Research Article

Statistics in Medicine

Received 12 June 2013, Accepted 3 April 2014 Published online 13 May 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6183

Flexible sequential designs for multi-arm clinical trials

D. Magirr,[a*†] N. Stallard[b] and T. Jaki[a]

Adaptive designs that are based on group-sequential approaches have the benefit of being efficient as stopping boundaries can be found that lead to good operating characteristics with test decisions based solely on sufficient statistics. The drawback of these so called 'pre-planned adaptive' designs is that unexpected design changes are not possible without impacting the error rates. 'Flexible adaptive designs' on the other hand can cope with a large number of contingencies at the cost of reduced efficiency. In this work, we focus on two different approaches for multi-arm multi-stage trials, which are based on group-sequential ideas, and discuss how these 'pre-planned adaptive designs' can be modified to allow for flexibility. We then show how the added flexibility can be used for treatment selection and sample size reassessment and evaluate the impact on the error rates in a simulation study. The results show that an impressive overall procedure can be found by combining a well chosen pre-planned design with an application of the conditional error principle to allow flexible treatment selection.

Keywords: adaptive designs; closed testing; combination test; conditional error; multi-arm

1. Introduction

Sequential methods that allow monitoring of accumulating data at a series of interim analyses are relatively common in clinical trials comparing a single experimental treatment with control. The interim analyses provide opportunities to stop the trial early if there is sufficient evidence of efficacy, or sufficient lack thereof. This is important for ethical reasons, because the trial should involve the smallest number of patients necessary for it to reach a firm conclusion, and possibly also for administrative and economic reasons [1].

Recently, there has been considerable interest in extending group-sequential methods to design multi-arm multi-stage trials where, alongside determining whether or not to stop the trial early, interim analyses are used to select only the most promising treatment(s) to be compared with the control in subsequent stages [2–7]. The proposed methods use different selection rules but all build on the idea of testing a family of null hypotheses by monitoring sufficient statistics for the corresponding parameters of interest, and comparing these with stopping boundaries. To control the familywise error rate, that is, the probability of rejecting at least one true null hypothesis, appropriate stopping boundaries can be determined from the joint null distribution.

One limitation of the proposed methodology is that once the trial has started, it must be conducted as specified. This has been described as *pre-planned* or *pre-specified adaptivity* [8]. Whilst specifying decision rules helps maintain the integrity of the trial and credibility of any conclusions, it is not possible to foresee all eventualities at the design stage. If circumstances arise that force investigators to deviate from the pre-specified design then, depending on the specific design change, the group-sequential methodology may be rendered invalid or lose statistical power.

An alternative to pre-planned adaptive methods is the use of *flexible adaptive designs* [9–11]. Such designs have the advantage that design changes can be made at interim analyses without pre-

[a]*Medical and Pharmaceutical Statistics Research Unit, Lancaster University, Lancaster LA1 4YF, U.K.*
[b]*Warwick Medical School, University of Warwick, Coventry CV4 7AL, U.K.*
**Correspondence to: D. Magirr, Medical and Pharmaceutical Statistics Research Unit, Lancaster University, Lancaster LA1 4YF, U.K.*
[†]*E-mail: d.magirr@gmail.com*

specification, yet the familywise error rate can still be controlled. This is usually achieved by calculating a p-value at each analysis based only on data from the preceding stage and combining these p-values using a pre-specified *combination function* to make test decisions. In general, however, the resulting tests are not based on sufficient statistics for the relevant parameters.

Brannath *et al*. [8] note that a fixed sample size or pre-planned adaptive design can be used as the basis for a flexible design, with flexibility introduced via the so-called *conditional error* approach [12]. Considering that flexible adaptive designs have been criticised for making inefficient use of the data [13, 14], this suggestion brings to mind an intuitively appealing strategy: start the trial with a pre-planned adaptive design that has good operating characteristics should the trial proceed as envisioned, and only apply the flexible adaptive methodology if necessary. Koenig *et al*. [15] apply such a strategy with a fixed sample Dunnett test [16] used as the initial design, leading to a design that has been shown to compare well in terms of power with competing methods [17]. Di Scala & Glimm [18] extend this approach to deal with time-to-event endpoints, and Friede *et al*. [19] show that the same techniques are applicable to trials involving subgroup selection.

In this article, we develop flexible sequential designs for multi-arm clinical trials with early stopping for efficacy and futility. More specifically, we will consider the design of Stallard & Todd [3] in which only a single experimental treatment is taken forward beyond the first stage, and the design of Magirr *et al*. [7] in which all sufficiently promising experimental treatments are carried forward. Flexible modifications of the Stallard & Todd design are proposed based on both the combination testing approach and the conditional error function approach, and the Magirr *et al*. design is modified using the conditional error function approach. All three modifications provide fully flexible multi-arm multi-stage designs that control the familywise error rate in the strong sense. The properties of the three new approaches are assessed in a simulation study.

2. Multi-arm multi-stage trial designs

Suppose that patients are randomised between a control treatment ($k = 0$) and K experimental treatments ($k = 1, \ldots, K$), with observations from treatment group k independently normally distributed with mean μ_k and variance 1. We denote the treatment effect size for treatment k by $\theta_k = \mu_k - \mu_0$ ($k = 1, \ldots, K$), and the vector of treatment responses by $\theta = (\theta_1, \ldots, \theta_K)$, and consider the null hypotheses $H_k : \theta_k \leqslant 0$ ($k = 1, \ldots, K$). We wish to control the familywise error rate in the strong sense, that is to have

$$P_\theta \left(\text{reject at least one } \textit{true } H_k \right) \leqslant \alpha \text{ for all } \theta. \tag{1}$$

Suppose that observations are made in J stages with $n_{k,j}$ observations in group k in the first j stages. Let $\mathcal{I}_{k,j}$ and $Z_{k,j}$ denote respectively the Fisher's information for θ_k and the standardised test statistic for H_k based on all data observed up to and including stage j ($k = 1, \ldots, K; j = 1, \ldots, J$), with Z_j denoting the vector of test statistics at stage j, $\left(Z_{1,j}, \ldots, Z_{K,j}\right)$. We note that $\mathcal{I}_{k,j} = \left(1/n_{0,j} + 1/n_{k,j}\right)^{-1}$ and $Z_{k,j} = \mathcal{I}_{k,j}^{1/2}\left(\bar{X}_{k,j} - \bar{X}_{0,j}\right)$ with $\bar{X}_{k,j}$ the sample mean of the observations in group k in the first j stages.

The standardised test statistics $Z_1, \ldots, Z_J$ are multivariate normal with

$$\begin{aligned} E\left(Z_{k,j}\right) &= \theta_k \mathcal{I}_{k,j}^{1/2}, \\ \text{var}\left(Z_{k,j}\right) &= 1, \\ \text{cov}\left(Z_{k,j}, Z_{k,j'}\right) &= \mathcal{I}_{k,\min\{j,j'\}}^{1/2} / \mathcal{I}_{k,\max\{j,j'\}}^{1/2}, \\ \text{cov}\left(Z_{k,j}, Z_{k',j'}\right) &= \mathcal{I}_{k,j}^{1/2} \mathcal{I}_{k',j'}^{1/2} / n_{0,\max\{j,j'\}} \qquad \left(k \neq k'\right). \end{aligned} \tag{2}$$

Cases with unknown variance and other distributional forms are discussed in the succeeding text.

Prior to each stage, a set of treatments are chosen to continue to the next stage. Before the first stage, this set includes all treatments. For later stages, treatments are selected on the basis of data already observed. Two possible approaches are described in detail in the succeeding text. After data from stage j have been observed, in order to test H_k for those treatments still in the trial, test statistic $Z_{k,j}$ is compared with upper and lower stopping boundaries, u_j and l_j. If $Z_{k,j} \notin \left(l_j, u_j\right)$, treatment k is dropped from the trial with H_k rejected if $Z_{k,j} \geqslant u_j$ and retained if $Z_{k,j} \leqslant l_j$. If any treatments remain in the

trial, a selection is made of which should continue to the next stage. In order to ensure the trial stops after J stages, we set $u_J = l_J$. The entire design is then specified by the sample sizes, $n_{k,j}$, the critical values $u_1, \ldots, u_J$ and $l_1, \ldots, l_J$ and the pre-specified selection rule.

Two specific examples of treatment selection rules, with corresponding methods for construction of $u_1, \ldots, u_J$ and $l_1, \ldots, l_J$ to control the familywise error rate, are given in [3] and [7]. Stallard & Todd [3] propose a multi-arm multi-stage design in which only the best performing experimental treatment in the first stage, that is, the treatment with the largest test statistic $Z_{k,1}$, may continue with the control treatment beyond the first interim analysis. Magirr *et al*. [7] describe an alternative selection procedure in which all treatments with $l_j < Z_{k,j} < u_j$ remain in the study to proceed to stage $(j+1)$. We will refer to these two selection rules as the 'select the best' and the 'keep all promising' selection rules.

It is shown in [4] and [7] that, in both designs, the familywise error rate is largest when $\theta_1 = \cdots = \theta_K = 0$, which will be written as $\theta = 0$. Assuming for simplicity that $l_1, \ldots, l_{J-1}$ are fixed, in both cases, critical values $u_1, \ldots, u_J$ to make (1) an equality can thus be found using (2) with $\theta = 0$ allowing for the selection rule used.

The single constraint (1) is insufficient to determine $u_1, \ldots, u_J$ uniquely, and a common approach in sequential analysis is to specify the type I error to be spent at each interim analysis. That is, to find $u_1, \ldots, u_J$ such that

$$P_0 \text{ (reject any } H_k \text{ at or before interim analysis } j; \theta = 0) = \alpha_j^*, \tag{3}$$

where $\alpha_1^* \leqslant \cdots \leqslant \alpha_J^* = \alpha$ are either specified in advance [20] or depend on the $\mathcal{I}_{k,j}$ in some pre-determined way [5, 21].

It is easy to show that with the 'keep all promising' design, dropping some experimental treatment(s) with a test statistic that exceeds the lower boundary with the testing procedure otherwise unaltered leads to a conservative procedure. Similarly, with the 'select the best' design, selecting some single experimental treatment other than that with the largest $Z_{k,1}$ with the testing procedure otherwise unaltered leads to a conservative procedure [22]. If an experimental treatment with test statistic below the lower boundary is retained in the 'keep all promising' design, however, or if more than one experimental treatment is allowed to continue beyond the first stage in the 'select the best' design, the familywise error rate will not be controlled at level α.

In addition to the constraint on the familywise error rate (1), it is desirable that a test satisfies a specified power requirement. Following a suggestion of Dunnett [23], we let δ denote a clinically relevant treatment effect size and let δ_0 $(0 \leqslant \delta_0 < \delta)$ denote a marginal improvement. We define power as the probability under $\theta = (\delta, \delta_0, \ldots, \delta_0)$ that H_1 is rejected with treatment 1 identified as the most promising treatment in the sense that if interim analysis j is the first at which some null hypothesis is rejected, then $Z_{1,j} > Z_{k,j}$ for all other experimental treatments k still in the trial at stage j.

3. New flexible methods for multi-arm multi-stage trials

The selection rules in [3] and [7] are easy to describe and they are useful at the design stage as they make it possible to evaluate the joint distribution of the test statistics. However, they may be unrealistic in clinical practice, where a judgement is made by investigators on which treatments should continue. Letting X_{j-1} denote all data available prior to stage j, suppose that the treatments selected to continue to stage j may now depend on X_{j-1} in some unspecified way. In addition to all test statistics observed in stages 1 to $j-1$, X_{j-1} may include secondary endpoint efficacy data, safety data, etc. It is assumed, however, that independent cohorts of patients are recruited at each stage so that X_{j-1} does not contain any information that is predictive for responses of patients recruited in stages $j, \ldots, J$, such as a correlated short term outcome when there is a delay in observing the primary outcome.

We now show how two different approaches may be used to develop new methodology to enable error rate control in this setting. In both cases, strong familywise error control is achieved through application of the closure principle [24]. To apply the closure principle to the family of null hypotheses, $H_1, \ldots, H_K$, we first perform local hypothesis tests for all intersection hypotheses of the form $H_I = \bigcap_{k \in I} H_k$ for I a (non-empty) subset of $\{1, \ldots, K\}$. We then reject H_k if and only if H_I is rejected at local level α for all I that include k. This controls the familywise error rate at level α.

It is possible to re-express the multiple testing procedures in Section 2 as closed testing procedures, whereby H_I is rejected whenever H_k is rejected for at least one k in I. Denote the global null $\bigcap_{k=1}^{K} H_k$ by H. As the point 0, where the familywise error rate in (1) is maximised, is contained in H, the local test

of H has size exactly equal to α whereas all other local tests are conservative. This conservatism means that the rejection regions of the local tests could be enlarged such that the familywise error rate remains controlled at level α. This can be achieved by simply decreasing the upper boundary. For each I, we will denote the upper boundary values corresponding to the sequential local test of H_I by $u_{I,1}, \ldots, u_{I,J}$. We will refer to the resulting closed test procedures as *step-down* versions of the [3] and [7] designs because they are analogous to the step-down Dunnett test [24] and Holm's step-down extension of the Bonferroni procedure [25].

Assuming a fixed lower boundary $l_1, \ldots, l_{J-1}$, critical values $u_{I,1}, \ldots, u_{I,J}$ can be found for each I following the steps in Section 2 considering only those treatments contained in I. That is, $u_{I,1}, \ldots, u_{I,J}$ are found to satisfy

$$\sup_{\theta \in H_I} P_\theta \left(\text{reject } H_k \text{ for some } k \in I\right) = \alpha \tag{4}$$

with the pre-specified selection rule restricted to select only treatments from the set I.

The supremum in (4), for both the 'select the best' and 'keep all promising' designs, is attained for $\theta_k = 0$ $(k \in I)$ and arbitrary θ_k $(k \notin I)$. In practice, the construction of $u_{I,1}, \ldots, u_{I,J}$ is equivalent to finding boundaries for the original design with $K = |I|$ and ignoring the presence of treatments $k \notin I$. Again, the single constraint in (4) is insufficient to determine $u_{1,I}, \ldots, u_{J,I}$ uniquely and a similar sequence of expressions to (3) giving the amount of error rate to be spent at each interim analysis is necessary.

3.1. A combination test modification of the 'select the best' design

Bauer & Köhne [10] propose allowing for data-dependent design changes in multi-stage clinical trials by combining p-values calculated from the data collected in each stage. In one possible approach the stage-wise p-values may be transformed into independent normal increments, which are then summed, emulating a standard group-sequential trial [26]. For the simple case of a two-arm, two-stage trial to test the single null hypothesis, $H_1 : \theta_1 \leqslant 0$, suppose (l_1, u_1, u_2) are critical values for a group-sequential trial such that the probability under H_1 that either $Z^{(1)} \geqslant u_1$ or $Z^{(2)} \geqslant u_2$ with $l_1 < Z^{(1)} < u_1$ is equal to α, where $Z^{(1)}$ and $Z^{(2)}$ are standardised z-statistics based on cumulative data at stages 1 and 2, respectively, with correlation ρ depending on the pre-planned sample sizes. A p-value combination test can then be performed as follows. A first-stage test of H is chosen with associated p-value, $p^{(1)}$, calculated at the interim analysis. If $\Phi^{-1}\left(1 - p^{(1)}\right) > u_1$, where Φ is the standard normal cumulative distribution function, H_1 is rejected at level-α and the trial is stopped. If $\Phi^{-1}\left(1 - p^{(1)}\right) \leqslant l_1$, the trial is stopped early without rejection of H_1. If $l_1 < \Phi^{-1}\left(1 - p^{(1)}\right) \leqslant u_1$, the trial continues, and a second stage p-value, $p^{(2)}$, for a test of H_1, is calculated. At the final analysis, H_1 is rejected at level-α if and only if $\rho\Phi^{-1}\left(1 - p^{(1)}\right) + \sqrt{1-\rho^2}\Phi^{-1}\left(1 - p^{(2)}\right) > u_2$.

The choice of the second stage test is allowed to depend on the first stage data and external factors, as long as the null distribution of $p^{(2)}$, conditional on the first stage data and choice of second stage test statistic, is stochastically larger than or equal to the uniform distribution. Following [27], we will henceforth describe a p-value satisfying this property as *p-clud*.

Returning to the multi-arm multi-stage setting considered earlier, the test statistics used in [3] can be expressed as a combination of *stage-wise* p-values, so that combining these p-values and using a closed testing procedure can allow flexibility in the number of experimental treatments continuing at each stage, as well as future sample sizes.

In detail, suppose that for each $I \subseteq \{1, \ldots, K\}$, boundary values $u_{I,1}, \ldots, u_{I,J}$ and $l_1, \ldots, l_{J-1}$ have been found for a test of H_I as described in the preceding text. We assume that, in the initial design, an equal number of patients is allocated to each experimental treatment arm, with the number of patients allocated to the control arm a constant multiple of this number so that $n_{1,j} = \cdots = n_{K,j}$ and $n_{0,j} = \lambda n_{1,j}$. Thus $\mathcal{I}_{k,j} = n_{1,j}\lambda/(1+\lambda)$ for $k = 1, \ldots, K$.

At the jth analysis , define $\tilde{Z}_I^{(j)} = \max_{k \in I} Z_{k,1}$ for $j = 1$ and

$$\tilde{Z}_I^{(j)} = \Phi^{-1}\left(1 - p_I^{(j)}\right)\left\{\left(\mathcal{I}_j - \mathcal{I}_{j-1}\right)/\mathcal{I}_j\right\}^{1/2} + \tilde{Z}_I^{(j-1)}\left(\mathcal{I}_{j-1}/\mathcal{I}_j\right)^{1/2}$$

for $j > 1$, where $p_I^{(j)}$ is obtained as described in the succeeding text. Hypothesis H_I is rejected if $\tilde{Z}_I^{(j)} \geqslant u_{I,j}$ and retained if $\tilde{Z}_I^{(j)} \leqslant l_j$. If $l_j < \tilde{Z}_I^{(j)} \leqslant u_{I,1}$ and the study continues to a $(j+1)th$

stage with treatments selected in any way and the experimenter chooses a $(j+1)$th stage test for H_I with associated p-value $p_I^{(j+1)}$. On termination of the trial, the closure principle is applied so that H_k is rejected at familywise level-α if and only if H_I is rejected at local level-α for all I including k.

The only requirement on $p_I^{(j)}$ is that it is p-clud given X_{j-1} for all $\theta \in H_I$. In particular, we are free to ignore the 'select the best' selection rule and can re-set future sample sizes. A possible choice is to take $p_I^{(j)}$ to be the p-value from a single stage Dunnett test [16] comparing treatments in I that remain in the trial with control based on the stage j data (if no treatments in I remain in the trial, $p_I^{(j)}$ can be set equal to 1). We will use this approach in the example and simulation study reported in the succeeding text.

For the design described in [7], it is not possible to express the statistic for testing H_I as a combination of stage-wise p-values (except when $|I| = 1$), and therefore the combination test method for adding flexibility is not available.

3.2. A conditional error approach modification of the 'select the best' and 'keep all promising' designs

Flexibility may alternatively be introduced using the conditional error approach. Suppose that we have some test of a null hypothesis, H_k, that controls the type I error rate at level α. This test can be modified to allow for adaptations made at an interim analysis, following observation of data X. Let $A(X)$ denote the *conditional error* defined as the conditional probability under H_k of rejecting H_k using the original design given the observed interim data X. Following adaptation, the type I error rate is controlled providing the conditional probability of rejecting H_k under H_k does not exceed $A(X)$.

For the test of H_I, suppose that we have obtained $u_{I,1}, \ldots, u_{I,J}$ as described earlier assuming that either the 'select the best' or the 'keep all promising' selection rule is to be used. In order to maintain the local type I error rate, at each stage, the upper boundary is updated based on the observed data and treatments selected (and possibly changes to future sample sizes). Suppose that at the jth interim analysis we have observed data X_j. The conditional error, which will be denoted $A_I\left(X_j\right)$, defined as the conditional probability given X_j of rejecting H_I under H_I, can be calculated on the basis of the current critical values, and the assumption that henceforth either the 'select the best' or the 'keep all promising' rule will be used. If $A_I\left(X_j\right) = 1$, we have already rejected H_I and if $A_I\left(X_j\right) = 0$, we have already retained H_I. Suppose $A_I\left(X_j\right) \in (0, 1)$ and that the trial continues to a $(j+1)$th stage with treatments selected depending on X_j in some arbitrary fashion, and/or with future sample sizes updated on the basis of X_j. Given the treatments selected and the sample sizes chosen, we update the upper boundary such that the conditional probability of rejecting H_I under H_I remains controlled at level $A_I\left(X_j\right)$. Additional details are given in the Appendix.

As was the case when finding the original boundaries, the constraint (4) is insufficient to determine critical values for all future interim analyses (unless $j = J-1$). Analogously to the α-spending approach, we will require that the conditional probability of rejecting H_I at each future interim analysis is maintained.

We do not consider continuing with experimental treatments that have crossed the binding futility boundary, that is for which $Z_{k,j} \leqslant l_j$. Whilst it is possible to apply the conditional error approach in this scenario, we note that in this case the conditional error for H_k is 0. Consequently, H_k cannot be rejected at familywise level α and there is little to be gained from including treatment k in the remaining intersection tests. If a flexible lower boundary is required, this can be obtained by removing the binding futility boundary, that is setting $l_j = -\infty$.

3.3. Example

Consider a three-stage trial comparing three experimental treatments with control. Suppose an α-spending approach is used with $\alpha_j^* = 0.025j/3$ $(j = 1, 2, 3)$. If there is no binding futility boundary, that is, $l_1 = l_2 = -\infty$, the upper boundaries for the step-down designs are as given in columns two to four of Table I. Note that the boundary values depend only on $|I|$.

We assume a sample size of $n = 34$ per arm per stage, that is, $n_{k,j} = jn$ $(k = 0, 1, 2, 3;\ j = 1, 2, 3)$, chosen to ensure that the power of the 'keep all promising' design as defined in Section 2 is at least 0.8 given $\delta = 0.5$ and $\delta_0 = 0.2$.

Suppose that $Z_{1,1} = 2,\ Z_{2,1} = 1.1$ and $Z_{3,1} = 1$ at the first interim analysis. Furthermore, suppose that the investigators decide to drop treatment 1 from the remainder of the study due to a safety endpoint but continue with both remaining treatments.

Table I. Original and modified upper stopping boundaries given a non-binding futility boundary with $J = K = 3, Z_{1,1} = 2, Z_{2,1} = 1.1, Z_{3,1} = 1$ and treatment 1 dropped after stage 1.

'Select the best' design

	Original boundary values			Conditional type I error spent		Modified boundary values	
I	Look 1	Look 2	Look 3	Look 2	Look 3	Look 2	Look 3
$\{1,2,3\}$	2.75	2.61	2.48	0.046	0.079	2.16	2.15
$\{1,2\}$	2.62	2.50	2.38	0.063	0.102	1.86	1.86
$\{1,3\}$	2.62	2.50	2.38	0.063	0.103	1.80	1.81
$\{2,3\}$	2.62	2.50	2.38	0.007	0.021	2.66	2.55
$\{1\}$	2.39	2.29	2.20	0.108	0.158	.	.
$\{2\}$	2.39	2.29	2.20	0.016	0.037	2.29	2.20
$\{3\}$	2.39	2.29	2.20	0.012	0.031	2.29	2.20

'Keep all promising' design

	Original boundary values			Conditional type I error spent		Modified boundary values	
I	Look 1	Look 2	Look 3	Look 2	Look 3	Look 2	Look 3
$\{1,2,3\}$	2.75	2.66	2.59	0.043	0.075	2.15	2.18
$\{1,2\}$	2.62	2.53	2.45	0.061	0.100	1.87	1.87
$\{1,3\}$	2.62	2.53	2.45	0.060	0.098	1.81	1.82
$\{2,3\}$	2.62	2.53	2.45	0.011	0.029	2.53	2.45
$\{1\}$	2.39	2.29	2.20	0.108	0.158	.	.
$\{2\}$	2.39	2.29	2.20	0.016	0.037	2.29	2.20
$\{3\}$	2.39	2.29	2.20	0.013	0.031	2.29	2.20

We consider first the modification of the 'select the best' design using the combination test approach as described in Section 3.1. To use this approach, we must choose our second stage tests with associated p-values $p_I^{(2)}$ for each $I \subseteq \{1,2,3\}$ such that I includes at least one of the continuing treatments, that is, $I \cap \{2,3\} \neq \varnothing$. For $I = \{1,2,3\}$ and $I = \{2,3\}$, we choose a Dunnett test comparing treatments 2 and 3 with control. For $I = \{j\}$ and $I = \{1,j\}$ $(j = 2,3)$, a one-sided z-test comparing treatment j with control is used. Additionally, $p_{\{1\}}^{(2)}$ is set to 1.

Suppose now that the second stage results are $Z_{2,2} = 2.55$ and $Z_{3,2} = 1$, or, equivalently; $p_{\{2,3\}}^{(2)} = 0.012$, $p_{\{2\}}^{(2)} = 0.006$ and $p_{\{3\}}^{(2)} = 0.34$. In order to apply the closed test, we find

$$\begin{aligned}
\tilde{Z}_{\{1,2,3\}}^{(2)} &= \left\{2 + \Phi^{-1}(1-0.012)\right\}/\sqrt{2} &&= 3.01\\
\tilde{Z}_{\{1,2\}}^{(2)} &= \left\{2 + \Phi^{-1}(1-0.006)\right\}/\sqrt{2} &&= 3.19\\
\tilde{Z}_{\{1,3\}}^{(2)} &= \left\{2 + \Phi^{-1}(1-0.34)\right\}/\sqrt{2} &&= 1.71\\
\tilde{Z}_{\{2,3\}}^{(2)} &= \left\{1.1 + \Phi^{-1}(1-0.012)\right\}/\sqrt{2} &&= 2.37\\
\tilde{Z}_{\{1\}}^{(2)} &= \left\{2 + \Phi^{-1}(1-1)\right\}/\sqrt{2} &&= -\infty\\
\tilde{Z}_{\{2\}}^{(2)} &= \left\{1.1 + \Phi^{-1}(1-0.006)\right\}/\sqrt{2} &&= 2.55\\
\tilde{Z}_{\{3\}}^{(2)} &= \left\{1 + \Phi^{-1}(1-0.34)\right\}/\sqrt{2} &&= 1.00.
\end{aligned}$$

By comparing these values with the boundaries given in Table I, we see that $H_{\{1,2,3\}}$, $H_{\{1,2\}}$ and $H_{\{2\}}$ can all be rejected at local level 0.025. However, there is not yet any k such that H_I has been rejected at local level 0.025 for all $I \subseteq \{1,2,3\}$ with $k \in I$. At this point, third stage tests could be defined and the study continued.

We consider next the modified designs using the conditional error approach. Suppose again we have $Z_{1,1} = 2, Z_{2,1} = 1.1$ and $Z_{3,1} = 1$ and select treatments 2 and 3. We must update the boundaries for all $I \subseteq \{1,2,3\}$ such that $I \cap \{2,3\} \neq \varnothing$ as described in Section 3.2. We first find the conditional probability given Z_1, of rejecting H_I at or before the second and third interim analyses assuming that either the 'select the best' or 'keep all promising' selection rule is to be used. These values are given

in the fifth and sixth columns of Table I. We then find new critical values $u_{2,I}$ and $u_{3,I}$ to give these probabilities if we continue with treatments 2 and 3. These values are given in the seventh and eighth columns in Table I. Because treatment 1 has been dropped from the trial, we do not test $H_{\{1\}}$ at stage two or three. Note that when $I = I \cap \{2, 3\}$, the boundaries are unchanged.

Suppose again that, at the second stage, we have observed $Z_{2,2} = 2.55$ and $Z_{3,2} = 1$. With the modified 'select the best' design, we have $Z_{2,2} < u_{\{2,3\},2}$ and $Z_{3,2} < u_{\{2,3\},2}$, so that neither H_2 nor H_3 can be rejected at this stage, whereas with the modified 'keep all promising' design, we have $Z_{2,2} > \max\left(u_{\{1,2,3\},2}, u_{\{1,2\},2}, u_{\{2,3\},2}, u_{\{2\},2}\right)$ so that H_2 can be rejected at this stage at familywise level of 0.025.

The properties of the test methods are explored further in the simulation study reported in the next section.

3.4. Simulation study

A flexible selection rule makes it difficult to compare the performance of the competing designs. Because the relative performance of the treatments with respect to the primary efficacy endpoint is of high importance, the following simple selection rule, which has been used in previous comparison studies [17, 28], will be used here. Assuming a non-binding futility boundary, for some suitable ε, treatment k remains in the trial if it has not previously been dropped and $Z_{k,j} \geqslant \max\{Z_{k',j}\} - \varepsilon$, the maximum being taken over all treatments still in the trial at stage j. The original designs [3] and [7] with non-binding futility boundaries are special cases with $\varepsilon = 0$ and $\varepsilon = \infty$, respectively. The methods of Sections 3.1 and 3.2 have been compared based on 100,000 simulations of the trial design specified in Section 3.3 using this treatment selection rule for a range of ε values. In the case of the 'keep all promising' design, a non-binding futility boundary means that no selection is taking place. We do not consider using a binding futility boundary as this would require substantially more computation and our main goal is to demonstrate the increase in flexibility that can be achieved by the new methods.

Also included in the comparison is the original step-down 'keep all promising' design with no adjustment of the upper boundaries on account of treatments being dropped. Because we cannot use the unadjusted step-down 'select the best' design for $\varepsilon > 0$ without error rate inflation, we do not consider it here.

The flexible modifications in Section 3 focus on protecting the familywise error rate with no explicit consideration of the effect on power. Here, power is defined as the probability $\theta = (\delta, \delta_0, \ldots, \delta_0)$ of rejecting H_1 and identifying treatment 1 as the most promising as defined in Section 3.

The results are given in Figure 1. As one would expect, the original step-down design, which does not reduce the upper boundaries on account of dropped treatments, looses power relative to the other methods at low values of ε. At high values of ε, the p-value combination test performs poorly relative to the other methods. The explanation for this is that if there are several experimental treatments in the later stages, then responses are weighted unequally in the final test statistics and this is an inefficient use of the data. The conditional error approach has strikingly good performance across all values of ε. It should be noted, however, that high values of ε are unlikely to be used in practice, and that the p-value combination test has the virtue of being simple to implement.

3.5. Sample size re-calculation

As a further demonstration of the methodology, we consider a four-arm two-stage design with futility boundary $l_1 = 0$ and upper boundaries such that $\alpha_1^* = 0.025/3$ and $\alpha_2^* = 0.025$, as presented in Table II. A sample size of 27 patients per arm per stage ensures that

$$P_{\theta=(0.5,0.5,0.5)}(\text{reject at least one } H_k) \geqslant 0.8,$$

when using the 'keep all promising' selection rule. Instead of following the pre-defined treatment selection rules, we will assume that each experimental treatment is declared unsafe after the first stage independently with probability p. In addition, only those treatments with a standardised test statistic that is within $\varepsilon = 0.5$ of the 'best' will be taken forward. Furthermore, provided that at least one experimental treatment continues into the second stage, the potential 144 second-stage patients will be randomised in equal proportions to the continuing treatment arms. In other words, the trial will either have total sample size 108 if it stops after at the interim analysis or a total sample size 216 if any treatments continue to the second stage.

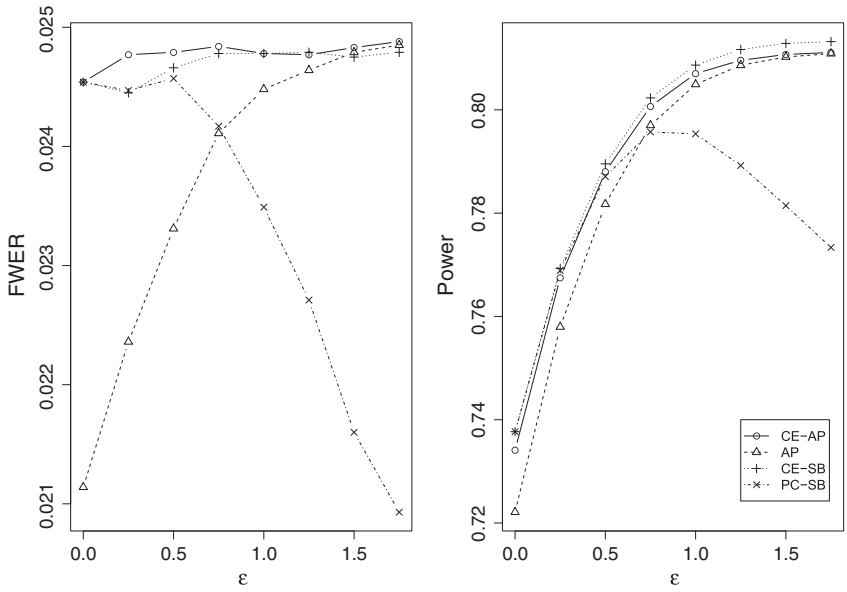

Figure 1. Empirical familywise error rate and power based on 100,000 simulations of the three-arm three-stage trial design described in Section 3.3, with selection rule described in the text, for a range of ε values using the conditional error modified 'keep all promising' design (CE-AP), the original 'keep all promising' design (AP), the conditional error modified 'select the best' design (CE-SB) and the p-value combination modified 'select the best' design (PC-SB).

Table II. Upper boundaries for a four-arm two-stage design with zero futility boundary under the 'select the best' and 'keep all promising' treatment selection rules.

	'Select the best'		'Keep all promising'	
$\|I\|$	$u_{I,1}$	$u_{I,2}$	$u_{I,1}$	$u_{I,2}$
3	2.75	2.37	2.75	2.43
2	2.62	2.26	2.62	2.30
1	2.39	2.04	2.39	2.04

Table III. Probability of rejecting at least one null hypothesis in a four-arm two-stage design with sample size re-allocation in the second stage.

	P_θ(reject at least one H_k)					
	$\theta = (0,0,0)$			$\theta = (0.5,0.5,0.5)$		
P(Unsafe)	PC-SB	CE-SB	CE-AP	PC-SB	CE-SB	CE-AP
0	0.0250	0.0250	0.0254	0.990	0.990	0.993
0.25	0.0210	0.0209	0.0216	0.958	0.958	0.960
0.5	0.0171	0.0172	0.0174	0.901	0.901	0.902
0.75	0.0129	0.0130	0.0133	0.815	0.815	0.815
1	0.0082	0.0082	0.0082	0.693	0.693	0.693

Methods: p-value combination modification of the 'select the best' rule (PC-SB), the conditional error modification of the 'select the best' rule (CE-SB) and the conditional error modification of the 'keep all promising' rule (CE-AP).

The probability of rejecting at least one null hypothesis, under the null configuration, and assuming all treatments are effective, is shown in Table III for various values of p. In this case, all three methods perform equally well in terms of power, with the familywise error rate controlled at the original level $\alpha = 0.025$.

4. Concluding remarks

In this work, we have shown how flexibility can be added to pre-planned adaptive multi-arm multi-stage clinical studies. We have focused on adding flexibility to the treatment selection rule, as well as updating sample sizes, but the same methodology could be used to, for example, add further interim analyses. In addition, the new methods would work equally well if applied to the problem of subgroup selection rather than treatment selection.

In Section 3, we made the assumption that independent cohorts of patients are recruited in each stage, as this ensures the joint distribution (2) holds. If X_{j-1} were to contain short-term endpoint information that would be predictive for responses in stages $j, \ldots, J$ then a naive application of the new methods could lead to an inflation of the familywise error rate [29]. Special methods are required to handle this situation, see [30, 31] and [32].

Both our example and simulation study demonstrate the potential gain in power that can be achieved by applying the conditional error principle. The methodology will be most useful when reacting to unforeseen events, for example, a safety issue on a particular treatment arm, and it is important that the original pre-planned adaptive design is well chosen. We have not attempted to find optimal designs in this paper, and the power criterion we have used for the comparison in Section 3.4 is one of many sensible criteria that could be applied in this setting.

It should be acknowledged that whilst the step-down procedures described in Section 3 are clearly more powerful than their original counterparts due to their reduced upper boundaries, this comes at the cost of increased difficulties in constructing useful confidence intervals [33] and therefore does not constitute a uniform improvement. See, for example, [34] or [35] for a discussion of this phenomenon in the context of a fixed sample design.

Finally, we have assumed that observations are normally distributed with known variance. Asymptotic results can be used for other endpoints in a similar fashion to the discussion in [36]. An implementation of the new procedures will be included in an upcoming version of the R package MAMS [37].

Appendix

In this appendix, we give details of the computation of the conditional error for H_I, $A_I(X_j)$.

Suppose that at the jth interim analysis we have observed data X_j. Let $T_{j'}$ denote the set of treatments selected to continue to stage j' for $j' = j, \ldots, J$, where T_j is chosen unrestrictedly based on the data from the previous stage, and $T_{j'}$ for $j' = j+1, \ldots, J$ is as given by the pre-specified selection rule. $A_I(X_j)$ is defined as the conditional probability of rejecting H_I given X_j. This is given by

$$A_I(X_j) = P_0\left(\bigcup_{k\in I}\bigcup_{j'\in\{j,\ldots,J\}} k \in T_{j'} \text{ and } Z_{k,j'} \geqslant u_{I,j'}^{(j-1)} \mid X_j, \mathcal{D}_I^{(j-1)}\right), \tag{A.1}$$

which depends on the design $\mathcal{D}_I^{(j-1)}$, comprising the pre-planned selection rule, critical values $l_j, \ldots, l_{J-1}$ and $u_{I,j}^{(j-1)}, \ldots, u_{I,J}^{(j-1)}$, and sample sizes, $n_j^{(j-1)}, \ldots, n_J^{(j-1)}$, where the superscripts indicate that the upper critical values and sample sizes may have been updated following looks, $1, \ldots, j-1$.

The conditional probability of rejecting H_I at or before interim analysis j' for $j' = j, \ldots, J$ is similarly given by

$$\alpha_{j',I}^{(j)} = P_0\left(\bigcup_{k\in I}\bigcup_{j''\in\{j,\ldots,j'\}} k \in T_{j''} \text{ and } Z_{k,j''} \geqslant u_{I,j''}^{(j-1)} \mid X_j, \mathcal{D}_I^{(j-1)}.\right) \tag{A.2}$$

The fact that independent cohorts of patients are to be recruited in stages $j+1,\ldots,J$ means that the conditional distribution of $Z_{j+1},\ldots,Z_J$ depends on X_j only through the observed value of z_j and is therefore multivariate normal with, by a standard calculation (see, e.g., Section 2.5 of [38]),

$$\begin{aligned} E\left(Z_{k,j'}\right) &= \theta_k\left(\mathcal{I}_{k,j'}-\mathcal{I}_{k,j}\right)/\mathcal{I}_{k,j'}^{1/2}+z_{k,j}\left(\mathcal{I}_{k,j}/\mathcal{I}_{k,j'}\right)^{1/2},\\ \operatorname{var}\left(Z_{k,j'}\right) &= \left(\mathcal{I}_{k,j'}-\mathcal{I}_{k,j}\right)/\mathcal{I}_{k,j'},\\ \operatorname{cov}\left(Z_{k,j'},Z_{k,j''}\right) &= \left(\mathcal{I}_{k,\min\{j',j''\}}-\mathcal{I}_{k,j}\right)/\left(\mathcal{I}_{k,j'}\mathcal{I}_{k,j''}\right)^{1/2},\\ \operatorname{cov}\left(Z_{k',j'},Z_{k'',j''}\right) &= \mathcal{I}_{k,j'}^{1/2}\mathcal{I}_{k',j''}^{1/2}\left(n_{0,\min\{j',j''\}}-n_{0,j}\right)/\left(n_{0,j'}n_{0,j''}\right) \qquad (k'\neq k). \end{aligned} \tag{A.3}$$

If the trial does not stop at the interim analysis j, T_{j+1}, the set of treatments to continue to the next stage, along with future sample sizes, may be chosen in some arbitrary fashion based on X_j. Given these choices, the boundary values are updated by finding $u_{j+1}^{(j)},\ldots,u_J^{(j)}$ such that

$$P_0\left(\bigcup_{k\in I}\bigcup_{j''\in\{j+1,\ldots,j'\}} k\in T_{j''} \text{ and } Z_{k,j''}\geqslant u_{I,j''}^{(j)} \mid X_j,\mathcal{D}_I^{(j)}\right)=\alpha_{j',I}^{(j)} \tag{A.4}$$

for $j'=j+1,\ldots,J$ where $\mathcal{D}_I^{(j)}$, comprises the critical values $l_{j+1},\ldots,l_{J-1}$ and $u_{I,j+1}^{(j)},\ldots,u_{I,J}^{(j)}$, the modified sample sizes and the selection rule now updated to allow for the choice of T_{j+1} but assuming that the pre-specified rule will be followed at all subsequent stages.

Upon observation of X_j, one knows $\arg\max\{Z_{k,j}\}$ and therefore finding $A_I(X_j)$, assuming use of the 'select the best' rule with a non-binding futility boundary, requires evaluation of a single tail area for the multivariate normal distribution defined by (A.3), for example, using the `pmvnorm` function from the R package `mvtnorm` [39]. However, to find the probability on the left-hand side of (A.4) requires $|I\cap T_{j+1}|$ calls to `pmvnorm` where one must first find, for each $k\in I\cap T_{j+1}$, the conditional joint distribution of $\left(Z_{k,j'},Z_{k,j+1}-Z_{k',j+1}\right)$ $\left(j'\in\{j+1,\ldots,J\};k'\in I\cap T_{j+1}\setminus\{k\}\right)$ via an appropriate transformation of (A.3). Using the 'keep all promising' rule, in the case of a non-binding futility boundary, finding $1-A_I(X_j)$ requires a single call to `pmvnorm`. If a binding futility boundary is used, up to $|J-j|^{|I\cap T_j|}$ calls to `pmvnorm` are necessary, corresponding to all possible combinations of treatments crossing the futility boundary.

Acknowledgements

This report is independent research arising from Dr Jaki's Career Development Fellowship (NIHR-CDF-2010-03-32) supported by the National Institute for Health Research and the MRC grant MR/J004979/1. The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health.

The authors are grateful to an anonymous reviewer and the associate editor for helpful comments to improve clarity of the manuscript.

References

1. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman and Hall: Boca Raton, FL, 2000.
2. Thall PF, Simon R, