Group-sequential response-adaptive designs for censored survival outcomes

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

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

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

Abstract

Previous work on two-treatment comparisons for immediate responses 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 error rates. In this paper, the application of the combined approach to censored survival responses is investigated and different optimal response-adaptive randomised procedures are compared. For a maximum duration trial, the information level at the final look is usually unpredictable. An approximate information time is defined. Group sequential tests and optimal allocations for two measures of treatment difference are given. Operating characteristics of the combined approach are investigated by simulation, including cases of exponential and Weibull survival responses and redesign of a clinical trial. The results reveal that the existing boundaries for standard group sequential designs derived based on the error-spending approach can be applied as approximate tests to control the overall type I error rate. Compared to the group sequential complete randomisation design, the combined approach is found to retain ethical advantages as in previous work on immediate responses while the power is not adversely affected.

Keywords: Error-spending function; Information time; Log hazard ratio; Optimal allocation; Power; Weibull response

1 Introduction

Periodic group sequential designs, in which a number of interim analyses are conducted after groups of observations, can require fewer patients than a fixed-sample design to achieve the same error probabilities (Jennison and Turnbull, 2000). Since early termination of trials is allowed, 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 cumulative responses, can further reduce the numbers of participants allocated to the inferior treatments compared to complete randomisation (Atkinson and Biswas, 2014). The use of the combined approach of group sequential analysis and response-adaptive randomisation can achieve both individual and collective ethics.

Few studies of the application of the combined approach to two-armed trials with immediate responses have been investigated. Jennison and Turnbull (2001) derived theory to support that the combined approach still maintains the overall error rates for two-armed normal trials with known variances. The authors proved that the joint distribution of the test statistics has a standard form similar to that for a group-sequential non-adaptive design, but with the additional feature that the information level can depend on previous test statistics. 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: an approximate confidence interval using a pivotal method and 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 urn-model type designs and the doubly-adaptive biased coin design (DBCD) (Eisele and Woodroofe, 1995). 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.

For normal and binary responses, Zhu and Hu (2010) considered monitoring the DBCD at continuous information time utilising critical boundaries derived by the error-spending approach (Lan and DeMets, 1983). They considered the α-spending function, which spends the type I error rate as a function of the information time. Their simulation results revealed that the use of the combined approach can preserve the advantages of both group sequential analysis and optimal response-adaptive randomisation.

In this paper, the combined approach generalised to censored survival responses is explored, which allows staggered entry and right-censoring. We initially assume an exponential survival model in which the arrival and censoring times are both uniformly distributed. Since such a model may be misclassified in practice, some of these assumptions are later relaxed and a more robust approach considered.

For survival responses, the information levels usually cannot be attained accurately, since they depend on the realised pattern of events and censoring (Jennison and Turnbull, 2000). There are maximum duration trials and maximum information trials. The former are more feasible in practice, since the maximum length of the trials is fixed, whereas the latter consider a pre-determined maximum information level. In practice, the 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. We consider maximum duration trials that use an approximate information time. The optimal response-adaptive randomisation procedures are used to target different optimal allocations derived based on some optimality criteria, for instance, minimising the total sample size or the expected number of failure events. In addition to the DBCD, we also consider the efficient randomised-adaptive design (ERADE) (Hu, Zhang and He, 2009).

The structure of the remaining sections is as follows. In Section 2, the parametric model for the responses, which characterises the staggered entry, the right-censoring and the survival time, is introduced. An approximate information time based on the model assumptions is defined. Group sequential tests for two measures of treatment difference, the simple difference and the log hazard ratio, are given in Section 3. In Section 4, the optimal allocations derived based on different optimality criteria for the two measures of treatment difference are shown. Then optimal response-adaptive randomisation procedures, which aim to target the pre-specified optimal allocations, are described. Simulation results comparing the designs are presented in Section 5, including the error probabilities, the expected number of patients, the expected number of failures and the average allocation proportion with its variability. In addition, the redesign of a clinical trial is investigated. Conclusions and further work are in Section 6. Supplementary material provides the derivation of the probability of an event for the model-based approach, comparison with a nonparametric approach based on the logrank test and a simulation study of model misspecification.

2 The model

2.1 Information time

Suppose that N is the planned number of patients for a trial with K group sequential analyses. The information level for survival responses at look k, $\mathcal{I}_k$, is proportional to the number of events. 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, a predicted value for the final information level evaluated at interim analysis k, $\hat{\mathcal{I}}_K^{(k)}$, is needed. Then the information time at group sequential test k can be expressed as

$$t_k = \frac{\mathcal{I}_k}{\hat{\mathcal{I}}_K^{(k)}} = \frac{e_k}{\hat{e}_K^{(k)}}, \quad k = 1, ..., K, \tag{1}$$

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, Boucher and Tsiatis (1995) considered

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

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 estimate of e_K. One is under the null hypothesis of no treatment difference and the other is based on the specified alternative hypothesis, which result in two information time scales. Kim, Boucher and Tsiatis (1995) showed that the overall type I error rate can be preserved by using either information time scale for a logrank test. The power depends on the actual information level obtained.

For parametric tests, (1) can be approximated by

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

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 model. For uniformly distributed arrival and censoring times, and exponentially distributed survival time with mean θ_j, the probability of

an event is

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

where D is the maximum duration of the trial. Details of the derivation are in Supplementary Material 1. 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. However, $\epsilon_{j,k}$ decreases when the mean survival time for treatment j is increased. Since $\epsilon_{j,k}$ is a function of an unknown parameter, initial estimates from a previous study or obtained in the learning phase of the trial can be used. The parameter estimates are then updated based on the cumulative responses.

