees of freedom. Under H_{G_a}, it is asymptotically noncentral chi-squared with noncentrality parameter

$$\eta_k = \boldsymbol{\mu}_G^T \Sigma_k^{-1} \boldsymbol{\mu}_G = \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2} (\mu_j - \mu_J)^2 - \frac{1}{\sum_{j=1}^{J} \frac{m_{j,k}}{\sigma_j^2}} \left\{ \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2} (\mu_j - \mu_J) \right\}^2. \quad (3.4)$$

The proof of (3.4) is given in Appendix B.

Binary responses

Suppose that we wish to test the global null hypothesis $H_{G_0} : \boldsymbol{p}_G = \mathbf{0}$ versus $H_{G_a} : \boldsymbol{p}_G \neq \mathbf{0}$, where $\boldsymbol{p}_G = (p_1 - p_J,\ p_2 - p_J, ...,\ p_{J-1} - p_J)^T$ is the vector of treatment contrasts of the probabilities of success. The testing problem for optimal response-adaptive randomisation in a fixed-sample design has been discussed by Tymofyeyev et al. (2007). For group sequential designs, the test statistic can be written as

$$S_k = \hat{\boldsymbol{p}}_{G_k}^T\ \hat{\Sigma}_k^{-1}\ \hat{\boldsymbol{p}}_{G_k}, \quad k = 1, ..., K, \quad (3.5)$$

where $\hat{\boldsymbol{p}}_{G_k}$ is the maximum likelihood estimator of $\boldsymbol{p}_G$ based on the responses obtained so far and

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

Since the test statistic (3.5) is a quadratic form in asymptotically normal variables, the marginal distribution of S_k is asymptotically χ^2 with $J-1$ degrees of freedom under H_{G_0}. Under H_{G_a}, the distribution is asymptotically noncentral chi-squared with noncentrality parameter

$$\eta_k = \boldsymbol{p}_G^T \, \Sigma_k^{-1} \, \boldsymbol{p}_G = \sum_{j=1}^{J-1} \frac{m_{j,k}}{p_j q_j}(p_j - p_J)^2 - \frac{1}{\sum_{j=1}^{J} \frac{m_{j,k}}{p_j q_j}} \left\{ \sum_{j=1}^{J-1} \frac{m_{j,k}}{p_j q_j}(p_j - p_J) \right\}^2 . \quad (3.6)$$

The derivation of (3.6) is similar to the normal responses case. In addition, the derived noncentrality parameter η_k has the same form as the one obtained by Tymofyeyev et al. (2007) in a fixed-sample design. Therefore, (3.6) possesses the same properties. For example, it is a concave function and $\partial\eta_k/\partial m_{j,k} \geq 0$. When the cumulative number of patients on any treatment arm, $m_{j,k}$, $j = 1, ..., J$, increases, η_k is increased.

Censored survival responses

Proschan et al. (1994) and Follmann et al. (1994) demonstrated group sequential monitoring multi-armed survival trials using a nonparametric logrank statistic. The work considered equal allocation or fixed unequal allocation determined before commencing the trial. Using a parametric approach, Sverdlov et al. (2011)

investigated the implementation of optimal response-adaptive randomisation in multi-armed censored survival trials in a fixed-sample design. In this section, an extension of the work of Sverdlov et al. (2011) to group sequential designs is studied.

Sverdlov et al. (2011) considered the same model assumptions as described in Section 2.3.2. The observed survival response for patient $i,\ i = 1, ..., m_{j,k}$, on treatment $j,\ j = 1, ..., J$, at look $k,\ k = 1, ..., K$, can be expressed as $Y_{i,j,k} = \min(S_{i,j},\ C_i,\ Dt_k - A_i)$, where D is the duration of the trial and t_k is the information time at look k. As shown in Section 2.3.2, we have the maximum likelihood estimate of the mean survival time for treatment j evaluated at look k, $\hat{\theta}_{j,k} = \sum_{i=1}^{m_{j,k}} y_{i,j,k}/r_{j,k}$, and $\text{var}(\hat{\theta}_{j,k}) = \theta_j^2/E(r_{j,k})$, where $r_{j,k}$ is the cumulative number of events on treatment j at look k.

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

$$S_k = \hat{\boldsymbol{\theta}}_{G_k}^T\ \hat{\Sigma}_k^{-1}\ \hat{\boldsymbol{\theta}}_{G_k},\quad k = 1, ..., K, \tag{3.7}$$

where $\hat{\boldsymbol{\theta}}_{G_k}$ is the current maximum likelihood estimator of $\boldsymbol{\theta}_G$ and

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

Again, the marginal distribution of S_k is asymptotically chi-squared with $J - 1$ degrees of freedom under H_{G_0}. Under H_{G_a}, the distribution is asymptotically non-central chi-squared with noncentrality parameter

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

which is derived in a similar way to the normal and binary responses cases.

3.1.3. Stopping boundaries

For the global tests, the joint distribution of the sequence of test statistics $\{S_1, ..., S_K\}$ does not have the standard canonical joint distribution (Jennison and Turnbull, 2000). However, for normal responses and equal variances with equal allocation, 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 $k \geq 1$. Based on the joint distribution, the exact critical boundaries can be obtained recursively. The boundaries analogous to Pocock's and the O'Brien and Fleming boundaries can be found in Jennison and Turnbull (1991, 2000). For the unknown variance case, sequential F statistics are used. The critical boundaries for this case can also be obtained from Jennison and Turnbull (1991).

For the cases of unequal variances and unbalanced sample sizes across treatments, the exact critical boundaries are not available. One approach is to compute them by simulation. The joint distribution of $\{S_1, ..., S_K\}$ can be obtained numerically because S_k is a quadratic form in asymptotically normal variables with known covariances. Consider the normal responses case as an example. Then the covariance of $X_{j,k}$ and $X_{j',k'}$, where $X_{j,k} = \bar{Y}_{j,k} - \bar{Y}_{J,k}$, for $1 \leq j \leq j' \leq J-1$ and $1 \leq k < k' \leq K$ is

$$\text{cov}(X_{j,k}, X_{j',k'}) = \begin{cases} \dfrac{\sigma_j^2}{m_{j,k'}} + \dfrac{\sigma_J^2}{m_{J,k'}} & \text{if } j = j', \\ \dfrac{\sigma_J^2}{m_{J,k'}} & \text{if } j \neq j'. \end{cases}$$

Alternatively, the significance level approach (Jennison and Turnbull, 2000) can be used to give an approximate test, as long as the imbalance in the sample sizes is not too severe. In this case, the critical boundaries derived based on equal variances and equal allocation (Jennison and Turnbull, 1991, 2000) are applied to the actual observed responses to give an approximate test.

The stopping rule for multi-armed clinical trials without allowing dropping of inferior treatments is given below.

- Stop the trial and reject H_{G_0} the first time that $S_k \geq d_k$.
- Accept H_{G_0} if $S_K < d_K$ at the final stage.

3.2. Optimal response-adaptive randomisation

3.2.1. Optimal allocations

In this section, two optimal allocations for multi-armed trials are introduced. One ensures the most efficient estimates of the treatment effects and the other aims to maximise 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.

D_A-optimal allocation

Most of the optimal allocations are derived for homoscedastic models which assume a constant variance for the responses across treatments. In this case, the equal allocation rule is optimal in terms of efficiency. However, for heteroscedastic

models, an unbalanced allocation can be more efficient. Wong and Zhu (2008) derived the D_A-optimal allocation for multi-armed normal trials using a fixed-sample design. Assume that the responses are based on the heteroscedastic model

$$\boldsymbol{Y} = f^T(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Here, $\boldsymbol{x} = (x_1, ..., x_J)^T$ and the jth element, x_j, indicates arm j. For treatment j, $f(x_j)$ is a zero vector with the jth element replaced by one, and $\boldsymbol{\beta}$ is the vector of parameters. The vector $\boldsymbol{\epsilon}$ is assumed to be multivariate normal with zero mean vector and a diagonal covariance matrix, under the assumption that the responses from different treatment arms are independent. Let the variance of the responses on treatment j be $\sigma^2/w(x_j)$, where $w(x_j)$ is inversely proportional to the variance of the responses on arm j and assumed known. Without loss of generality, the case $\sigma^2 = 1$ is considered. The variance for treatment j then becomes $1/w(x_j)$.

Let ρ_j be the allocation proportion for treatment j. A randomisation design ξ is determined by $\{(x_j,\ \rho_j),\ j = 1, ..., J\}$. Given ξ, Wong and Zhu (2008) derived the information matrix

$$\begin{aligned} M(\xi) &= \sum_{j=1}^{J} w(x_j)\rho_j f(x_j) f(x_j)^T \\ &= \begin{pmatrix} w(x_1)\rho_1 & 0 & ... & 0 \\ 0 & w(x_2)\rho_2 & ... & 0 \\ & & \ddots & \\ & & & \ddots \\ 0 & ... & & w(x_J)\rho_J \end{pmatrix}. \end{aligned} \tag{3.8}$$

Suppose that the parameter of interest is the vector of treatment contrasts $A^T\boldsymbol{\beta} = (\beta_1 - \beta_J, ..., \beta_{J-1} - \beta_J)^T$, where A^T is a $(J-1) \times J$ matrix. Then the D_A-optimal allocation ensures the most precise estimate of $A^T\boldsymbol{\beta}$. The solution to the optimal allocation is obtained 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 randomisa-

tion designs ξ, or, equivalently, by minimising its logarithm. Then the optimal rule is to minimise $\Phi(\xi) = \log|A^T M^{-1}(\xi)A|$. The D_A-optimal allocation yields the smallest confidence ellipsoid for $A^T\boldsymbol{\beta}$.

In practice, one can use the general equivalence theorem (GET) (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 $\Phi(\xi^*)$, ξ^* is the optimum value of ξ and $J - 1$ is the rank of A. For normal responses, we obtain

$$d_A(x_j, \xi^*) = \frac{1}{\rho_j} - \frac{w(x_j)}{\sum_{l=1}^{J} w(x_l)\rho_l}, \quad j = 1, ..., J,$$

as derived by Wong and Zhu (2008).

Sverdlov et al. (2011) generalised the D_A-optimal allocation to censored survival responses. For exponentially distributed survival responses, the solution is obtained by solving the equations

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

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.

For multi-armed binary trials, the solution is obtained by solving the equations

$$d_A(x_j, \xi^*) = \frac{1}{\rho_j} - \frac{1/(p_j q_j)}{\sum_{l=1}^{J} 1/(p_l q_l)\rho_l} = J - 1, \quad j = 1, ..., J, \tag{3.9}$$

where $1/(p_j q_j)$ is inversely proportional to the variance for treatment 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 $\theta_j^2/E(r_j)$, as shown in Section 2.3.2. In this case, the variance increases when the mean survival time for treatment j, θ_j, is increased. Therefore, for exponential survival responses, the D_A-optimal allocation is always ethical (Sverdlov et al., 2011). However, for normal and binary responses, the most efficient design may assign more patients to the inferior treatments. The optimal allocation depends on the unknown parameters. In practice, the current parameter estimates are updated after each response.

The D_A-optimal allocation proportion, $\rho_j,\ j = 1, ..., J$, satisfies $0 \leq \rho_j \leq 1/(J-1)$. When the number of treatments is $J = 2$, the D_A-optimal allocation reduces to Neyman allocation. For $J \geq 3$, the solution has no closed form. However, it can be computed numerically by finding the root of the system of nonlinear equations. The R package *rootSolve*, which utilises the Newton-Raphson method, can be used.

Optimal allocations based on nonlinear programming

The process for solving an optimality problem, which maximises or minimises an objective function with equalities or inequalities as constraints, where the objective function or some of the constraints are nonlinear, is called nonlinear programming. To find a design that maximises the power of test, one can consider maximising the noncentrality parameter η (as shown in Section 3.1.2) since the power increases as η is increased. In this case, we have nonlinear objective function because η is a nonlinear function of the sample size.

Tymofyeyev et al. (2007) considered the following optimality problems for multi-armed binary trials with a fixed-sample design: (a) maximising the power of the test of homogeneity, subject to the constraint that the weighted sum of the sample

sizes does not exceed a positive constant C, and also subject to a user-specified lower bound, B, for the allocation proportion $\rho_j,\ j = 1, ..., J$; (b) minimising the weighted sum of the sample sizes for a fixed power. The authors showed that the two optimality problems yield the same solution. The optimality rule can be expressed as

$$\begin{gathered}(a) \text{ maximise the noncentrality parameter } \eta \text{ such that} \\ \sum_{j=1}^{J} v_j M_j \leq C \quad \text{and} \quad \frac{M_j}{N} \geq B, \quad \text{for } j = 1, ..., J,\end{gathered} \tag{3.10}$$

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 $B \in [0, 1/J]$. By selecting $B > 0$, one can ensure that every treatment arm is allocated. Similarly, we can write

$$\begin{gathered}(b) \text{ minimise } \sum_{j=1}^{J} v_j M_j \text{ such that} \\ \eta \geq C \quad \text{and} \quad \frac{M_j}{N} \geq B, \quad \text{for } j = 1, ..., J.\end{gathered}$$

When the vector of weights $(v_1,\ ..., v_J) = (1, ..., 1)$ 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. Tymofyeyev et al. (2007) provided a closed form for the solution for the case when $(v_1,\ ..., v_J) = (1, ..., 1)$, which is given below. For other cases, numerical methods are required to obtain the solution.

Let the success probabilities be $p_1 = ... = p_s > p_{s+1} \geq ... \geq p_{J-g} > p_{J-g+1} = ... = p_J$ for some positive integers s and g. For instance, for three-armed trials,

$s = g = 1$. This assumption is needed to specify the multiplicities of the highest and lowest values of the underlying success probabilities.

(i) When $B \in [0, \tilde{B}]$, $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_J, 1/J)$, the optimal allocation based on nonlinear programming (NP) is $(\rho_1, ..., \rho_J)$ with

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

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

$$\begin{aligned}
Q = &\frac{\sqrt{p_1 q_1}}{\sqrt{p_1 q_1} + \sqrt{p_J q_J}}\left\{\sum_{j=s+1}^{J-g}\frac{p_J q_J}{p_j q_j} - (J - s - g)\right\} \\
&- \frac{\sqrt{p_1 q_1 p_J q_J}}{p_1 - p_J}\sum_{j=s+1}^{J-g}\frac{p_j - p_J}{p_j q_j}.
\end{aligned} \tag{3.11}$$

(ii) When $B > \tilde{B}$ and $\tilde{B} = \tilde{B}_1$, the vector of optimal allocation proportions is

$$\begin{aligned}
\boldsymbol{\rho} =&(B, ..., B,\ \rho_{J-g+1}, ..., \rho_J) \\
&\text{with}\ \ \rho_{J-g+1} = ... = \rho_J = \frac{1 - (J - g)B}{g}.
\end{aligned}$$

(iii) When $B > \tilde{B}$ and $\tilde{B} = \tilde{B}_J$, the solution is

$$\begin{aligned}
\boldsymbol{\rho} =&(\rho_1, ..., \rho_s,\ B, ..., B) \\
&\text{with}\ \ \rho_1 = ... = \rho_s = \frac{1 - (J - s)B}{s}.
\end{aligned}$$

In cases (ii) and (iii), the optimal allocation proportions are fixed as shown above. In case (i), the optimal allocation proportions for treatments other than the best and the worst ones are fixed at the lower bound, that is, $\rho_{s+1} = ... = \rho_{J-g} = B$,

whereas the best and worst treatment allocation proportions depend on the unknown parameters. In practice, the parameter estimates are used.

When $B = 0$, the solution maximises the power. However, patients are assigned to the best and the worst treatments only. When $B = 1/J$, the solution becomes equal allocation by definition. Tymofyeyev et al. (2007) pointed out that, although theoretically the power increases as B decreases, in practice, the increase in the variability of the allocation proportions has an adverse effect on the power for small values of B. For trials with $J = 3$ arms, the authors took $B = 0.25$ to give a reasonably small variability in the allocation proportions while preserving the power. In addition, they observed by simulation that a choice of $B = 0.1,\ 0.2$ and 0.25 does not affect the conclusions when comparing the power of the response-adaptive design to the CR design.

Sverdlov et al. (2011) generalised NP to censored survival responses. Assume that the order of the J mean survival times is $\theta_1 = ... = \theta_s > \theta_{s+1} \geq ... \geq \theta_{J-g} > \theta_{J-g+1} = ... = \theta_J$ for some positive integers s and g.

(i) When $B \in [0, \tilde{B}]$, $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_J, 1/J)$, the solution becomes

$$\rho_1 = ... = \rho_s = \frac{1}{s}\left(QB + \frac{\theta_1/\sqrt{\epsilon_1}}{\theta_1/\sqrt{\epsilon_1} + \theta_J/\sqrt{\epsilon_J}}\right),$$
$$\rho_{s+1} = ... = \rho_{J-g} = B,$$
$$\rho_{J-g+1} = ... = \rho_J = \frac{1}{g}\{1 - B(K - s - g) - s\rho_1\},$$

where

$$\tilde{B}_1 = \frac{1}{s - Q}\frac{\theta_1/\sqrt{\epsilon_1}}{\theta_1/\sqrt{\epsilon_1} + \theta_J/\sqrt{\epsilon_J}},$$
$$\tilde{B}_J = \frac{1}{J + Q - s}\frac{\theta_J/\sqrt{\epsilon_J}}{\theta_1/\sqrt{\epsilon_1} + \theta_J/\sqrt{\epsilon_J}}$$

and

$$Q = \frac{\theta_1/\sqrt{\epsilon_1}}{\theta_1/\sqrt{\epsilon_1} + \theta_J/\sqrt{\epsilon_J}} \left(\sum_{j=s+1}^{J-g} \frac{\theta_J^2/\epsilon_J}{\theta_j^2/\epsilon_j} - (J-s-g) \right) - \frac{(\theta_1/\sqrt{\epsilon_1})(\theta_J/\sqrt{\epsilon_J})}{\theta_1 - \theta_J} \sum_{j=s+1}^{J-g} \frac{\theta_j - \theta_J}{\theta_j^2/\epsilon_j}. \tag{3.12}$$

For case (ii) when $B > \tilde{B}$ and $\tilde{B} = \tilde{B}_1$, and case (iii) when $B > \tilde{B}$ and $\tilde{B} = \tilde{B}_J$, the optimal allocation proportions are fixed and identical to the binary case.

When $J = 2$, both the D_A-optimal allocation and the optimal allocation based on NP with $(v_1, ..., v_J) = (1, ..., 1)$ reduce to Neyman allocation. However, when $J \geq 3$, they are quite different (Sverdlov et al., 2011). The authors compared the DBCD targeting D_A-optimal allocation and NP allocation by simulation for three-armed censored survival trials in a fixed-sample design. It was found that the design targeting NP allocation attained a higher power than those using the D_A-optimal allocation and complete randomisation. Nevertheless, the response-adaptive designs targeting D_A-optimal allocation and NP allocation led to a lower total expected hazard, which means that they are more ethical compared with the CR design.

The solution to NP is derived using the Lagrange multiplier method generalised to inequality constraints. Details of the calculations can be found in Tymofyeyev et al. (2007). As shown above, the closed form for the solution assumes the order of the unknown parameters as $p_1 = ... = p_s > p_{s+1} \geq ... \geq p_{J-g} > p_{J-g+1} = ... = p_J$ for binary responses and $\theta_1 = ... = \theta_s > \theta_{s+1} \geq ... \geq \theta_{J-g} > \theta_{J-g+1} = ... = \theta_J$ for exponential survival responses. Then the solution can be obtained by the Lagrange multiplier method for binary and survival responses, which involve one parameter. However, for normal responses, there is also a nuisance parameter. A closed form for the solution to NP for multi-armed normal trials is not yet available.

3.2.2. Optimal response-adaptive randomisation procedures

The optimal response-adaptive randomisation procedures described in Section 2.1.2 have been generalised to multi-armed clinical trials. Hu and Zhang (2004) studied the DBCD and Zhang (2016) developed the ERADE for multi-armed trials, which aim to target the pre-specified optimal allocations such as the D_A-optimal allocation and the NP allocation described in Section 3.2.1. Similar to two-armed trials, the optimal allocations usually depend on the unknown parameters. Therefore, a learning phase is required until initial parameter estimates are available. Here, permuted-block randomisation is applied to the first 10% of the N patients. Permuted-block randomisation can guarantee that the sample sizes on each of the treatments are equal. Moreover, it can prevent the case where some arms receive no patients, which can occur in the complete randomisation design.

After obtaining initial parameter estimates, the optimal response-adaptive randomisation procedures can be implemented. 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 allocation probability functions for the DBCD and the ERADE are given below, which depend on the current and the optimal allocation proportions.

Doubly-adaptive biased coin design

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 randomness. 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.

Efficient randomised-adaptive design

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

$$g_j = \begin{cases} \dfrac{\hat{\rho}_j^{(i)} \ \psi\left(\dfrac{\hat{\rho}_j^{(i)}}{m_j^{(i)}/i}\right)}{\sum_{l=1}^{J} \hat{\rho}_l^{(i)} \ \psi\left(\dfrac{\hat{\rho}_l^{(i)}}{m_l^{(i)}/i}\right)} & \text{if } 0 < m_j^{(i)}/i < 1, \\ 1 - m_j^{(i)}/i & \text{if } m_j^{(i)}/i = 0,\ 1. \end{cases} \tag{3.13}$$

Here, the tuning parameter γ' can be any positive number. Through personal communication, the author suggests a value $2 \leq \gamma' \leq 4$ to achieve a high power while allowing a reasonable degree of randomness. Notice that, for multi-treatment trials, g_j in (3.13) is a continuous function, whereas, for two-armed trials, the allocation probability function in (2.8) is discrete.

The generalisation of the ERADE to multiple treatments has also been shown to attain the Cramér-Rao lower bound for the variance of the allocation propor-

tions (Zhang, 2016). In other words, use of the ERADE guarantees the least variability in the randomisation compared to other response-adaptive designs.

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 NP allocation is discontinuous when the parameters are all equal. As can be seen in (3.11) for binary responses, Q is discontinuous when $p_1 = p_J$. In practice, the parameter estimates are used to obtain the optimal allocation. Problems can also occur when the denominator $p_1 - p_J$ in Q has a value close to zero, that is, when the value of $\hat{p}_1 - \hat{p}_J$ is extremely small. The issue of discontinuity can be addressed by replacing the denominator in Q by $p_1 - p_J + 1$, where $p_1 > p_J$. This ensures that $\hat{p}_1 - \hat{p}_J + 1$ is a positive value, since $\hat{p}_1,\ \hat{p}_J \in (0, 1)$. The same approach can be used in (3.12) for censored survival responses to avoid the denominator in Q being zero.

3.3. Simulation studies

3.3.1. Three-armed normal trials

Consider $J = 3$ treatments and testing $H_{G_0} : \boldsymbol{\mu}_G = \mathbf{0}$ versus $H_{G_a} : \boldsymbol{\mu}_G \neq \mathbf{0}$, where $\boldsymbol{\mu}_G = (\mu_{E1} - \mu_C,\ \mu_{E2} - \mu_C)^T$, $E1$ and $E2$ refer to two experimental treatments, and C denotes the control. The global test statistic (3.2) is used. The adaptive designs without dropping inferior treatments are investigated by simulation in terms of the error probabilities, the expected number of patients (ENP), the allocation proportions and the corresponding standard deviations. The nominal type I error rate was 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) obtained from Table 16.1 in Jennison and Turnbull

(2000) were used as an approximation. The boundaries were derived based on normal responses with equal allocation and equal increments in information. For optimal response-adaptive randomisation, the tuning parameters $\gamma = \gamma^{'} = 2$ were set for the DBCD and the ERADE functions. D_A-optimal allocation was used as the targeted optimal allocation for multi-armed normal trials. 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.

Table 3.1.: Simulated type I error rate for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = \mu_{E2} = \mu_C = 16,\ \sigma_{E1} = \sigma_{E2} = \sigma_C = 10,\ N = 138$.

Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.038	137.8	(3.3)	0.334	(0.039)	0.333	(0.039)	0.334	(0.039)
DBCD_{D_A}	0.037	137.7	(3.4)	0.333	(0.025)	0.334	(0.025)	0.333	(0.025)
ERADE_{D_A}	0.038	137.7	(3.8)	0.333	(0.022)	0.334	(0.022)	0.333	(0.023)

Table 3.2.: Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = 20,\ \mu_{E2} = 16,\ \mu_C = 13,\ \sigma_{E1} = \sigma_{E2} = \sigma_C = 10,\ N = 138$.

Procedure	power	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.790	123.6	(21.5)	0.334	(0.041)	0.333	(0.041)	0.334	(0.042)
DBCD_{D_A}	0.814	122.0	(22.3)	0.342	(0.026)	0.333	(0.027)	0.326	(0.029)
ERADE_{D_A}	0.819	121.5	(22.6)	0.342	(0.024)	0.332	(0.025)	0.326	(0.026)

The target D_A-optimal allocation is (0.333 0.333 0.333), since the variances are equal.

Tables 3.1 and 3.2 consider the case where the variance of the responses is equal across treatments. In this case, the D_A-optimal rule is equal allocation. Under H_{G_0}, from Table 3.1, the results for the adaptive designs and the group sequential CR design are similar. However, the CR design consistently has a higher variability in the allocation proportions. Here, a conservative type I error rate is obtained

for all of the designs. This may be due to the fact that the test statistic (3.2) for the unequal variances case is applied, $\hat{\sigma}^2_{j,k},\ j = 1, 2, 3$, is used in (3.3) and the sample size is small in this case. If σ^2_j is used instead of $\hat{\sigma}^2_{j,k}$, $\tilde{\alpha}$ becomes 0.046 for CR and the ERADE, and 0.047 for the DBCD.

Under the alternative hypothesis, from Table 3.2, the average allocation proportion for experimental treatment 1, $\tilde{\rho}_{E1}$, using the adaptive designs is 0.342, which is within one standard deviation of the target allocation proportion, as are $\tilde{\rho}_{E2}$ and $\tilde{\rho}_C$. The adaptive designs have a lower variability in the allocation proportions than the CR design. In addition, compared with the CR design, the power for the adaptive designs is around 3% higher.

Table 3.3.: Simulated type I error rate for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = \mu_{E2} = \mu_C = 1,\ \sigma_{E1} = 4,\ \sigma_{E2} = 2,\ \sigma_C = 1,\ N = 300$.

Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.051	299.0	(10.3)	0.333	(0.026)	0.333	(0.025)	0.334	(0.026)
DBCD_{D_A}	0.053	298.9	(10.3)	0.453	(0.015)	0.355	(0.019)	0.192	(0.018)
ERADE_{D_A}	0.048	299.2	(9.0)	0.451	(0.011)	0.355	(0.016)	0.194	(0.017)

The target D_A-optimal allocation is (0.454, 0.356, 0.191).

Table 3.4.: Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = 2,\ \mu_{E2} = 1.5,\ \mu_C = 1,\ \sigma_{E1} = 4,\ \sigma_{E2} = 2,\ \sigma_C = 1,\ N = 300$.

Procedure	power	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.792	261.1	(51.7)	0.333	(0.028)	0.334	(0.028)	0.333	(0.028)
DBCD_{D_A}	0.807	259.2	(52.5)	0.453	(0.016)	0.356	(0.020)	0.192	(0.020)
ERADE_{D_A}	0.811	257.7	(53.1)	0.451	(0.013)	0.355	(0.018)	0.193	(0.019)

Tables 3.3 - 3.6 consider the unequal variances case. Table 3.4 considers the case where treatment arms that have a greater variance in response also have a higher mean response, whereas Table 3.6 considers the case where the orders of the means and the variances are not the same. For instance, $E1$ has the smallest mean re-

sponse, yet the highest variance.

As can be seen from Tables 3.3 - 3.6, the adaptive designs targeting D_A-optimal allocation assign more patients to the treatment arm that has the larger variance under both the null and the alternative hypotheses. Both the DBCD and the ERADE target the D_A-optimal allocation well, with a difference of less than one standard deviation from the target allocation proportions. In addition, the ERADE consistently has the lowest variability in the allocation proportions.

Table 3.5.: Simulated type I error rate for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = \mu_{E2} = \mu_C = 1,\ \sigma_{E1} = 4,\ \sigma_{E2} = 2,\ \sigma_C = 1,\ N = 410$.

		$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.053	408.3	(15.2)	0.333	(0.022)	0.333	(0.022)	0.333	(0.022)
DBCD_{D_A}	0.051	408.5	(14.6)	0.453	(0.012)	0.355	(0.015)	0.192	(0.016)
ERADE_{D_A}	0.047	408.7	(13.0)	0.452	(0.009)	0.355	(0.013)	0.193	(0.014)
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.043	409.1	(8.5)	0.333	(0.022)	0.334	(0.022)	0.333	(0.022)
DBCD_{D_A}	0.052	409.1	(8.3)	0.453	(0.012)	0.355	(0.016)	0.191	(0.016)
ERADE_{D_A}	0.045	409.2	(8.3)	0.452	(0.009)	0.355	(0.013)	0.193	(0.014)
		Fixed-sample design							
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.050	410	(0)	0.334	(0.022)	0.333	(0.022)	0.333	(0.022)
DBCD_{D_A}	0.050	410	(0)	0.453	(0.012)	0.355	(0.016)	0.192	(0.016)
ERADE_{D_A}	0.050	410	(0)	0.452	(0.009)	0.355	(0.013)	0.193	(0.014)

The target D_A-optimal allocation is (0.454, 0.356, 0.191).

Under the null hypothesis, from Tables 3.3 and 3.5, the type I error rate for all of the designs are close to the nominal value. For the fixed-sample designs and the group sequential designs with equally spaced information times, $\tilde{\alpha}$ is within one standard error of 0.05. For group sequential analysis with $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), $\tilde{\alpha}$ for the DBCD, the ERADE and the CR designs is within one, two and three standard errors, respectively. Since the parameters are all equal under H_{G_0}, the ENP for all of the designs is similar.

Table 3.6.: Simulated power for three-armed normal trials using complete randomisation and response-adaptive randomisation, $\mu_{E1} = 1,\ \mu_{E2} = 2,\ \mu_C = 1.5,\ \sigma_{E1} = 4,\ \sigma_{E2} = 2,\ \sigma_C = 1,\ N = 410$.

		$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)							
Procedure	power	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.798	356.7	(70.5)	0.332	(0.024)	0.334	(0.024)	0.334	(0.024)
DBCD_{D_A}	0.841	348.6	(72.8)	0.452	(0.014)	0.354	(0.017)	0.194	(0.017)
ERADE_{D_A}	0.839	350.6	(72.6)	0.450	(0.010)	0.354	(0.014)	0.196	(0.016)
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	power	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.797	368.0	(48.4)	0.334	(0.023)	0.334	(0.023)	0.333	(0.023)
DBCD_{D_A}	0.841	361.6	(50.8)	0.452	(0.013)	0.354	(0.016)	0.194	(0.016)
ERADE_{D_A}	0.837	362.1	(50.6)	0.451	(0.010)	0.354	(0.014)	0.195	(0.015)
		Fixed-sample design							
Procedure	power	ENP	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.792	410	(0)	0.334	(0.022)	0.333	(0.022)	0.333	(0.022)
DBCD_{D_A}	0.838	410	(0)	0.452	(0.012)	0.354	(0.015)	0.193	(0.015)
ERADE_{D_A}	0.841	410	(0)	0.451	(0.009)	0.354	(0.013)	0.195	(0.014)

The target D_A-optimal allocation is (0.454, 0.356, 0.191).

Under the alternative hypothesis, from Tables 3.4 and 3.6, the ENP is about 13% lower under group sequential analysis compared to a fixed-sample design. For example, in Table 3.4, the maximum number of patients is 300. However, the ENP is around 260, due to the fact that early termination is allowed. In addition, use of the adaptive designs can reduce the ENP slightly, yet yield a higher power compared with the group sequential CR design. For instance, for $(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1) in Table 3.6, the DBCD design can increase the power by around 4% while using eight fewer patients on average compared to the CR design.

3.3.2. Three-armed binary trials

Consider testing $H_{G_0} : \boldsymbol{p}_G = \mathbf{0}$ versus $H_{G_a} : \boldsymbol{p}_G \neq \mathbf{0}$ with $\boldsymbol{p}_G = (p_{E1} - p_C,\ p_{E2} - p_C)^T$ using test statistic (3.5), where the ps are the success rates for the treatments. Tables 3.7 - 3.9 compare the adaptive designs with the group sequential

CR design in terms of the error probabilities, the expected number of patients (ENP) and the expected number of failures (ENF) for different scenarios of the success rates. For binary responses, the D_A-optimal allocation and the optimal allocation based on nonlinear programming (NP) were used as the target allocations for the response-adaptive designs. For the NP allocation, the user-specified lower bound for the allocation proportions was set to be $B = 0.25$. There are $K = 3$ group sequential tests planned at equally spaced information times. The O'Brien and Fleming critical boundaries derived based on normal responses with equal allocation were used as an approximation. The type I error rate and other settings were the same as in Section 3.3.1.

Table 3.7.: Simulated type I error rate for three-armed binary trials using complete randomisation and response-adaptive randomisation.

p_{E1}	p_{E2}	p_C	N	Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)
0.2	0.2	0.2	210	CR	0.050	209.0	(8.5)	167.4	(8.9)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.069	208.2	(11.3)	166.5	(10.7)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.061	208.6	(10.3)	166.9	(10.1)
				DBCD_{NP}	0.076	207.8	(12.2)	166.3	(11.2)
				ERADE_{NP}	0.073	208.0	(11.7)	166.4	(10.9)
0.5	0.5	0.5	600	CR	0.053	597.7	(21.4)	298.9	(10.7)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.047	598.2	(18.7)	299.1	(9.4)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.051	597.5	(22.5)	298.8	(11.2)
				DBCD_{NP}	0.051	598.0	(20.3)	299.0	(10.2)
				ERADE_{NP}	0.049	597.2	(23.6)	298.6	(11.8)
0.6	0.6	0.6	81	CR	0.062	80.4	(4.6)	32.1	(4.7)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.071	80.2	(5.0)	32.0	(4.9)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.072	80.3	(4.6)	32.1	(4.8)
				DBCD_{NP}	0.063	80.4	(4.5)	32.1	(4.8)
				ERADE_{NP}	0.056	80.4	(4.7)	32.1	(4.8)
0.7	0.7	0.7	289	CR	0.053	287.4	(12.9)	86.0	(8.7)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.059	287.0	(16.1)	86.0	(9.2)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.059	287.4	(13.4)	86.1	(8.7)
				DBCD_{NP}	0.048	287.9	(10.9)	86.1	(8.5)
				ERADE_{NP}	0.047	287.9	(11.2)	86.3	(8.5)

Table 3.7 shows that the ENP is close to the maximum number of patients, N, and that the ENF is close to qN under the null hypothesis where the parameters are all equal, where q is the probability of failure. The type I error rate is close to the nominal value for large sample sizes and when the success probabilities are close

to 0.5. For $p_{E1} = p_{E2} = p_C = 0.5$ and $N = 600$, $\tilde{\alpha}$ is within one standard error of 0.05. For $p_{E1} = p_{E2} = p_C = 0.7$ and $N = 289$, $\tilde{\alpha}$ for CR and the response-adaptive designs targeting the NP allocation is within one standard error, whereas, for the D_A-optimal designs, $\tilde{\alpha} = 0.059$ is three standard errors away. However, for small sample sizes or when the success rates are close to zero or one, $\tilde{\alpha}$ for the adaptive designs can be inflated by around 1-2%, as shown in the first and third scenarios. For $p_{E1} = p_{E2} = p_C = 0.6$ and $N = 81$, $\tilde{\alpha}$ for CR is also inflated to 0.062, which is more than three standard errors above 0.05.

Table 3.8.: Simulated power for three-armed binary trials using complete randomisation and response-adaptive randomisation.

p_{E1}	p_{E2}	p_C	N	Procedure	Power	ENP	(s.d.)	ENF	(s.d.)
0.3	0.2	0.1	210	CR	0.783	183.9	(35.9)	147.2	(30.0)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.795	180.8	(36.4)	143.0	(30.5)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.785	181.9	(36.3)	143.9	(30.4)
				DBCD_{NP}	0.784	180.8	(36.5)	140.0	(30.0)
				ERADE_{NP}	0.783	181.3	(36.3)	140.3	(29.9)
0.65	0.55	0.5	600	CR	0.805	518.1	(104.3)	224.5	(45.2)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.800	517.5	(103.9)	224.7	(45.0)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.806	517.2	(105.6)	224.5	(45.7)
				DBCD_{NP}	0.829	513.5	(105.7)	214.1	(44.0)
				ERADE_{NP}	0.823	511.8	(106.0)	213.6	(44.2)
0.8	0.6	0.4	81	CR	0.790	68.7	(15.7)	27.4	(7.2)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.810	67.6	(16.0)	27.6	(7.0)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.805	68.0	(15.8)	27.8	(6.9)
				DBCD_{NP}	0.823	66.7	(16.8)	25.5	(6.6)
				ERADE_{NP}	0.819	67.2	(16.5)	25.7	(6.5)
0.8	0.7	0.6	289	CR	0.795	249.3	(51.3)	74.9	(16.6)
				$\text{DBCD}_{\text{D}_\text{A}}$	0.803	246.7	(54.2)	75.1	(17.5)
				$\text{ERADE}_{\text{D}_\text{A}}$	0.801	247.1	(54.4)	75.3	(17.4)
				DBCD_{NP}	0.800	247.8	(52.8)	71.2	(15.7)
				ERADE_{NP}	0.796	248.0	(52.3)	71.4	(15.6)

Under the alternative hypothesis, from Table 3.8, since the type I error rate is not comparable for the designs in the first and third scenarios, the power can not be compared directly. The increase in the power for the adaptive designs may be because of the inflation in the type I error rate. Looking at the second scenario with a large sample size, it is clear that the power for CR and the adaptive designs using the D_A-optimal allocation is similar at about 80%. This is because, for a

small difference in the variances of the responses for the treatments, D_A-optimal allocation is close to equal allocation. However, for the NP allocation, the power is increased to around 82% while using fewer patients on average even in such a case where the treatment contrasts are quite small. Moreover, the adaptive designs using the NP allocation can also reduce the ENF by about 11 compared with the other designs.

Table 3.9.: Simulated allocation proportions for three-armed binary trials using complete randomisation and response-adaptive randomisation.

p_{E1}	p_{E2}	p_C	N	Procedure	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)	ρ_{E1}	ρ_{E2}	ρ_C
0.3	0.2	0.1	210	CR	0.333	(0.033)	0.333	(0.033)	0.334	(0.033)	0.333	0.333	0.333
				$\text{DBCD}_{\text{D}_\text{A}}$	0.376	(0.026)	0.347	(0.031)	0.277	(0.040)	0.374	0.346	0.280
				$\text{ERADE}_{\text{D}_\text{A}}$	0.375	(0.023)	0.347	(0.029)	0.278	(0.037)	0.374	0.346	0.280
				DBCD_{NP}	0.496	(0.085)	0.281	(0.081)	0.223	(0.029)	0.520	0.250	0.230
				ERADE_{NP}	0.491	(0.086)	0.285	(0.082)	0.224	(0.030)	0.520	0.250	0.230
0.65	0.55	0.5	600	CR	0.333	(0.020)	0.333	(0.020)	0.333	(0.020)	0.333	0.333	0.333
				$\text{DBCD}_{\text{D}_\text{A}}$	0.326	(0.011)	0.336	(0.010)	0.338	(0.010)	0.327	0.336	0.337
				$\text{ERADE}_{\text{D}_\text{A}}$	0.326	(0.008)	0.336	(0.007)	0.337	(0.007)	0.327	0.336	0.337
				DBCD_{NP}	0.464	(0.035)	0.266	(0.035)	0.270	(0.017)	0.479	0.250	0.271
				ERADE_{NP}	0.461	(0.039)	0.269	(0.040)	0.270	(0.018)	0.479	0.250	0.271
0.8	0.6	0.4	81	CR	0.335	(0.054)	0.332	(0.054)	0.333	(0.054)	0.333	0.333	0.333
				$\text{DBCD}_{\text{D}_\text{A}}$	0.302	(0.042)	0.349	(0.035)	0.349	(0.036)	0.303	0.349	0.349
				$\text{ERADE}_{\text{D}_\text{A}}$	0.303	(0.040)	0.349	(0.032)	0.348	(0.032)	0.303	0.349	0.349
				DBCD_{NP}	0.395	(0.060)	0.281	(0.053)	0.324	(0.052)	0.420	0.250	0.330
				ERADE_{NP}	0.395	(0.058)	0.282	(0.054)	0.322	(0.051)	0.420	0.250	0.330
0.8	0.7	0.6	289	CR	0.333	(0.029)	0.334	(0.029)	0.333	(0.029)	0.333	0.333	0.333
				$\text{DBCD}_{\text{D}_\text{A}}$	0.305	(0.027)	0.340	(0.021)	0.355	(0.019)	0.308	0.339	0.353
				$\text{ERADE}_{\text{D}_\text{A}}$	0.305	(0.025)	0.340	(0.019)	0.355	(0.016)	0.308	0.339	0.353
				DBCD_{NP}	0.420	(0.050)	0.279	(0.050)	0.301	(0.034)	0.446	0.250	0.304
				ERADE_{NP}	0.418	(0.050)	0.281	(0.052)	0.300	(0.034)	0.446	0.250	0.304

As shown in Table 3.9, the adaptive designs using the D_A-optimal rule perform well in terms of targeting the optimal allocation for both the accuracy and efficiency. For the NP allocation, treatment assignment using the closed form solution (3.11) is based on the order of the current parameter estimates. When the sample size is small early in the trials, the parameter estimates are less accurate. The order of the parameter estimates can also be inconsistent with respect to the true values. Hence, the variability in the allocation proportions for the NP allocation seems to be higher than for the other designs. These findings support the theory that

the D_A-optimal rule ensures the most efficient estimates, while the NP allocation maximises the power of the test of homogeneity subject to a fixed total number of patients.

Focusing on the second scenario with a large sample size, Tables 3.10 and 3.11 compare the designs under group sequential monitoring with equal and unequal increments in information time. Also, fixed-sample designs are provided alongside. Under the null hypothesis in Table 3.10, the type I error rate can be well preserved for the case when $(t_1,\ t_2,\ t_3) = (0.5,\ 0.8,\ 1)$, where $\tilde{\alpha}$ is within two standard errors of the nominal value. This indicates that the sequential critical boundaries can also be applied as an approximate result to the case of unequally-spaced information times for large sample sizes.

Table 3.10.: Simulated type I error rate for three-armed binary trials using complete randomisation and response-adaptive randomisation, $p_{E1} = p_{E2} = p_C = 0.5,\ N = 600$.

				$(t_1,\ t_2,\ t_3)$=(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	0.053	597.7	(21.4)	298.9	(10.7)	0.333	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD_{D_A}	0.047	598.2	(18.7)	299.1	(9.4)	0.333	(0.009)	0.333	(0.009)	0.333	(0.009)
ERADE_{D_A}	0.051	597.5	(22.5)	298.8	(11.2)	0.333	(0.006)	0.333	(0.006)	0.333	(0.006)
DBCD_{NP}	0.051	598.0	(20.3)	299.0	(10.2)	0.332	(0.093)	0.334	(0.094)	0.334	(0.094)
ERADE_{NP}	0.049	597.2	(23.6)	298.6	(11.8)	0.331	(0.094)	0.336	(0.095)	0.333	(0.094)
				$(t_1,\ t_2,\ t_3)$=(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	0.050	598.9	(11.2)	299.5	(5.6)	0.333	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD_{D_A}	0.050	598.7	(12.4)	299.4	(6.2)	0.333	(0.009)	0.333	(0.009)	0.333	(0.009)
ERADE_{D_A}	0.049	598.6	(13.8)	299.3	(6.9)	0.333	(0.006)	0.333	(0.006)	0.333	(0.006)
DBCD_{NP}	0.048	598.5	(13.7)	299.3	(6.9)	0.333	(0.093)	0.335	(0.093)	0.332	(0.092)
ERADE_{NP}	0.044	598.7	(12.3)	299.4	(6.1)	0.333	(0.095)	0.331	(0.094)	0.336	(0.095)
				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	0.051	600	(0)	300	(0)	0.333	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD_{D_A}	0.052	600	(0)	300	(0)	0.333	(0.009)	0.333	(0.009)	0.333	(0.009)
ERADE_{D_A}	0.052	600	(0)	300	(0)	0.333	(0.006)	0.333	(0.006)	0.333	(0.006)
DBCD_{NP}	0.047	600	(0)	300	(0)	0.332	(0.093)	0.333	(0.094)	0.336	(0.094)
ERADE_{NP}	0.045	600	(0)	300	(0)	0.332	(0.095)	0.333	(0.095)	0.335	(0.096)

Table 3.11.: Simulated power for three-armed binary trials using complete randomisation and response-adaptive randomisation, $p_{E1} = 0.65,\ p_{E2} = 0.55,\ p_C = 0.5,\ N = 600$.

					$(t_1,\ t_2,\ t_3)$=(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	0.805	518.1	(104.3)	244.5	(45.2)	253.4	(9.6)	0.333	(0.020)	0.333	(0.020)	0.333	(0.020)
DBCD$_{D_A}$	0.800	517.5	(103.9)	224.7	(45.0)	253.7	(9.5)	0.326	(0.011)	0.336	(0.010)	0.338	(0.010)
ERADE$_{D_A}$	0.806	517.2	(105.6)	224.5	(45.7)	253.7	(9.6)	0.326	(0.008)	0.336	(0.007)	0.337	(0.007)
DBCD$_{NP}$	0.829	513.5	(105.7)	214.1	(44.0)	244.5	(7.4)	0.464	(0.035)	0.266	(0.035)	0.270	(0.017)
ERADE$_{NP}$	0.823	511.8	(106.0)	213.6	(44.2)	244.6	(7.7)	0.461	(0.039)	0.269	(0.040)	0.270	(0.018)
					$(t_1,\ t_2,\ t_3)$=(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	0.809	536.0	(74.4)	232.2	(32.3)	254.8	(7.4)	0.333	(0.019)	0.333	(0.019)	0.333	(0.019)
DBCD$_{D_A}$	0.809	535.7	(74.3)	232.6	(32.1)	255.2	(7.2)	0.326	(0.010)	0.336	(0.010)	0.337	(0.009)
ERADE$_{D_A}$	0.813	534.8	(74.2)	232.2	(32.1)	255.1	(7.2)	0.326	(0.008)	0.336	(0.007)	0.337	(0.006)
DBCD$_{NP}$	0.832	529.8	(75.5)	220.9	(31.5)	245.6	(5.6)	0.465	(0.035)	0.266	(0.035)	0.270	(0.017)
ERADE$_{NP}$	0.814	530.6	(76.4)	221.3	(31.9)	245.7	(5.7)	0.463	(0.037)	0.267	(0.038)	0.270	(0.018)
					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	0.812	600	(0)	260.0	(1.4)	-	-	0.333	(0.018)	0.333	(0.018)	0.333	(0.018)
DBCD$_{D_A}$	0.813	600	(0)	260.5	(0.7)	-	-	0.327	(0.009)	0.336	(0.009)	0.337	(0.009)
ERADE$_{D_A}$	0.813	600	(0)	260.5	(0.5)	-	-	0.327	(0.007)	0.336	(0.006)	0.337	(0.006)
DBCD$_{NP}$	0.833	600	(0)	250.2	(2.3)	-	-	0.465	(0.036)	0.265	(0.036)	0.270	(0.017)
ERADE$_{NP}$	0.828	600	(0)	250.2	(2.3)	-	-	0.464	(0.035)	0.266	(0.036)	0.270	(0.017)

The target D_A-optimal and NP allocations are (0.327, 0.336, 0.337) and (0.479, 0.25, 0.271), respectively.

Under the alternative hypothesis, in Table 3.11, the ENP and hence the ENF are lower for the group sequential designs than for the fixed-sample designs. To see the advantages of using response-adaptive designs over CR using the same number of subjects, the ENF is calculated based on the maximum number of patients. 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. Based on the same sample size, the ENF$'$ for the group sequential designs can be used to compare with the ENF for the fixed-sample designs. In practice, trials stop when a decision is made.

3.3.3. 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 $\boldsymbol{\theta}_G = (\theta_{E1} - \theta_C,\ \theta_{E2} - \theta_C)^T$ using test statistic (3.7), 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 the 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 optimal allocation based on nonlinear programming (NP) were 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 was set to be 0.20 to satisfy $B \in [0, \tilde{B}]$, $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_3, 1/3)$, in (3.12). There are $K = 3$ group sequential tests at equally spaced information times. The type I error rate, the approximate critical boundaries and other settings are the same as in Section 3.3.1.

From Table 3.12, we find that the critical boundaries derived based on normal responses 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}.

Table 3.12.: Simulated type I error rate for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = \theta_{E2} = \theta_C = 24,\ N = 312$.

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)

Under the alternative hypothesis, from Table 3.13, 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 were 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.

Table 3.13.: Simulated power for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = 34,\ \theta_{E2} = 24,\ \theta_C = 20,\ N = 312$.

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, 0.2, 0.256), respectively.

Tables 3.14 and 3.15 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 3.12 and 3.13, the mean survival time for each treatment is increased. When the mean survival time is 24 months as in Table 3.12, the probability of an event computed by (2.11) is 0.62, whereas, when the mean survival time is 45 months, as in Table 3.14, the probability is 0.45. Thus, the chance of observing a non-censored event is high when survival is poor. Therefore, the probability of a censored response is increased for longer mean survival times. In other words, more patients would not have responded by the end of the study.

Tables 3.14 and 3.15 demonstrate similar behaviour as Tables 3.12 and 3.13. 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 3.14.: Simulated type I error rate for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = \theta_{E2} = \theta_C = 45,\ N = 600$.

				$(t_1,\ t_2,\ t_3)$=(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)

It is worth mentioning that the D_A-optimal allocation is always ethical for censored survival responses. More specifically, the D_A-optimal rule assigns more patients to the treatment with a higher variance in the responses, which corresponds to the treatment with the higher mean survival time. Sverdlov et al. (2011) proved in theory that, if $\theta_1 \geq ... \geq \theta_J$, then $\rho_1 \geq ... \geq \rho_J$ for the D_A-optimal allocation. The results displayed in Tables 3.13 and 3.15 support the theory. However, the

property does not hold for normal and binary responses, where the most efficient allocation may assign more patients to the less promising treatments.

Table 3.15.: Simulated power for three-armed censored survival trials using complete randomisation and response-adaptive randomisation, $\theta_{E1} = 59,\ \theta_{E2} = 45,\ \theta_C = 37,\ N = 600$.

					$(t_1,\ t_2,\ t_3)$=(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, 0.2, 0.281), respectively.

3.3.4. Redesigning a four-armed binary trial

In the previous sections, three-treatment comparisons were considered. The adaptive designs can also be applied to multi-armed clinical trials with $J > 3$ treatments. In this section, a four-armed binary trial is redesigned. NeoSphere (Gianni et al., 2012) is a phase II randomised clinical 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 ei-

ther 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 numbers of patients in the control group and on the experimental treatment arms $E1$ and $E2$ were 107, and 96 patients were assigned to $E3$ since this was added to the study after a protocol amendment. The endpoint considered in the study was pathological complete response, which is 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.

Here, the trial is redesigned using the adaptive designs without dropping inferior treatments. 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. For the combined approach, the first 40 patients were randomly assigned using permuted-block randomisation with ratio 1:1:1:1 to obtain initial parameter estimates. Then optimal response-adaptive randomisation using the D_A-optimal allocation and the optimal allocation based on nonlinear programming (NP) was performed.

For the NP allocation, the user-specified lower bound for the allocation proportions B was set to be 0.20 to satisfy $B \in [0, \tilde{B}]$, where $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_4, 1/4)$ in (3.11). The closed form solution for 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 the optimal allocation proportions based on nonlinear programming which maximise the power subject to the total sample size not exceeding a

fixed value are

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

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 $B = 1/4$, the solution becomes equal allocation.

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 (3.9), where $J - 1 = 3$.

The nominal type I error rate was set to 5% and $K = 3$ group sequential tests were planned at equally spaced information times. The O'Brien and Fleming critical boundaries (23.76, 11.88, 7.92) for $J = 4$ treatments obtained from Table 16.1 in Jennison and Turnbull (2000) were used as an approximation. The boundaries were derived based on normal responses with equal variances. Results for the group sequential CR design and the fixed-sample CR and response-adaptive designs are also provided for comparison. For the fixed-sample designs, the critical boundary is 7.81. The simulation results are based on 5,000 replicates.

Under the null hypothesis, from Table 3.16, the type I error rate 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.

Table 3.16.: Simulated type I error rate for redesigning NeoSphere trial using complete randomisation and response-adaptive randomisation, $p_C = 0.29,\ p_{E1} = 0.29,\ p_{E2} = 0.29,\ p_{E3} = 0.29,\ N = 417$.

					$(t_1,\ t_2,\ t_3)$=(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)

Since there are significant differences in the treatment effects, a high probability of rejecting H_{G_0} under the alternative hypothesis is obtained for all of the designs: see Table 3.17. This agrees with the finding of 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 (Gianni et al., 2012).

A similar figure for the expected number of failures (ENF), 296.5, is found for the fixed-sample CR design. If fixed-sample response-adaptive designs are 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 are used, a further reduction in the ENF could be obtained. Since the expected number of patients (ENP) for the group sequential designs is substantially lower than the fixed-sample designs, the ENF is also decreased. If the group-sequential response-adaptive design with D_A-optimal allocation is applied, around 86 failures could be avoided. If the NP allocation is used, about 109 fewer failures can be achieved.

Table 3.17.: Simulated power for redesigning NeoSphere trial using complete randomisation and response-adaptive randomisation, $p_C = 0.29,\ p_{E1} = 0.458,\ p_{E2} = 0.168,\ p_{E3} = 0.24,\ N = 417$.

						$(t_1,\ t_2,\ t_3)$=(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 (0.2, 0.479, 0.121, 0.2), respectively.

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

3.4. Conclusions

Based on the results obtained in this chapter, the critical boundaries derived based on normal responses with equal allocation and equally spaced information times (Jennison and Turnbull, 1991, 2000) can be used as an approximation for the adaptive design with equal and unequally spaced information times for multi-armed normal trials. For trials with binary and censored survival responses, the overall type I error rate can also be controlled when the sample size is large.

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.

For the NP allocation, more patients are assigned to the best and the worst treatments, while the allocation proportions for the other treatment(s) are fixed. Such allocation is according to the order of the current parameter estimates. The closed form solution for NP allocation is available for binary and survival responses with one parameter. However, for normal responses, which involve a nuisance parameter, a solution has not yet been derived. Here, the NP allocation that maximises the power subject to the total sample size not exceeding a fixed value is considered. For NP allocation which maximises the power subject to a fixed total number of failures, there is no closed form for the solution. Numerical methods are required. The D_A-optimal allocation, which depends on the variance of the responses for each treatment, can be used for all types of responses.

4. Group-sequential response-adaptive designs for multi-armed trials with dropping of inferior arm(s)

4.1. Global and pairwise tests

As can be seen from the simulation results in Chapter 3, the combined approach can be used for global tests of homogeneity. However, when the global null hypothesis is rejected at an interim analysis, one probably does not want to end the entire trial. Testing which treatments differ is usually of concern. The group sequential Fisher's least significant difference (LSD) method, which consists of a global test and pairwise comparisons, has been studied (Proschan et al., 1994). 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. However, the authors considered the cases of equal allocation and fixed unequal allocation determined before the commencement of the experiments.

In this chapter, an analogue of Fisher's LSD method that generalises to group-

sequential response-adaptive designs is investigated. The test statistics described in Chapters 2 and 3 can be used for the pairwise and global tests, respectively. For simplicity, the design for three-armed normal trials is illustrated, which can be applied to other types of responses and extended to more than three treatment arms.

Let K be the number of group sequential analyses. First, we wish to test the global null hypothesis $H_{G_0} : \boldsymbol{\mu}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{\mu}_G \neq \mathbf{0}$, where $\boldsymbol{\mu}_G = (\mu_{E1}-\mu_C,\ \mu_{E2}-\mu_C)^T$ is a vector of treatment contrasts and C refers to the control group. The test statistics S_k and the corresponding critical boundaries d_k at look $k,\ k = 1, ..., K$, can be found in Chapter 3. If $S_k < d_k,\ k = 1, ..., K-1$, 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 $S_k \geq d_k,\ k = 1, ..., K$, we reject H_{G_0} and start pairwise comparisons.

For all pairwise comparisons, three pairwise tests are conducted at each look. For comparisons with a control, there are two pairwise tests. Here, comparisons with a control are considered. 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 shown in Chapter 2 can be used. These critical boundaries do not take into account the number of pairwise tests conducted, as for Fisher's LSD method in a fixed-sample design. 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 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. There are sev-

eral outcomes of the pairwise tests for $J = 3$ treatments, which include the control.

a. 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.

b. 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.

c. 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.

d. 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.

e. 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.

f. 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.

g. If $Z_{1C,k} \leq -c_k$ and $-c_k < Z_{2C,k} < c_k$, $k = 1, ..., K-1$, 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 $Z_{2C,k} \leq -c_k$, $k = 1, ..., K-1$, 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 $-c_k < Z_{2C,k} < c_k$, $k = 1, ..., K-1$, 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.

For cases a, b, c and d, where both pairwise null hypotheses are rejected, the trial stops. For cases e, f, g and h, where one of the two null hypotheses is rejected, since a superior treatment has been found for cases e and f, the trial stops. For cases g and h, where an inferior treatment is identified, the design allows dropping of the inferior arm and proceeds to the next look with more information to test the difference between the other experimental treatment and the control. 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 treatments J increases, there are more possible outcomes. For comparisons with a control, there are 3^{J-1} outcomes. For all pairwise comparisons, there are $3^{J(J-1)/2}$ outcomes.

4.2. Information time

Immediate responses

The formulae for the information time for trials that allow dropping of inferior treatment(s) are different from those described in previous chapters. Suppose that N is the planned maximum number of patients. Recall that the information time for immediate responses is a ratio of the current sample size to the maximum sample size N. For trials with dropping of inferior treatments, Follmann et al. (1994) defined the information time as 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,

$$t_k = \frac{\sum_{j\in\mathcal{C}} m_{j,k}}{\sum_{j\in\mathcal{C}} M_j} \in (0,1], \quad k = 1, ..., K,$$

where $\mathcal{C}$ represents the current set of treatments remaining in the study, which is a subset of $\{1, ..., J\}$, $m_{j,k}$ is the cumulative number of patients on treatment j at look k and M_j is the total sample size on treatment j. The authors proved that

$$t_k = \frac{n_k}{N} = \frac{\sum_{j=1}^{J} m_{j,k}}{\sum_{j=1}^{J} M_j} = \frac{\sum_{j\in\mathcal{C}} m_{j,k}}{\sum_{j\in\mathcal{C}} M_j},$$

where n_k is the cumulative sample size at look k. This shows that the information time at look k remains the same if some treatments are dropped at this look. The information time increases when the trial proceeds, with $t_0 = 0$ and $t_K = 1$.

For optimal response-adaptive randomisation, the total sample size on treatment j, M_j, is random. However, M_j can be approximated by $\rho_j N$, since M_j/N converges almost surely to the pre-specified desired optimal allocation proportion for treatment j, ρ_j (Hu and Zhang, 2004). The approximate information time becomes

$$t_k = \frac{\sum_{j\in\mathcal{C}} m_{j,k}}{\sum_{j\in\mathcal{C}} \hat{\rho}_j N} \in (0,1], \quad k = 1, ..., K, \tag{4.1}$$

where $\hat{\rho}_j$ is the current estimate of ρ_j, since the optimal allocation proportions usually depend on the unknown parameters. Then the optimal allocation proportions $(\rho_1, ..., \rho_J)$ for multi-armed clinical trials described in Section 3.2.1 can be used. More specifically, the optimal allocation proportions without treatment dropping are used in (4.1), so that $\hat{\rho}_j N$ is an estimate of the original planned total sample size on treatment j, M_j. Note that $\sum_{j\in C} \hat{\rho}_j \leq 1$, since $\sum_{j=1}^{J} \hat{\rho}_j = 1$.

Applying (4.1) with the error-spending approach, the number of interim tests, K, is not required to be pre-specified, and interim analysis can be planned at any continuous information time $t_k \in [0, 1)$. In addition, one can calculate the total expected number of patients needed at look k when some inferior treatment(s)

have been dropped as

$$\sum_{j\in C} m_{j,k} = \lceil t_k \sum_{j\in C} \hat{\rho}_j N \rceil,$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x.

For equal allocation, $\hat{\rho}_j = 1/J,\ j = 1, ..., J$. Let $m_{.k}$ be the cumulative sample size for each treatment at look k and et r be the number of arms remaining in the study. Then the approximate information time reduces to

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

So the number of patients needed for each treatment at look k is $m_{.k} = \lceil t_k N/J \rceil$.

Censored survival responses

Recall that, for censored survival responses, the information time is proportional to the number of events. For censored survival trials with treatment dropping, 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,

$$t_k = \frac{\hat{e}_k}{\hat{e}_K^{(k)}} = \frac{\sum_{j\in \mathcal{C}} m_{j,k} \hat{\epsilon}_{j,k}}{\sum_{j\in \mathcal{C}} M_j \hat{\epsilon}_{j,K}},$$

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 evaluated based on the responses obtained so far and $\hat{\epsilon}_{j,k}$ is the estimated probability of an event on arm j at look k. As mentioned in Sections 2.3.1 and 3.1.1, 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. However, both can be used with the error-spending function to control the overall type I error rate (Kim et al., 1995).

For simplicity, we consider the information time scale under H_{G_0}, where $\theta_1 = ... = \theta_J$, 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

$$t_k = \frac{\sum_{j \in \mathcal{C}} m_{j,k} \hat{\epsilon}_k}{\sum_{j \in \mathcal{C}} M_j \hat{\epsilon}_K} = \frac{\sum_{j \in \mathcal{C}} m_{j,k} \hat{\epsilon}_k}{\sum_{j \in \mathcal{C}} \hat{\rho}_j N \hat{\epsilon}_K} \in (0, 1], \quad k = 1, ..., K. \tag{4.2}$$

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 the O'Brien and Fleming critical boundaries, which allocate little type I error rate to the early group sequential tests, seems to be more sensible.

4.3. Optimal response-adaptive randomisation

Permuted-block randomisation is used for the first 10% of the maximum number of patients, N, to obtain initial parameter estimates. Then the optimal allocation proportions $(\rho_1, ..., \rho_J)$ for multi-armed clinical trials, the D_A-optimal allocation and the optimal allocation based on nonlinear programming (NP), described in Section 3.2.1, can be used when no treatment has been dropped. However, if the number of arms remaining in the study decreases as one or more inferior treatments are dropped, the allocation proportions for the remaining arms will increase (Follmann et al., 1994).

The new optimal allocation proportion for treatment j, $\rho_j^{'}$, after some arm(s)

have been dropped becomes

$$\rho_j^{'} = \frac{M_j}{\sum_{l\in\mathcal{C}} M_l} = \frac{\hat{\rho}_j N}{\sum_{l\in\mathcal{C}} \hat{\rho}_l N} = \frac{\hat{\rho}_j}{\sum_{l\in\mathcal{C}} \hat{\rho}_l}, \qquad \sum_{j\in\mathcal{C}} \rho_j^{'} = 1, \tag{4.3}$$

where $\mathcal{C}$ 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 treatment(s) based on the responses available.

The optimal response-adaptive randomisation procedures DBCD and ERADE for multi-armed trials described in Section 3.2.2 can be applied to target the optimal allocation proportion ρ_j or $\rho_j^{'}$ if some treatments have been dropped.

4.4. Simulation studies

4.4.1. Three-armed normal trials

Consider comparing $J = 3$ treatments using the analogue of Fisher's LSD method. Let $E1$ and $E2$ denote the experimental treatments and C refers to the control. First, for the test of homogeneity, the global null hypothesis $H_{G_0} : \boldsymbol{\mu}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{\mu}_G \neq \mathbf{0}$, where $\boldsymbol{\mu}_G = (\mu_{E1} - \mu_C,\ \mu_{E2} - \mu_C)^T$, 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 three equally spaced group sequential analyses were used. The critical values for the chi-squared statistics are obtained from Table 16.1 in Jennison and Turnbull (2000).

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 for each

pairwise Z test. These critical values were computed for the equally spaced information times (0.33, 0.67, 1) using a program provided by Proschan et al. (2006).

The two group-sequential response-adaptive designs DBCD and ERADE are compared with the group-sequential non-adaptive design CR in terms of the error probabilities (I: the probability of rejecting at least one of the pairwise null hypotheses and II: the probability of rejecting all pairwise null hypotheses), the expected number of patients (ENP) and the average allocation proportions with standard deviations. For the optimal response-adaptive designs, permuted-block randomisation is used for the first 10% of the N patients to obtain initial parameter estimates. Then the DBCD and ERADE functions with tuning parameters $\gamma = \gamma^{'} = 2$ are used to compute the allocation probability for the next patient. For normal responses, the D_A-optimal allocation is used as the desired allocation. The results are based on 5,000 replicates.

Treatment(s) inferior to the control are allowed to be dropped after the pairwise tests. Taking different reference groups may influence the decisions of treatment dropping and/or trial termination. More precisely, when $\mu_{E1} > \mu_{E2} > \mu_C$ or $\mu_{E2} > \mu_{E1} > \mu_C$, the test are unlikely to identify and drop an experimental treatment inferior to the control. However, they may result in early stopping of the trial with the claim that one or more experimental treatments are superior to the control. If both experimental treatments are found to be significantly inferior to the control ($\mu_C > \mu_{E1} > \mu_{E2}$ or $\mu_C > \mu_{E2} > \mu_{E1}$), the trial stops. If one experimental treatment has shown significant inefficacy but the other one has not yet shown any difference in efficacy compared to the control ($\mu_{E1} > \mu_C > \mu_{E2}$ or $\mu_{E2} > \mu_C > \mu_{E1}$), we drop the inferior arm and continue the others to the next look. The order of $E1$ and $E2$ does not affect the results. In this section, one of each case is investigated.

As shown in Table 4.1, 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 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 4.1.: 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	299.3	(8.0)	0.334	(0.026)	0.333	(0.026)	0.334	(0.026)
DBCD_{D_A}	0.048	0.002	299.3	(7.3)	0.453	(0.016)	0.355	(0.019)	0.192	(0.019)
ERADE_{D_A}	0.047	0.004	299.4	(6.7)	0.451	(0.013)	0.355	(0.016)	0.194	(0.017)

Table 4.2.: 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	0.777	0.230	264.2	(49.1)	0.333	(0.027)	0.334	(0.028)	0.334	(0.028)
DBCD_{D_A}	0.808	0.276	260.2	(51.5)	0.453	(0.016)	0.355	(0.020)	0.192	(0.020)
ERADE_{D_A}	0.815	0.273	259.3	(51.0)	0.451	(0.013)	0.355	(0.017)	0.194	(0.019)

Table 4.2 considers the case where the control is inferior to the two experimental therapies. In this case, the ENP is about 264 for group sequential CR design and 260 for the 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 with-

out dropping inferior treatments is (0.454, 0.356, 0.191). Both adaptive designs target the D_A-optimal allocation proportion well, with the ERADE consistently having lower standard deviations for the allocation proportions.

Table 4.3.: 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	0.614	0.086	280.0	(39.2)	0.331	(0.028)	0.335	(0.027)	0.334	(0.027)
DBCD$_{D_A}$	0.653	0.092	279.8	(39.0)	0.446	(0.028)	0.358	(0.024)	0.197	(0.020)
ERADE$_{D_A}$	0.652	0.086	280.6	(38.5)	0.445	(0.025)	0.357	(0.022)	0.198	(0.019)

Table 4.3 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, $\rho_{E2}^{'}$, $\rho_C^{'}$), 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 4.4 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 arm $E1$ is dropped, the target D_A-optimal allocation becomes (0, 0.651, 0.349). If arm $E2$ 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 arm(s) 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 the 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.

Table 4.4.: 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	0.983	0.221	264.9	(26.7)	0.363	(0.034)	0.273	(0.049)	0.364	(0.035)
DBCD$_{D_A}$	0.976	0.250	270.1	(26.3)	0.490	(0.039)	0.292	(0.053)	0.218	(0.025)
ERADE$_{D_A}$	0.969	0.253	270.5	(26.1)	0.488	(0.038)	0.293	(0.053)	0.219	(0.025)

4.4.2. Three-armed binary trials

Consider testing the contrasts of the probabilities of success between two experimental treatments and a control. First, we wish to test the global null hypothesis $H_{G_0} : \boldsymbol{p}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{p}_G \neq \mathbf{0}$, where $\boldsymbol{p}_G = (p_{E1} - p_C,\ p_{E2} - p_C)^T$. If H_{G_0} is rejected, then pairwise comparisons with the null hypotheses $H_0^{(j)} : p_{Ej} = p_C$ versus the alternative hypotheses $H_a^{(j)} : p_{Ej} \neq p_C,\ j = 1, 2$, are tested subsequently. For the response-adaptive designs, the D_A-optimal allocation and the optimal allocation based on nonlinear programming (NP) are used and compared. For the NP allocation, the lower bound for the allocation proportions B is set to be 0.25. This is chosen as $B \in [0, \tilde{B}],\ \tilde{B} = \min(\tilde{B}_1, \tilde{B}_3, 1/3)$, where $\tilde{B}_1$ and $\tilde{B}_3$ are obtained from (3.11). The other simulation settings are the same as in Section 4.4.1.

Under H_{G_0}, from Table 4.5, there is a conservative type I error rate $\tilde{\alpha}^I$ for all of the designs. This may be because small treatment contrasts with large variances are obtained when $p_{E1} = p_{E2} = p_C = 0.5$. Under H_{G_0}, the probabilities of early termination and dropping arms are small. The ENP is about 600 and the ENF is about 300. In addition, the average allocation proportions are close to equal allocation for all of the designs. The use of the D_A-optimal allocation yields the smallest standard deviations for the allocation proportions, whereas the use of the NP allocation gives the largest variability, since this assigns patients according to the order of the parameter estimates, which is random under H_{G_0}.

Table 4.5.: Simulated type I error rates for three-armed binary trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $p_{E1} = p_{E2} = p_C = 0.5,\ N = 600$.

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	0.040	0.008	599.0	(12.4)	299.7	(13.6)	0.333	(0.019)	0.333	(0.019)	0.333	(0.018)
DBCD_{D_A}	0.040	0.008	598.8	(14.3)	299.6	(14.0)	0.333	(0.010)	0.333	(0.010)	0.333	(0.009)
ERADE_{D_A}	0.042	0.007	598.8	(15.1)	299.7	(14.3)	0.333	(0.007)	0.333	(0.007)	0.333	(0.006)
DBCD_{NP}	0.039	0.006	598.8	(14.1)	299.5	(14.2)	0.330	(0.092)	0.335	(0.094)	0.335	(0.094)
ERADE_{NP}	0.036	0.006	599.3	(10.5)	299.9	(13.2)	0.332	(0.095)	0.333	(0.095)	0.334	(0.095)

Table 4.6.: Simulated powers for three-armed binary trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $p_{E1} = 0.65,\ p_{E2} = 0.55,\ p_C = 0.5,\ N = 600$.

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	0.792	0.103	522.7	(103.2)	226.2	(45.9)	0.333	(0.020)	0.333	(0.019)	0.333	(0.019)
DBCD_{D_A}	0.786	0.102	521.9	(103.6)	226.3	(46.1)	0.326	(0.011)	0.336	(0.010)	0.337	(0.010)
ERADE_{D_A}	0.794	0.098	522.8	(101.9)	226.7	(45.4)	0.326	(0.008)	0.336	(0.007)	0.337	(0.007)
DBCD_{NP}	0.801	0.063	518.4	(103.0)	216.1	(44.2)	0.463	(0.038)	0.266	(0.037)	0.270	(0.018)
ERADE_{NP}	0.791	0.071	521.4	(102.5)	217.2	(43.7)	0.463	(0.036)	0.267	(0.037)	0.270	(0.017)

Table 4.6 considers the case where the control is inferior to the two experimental treatments. For the response-adaptive designs targeting the D_A-optimal allocation, the target allocation is (0.327, 0.336, 0.337), which is close to equal allocation, since the difference in the variances of the responses is small among the treatment arms in this case. As a result, the simulation results for the response-adaptive designs targeting the D_A-optimal allocation are similar to CR. For the response-adaptive designs targeting the NP allocation, the target allocation is (0.479, 0.25,

0.271). These designs require slightly fewer patients on average and reduce the number of failures by about 10 compared to other designs.

Table 4.7 considers the case where one experimental treatment is superior and one is inferior to the control group. For the DBCD and the ERADE targeting the D_A-optimal allocation, a higher power is achieved with a lower ENP compared to other designs, whereas, for the response-adaptive designs targeting the NP allocation, a lower ENF is obtained, since, in this case, the average allocation proportion for $E1$, $\tilde{\rho}_{E1}$, is 0.466. Without dropping treatments, the target D_A-optimal allocation is (0.327, 0.337, 0.336) and the target NP allocation is (0.479, 0.271, 0.25). If $E2$ is dropped, the target allocation becomes (0.493, 0, 0.507) for the D_A-optimal allocation and (0.657, 0, 0.343) for the NP allocation.

Table 4.7.: Simulated powers for three-armed binary trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $p_{E1} = 0.65,\ p_{E2} = 0.5,\ p_C = 0.55,\ N = 600$.

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	0.641	0.029	560.1	(79.6)	242.3	(36.5)	0.335	(0.021)	0.330	(0.026)	0.335	(0.021)
DBCD$_{D_A}$	0.648	0.028	558.3	(80.9)	242.2	(36.9)	0.328	(0.014)	0.334	(0.020)	0.338	(0.013)
ERADE$_{D_A}$	0.654	0.033	556.9	(83.0)	241.7	(37.7)	0.328	(0.012)	0.333	(0.020)	0.338	(0.011)
DBCD$_{NP}$	0.626	0.017	560.3	(79.8)	233.3	(35.3)	0.466	(0.035)	0.267	(0.019)	0.266	(0.036)
ERADE$_{NP}$	0.625	0.015	562.4	(77.7)	234.4	(34.6)	0.466	(0.035)	0.267	(0.019)	0.267	(0.037)

Table 4.8 considers the case where the experimental treatments are inferior to the control. In this case, a slightly higher power^I yet lower power^{II} are obtained for the response-adaptive designs using NP allocation compared to the other designs. In addition, these designs yield around eight fewer failures. Since more patients are assigned to the best treatment, which is the control group in this case, $\tilde{\rho}_C$=0.478. Without dropping treatments, the target allocations for the D_A-optimal and the NP allocations are (0.336, 0.337, 0.327) and (0.25, 0.271, 0.479), respectively. If $E1$ is dropped, the target allocations become (0, 0.508, 0.492) for the D_A-optimal allocation and (0, 0.361, 0.639) for the NP allocation. If $E2$ is dropped, the target allocations are (0.507, 0, 0.493) for the D_A-optimal allocation and (0.343, 0, 0.657)

for the NP allocation. In addition, the difference in the simulation results between the DBCD and the ERADE is small.

Table 4.8.: Simulated powers for three-armed binary trials using complete randomisation and response-adaptive randomisation with dropping of inferior treatment, $p_{E1} = 0.55,\ p_{E2} = 0.5,\ p_C = 0.65,\ N = 600$.

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	0.795	0.490	551.9	(73.7)	238.0	(34.2)	0.341	(0.029)	0.315	(0.042)	0.344	(0.027)
DBCD$_{D_A}$	0.797	0.493	551.0	(75.1)	238.2	(34.6)	0.344	(0.026)	0.320	(0.039)	0.336	(0.021)
ERADE$_{D_A}$	0.791	0.496	549.8	(76.2)	237.6	(35.2)	0.344	(0.025)	0.320	(0.038)	0.336	(0.020)
DBCD$_{NP}$	0.810	0.474	555.5	(70.0)	230.4	(31.8)	0.270	(0.038)	0.252	(0.032)	0.478	(0.043)
ERADE$_{NP}$	0.819	0.472	555.9	(69.0)	230.4	(31.3)	0.270	(0.039)	0.252	(0.031)	0.478	(0.042)

4.4.3. Three-armed censored survival trials

Consider testing the contrasts of the mean survival times between two experimental treatments and the control. We wish to test the global null hypothesis $H_{G_0} : \boldsymbol{\theta}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{\theta}_G \neq \mathbf{0}$, where $\boldsymbol{\theta}_G = (\theta_{E1} - \theta_C,\ \theta_{E2} - \theta_C)^T$. 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 B is set to be 0.2. The value is chosen such that $B \in [0, \tilde{B}]$, $\tilde{B} = \min(\tilde{B}_1, \tilde{B}_3, 1/3)$, where $\tilde{B}_1$ and $\tilde{B}_3$ are obtained from (3.12), which depend on the probability of having a non-censored event. The nominal type I error rate, the approximate critical boundaries and the other simulation settings are the same as in Section 4.4.1.

Under the null hypothesis, from Table 4.9, 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 designs targeting the NP allocation having larger standard deviations for the allocation proportions compared to other designs.

Table 4.9.: 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	0.005	311.7	(4.2)	203.5	(8.6)	0.333	(0.026)	0.333	(0.025)	0.333	(0.025)
DBCD$_{D_A}$	0.050	0.006	311.4	(6.3)	203.2	(9.4)	0.334	(0.034)	0.334	(0.034)	0.333	(0.034)
ERADE$_{D_A}$	0.050	0.008	311.3	(7.1)	203.2	(9.8)	0.334	(0.032)	0.333	(0.032)	0.333	(0.031)
DBCD$_{NP}$	0.042	0.007	311.3	(7.1)	203.2	(9.6)	0.333	(0.091)	0.334	(0.091)	0.333	(0.092)
ERADE$_{NP}$	0.044	0.005	311.5	(6.1)	203.2	(9.2)	0.332	(0.088)	0.334	(0.089)	0.334	(0.090)

Under the alternative hypothesis, in Table 4.10, the target D_A-optimal allocation without dropping 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.

Table 4.10.: 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	0.716	0.094	295.7	(32.2)	186.5	(25.4)	0.333	(0.026)	0.333	(0.026)	0.333	(0.026)
DBCD$_{D_A}$	0.773	0.112	286.0	(37.8)	176.3	(29.9)	0.408	(0.029)	0.325	(0.038)	0.267	(0.040)
ERADE$_{D_A}$	0.779	0.107	285.5	(37.8)	175.9	(29.6)	0.407	(0.026)	0.325	(0.034)	0.269	(0.036)
DBCD$_{NP}$	0.799	0.046	280.2	(40.1)	168.3	(32.2)	0.530	(0.092)	0.230	(0.062)	0.240	(0.051)
ERADE$_{NP}$	0.802	0.050	279.3	(40.3)	167.7	(32.3)	0.526	(0.091)	0.233	(0.064)	0.240	(0.049)

For Table 4.11, 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 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 other designs. However, the designs using the NP sampling rule achieve around seven fewer failures compared to the designs 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 designs 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.

Table 4.11.: 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	0.538	0.021	306.6	(18.3)	197.2	(15.9)	0.331	(0.027)	0.334	(0.026)	0.334	(0.026)
DBCD$_{D_A}$	0.615	0.039	300.3	(26.2)	189.4	(21.8)	0.268	(0.043)	0.409	(0.029)	0.324	(0.038)
ERADE$_{D_A}$	0.620	0.040	300.5	(25.9)	189.7	(21.3)	0.269	(0.042)	0.407	(0.027)	0.324	(0.036)
DBCD$_{NP}$	0.593	0.011	296.3	(30.0)	182.7	(25.3)	0.241	(0.050)	0.529	(0.090)	0.230	(0.061)
ERADE$_{NP}$	0.589	0.015	296.2	(30.0)	182.7	(25.3)	0.243	(0.049)	0.524	(0.088)	0.232	(0.062)

Table 4.12.: 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.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_C$	(s.d.)
CR	0.705	0.382	304.2	(18.8)	194.1	(15.3)	0.324	(0.034)	0.338	(0.028)	0.338	(0.028)
DBCD$_{D_A}$	0.763	0.466	299.1	(25.1)	187.1	(20.5)	0.259	(0.052)	0.328	(0.042)	0.413	(0.034)
ERADE$_{D_A}$	0.768	0.465	299.6	(24.7)	187.4	(20.3)	0.261	(0.050)	0.328	(0.040)	0.411	(0.031)
DBCD$_{NP}$	0.809	0.480	293.4	(28.7)	179.8	(23.9)	0.229	(0.054)	0.230	(0.060)	0.540	(0.093)
ERADE$_{NP}$	0.807	0.480	293.0	(29.2)	179.4	(24.4)	0.231	(0.053)	0.234	(0.061)	0.535	(0.093)

For Table 4.12, 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, 0.443, 0.557) for the D_A-optimal allocation and (0, 0.269, 0.731) for the NP allo-

cation. 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 using the D_A-optimal allocation.

4.4.4. Redesigning a four-armed binary trial

The NeoSphere trial (Gianni et al., 2012) is redesigned using the adaptive designs with dropping of inferior treatments during the course of the trial. The probability of success for each treatment is $p_C = 0.29$, $p_{E1} = 0.458$, $p_{E2} = 0.168$ and $p_{E3} = 0.24$. First, a chi-squared test statistic is monitored for tests of homogeneity. The global null hypothesis is $H_{G_0} : \boldsymbol{p}_G = \mathbf{0}$ versus the alternative hypothesis $H_{G_a} : \boldsymbol{p}_G \neq \mathbf{0}$ with $\boldsymbol{p}_G = (p_{E1} - p_C,\ p_{E2} - p_C,\ p_{E3} - p_C)^T$. 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 chi-squared statistics, the sequence of critical boundaries at information times $(t_1, t_2, t_3) = (0.33, 0.67, 1)$ is (23.76, 11.88, 7.92).

If the global null hypothesis is rejected, then $J - 1 = 4 - 1 = 3$ pairwise Z tests are carried out at the current and subsequent looks. The 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). The following two error probabilities are considered: the probability of rejecting at least one of the three null hypotheses, that is, $\tilde{\alpha}^I$ under the null hypotheses and power^I under the alternative hypotheses; and the probability of rejecting all three null hypotheses, that is, $\tilde{\alpha}^{II}$ under the null hypotheses and power^{II} under the alternative hypotheses.

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. More specifically, trials stop when one or more of the treatments have been found superior to the control or when all of the experimental treatments are inferior to the control. Trials proceed when more information is needed on all of the treatments or when some treatment(s) have been found inferior to the control while others have not been shown to be significantly different from the control. 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 $|Z_{3C,k}| < c_k$, $k = 1, ..., K-1$, 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 $|Z_{3C,k}| < c_k$, $k = 1, ..., K-1$, 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 $Z_{3C,k} \leq -c_k$, $k = 1, ..., K-1$, 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 $|Z_{3C,k}| < c_k$, $k = 1, ..., K-1$, then $E1$ and $E2$ are dropped. $E3$ and the control are continued to the next look.

(v). If $Z_{1C,k} \leq -c_k$, $|Z_{2C,k}| < c_k$ and $Z_{3C,k} \leq -c_k$, $k = 1, ..., K-1$, then $E1$ and $E3$ are dropped. $E2$ and the control are continued to the next look.

(vi). If $|Z_{1C,k}| < c_k$, $Z_{2C,k} \leq -c_k$ and $Z_{3C,k} \leq -c_k$, $k = 1, ..., K-1$, then $E2$ and $E3$ are dropped. $E1$ and the control are continued to the next look.

(vii). If $|Z_{1C,k}| < c_k$, $|Z_{2C,k}| < c_k$ and $|Z_{3C,k}| < c_k$, $k = 1, ..., K-1$, then all treatments are continued to the next look.

In addition, fixed-sample designs are provided alongside for comparison. For these, the critical boundary for the chi-squared test is 7.81. The critical value for all three pairwise test statistics is 1.96. For optimal response-adaptive randomisation, the D_A-optimal allocation and the optimal allocation based on nonlinear programming (NP) described in Section 3.2.1 can be used if no arm has been dropped. For the NP optimal allocation, the lower bound for the allocation proportions B is set to be 0.2. After dropping treatments, (4.3) is used to obtain the optimal allocation proportions for the remaining arms.

As can be seen in Table 4.13, 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 hypotheses, 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 than for the other designs.

Under the alternative hypothesis, in Table 4.14, a similar power is obtained for

Table 4.13.: Simulated type I error rates for redesigning NeoSphere trial using complete randomisation and response-adaptive randomisation with dropping of inferior treatment(s), $p_C = 0.29,\ p_{E1} = 0.29,\ p_{E2} = 0.29,\ p_{E3} = 0.29,\ N = 417$.

					$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)									
Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.043	0.003	416.3	(8.8)	295.6	(6.2)	0.250	(0.020)	0.250	(0.020)	0.250	(0.021)	0.250	(0.021)
$DBCD_{D_A}$	0.049	0.003	415.7	(15.6)	295.1	(11.1)	0.250	(0.012)	0.250	(0.013)	0.250	(0.013)	0.250	(0.012)
$ERADE_{D_A}$	0.047	0.004	415.9	(12.9)	295.3	(9.2)	0.250	(0.010)	0.250	(0.010)	0.250	(0.010)	0.250	(0.010)
$DBCD_{NP}$	0.047	0.002	416.0	(13.1)	295.3	(9.3)	0.241	(0.108)	0.247	(0.117)	0.249	(0.118)	0.263	(0.104)
$ERADE_{NP}$	0.050	0.002	415.9	(13.0)	295.3	(9.3)	0.243	(0.109)	0.246	(0.115)	0.244	(0.116)	0.267	(0.103)
					Fixed-sample design									
Procedure	$\tilde{\alpha}^I$	$\tilde{\alpha}^{II}$	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	0.042	0.002	417	(0)	296.1	(0)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
$DBCD_{D_A}$	0.045	0.002	417	(0)	296.1	(0)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)	0.250	(0.011)
$ERADE_{D_A}$	0.047	0.004	417	(0)	296.1	(0)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)	0.250	(0.009)
$DBCD_{NP}$	0.044	0.002	417	(0)	296.1	(0)	0.242	(0.108)	0.245	(0.117)	0.246	(0.116)	0.266	(0.104)
$ERADE_{NP}$	0.044	0.003	417	(0)	296.1	(0)	0.241	(0.107)	0.246	(0.115)	0.245	(0.115)	0.268	(0.105)

Table 4.14.: Simulated powers for redesigning NeoSphere trial using complete randomisation and response-adaptive randomisation with dropping of inferior treatment(s), $p_C = 0.29,\ p_{E1} = 0.458,\ p_{E2} = 0.168,\ p_{E3} = 0.24,\ N = 417$.

					$(t_1,\ t_2,\ t_3)$=(0.33, 0.67, 1)									
Procedure	power^I	power^{II}	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_C$	(s.d.)	$\tilde{\rho}_{E1}$	(s.d.)	$\tilde{\rho}_{E2}$	(s.d.)	$\tilde{\rho}_{E3}$	(s.d.)
CR	0.944	0.007	364.7	(63.3)	258.5	(45.0)	0.255	(0.024)	0.255	(0.024)	0.238	(0.034)	0.252	(0.026)
$DBCD_{D_A}$	0.944	0.005	360.3	(66.6)	253.4	(47.0)	0.260	(0.020)	0.272	(0.017)	0.218	(0.036)	0.250	(0.022)
$ERADE_{D_A}$	0.948	0.009	361.7	(65.7)	254.5	(46.3)	0.260	(0.019)	0.272	(0.016)	0.219	(0.034)	0.249	(0.022)
$DBCD_{NP}$	0.942	0.005	356.6	(71.6)	234.4	(47.3)	0.198	(0.035)	0.472	(0.033)	0.143	(0.033)	0.187	(0.033)
$ERADE_{NP}$	0.939	0.005	356.8	(70.8)	234.9	(46.8)	0.198	(0.033)	0.470	(0.032)	0.145	(0.032)	0.187	(0.031)
					Fixed-sample design									
Procedure	power^I	power^{II}	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	0.946	0	417	(0)	296.5	(2.1)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)	0.250	(0.020)
$DBCD_{D_A}$	0.953	0	417	(0)	294.3	(1.3)	0.256	(0.012)	0.267	(0.011)	0.230	(0.015)	0.248	(0.012)
$ERADE_{D_A}$	0.948	0	417	(0)	294.3	(1.0)	0.256	(0.009)	0.266	(0.008)	0.230	(0.013)	0.248	(0.010)
$DBCD_{NP}$	0.947	0	417	(0)	274.4	(2.8)	0.198	(0.030)	0.470	(0.033)	0.145	(0.029)	0.187	(0.030)
$ERADE_{NP}$	0.947	0	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 allocation $(\rho_C, \rho_{E1}, \rho_{E2}, \rho_{E3})$ is (0.256, 0.266, 0.230, 0.248) and the NP optimal allocation is (0.2, 0.479, 0.121, 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.

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 attains 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.

4.5. Conclusions

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 inferior arms at interim looks are allowed.

As shown in Section 3.3.4 where the NeoSphere trial is redesigned using the combined approach without dropping treatments, the type I error rate is slightly inflated. For the tests of homogeneity considered in Chapter 3, the critical boundaries under the assumption of equal variances for all treatments are used as an approximation. However, heterogeneity increases when the treatments have unequal variances, which can result in a higher probability of rejecting the global null hypothesis.

Nevertheless, if the global null hypothesis is rejected, pairwise comparisons are conducted subsequently using Fisher's LSD method in Chapter 4. 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 (Jennison and Turnbull, 2000). Therefore, although a false rejection of the global null hypothesis could be made using the critical boundaries derived under the assumption of equal variances, 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.

5. Discussion

5.1. Conclusions

Group sequential monitoring has become a standard procedure in clinical trials. In addition, the use of response-adaptive randomisation can be more ethical, since the probability that a newly-arrived patient will receive the more promising treatment is increased. However, few researches have explored the properties of the combined approach of group sequential analysis with response-adaptive randomisation. The application of this approach to two-armed and multi-armed clinical trials with different types of responses is investigated in this thesis. Simulation results show that the approach can control the overall type I error rate and that the power is not adversely affected by the adaptive sampling rules. Furthermore, with the use of an optimal response-adaptive randomisation procedure, the approach can target any optimal allocation derived based on some optimality criterion. The ERADE in particular has the smallest variability in the allocation proportions. The combined approach has the advantage of being ethical in terms of reducing the expected number of patients and the expected number of failures for binary and censored survival responses. Moreover, the approach does not require numerical iterative methods to obtain the optimal allocation proportions. Existing critical boundaries can be used as an approximation to control the overall type I error rate, which facilitates the use of the combined approach.

The designs allow interim looks to be taken at any continuous information time.

For two-treatment comparisons, the error-spending approach can be used to obtain the critical boundaries for unequally spaced information times. These boundaries can be applied to response-adaptive designs, as long as the imbalance in the allocation is not too severe. For comparing multiple treatments, an analogue of Fisher's LSD method, which consists of a global test and pairwise tests, is considered. In a fixed-sample design, Fisher's LSD method has been shown to control the family-wise type I error rate, although not strongly. However, it is considered as one of the most powerful multiple comparison approaches (Christensen, 2002).

For the global tests, the critical boundaries derived under the assumptions of equal allocation and equally spaced information times are used as an approximation. This may result in a higher probability of rejecting the global null hypothesis when the assumptions are violated. Nevertheless, simulation results in Chapter 3 show that the inflation in the type I error rate seems to be modest, except for binary responses with small sample sizes and when the probabilities of success for all treatments are close to zero or one. When the global null hypothesis is rejected, pairwise tests are conducted subsequently. The above-mentioned boundaries for two-treatment comparisons can then be used to control the error probabilities. As can be seen from the simulation results in Chapter 4, the probability of rejecting one or more pairwise null hypotheses is generally less than or equal to the specified significance level.

For optimal allocations, the D_A-optimal design was derived based on the general linear model for multi-armed normal trials with unequal variances across treatments. The optimal rule was generalised to censored survival responses (Sverdlov et al., 2011). Here, the D_A-optimal allocation is also applied to binary responses. The allocation rule depends on the variances of the responses for each treatment arm. Computation of the target optimal allocation requires that the estimates for the variances are reliable. The D_A-optimal allocation is ethical for censored

survival responses. If the mean survival times are $\theta_1 > \theta_2 \geq ... \geq \theta_{J-1} > \theta_J$, then the D_A-optimal allocation proportions satisfy $\rho_1 > \rho_2 \geq ... \geq \rho_{J-1} > \rho_J$.

The optimal allocation based on nonlinear programming (NP) depends on the order of the unknown parameters and the user-specified lower bound for the allocation proportions B. The NP allocation assigns more patients to the best and the worst treatments and fixes the allocation proportions for the other treatments to be B. The simulation results show that the NP allocation can be more ethical in terms of reducing the total number of failures for binary and survival responses. However, the least promising treatment arm usually receives more patients than the arms with medium efficacy. As discussed in Section 3.2.1, for normal responses with a nuisance parameter, a closed form solution has not yet been derived and further research is needed. The closed form solution for binary and censored survival responses can be used. With a closed form, the simulation results for the NP allocation usually take less time than for the D_A-optimal allocation.

Simulation results on comparing two, three and four treatments are shown. These are often the cases for phase III trials. Nevertheless, the combined approach is also applicable for $J > 4$ arms. Both optimal response-adaptive randomisation procedures, the DBCD and the ERADE, have been generalised to multi-armed clinical trials. In addition, the D_A-optimal allocation and the NP optimal allocation can be applied. However, when the number of treatments increases, the computation time for the D_A-optimal allocation, which solves a system of equations, will increase. For NP optimal allocation, the variability in the order of the parameter estimates may be higher when the number of treatments increases. Consequently, a greater variance for the treatment allocation proportions may be obtained. Also, the number of possibilities after the pairwise comparisons increases significantly. For comparisons with a control, there are nine possibilities for three-armed trials and 27 possibilities for four-armed trials. An increase in the number of treatments

would make the program for the design more complicated.

5.2. Future work

Throughout the thesis, the parameter of interest considered is the simple difference, which determines the optimal allocation proportions and the test statistics used. Application of the combined approach to other parameters of interest, such as the log odds ratio for binary responses and the log hazard ratio for survival responses, is also worth investigating. Some optimal allocations based on these parameters for two-armed trials have been proposed. For two-armed binary trials, Morgan and Coad (2007) studied group-sequential response-adaptive designs using the log odds ratio. It is shown that, for the test statistic in this case, the normal approximation is more accurate. For multi-armed trials, the optimal allocations were derived based on the vector of simple difference parameters, including the D_A-optimal allocation and the NP allocation. A discussion of multi-treatment optimal allocations can be found in Biswas et al. (2011). There appear to be no optimal allocations for multi-armed trials derived based on parameters other than the simple difference.

This thesis focuses on maximum duration trials rather than maximum information ones. For a maximum duration trial with censored survival responses, the information level at the final look is usually unpredictable. An approximate information time is defined in Chapter 2. For a maximum information trial, the trial stops when the specified information is reached. This ensures attainment of the specified power. However, trials may reach the specified information well before the planned end of the study, or, at the end of the trial, the target information level may not be achieved. This increases the difficulty in utilising the error-spending approach to control the overall type I error rate. Further research is required to apply the combined approach to maximum information trials.

For censored survival responses, the optimal allocation usually depends on the probability of an event, $\epsilon_{j,k}$. For the combined approach, $\epsilon_{j,k}$ derived in Appendix A is based on the model assumptions of uniform arrival and censoring times, and exponential survival times. Although the assumptions are strong, it is a natural starting point to begin generalising the design to a group sequential setting. However, exponential survival times have a constant hazard, which is unrealistic in practice. Weibull survival times are considered by Zhang and Rosenberger (2007) for a fixed-sample response-adaptive design comparing two treatments. For multi-armed trials, Sverdlov et al. (2011) compared fixed-sample response-adaptive designs under the assumption of censored exponential survival times. An interesting open problem is to generalise the Weibull case to the group sequential setting. For example, the calculation of $\epsilon_{j,k}$ will be more difficult in the Weibull case, as it requires integration of the joint distribution of the arrival, survival and censoring times.

This thesis also focuses on testing. Development of inferential methods, such as a bias-adjusted maximum likelihood estimate and an approximate confidence interval following group sequential tests, has been studied (Whitehead, 1986; Morgan, 2003a). Research on estimation following the adaptive designs presented here may also be of interest.

In this thesis, the randomisation procedures depend on the previous treatment allocations and responses, but they do not take into account the covariates of the patients. Covariate-adjusted response-adaptive (CARA) designs have been proposed for fixed-sample designs. Designs that combine group sequential monitoring with CARA designs would be more complicated, yet may be of interest as well.

A. Calculation of the probability of an event

Consider the model assumptions described in Section 2.3.2. The observed survival time for patient i on treatment j at interim analysis k can be expressed as $Y_{i,j,k} = \min(S_{i,j},\ C_i,\ Dt_k - A_i)$, where the survival time $S_{i,j} \sim \text{Exp}(\theta_j^{-1})$, the censoring time $C_i \sim U(0, D)$ and the arrival time $A_i \sim U(0, Dt_k)$. Here, θ_j is the mean survival time for treatment j, D is the maximum duration of the trial and $t_k \in (0, 1]$ is the information time at look k.

The probability of having an event on treatment j at look k is

$$\begin{aligned}\epsilon_{j,k} &= P(Y_{i,j,k} = S_{i,j}) = P\{S_{i,j} \leq \min(C_i,\ Dt_k - A_i)\} \\ &= \int_0^{Dt_k} \int_0^{D} \int_0^{\min(c,\ Dt_k - a)} \frac{1}{Dt_k} \frac{1}{D} \frac{1}{\theta_j} \exp\left(-\frac{s}{\theta_j}\right) ds\ dc\ da.\end{aligned}$$

By integrating out s, we obtain

$$\begin{aligned}\epsilon_{j,k} &= \int_0^{Dt_k} \int_0^{D} \frac{1}{D^2 t_k} \left\{1 - \exp\left(-\frac{\min(c,\ Dt_k - a)}{\theta_j}\right)\right\} dc\ da \\ &= 1 - \frac{1}{D^2 t_k} \int_0^{Dt_k} \left\{\int_0^{Dt_k - a} \exp\left(-\frac{c}{\theta_j}\right) dc + \int_{Dt_k - a}^{D} \exp\left(-\frac{Dt_k - a}{\theta_j}\right) dc\right\} da,\end{aligned}$$

where the first term inside the parentheses refers to the case $c < Dt_k - a$ and the second term represents the case $c > Dt_k - a$. Then, by integrating out c, we have

$$\begin{aligned}\epsilon_{j,k} &= 1 - \frac{1}{D^2 t_k}\int_0^{Dt_k}\left[-\theta_j \exp\left(-\frac{c}{\theta_j}\right)\right]_0^{Dt_k-a} da \\ &\quad - \frac{1}{D^2 t_k}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right)(D - Dt_k + a)da \\ &= 1 - \frac{1}{D^2 t_k}\int_0^{Dt_k}\left\{\theta_j - \theta_j \exp\left(-\frac{Dt_k-a}{\theta_j}\right)\right\} da \\ &\quad - \frac{1}{D^2 t_k}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right)(D - Dt_k + a)da.\end{aligned}$$

After rearrangement, we find that

$$\begin{aligned}\epsilon_{j,k} &= 1 - \frac{\theta_j}{D} + \frac{\theta_j}{D^2 t_k}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right) da \\ &\quad - \frac{1}{Dt_k}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right) da \\ &\quad + \frac{1}{D}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right) da \\ &\quad - \frac{1}{D^2 t_k}\int_0^{Dt_k}\exp\left(-\frac{Dt_k-a}{\theta_j}\right) a\ da\end{aligned}$$

Then we integrate out a to obtain

$$\begin{aligned}\epsilon_{j,k} &= 1 - \frac{\theta_j}{D} + \frac{\theta_j}{D^2 t_k}\left\{\theta_j - \theta_j \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\} \\ &\quad - \frac{1}{Dt_k}\left\{\theta_j - \theta_j \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\} \\ &\quad + \frac{1}{D}\left\{\theta_j - \theta_j \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\} \\ &\quad - \frac{1}{D^2 t_k}\left\{\left[a\theta_j \exp\left(-\frac{Dt_k-a}{\theta_j}\right)\right]_0^{Dt_k} - \int_0^{Dt_k}\theta_j \exp\left(-\frac{Dt_k-a}{\theta_j}\right)\ da\right\}.\end{aligned}$$

Here, integration by parts is used. Finally, after rearrangement, we have

$$\epsilon_{j,k} = 1 - \frac{\theta_j}{D}\left\{1 + \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\} - \frac{\theta_j}{Dt_k}\left(1 - \frac{2\theta_j}{D}\right)\left\{1 - \exp\left(-\frac{Dt_k}{\theta_j}\right)\right\}.$$

B. Derivation of the noncentrality parameter

First, we utilise the following formula in Atkinson (1982) to obtain the inverse of Σ_k in (3.3):

$$\left(C_k + U_k V_k^T\right)^{-1} = C_k^{-1} - \frac{C_k^{-1} U_k V_k^T C_k^{-1}}{1 + V_k^T C_k^{-1} U_k},$$

where C_k is a $(J-1) \times (J-1)$ matrix and U_k and V_k are $(J-1) \times 1$ vectors. Let $\Sigma_k^{-1} = \left(C_k + U_k V_k^T\right)^{-1}$. Then we have

$$C_k = \begin{pmatrix} \frac{\sigma_1^2}{m_{1,k}} & 0 & \dots & 0 \\ 0 & \frac{\sigma_2^2}{m_{2,k}} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \frac{\sigma_{J-1}^2}{m_{J-1,k}} \end{pmatrix},$$

$$U_k = \sqrt{\frac{\sigma_J^2}{m_{J,k}}}\ \mathbf{1} \quad \text{and} \quad V_k^T = \sqrt{\frac{\sigma_J^2}{m_{J,k}}}\ \mathbf{1}^T,$$

where $\mathbf{1}$ is the $(J-1) \times 1$ vector of ones. Thus, we obtain

$$C_k^{-1} U_k = \sqrt{\frac{\sigma_J^2}{m_{J,k}}} \begin{pmatrix} \frac{m_{1,k}}{\sigma_1^2} \\ \vdots \\ \frac{m_{J-1,k}}{\sigma_{J-1}^2} \end{pmatrix}$$

and

$$V_k^T C_k^{-1} = \sqrt{\frac{\sigma_J^2}{m_{J,k}}} \left(\frac{m_{1,k}}{\sigma_1^2}, ..., \frac{m_{J-1,k}}{\sigma_{J-1}^2} \right).$$

By multiplying $C_k^{-1} U_k$ and $V_k^T C_k^{-1}$, we have

$$C_k^{-1} U_k V_k^T C_k^{-1} = \frac{\sigma_J^2}{m_{J,k}} \begin{pmatrix} \left(\frac{m_{1,k}}{\sigma_1^2}\right)^2 & \frac{m_{1,k}m_{2,k}}{\sigma_1^2\sigma_2^2} & \cdots & \frac{m_{1,k}m_{J-1,k}}{\sigma_1^2\sigma_{J-1}^2} \\ \frac{m_{2,k}m_{1,k}}{\sigma_2^2\sigma_1^2} & \left(\frac{m_{2,k}}{\sigma_2^2}\right)^2 & & \\ \vdots & & \ddots & \\ \frac{m_{J-1,k}m_{1,k}}{\sigma_{J-1}^2\sigma_1^2} & & & \left(\frac{m_{J-1,k}}{\sigma_{J-1}^2}\right)^2 \end{pmatrix}.$$

In addition, after multiplying V_k^T and $C_k^{-1} U_k$, we find that

$$V_k^T C_k^{-1} U_k = \frac{\sigma_J^2}{m_{J,k}} \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2}.$$

It follows that

$$\Sigma_k^{-1} = \begin{pmatrix} \frac{m_{1,k}}{\sigma_1^2} & 0 & ... & 0 \\ 0 & \frac{m_{2,k}}{\sigma_2^2} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \frac{m_{J-1,k}}{\sigma_{J-1}^2} \end{pmatrix} - \frac{\frac{\sigma_J^2}{m_{J,k}}}{1 + \frac{\sigma_J^2}{m_{J,k}} \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2}} \begin{pmatrix} \left(\frac{m_{1,k}}{\sigma_1^2}\right)^2 & \frac{m_{1,k}m_{2,k}}{\sigma_1^2\sigma_2^2} & \cdots & \frac{m_{1,k}m_{J-1,k}}{\sigma_1^2\sigma_{J-1}^2} \\ \frac{m_{2,k}m_{1,k}}{\sigma_2^2\sigma_1^2} & \left(\frac{m_{2,k}}{\sigma_2^2}\right)^2 & & \\ \vdots & & \ddots & \\ \frac{m_{J-1,k}m_{1,k}}{\sigma_{J-1}^2\sigma_1^2} & & & \left(\frac{m_{J-1,k}}{\sigma_{J-1}^2}\right)^2 \end{pmatrix}$$

$$
\begin{aligned}
&= \begin{pmatrix} \frac{m_{1,k}}{\sigma_1^2} & 0 & ... & 0 \\ 0 & \frac{m_{2,k}}{\sigma_2^2} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \frac{m_{J-1,k}}{\sigma_{J-1}^2} \end{pmatrix} \\
&\quad - \frac{1}{\sum_{j=1}^{J} \frac{m_{j,k}}{\sigma_j^2}} \begin{pmatrix} \left(\frac{m_{1,k}}{\sigma_1^2}\right)^2 & \frac{m_{1,k}m_{2,k}}{\sigma_1^2\sigma_2^2} & ... & \frac{m_{1,k}m_{J-1,k}}{\sigma_1^2\sigma_{J-1}^2} \\ \frac{m_{2,k}m_{1,k}}{\sigma_2^2\sigma_1^2} & \left(\frac{m_{2,k}}{\sigma_2^2}\right)^2 & & \\ \vdots & & \ddots & \\ \frac{m_{J-1,k}m_{1,k}}{\sigma_{J-1}^2\sigma_1^2} & & & \left(\frac{m_{J-1,k}}{\sigma_{J-1}^2}\right)^2 \end{pmatrix}.
\end{aligned}
$$

Finally, we have the noncentrality parameter

$$
\begin{aligned}
\eta_k &= \boldsymbol{\mu}_G^T \, \Sigma_k^{-1} \, \boldsymbol{\mu}_G \\
&= (\mu_1 - \mu_J, \ \mu_2 - \mu_J, ..., \ \mu_{J-1} - \mu_J) \, \Sigma_k^{-1} \, (\mu_1 - \mu_J, \ \mu_2 - \mu_J, ..., \ \mu_{J-1} - \mu_J)^T \\
&= \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2} (\mu_j - \mu_J)^2 - \frac{1}{\sum_{j=1}^{J} \frac{m_{j,k}}{\sigma_j^2}} \left\{ \sum_{j=1}^{J-1} \frac{m_{j,k}}{\sigma_j^2} (\mu_j - \mu_J) \right\}^2 .
\end{aligned}
$$

Bibliography

Antognini, A. B. and Giovagnoli, A. (2015). *Adaptive Designs for Sequential Treatment Allocation*. Chapman and Hall/CRC: London.

Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69:61–67.

Atkinson, A. C. and Biswas, A. (2014). *Randomised Response-Adaptive Designs in Clinical Trials*. Chapman and Hall/CRC: London.

Biswas, A., Mandalb, S., and Bhattacharya, R. (2011). Multi-treatment optimal response-adaptive designs for phase III clinical trials. *J. Korean Statist. Soc.*, 40:33–44.

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*, 17:309.

Christensen, R. (2002). *Plain Answers to Complex Questions: The Theory of Linear Models*. Springer-Verlag: New York.

Connor, E., Sperling, R. S., Gelber, R., and Kiselev, P. (1994). Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *N. Engl. J. Med.*, 331:1173–1180.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall: London.

Dimairo, M., Boote, J., Julious, S. A., Nicholl, J. P., and Todd, S. (2015). Missing steps in a staircase: A qualitative study of the perspectives of key stakeholders on the use of adaptive designs in confirmatory trials. *Trials*, 16:430.

Eisele, J. R. and Woodroofe, M. B. (1995). Central limit theorems for doubly adaptive biased coin designs. *Ann. Statist.*, 23:234–254.

Follmann, D. A., Proschan, M. A., and Geller, N. L. (1994). Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics*, 50:325–336.

Fountzilas, G., Ciuleanu, E., Dafni, U., and Plataniotis, G. (2004). Concomitant radiochemotherapy vs radiotherapy alone in patients with head and neck cancer. *Med. Oncol.*, 21:95–107.

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 Oncol.*, 13:25–32.

Hatfield, I., Allison, A., Flight, L., Julious, S. A., and Dimairo, M. (2016). Adaptive designs undertaken in clinical research: A review of registered clinical trials. *Trials*, 17:150.

Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *J. Amer. Statist. Assoc.*, 98:671–678.

Hu, F. and Rosenberger, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. Wiley: New York.

Hu, F., Rosenberger, W. F., and Zhang, L.-X. (2006). Asymptotically best response-adaptive randomization procedure. *J. Statist. Plann. Inf.*, 136:1911–1922.

Hu, F. and Zhang, L.-X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Ann. Statist.*, 32:268–301.

Hu, F., Zhang, L.-X., and He, X. (2009). Efficient randomized-adaptive designs. *Ann. Statist.*, 37:2543–2560.

Ivanova, A. V. (2003). A play-the-winner type urn model with reduced variability. *Metrika*, 58:1–13.

Jennison, C. and Turnbull, B. W. (1991). Exact calculation for sequential t, χ^2 and F tests. *Biometrika*, 78:133–141.

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

Jennison, C. and Turnbull, B. W. (2001). Group sequential tests with outcome-dependent treatment assignment. *Sequential Anal.*, 20:209–234.

Kiefer, J. and Wolfowitz, J. (1960). The equivalence of two extremum problems. *Canad. J. Math.*, 12:363–366.

Kim, K., Boucher, H., and Tsiatis, A. A. (1995). Design and analysis of group sequential logrank tests in maximum duration versus information trials. *Biometrics*, 51:988–1000.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663.

Magirr, D., Jaki, T., and Whitehead, J. (2012). A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika*, 99:494–501.

Melfi, V. and Page, C. (1998). Variability in adaptive designs for estimation of success probabilities. In Flournoy, N., Rosenberger, W. F., and Wong, W. K., editors, *New Developments and Applications in Experimental Design*. pp. 106-114. Institute of Mathematical Statistics: Hayward, California.

Morgan, C. C. (2003a). Estimation following group-sequential response-adaptive clinical trials. *Control. Clin. Trials*, 24:523–543.

Morgan, C. C. (2003b). Sample size re-estimation in group-sequential response-adaptive clinical trials. *Statist. Med.*, 22:3843–3857.

Morgan, C. C. and Coad, D. S. (2007). A comparison of adaptive allocation rules for group-sequential binary response clinical trials. *Statist. Med.*, 26:1937–1954.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35:549–556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64:191–199.

Proschan, M. A., Follmann, D. A., and Geller, N. L. (1994). Monitoring multi-armed trials. *Statist. Med.*, 13:1441–1452.

Proschan, M. A., Follmann, D. A., and Waclawiw, M. A. (1992). Effects of assumption violations on type I error rate in group sequential monitoring. *Biometrics*, 48:1131–1143.

Proschan, M. A., Lan, K. K. G., and Wittes, J. T. (2006). *Statistical Monitoring of Clinical Trials: A Unified Approach.* Springer: New York.

Rosenberger, W. F. and Lachin, J. L. (2002). *Randomization in Clinical Trials: Theory and Practice.* Wiley: New York.

Rosenberger, W. F. and Seshaiyer, P. (1997). Adaptive survival trials. *J. Biopharm. Statist.*, 7:617–624.

Rosenberger, W. F., Stallard, N., Ivanova, A., Harper, C. N., and Ricks, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics*, 57:909–913.

Sverdlov, O., Tymofyeyev, Y., and Wong, W. K. (2011). Optimal response-adaptive randomised designs for multi-armed survival trials. *Statist. Med.*, 30:2890–2910.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:275–294.

Tymofyeyev, Y., Rosenberger, W. F., and Hu, F. (2007). Implementing optimal allocation in sequential binary response experiments. *J. Amer. Statist. Assoc.*, 102:224–234.

Wason, J., Stallard, N., Bowden, J., and Jennison, C. (2016). A multi-stage drop-the-losers design for multi-arm clinical trials. *Statist. Meth. Med. Res.*, 25 in press.

Wei, L. J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.*, 73:840–843.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 79:347–353.

Wong, W. K. and Zhu, W. (2008). Optimum treatment allocation rules under a variance heterogeneity model. *Statist. Med.*, 27:4581–4595.

Zhang, L. and Rosenberger, W. F. (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics*, 62:562–569.

Zhang, L. and Rosenberger, W. F. (2007). Response-adaptive randomization for survival trials: The parametric approach. *Appl. Statist.*, 56:153–165.

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. *Ann. Statist.*, 38:2218–2241.