The type I error rate can be guaranteed using either information time scale (Kim, Boucher and Tsiatis, 1995). For simplicity, we consider the information time scale under the null hypothesis where $\theta_1 = \theta_2$. Then the subscript j for $\epsilon_{j,k}$ denoting treatment can be suppressed. The approximate information time at look k becomes

$$t_k = \frac{\sum_{j=1}^{2} m_{j,k}\hat{\epsilon}_k}{\sum_{j=1}^{2} m_{j,K}\hat{\epsilon}_K} = \frac{n_k\hat{\epsilon}_k}{N\hat{\epsilon}_K} \in (0, 1], \quad k = 1, ..., K, \tag{3}$$

where $n_k = \sum_{j=1}^{2} m_{j,k}$ is the cumulative sample size at look k and $n_K = N$. 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 it is planned at $t_1 = 1/3$, and, from (3), $n_1 = \lceil t_1 N\hat{\epsilon}_K/\hat{\epsilon}_1 \rceil$ is the approximate number of patients needed at the first look, where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x.

2.2 Model assumptions

Let D be the length of the 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 $A_i \sim U(0, Dt_k)$. Also, assume that the survival time for patient i on treatment j, $S_{i,j}$, follows an exponential distribution with mean $\theta_j > 0$. Then the density function of $S_{i,j}$ is

$$f(s_{i,j}; \theta_j) = \frac{1}{\theta_j} \exp\left(-\frac{s_{i,j}}{\theta_j}\right)$$

for $s_{i,j} > 0$. The survival function is $P(S_{i,j} > s_{i,j}) = \exp(-s_{i,j}/\theta_j)$ 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 outcome 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 $Y_{i,j,k} = \min(S_{i,j}, C_i, Dt_k - A_i)$. Note that 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 $A_i \sim U(0, Dt_1)$. If $Y_{i,j,1} = S_{i,j}$, then the patient's outcome is an event. If $Y_{i,j,1} = C_i$, then the patient's outcome 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. The patient's outcome will then be followed up at later looks. Let $E_{i,j} = \min(S_{i,j}, C_i)$. Then, if the outcome 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 1, where $E_{i,j} = S_{i,j}$ and $Y_{i,j,2} = S_{i,j}$.

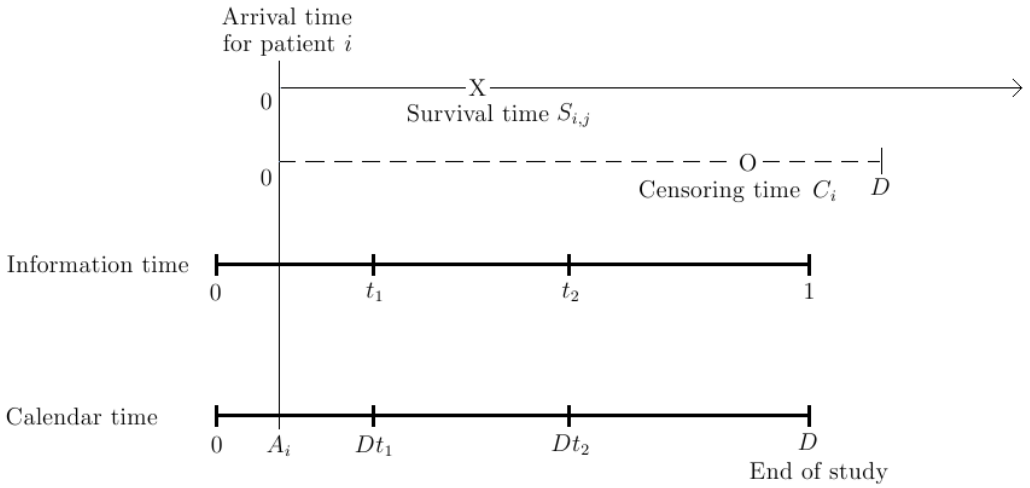

Figure 1: An example of a patient's arrival, survival and censoring times.

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 outcome 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 outcome will be followed up.

2.3 Sequential maximum likelihood estimation

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 outcome of patient i on arm j at look k is an event and $\delta_{i,j,k} = 0$ if the outcome is censored. Under the above model, the likelihood function for treatment j based on the outcomes obtained so far can be expressed as

$$L_k(\theta_j) = \prod_{i=1}^{m_{j,k}} \left\{ \frac{1}{\theta_j} \exp\left(-\frac{y_{i,j,k}}{\theta_j} \right) \right\}^{\delta_{i,j,k}} \left\{ \exp\left(-\frac{y_{i,j,k}}{\theta_j} \right) \right\}^{1-\delta_{i,j,k}}, \qquad (4)$$

where the first part of the product in (4) refers to the survival times and the second part represents the censoring times. The log-likelihood function is

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

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

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

for a maximum. Hence, we obtain the maximum likelihood estimate of the mean survival time θ_j to be $\hat{\theta}_{j,k} = \sum_{i=1}^{m_{j,k}} y_{i,j,k}/r_{j,k}$, which is the sum of the observed survival times divided by the number of events obtained so far.

The Fisher information for θ_j is

$$I_k(\theta_j) = -E\left\{ \frac{d^2 l_k(\theta_j)}{d\theta_j^2} \right\} = -E(r_{j,k})\theta_j^{-2} + 2E\left(\sum_{i=1}^{m_{j,k}} Y_{i,j,k} \right) \theta_j^{-3}.$$

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 for θ_j is approximately

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

Consequently, we have $\text{var}(\hat{\theta}_{j,k}) = I_k(\theta_j)^{-1} = \theta_j^2/E(r_{j,k})$ as the approximate variance of $\hat{\theta}_{j,k}$.

3 Group sequential analysis

3.1 Form of tests

Suppose that the parameter of interest is ϕ. The null hypothesis is $H_0 : \phi = 0$ versus the alternative hypothesis $H_a : \phi \neq 0$. Let $\hat{\phi}_k$ be the parameter

estimate at look k and let $\widehat{\text{var}}(\hat{\phi}_k)$ be its estimated variance. Then the test statistic is

$$Z_k = \frac{\hat{\phi}_k}{\sqrt{\widehat{\text{var}}(\hat{\phi}_k)}}, \qquad k = 1, ..., K, \tag{5}$$

which is approximately normal for large sample sizes.

3.1.1 Simple difference

If the parameter of interest is the difference in the two mean survival times, we have $\phi = \theta_1 - \theta_2$. Then $\hat{\phi}_k = \hat{\theta}_{1,k} - \hat{\theta}_{2,k}$ is the maximum likelihood estimate of ϕ at look k and

$$\text{var}(\hat{\phi}_k) = \text{var}(\hat{\theta}_{1,k} - \hat{\theta}_{2,k}) = I_k(\theta_1)^{-1} + I_k(\theta_2)^{-1} = \frac{\theta_1^2}{E(r_{1,k})} + \frac{\theta_2^2}{E(r_{2,k})}. \tag{6}$$

Based on the assumed model, $E(r_{j,k})$ can be approximated by $m_{j,k}\epsilon_{j,k}$.

3.1.2 Log hazard ratio

If the parameter of interest is the log hazard ratio, we have $\phi = \log(\theta_1/\theta_2)$. Then $\hat{\phi}_k = \log(\hat{\theta}_{1,k}/\hat{\theta}_{2,k})$ is the maximum likelihood estimate of ϕ at look k, and, by the δ-method,

$$\text{var}(\hat{\phi}_k) = \text{var}\{\log(\hat{\theta}_{1,k}/\hat{\theta}_{2,k})\} \approx \frac{1}{E(r_{1,k})} + \frac{1}{E(r_{2,k})}. \tag{7}$$

In practice, the observed number of events on arm j at look k, $r_{j,k}$, is used in (6) and (7). In addition, it is required that there is at least one event on both treatment arms, so that the denominator $r_{j,k} > 0$ for $j = 1, 2$.

3.2 Joint distribution of test statistics

As the number of interim analyses increases, the probability of falsely rejecting H_0 is increased. Critical boundaries that control the overall type I error rate are required. Derivation of the critical boundaries relies on the joint distribution of the sequential test statistics.

For a group-sequential non-adaptive randomised design, 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. This form applies exactly for normal responses with known variances and approximately for other types of endpoints.

Jennison and Turnbull (2001) showed that the combined approach of group sequential analysis with response-adaptive randomisation still maintained the overall error rates by proving that the joint distribution of the test statistics has a standard form similar to that for a group-sequential non-adaptive design, but with the additional feature that the information level can depend on previous test statistics. In the study of Morgan and Coad (2007), where several response-adaptive designs for two-armed binary trials were compared, it was shown that the error rates were preserved for all of the designs. By considering monitoring a response-adaptive design at a continuous information time, Zhu and Hu (2010) also proved that the sequence of test statistics converges in distribution to a Brownian motion and asymptotically satisfies the canonical joint distribution. Hence, the required error probabilities for the combined approach can be achieved approximately using the same critical boundaries for standard group sequential designs.

To see why the joint distribution of $\{Z_1, \ldots, Z_K\}$ in (5) is approximately the canonical form, let $I_k(\phi)$ be the Fisher information for ϕ at look k. Then, again by the δ-method,

$$E(Z_k) \approx \frac{E(\hat{\phi}_k)}{\sqrt{\text{var}(\hat{\phi}_k)}} \approx \sqrt{I_k(\phi)}\phi,$$

and, for $k_1 \leq k_2$,

$$\text{cov}(Z_{k_1}, Z_{k_2}) \approx \frac{\text{cov}(\hat{\phi}_{k_1}, \hat{\phi}_{k_2})}{\sqrt{\text{var}(\hat{\phi}_{k_1})\text{var}(\hat{\phi}_{k_2})}} \approx \sqrt{\frac{I_{k_1}(\phi)}{I_{k_2}(\phi)}}.$$

The section below describes the derivation of the critical boundaries based on the asymptotic joint distribution. The use of the critical boundaries as an approximate test for the combined approach with censored survival responses will be investigated by simulation.

3.3 Critical boundaries

The α-spending approach controls the overall type I error rate and allows interim analyses to be taken at any continuous information time. 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. It is a continuous and monotonically non-decreasing function with $\alpha(0) = 0$ and $\alpha(1) = \alpha$. For the O'Brien and Fleming boundaries (O'Brien and Fleming, 1979), we have

$$\alpha_{O-F}(t_k) = 2\{1 - \Phi(z_{\alpha/2}/\sqrt{t_k})\},$$

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.

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. For computing $c_k,\ k \geq 2$, integration of a multivariate normal distribution is required. The error-spending approach does not require the number of group sequential analyses, K, to be pre-specified.

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

we reject H_0 if $|Z_k| \geq c_k$, and accept H_0 otherwise.

4 Optimal response-adaptive randomisation

4.1 Optimal allocations

Optimal response-adaptive randomised designs aim to target the pre-specified optimal allocation derived based on some optimality criterion. These optimal allocations were derived using a fixed-sample design with one analysis conducted at the end of the trial, which is the case $K = 1$. However, they are updated sequentially. Here, some optimal allocations for censored survival responses are introduced.

Minimising the total sample size (Neyman allocation)
Let M_j be the sample size for treatment j, $j = 1, 2$, at the end of the trial and $N = M_1 + M_2$ be the total sample size. The allocation proportion for treatment $j, M_j/N$, is random for response-adaptive randomised designs, as M_j is not pre-determined. Neyman allocation is found by minimising the total sample size $M_1 + (N - M_1)$ with respect to M_1 under a variance constraint. For the simple difference test statistic, the constraint is

$$V = \frac{\theta_1^2}{E(r_1)} + \frac{\theta_2^2}{E(r_2)} = \frac{\theta_1^2}{M_1\epsilon_1} + \frac{\theta_2^2}{(N - M_1)\epsilon_2} = C,$$

where $C > 0$ is a constant. The solution is

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

For the log hazard ratio test statistic, the constraint is

$$V = \frac{1}{E(r_1)} + \frac{1}{E(r_2)} = \frac{1}{M_1\epsilon_1} + \frac{1}{(N - M_1)\epsilon_2} = C.$$

The solution is

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

Minimising the total expected hazard

Optimal allocation for survival responses 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 constraints. The solution for the simple difference test statistic is now

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

and that for the log hazard ratio becomes

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

Here, $\rho_j,\ j = 1, 2$, is a function of the unknown parameters. In addition, the probability of an event on treatment j, ϵ_j, which depends on the assumed model, is also a function of the unknown parameters, as shown in (2). In practice, the current parameter estimates based on the responses available are used.

4.2 Randomisation procedures

Response-adaptive randomisation assigns patients according to previous treatment allocations and outcomes. Permuted-block randomisation can be used early on to obtain initial parameter estimates. For a two-treatment comparison, varying block sizes of $\{2, 4, 6, ...\}$ can be chosen. This method balances the sample sizes across the treatment groups. Then 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 arm $j,\ j = 1, 2$, evaluated based on the outcomes available.

The probability that the $(i+1)^{th}$ patient will be assigned to arm 1 is

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

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 $\gamma = 0$ and is the most deterministic when $\gamma \to \infty$. In the study of Hu and Rosenberger (2003) for binary responses, $\gamma = 2$ is recommended, which can achieve a high power while allowing a reasonable degree of randomness. Hu, Rosenberger and Zhang (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 $\gamma \to \infty$. It has been applied to censored survival outcomes in a fixed-sample design by Zhang and Rosenberger (2007).

Efficient randomised-adaptive design (ERADE)

Similar to the DBCD function, the allocation probability function for the ERADE depends on both the current and the estimated target allocation proportions. However, the ERADE function is discontinuous. The probability that the next patient will be assigned to treatment 1 is

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

where $0 \leq \gamma' < 1$ is a constant that controls the degree of randomisation.

The ERADE allocation procedure becomes more deterministic when $\gamma' \to 0$. A value of γ' between 0.4 and 0.7 is recommended, and the ERADE has been proved to always attain the Cramér-Rao lower bound for the variance of the

allocation proportions (Hu, Zhang and He, 2009).

The allocation probability for treatment 1, g_1, is updated sequentially after each outcome observed. Although survival outcomes are usually not immediately available, the optimal response-adaptive randomisation procedures can be used as long as some outcomes have been obtained. It is shown that a moderate delay in censored survival responses has only a modest effect on the asymptotic properties of the DBCD (Zhang and Rosenberger, 2007) and similarly for the ERADE (Hu, Zhang and He, 2009).

5 Simulation study

5.1 Exponential survival times

Consider comparing the mean survival time for the two treatment groups. Similar parameters were used as in Zhang and Rosenberger (2007). However, our simulation study extends this to investigate different response-adaptive randomisation methods incorporated with group sequential designs using the simple difference and log hazard ratio test statistics. The designs are compared in terms of the error probabilities, the expected number of patients (ENP), the expected number of failures (ENF) calculated when the trial stops, the average allocation proportion for treatment 1 ($\tilde{\rho}_1$) and the corresponding variability. The operating characteristics for complete randomisation (CR) were provided to compare with the response-adaptive randomisation procedures and the results for the fixed-sample designs were displayed to compare with the group sequential designs. The simulation study was based on 10,000 replicates.

At subject level, a random sample of arrival times for n_k patients recruited at look k was drawn using $U \sim U(Dt_{k-1}, Dt_k)$, where $t_0 = 0$ and $t_K = 1$. The arrival times were then sorted and we obtained the arrival time for patient i, A_i. For each patient, a survival time $S_{i,j}$ for those allocated to treatment j was simulated together with a competing censoring time C_i, where $S_{i,j}$ and

C_i both began at time A_i. The observed survival time for patient i allocated to treatment j at look k is the minimum of $S_{i,j}$, C_i and the length of time that the patient had been in the trial by group sequential test k, $Dt_k - A_i$. The form of the tests is in Section 3.1. To achieve around 80% power to detect a hazard ratio of 4 using the log hazard ratio test statistic for the group sequential designs, we considered the mean survival times $\theta_1 = 1.4$ and $\theta_2 = 1$ for the two treatments and the maximum number of patients $N = 800$. As in Zhang and Rosenberger (2007), the duration of the trial $D = 1.5936$ and an overall type I error rate of 5% were set.

Two unequally-spaced information times (0.2, 0.5, 1) and (0.5, 0.8, 1) are planned for the group sequential designs with $K = 3$ analyses. The former conducts early interim looks, whereas the latter has interim analyses after reaching one half of the maximum information level. For the simple difference and log hazard ratio test statistics $Z_k,\ k = 1, 2, 3$, the O'Brien and Fleming boundaries derived by the error-spending approach, which were originally derived based on a non-adaptive randomisation setting, were utilised as approximate tests. The 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.

For the response-adaptive randomisation procedures, treatment allocation also depends on the parameter estimates. In our simulation study, permuted block randomisation, which ensured balanced allocation, was used for the first 10% of patients to obtain initial parameter estimates. Then the response-adaptive randomisation procedures were applied with the aim of targeting the optimal allocations that minimise the total expected hazard. The tuning parameters for the DBCD and the ERADE were $\gamma = 2$ and $\gamma^{'} = 0.5$, respectively, which were set to achieve a high power while allowing a reasonable degree of randomness.

5.1.1 Simple difference

In Table 1, a slightly conservative type I error rate was obtained for CR. Compared with CR, the response-adaptive randomised designs had a slightly higher $\tilde{\alpha}$, which was within the range of three standard errors from $\alpha = 0.05$, that is, (0.043, 0.057). This shows that the critical boundaries can be used approximately for the combined approach with censored survival responses. Under H_0, the ENP and the ENF were similar across all of the designs. However, the use of the response-adaptive designs increased the variability in the allocation proportion compared to CR.

Under H_a, from Table 2, the use of the response-adaptive randomised designs was more ethical. For instance, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), compared to CR, the use of optimal response-adaptive randomisation on average required 15 fewer patients and there were 18 fewer failures without a loss of power. Compared with the fixed-sample designs, the ENP and the ENF were both significantly decreased in the group sequential designs. In the case of $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), a 12% reduction in the ENP for the response-adaptive designs was attained. For CR, there was a 10% reduction in the ENP. The standard deviation of the ENP for the group sequential designs was large, since there were only three possible values for the number of patients. For instance, for the case $(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1), the number of patients was $n_1 = 442$, $n_2 = 561$ or $n_3 = 800$ when the trial terminated at the first, the second or the final stage, respectively. For the case $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), the number of patients could be $n_1 = 561$, $n_2 = 698$ or $n_3 = 800$. The case $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1) had the least ENP, and hence the ENF due to the fact that there was more early stopping in this case. The optimal allocation proportion for treatment 1, ρ_1, was well targeted for both of the response-adaptive randomised designs, with the ERADE consistently having a lower variability in the allocation proportion.

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

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.041	799.7	(8.6)	338.7	(14.8)	0.500	(0.017)
DBCD	0.055	798.5	(18.6)	338.1	(17.8)	0.500	(0.059)
ERADE	0.054	799.0	(15.5)	338.4	(16.5)	0.500	(0.051)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.045	797.6	(16.9)	338.6	(16.6)	0.500	(0.017)
DBCD	0.052	796.7	(22.0)	337.9	(18.6)	0.500	(0.057)
ERADE	0.056	796.6	(21.5)	338.1	(18.3)	0.501	(0.052)
	Fixed-sample designs						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.044	800	(0)	338.9	(13.7)	0.500	(0.017)
DBCD	0.054	800	(0)	338.9	(13.7)	0.501	(0.053)
ERADE	0.050	800	(0)	338.9	(13.7)	0.501	(0.050)

Table 2: Simulated power for two-armed censored survival trials with optimal allocation in group sequential and fixed-sample designs using tests of the simple difference, $\theta_1 = 1.4,\ \theta_2 = 1$ and $N = 800$.

		$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.830	767.6	(81.8)	285.9	(47.4)	0.500	(0.017)	0	1,356
DBCD	0.825	750.6	(96.8)	266.9	(56.7)	0.657	(0.067)	0	2,069
ERADE	0.821	751.7	(96.0)	267.9	(56.0)	0.654	(0.061)	0	2,023
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.810	718.2	(76.5)	263.1	(43.8)	0.500	(0.018)	1,283	5,012
DBCD	0.826	703.3	(86.6)	245.4	(51.8)	0.658	(0.064)	2,133	4,483
ERADE	0.823	704.5	(87.2)	245.7	(52.1)	0.661	(0.061)	2,135	4,363
		Fixed-sample designs							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.828	800	(0)	303.9	(13.5)	0.500	(0.017)	-	-
DBCD	0.836	800	(0)	293.2	(14.0)	0.653	(0.053)	-	-
ERADE	0.833	800	(0)	293.3	(14.0)	0.653	(0.050)	-	-

Here, the optimal allocation for treatment 1, ρ_1, is 0.652 for the DBCD and the ERADE. T1 and T2 denote the numbers of trials terminated early at information times t_1 and t_2, respectively, based on 10,000 replicates.

5.1.2 Log hazard ratio

From Table 3, the type I error rate was within two standard errors of 0.05 for all designs. Similar conclusions to those for the simple difference case were obtained. However, the standard deviations of $\tilde{\rho}_1$ for the DBCD and the ERADE are significantly reduced using the log hazard ratio test. Under H_a, from Table 4, the optimal allocation for the response-adaptive randomised designs was closer to equal allocation compared to the simple difference case. Hence, the results for the DBCD and the ERADE were similar to those for CR and the reduction in the number of failures by using the response-adaptive randomisation procedures was less compared to Table 2.

Table 3: Simulated type I error rate for two-armed censored survival trials with optimal allocation in group sequential and fixed-sample designs using tests of the log hazard ratio, $\theta_1 = \theta_2 = 1$ and $N = 800$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.046	799.3	(12.6)	338.6	(15.6)	0.500	(0.017)
DBCD	0.047	799.1	(14.3)	338.5	(16.3)	0.500	(0.028)
ERADE	0.052	799.0	(15.6)	338.3	(16.8)	0.500	(0.023)
	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.042	797.6	(16.8)	338.3	(16.4)	0.500	(0.017)
DBCD	0.048	797.1	(20.1)	338.4	(17.9)	0.500	(0.027)
ERADE	0.046	797.4	(18.9)	338.3	(17.5)	0.500	(0.024)
	Fixed-sample designs						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.047	800	(0)	338.9	(13.7)	0.500	(0.017)
DBCD	0.051	800	(0)	338.9	(13.7)	0.500	(0.025)
ERADE	0.049	800	(0)	338.9	(13.7)	0.500	(0.023)

Table 4: Simulated power for two-armed censored survival trials with optimal allocation in group sequential and fixed-sample designs using tests of the log hazard ratio, $\theta_1 = 1.4,\ \theta_2 = 1$ and $N = 800$.

		$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.832	749.1	(97.9)	275.7	(56.4)	0.500	(0.017)	0	2,130
DBCD	0.828	750.5	(96.9)	271.6	(56.4)	0.577	(0.036)	1	2,070
ERADE	0.832	749.8	(97.4)	271.2	(56.5)	0.577	(0.031)	3	2,096
		$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.825	705.2	(85.4)	255.4	(49.6)	0.500	(0.018)	2,017	4,566
DBCD	0.826	703.2	(86.1)	249.9	(50.8)	0.577	(0.033)	2,110	4,543
ERADE	0.827	702.2	(86.5)	249.2	(51.0)	0.579	(0.031)	2,159	4,533
		Fixed-sample designs							
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)	T1	T2
CR	0.838	800	(0)	303.9	(13.5)	0.500	(0.017)	-	-
DBCD	0.839	800	(0)	298.9	(13.6)	0.572	(0.026)	-	-
ERADE	0.839	800	(0)	298.9	(13.6)	0.572	(0.024)	-	-

Here, the optimal allocation for treatment 1, ρ_1, is 0.572 for the DBCD and the ERADE. T1 and T2 denote the numbers of trials terminated early at information times t_1 and t_2, respectively, based on 10,000 replicates.

5.2 Redesigning a clinical trial

Jones *et al.* (2005) is a study of a phase III survival trial in breast cancer comparing docetaxel and paclitaxel using a fixed-sample equal allocation design. The total sample size is 449 and the duration of the trial is 102 months. Zhang and Rosenberger (2007) investigated the redesign of the trial using the DBCD in a fixed-sample design. Based on the area under the Kaplan-Meier plot of survival in Jones *et al.* (2005), they used the mean overall survival times for docetaxel and paclitaxel as θ_1=16.1 and θ_2=12.7 months, respectively. They obtained $\tilde{\rho}_1 = 0.59$ and suggested that 265 patients would be randomised to the better treatment docetaxel using the fixed-sample DBCD design compared to 225 patients in the original equal allocation design. In this section, we investigate the redesign of the trial in a group sequential setting with $(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1), since the events were obtained quickly in this case. The other parameters used are the same as in Section 5.1.

Compared to CR, the combined approach reduced the expected numbers of patients and failures while not adversely affecting the error rates: see Tables 5 and 6. The inflated type I error rate for the DBCD and the ERADE, which was due to the use of critical boundaries for standard group sequential designs approximately, was within the range of three standard errors of 0.05. The average allocation proportion for the docetaxel arm of $\tilde{\rho}_1 = 0.60$ was obtained for the combined approach, and hence 258 out of the 430 expected number of patients would be allocated to the better treatment if the combined approach was used.

Table 5: Simulated type I error rate for the redesign of a clinical trial with optimal allocation in a group sequential design using tests of the simple difference, $\theta_1 = \theta_2 = 12.7$ and $N = 449$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.046	448.7	(7.0)	356.5	(10.4)	0.500	(0.023)
DBCD	0.056	447.7	(16.1)	355.6	(16.3)	0.500	(0.050)
ERADE	0.057	447.6	(16.7)	355.4	(16.8)	0.500	(0.046)

Table 6: Simulated power for the redesign of a clinical trial with optimal allocation in a group sequential design using tests of the simple difference, $\theta_1 = 16.1,\ \theta_2 = 12.7$ and $N = 449$.

	$(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1)						
Procedure	Power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.586	437.3	(46.3)	336.5	(40.4)	0.500	(0.023)
DBCD	0.611	430.1	(57.7)	328.3	(50.4)	0.602	(0.056)
ERADE	0.615	429.3	(58.7)	327.6	(51.5)	0.603	(0.052)

For the DBCD and the ERADE, the target optimal allocation for treatment 1, ρ_1, is 0.596 for the simple difference test.

5.3 Weibull survival times

Exponential survival responses have a constant hazard rate which may be unrealistic in practice. Other distributions of survival times, such as the Weibull

distribution, can be used when the hazard rate is changing over time. In this section, the application of the group-sequential response-adaptive design to Weibull survival responses is demonstrated.

Let $S_{i,j}$ be the survival time for patient i on treatment j that follows a Weibull distribution. Then $\log(S_{i,j})$ has an extreme value distribution with two parameters μ_j and b_j determining its scale and shape. As shown in Section 2.2, the observed survival response for patient i on treatment j at group sequential test k is $T_{i,j,k} = \min(S_{i,j}, C_i, Dt_k - A_i)$. Let $Y_{i,j,k} = \log(T_{i,j,k})$. We have two independent random samples $\{y_{i,j,k}, \delta_{i,j,k},\ i = 1, \ldots, m_{j,k}\}$ for $j = 1, 2$, where $\delta_{i,j,k} = 1$ indicates an event and $\delta_{i,j,k} = 0$ refers to right-censoring. Then the likelihood function is given by

$$L_k(\mu_j, b_j) = \prod_{i=1}^{m_{j,k}} \left[\frac{1}{b_j} \exp\left(\frac{y_{i,j,k} - \mu_j}{b_j}\right) \exp\left\{-\exp\left(\frac{y_{i,j,k} - \mu_j}{b_j}\right)\right\}\right]^{\delta_{i,j,k}} \times \left[\exp\left\{-\exp\left(\frac{y_{i,j,k} - \mu_j}{b_j}\right)\right\}\right]^{1-\delta_{i,j,k}}.$$

The log-likelihood function is

$$l_k(\mu_j, b_j) = -r_{j,k}\log(b_j) + \sum_{i=1}^{m_{j,k}} \{\delta_{i,j,k} z_{i,j,k} - \exp(z_{i,j,k})\},$$

where $r_{j,k} = \sum_{i=1}^{m_{j,k}} \delta_{i,j,k}$ is the cumulative number of events and $z_{i,j,k} = (y_{i,j,k} - \mu_j)/b_j$. The maximum likelihood estimates of the two parameters can be obtained numerically by solving

$$\frac{\partial l_k(\mu_j, b_j)}{\partial \mu_j} = -\frac{1}{b_j} \sum_{i=1}^{m_{j,k}} \{\delta_{i,j,k} - \exp(z_{i,j,k})\} = 0$$

and

$$\frac{\partial l_k(\mu_j, b_j)}{\partial b_j} = -\frac{r_{j,k}}{b_j} - \frac{1}{b_j} \sum_{i=1}^{m_{j,k}} \{\delta_{i,j,k} - \exp(z_{i,j,k})\} z_{i,j,k} = 0.$$

The approximate variance of the maximum likelihood estimator $\hat{\mu}_{j,k}$ is $\text{var}(\hat{\mu}_{j,k}) = b_j^2 G_{j,k}/m_{j,k}$, where

$$G_{j,k} = \frac{\epsilon_{j,k} + E\{z_{i,j,k}^2 \exp(z_{i,j,k})\}}{\epsilon_{j,k}^2 + \epsilon_{j,k} E\{z_{i,j,k}^2 \exp(z_{i,j,k})\} - E\{z_{i,j,k} \exp(z_{i,j,k})\}^2}. \tag{8}$$

Here, the probability of an event, $\epsilon_{j,k}$, is a function of the unknown parameters. For simplicity, we considered the simple nonparametric estimate $\hat{\epsilon}_{j,k} = r_{j,k}/m_{j,k}$. The expectations in (8) are estimated similarly.

For response-adaptive randomisation, consider the average hazard as the reciprocal of the expected mean survival time as for the exponential outcomes case. Then the average hazard rate can be expressed as $1/E(T_{i,j}) = \exp(-\mu_j)\Gamma^{-1}(1 + b_j)$, where Γ denotes the gamma function. The optimal allocation is found by minimising

$$M_1 \exp(-\mu_1)\Gamma^{-1}(1 + b_1) + (N - M_1) \exp(-\mu_2)\Gamma^{-1}(1 + b_2)$$

under the variance constraint $b_1^2 G_1/M_1 + b_2^2 G_2/(N - M_1) = C$, where $C > 0$ is a constant. The solution is

$$\rho_1 = \frac{b_1 \sqrt{\exp(-\mu_2)\Gamma(1 + b_1)G_1}}{b_1 \sqrt{\exp(-\mu_2)\Gamma(1 + b_1)G_1} + b_2 \sqrt{\exp(-\mu_1)\Gamma(1 + b_2)G_2}}$$

and $\rho_2 = 1 - \rho_1$. Then the optimal response-adaptive randomisation procedures described in Section 4.2 can be used to target the optimal allocation. However, the asymptotic variance of the allocation proportions for the Weibull case cannot be obtained, since there is no closed-form solution for the maximum likelihood estimators.

We now revisit the clinical trial of Jones *et al.* (2005) and consider the time to progression outcome. The summary results showed that the median times to progress were 5.7 and 3.6 months for docetaxel and paclitaxel, respectively. The hazard ratio was 1.64 with 95% confidence interval (1.33, 2.02) and p<0.0001. Assume that the shape parameter for the Weibull sur-

vival time is the same for both treatments. Then the parameters of the two Weibull survival times $\mu_1 = 1.10$ and $\mu_2 = 0.64$ and $b_1 = b_2 = 0.93$ were obtained. In this case, the probability of an event was high throughout the trial. From (3), we used the approximate information time

$$t_k \approx \frac{n_k}{N} \in (0, 1], \quad k = 1, ..., K,$$

since $\hat{\epsilon}_k \approx \hat{\epsilon}_K$. Consider tests of the simple difference. We have $\hat{\phi}_k = \hat{\mu}_{1,k} - \hat{\mu}_{2,k}$ and $\text{var}(\hat{\phi}_k) = b_1^2 G_{1,k}/m_{1,k} + b_2^2 G_{2,k}/m_{2,k}$ to construct the test statistic in (5). The marginal distribution of the test statistic is approximately normal for large sample sizes. The O'Brien and Fleming critical boundaries derived based on group sequential non-adaptive designs using normal responses are applied as approximate tests. Group sequential designs with $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1) were considered, as the simple nonparametric estimate of the probability of an event is more accurate for large sample sizes.

Tables 7 and 8 show that the standard critical boundaries can be used as an approximate test for the combined approach with Weibull survival responses to maintain the error rates. Due to the large effect size, a high power was obtained. The operating characteristics were similar for the designs in this case. As the difference in treatment effects was large, there was sufficient evidence for early stopping. The ENP was 355 on average, which was much lower than the maximum sample size.

Table 7: Simulated type I error rate for two-armed censored survival trials with Weibull responses using tests of the simple difference test statistic, $\mu_1 = \mu_2 = 0.64,\ b_1 = b_2 = 0.93$ and $N = 449$.

	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	$\tilde{\alpha}$	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.042	448.9	(3.6)	437.6	(4.8)	0.500	(0.022)
DBCD	0.042	448.8	(4.5)	437.5	(5.4)	0.500	(0.016)
ERADE	0.040	448.8	(3.9)	437.6	(5.0)	0.500	(0.012)

Table 8: Simulated power for two-armed censored survival trials with Weibull responses using tests of the simple difference test statistic, $\mu_1 = 1.10,\ \mu_2 = 0.64,\ b_1 = b_2 = 0.93$ and $N = 449$.

	$(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1)						
Procedure	power	ENP	(s.d.)	ENF	(s.d.)	$\tilde{\rho}_1$	(s.d.)
CR	0.999	355.0	(43.1)	346.5	(41.7)	0.500	(0.025)
DBCD	0.999	355.6	(42.2)	346.7	(40.7)	0.558	(0.019)
ERADE	0.999	355.6	(42.2)	346.8	(40.8)	0.559	(0.014)

6 Discussion

6.1 Conclusions

A generalisation of the combined approach to censored survival responses using an approximate information time is investigated. Standard non-adaptive group sequential critical boundaries derived using the error-spending approach can be applied to the designs. The inflated type I error rate for the response-adaptive designs may be due to the use of the approximate test. Alternatively, one may obtain the critical boundaries by simulation.

Incorporating adaptive sampling rules in group sequential designs preserves the error rates while assigning more patients to the more promising treatment. Moreover, both of the optimal response-adaptive randomisation procedures target the pre-specified optimal allocation well with reasonably small variability. The ERADE consistently has a lower standard deviation for the allocation proportion compared to the DBCD.

Compared with the group sequential logrank test based on a simple non-parametric estimate of the probability of an event in Supplementary Material 2, the proposed model-based combined approach has less early stopping in the case $(t_1,\ t_2,\ t_3)$=(0.2, 0.5, 1). In addition, for $(t_1,\ t_2,\ t_3)$=(0.5, 0.8, 1), the model-based approach has lower ENP and ENF. In practice, since there might be greater bias or variance in the simple empirical estimate of the probability of an event when the sample size is small, one needs to make

sure interim analyses are conducted when there is sufficient information.

An extension of the combined approach to Weibull survival times is achievable. However, the derivation of the probability of an event is more difficult. The simple nonparametric approach was used in Section 5.3. If the model assumptions are wrong, then estimation might be biased and this could negatively impact the design performance. A case of model misspecification, which assumed an exponential survival model when the true survival time had a Weibull distribution, was explored in Supplementary Material 3. The simulation results show that the error rates are robust in this case. However, more patients are allocated to the better treatment when an exponential survival model is assumed instead of the true Weibull survival model.

6.2 Further work

The O'Brien and Fleming boundaries are used in this paper. Different critical boundaries can also be considered. Simulation results for Pocock's boundaries (Pocock, 1977), which use the same nominal type I error rate at each group sequential analysis, lead to the same conclusions. However, Pocock's test is more likely to reject the null hypothesis at early looks than the O'Brien and Fleming boundaries.

In this paper, 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 (Biswas, Bhattacharya and Park, 2016). Designs that combine group sequential monitoring with CARA designs would be more complicated, yet may be of interest for future study.

Patients usually enter clinical trials sequentially, which may incur the risk of chronological bias due to unobserved time trends. An appropriate randomisation procedure should be chosen to minimise the risk of chronological bias (Tamm and Hilgers, 2014). For permuted block randomisation, small to

medium block sizes are sufficient to limit such bias. Exploring the issue for adaptive randomisation could provide additional information on the feasibility of the use of adaptive randomisation in practice. For some recent work in this direction, see Villar, Bowden and Wason (2018).

Throughout, we focus on the case of two-armed clinical trials. Several treatments are often compared in a clinical trial nowadays. Some of the ideas presented here can be generalised to more than two treatments, with all treatments continuing to the end of the trial or with treatment selection. A generalisation of Fisher's least significant difference method to group-sequential response-adaptive designs has been explored. This method is considered to be one of the most powerful multiple comparison approaches in fixed-sample designs (Christensen, 2002). Details of the group-sequential response-adaptive designs for multi-armed clinical trials will be reported separately.

Acknowledgements

This work was carried out whilst the first author was in receipt of funding from the Ministry of Education in Taiwan at Queen Mary, University of London. A paper based on this work was selected for one of the Student Conference Awards at the 35th Annual Conference of the International Society for Clinical Biostatistics in Vienna, Austria, during 24-28 August 2014. The authors also wish to thank two referees for their comments, which have led to an improved paper.

References

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

Biswas, A., Bhattacharya, R., Park, E., 2016. On a class of optimal covariate-adjusted response adaptive designs for survival outcomes. Statistical Methods in Medical Research 25, 2444-2456.

Christensen, R., 2002. Plane Answers to Complex Questions: The Theory of Linear Models, third ed. Springer, New York.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Chapman and Hall, London.

Eisele, J.R., Woodroofe, M.B., 1995. Central limit theorems for doubly adaptive biased coin designs. Annals of Statistics 23, 234-254.

Hu, F., Rosenberger, W.F., 2003. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. Journal of the American Statistical Association 98, 671-678.

Hu, F., Rosenberger, W.F., Zhang, L.-X., 2006. Asymptotically best response-adaptive randomization procedures. Journal of Statistical Planning and Inference 136, 1911-1922.

Hu, F., Zhang, L.-X., He, X., 2009. Efficient randomized-adaptive designs. Annals of Statistics 37, 2543-2560.

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

Jennison, C., Turnbull, B.W., 1991. Exact calculations for sequential t, χ^2 and F tests. Biometrika 78, 133-141.

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

Jennison, C., Turnbull, B.W., 2001. Group sequential tests with outcome-dependent treatment assignment. Sequential Analysis 20, 209-234.

Jones, S.E., Erban, J., Overmoyer, B., Budd, G.T., Hutchins, L., Lower, E., Laufman, L., Sundaram, S., Urba, W.J., Pritchard, K.I., Mennel, R., Richards, D., Olsen, S., Meyers, M.L., Ravdin, P.M., 2005. Randomized phase III study of docetaxel compared with paclitaxel in metastatic breast cancer. Journal of Clinical Oncology 23, 5542-5551.

Kim, K., Boucher, H., 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., DeMets, D.L., 1983. Discrete sequential boundaries for clinical trials. Biometrika 70, 659-663.

Morgan, C.C., 2003a. Estimation following group-sequential response-adaptive clinical trials. Controlled Clinical Trials 24, 523-543.

Morgan, C.C., 2003b. Sample size re-estimation in group-sequential response-adaptive clinical trials. Statistics in Medicine 22, 3843-3857.

Morgan, C.C., Coad, D.S., 2007. A comparison of adaptive allocation rules for group-sequential binary response clinical trials. Statistics in Medicine 26, 1937-1954.

O'Brien, P.C., 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.

Tamm, M.T., Hilgers, R.-D., 2014. Chronological bias in randomized clinical trials arising from different types of unobserved time trends. Methods of Information in Medicine 53, 501-510.

Villar, S.S., Bowden, J., Wason, J., 2018. Response-adaptive designs for binary responses: how to offer patient benefit while being robust to time trends? Pharmaceutical Statistics 17, 182-197.

Zhang, L., Rosenberger, W.F., 2007. Response-adaptive randomization for survival trials: the parametric approach. Applied Statistics 56, 153-165.

Zhu, H., Hu, F., 2010. Sequential monitoring of response-adaptive randomized clinical trials. Annals of Statistics 38, 2218-2241.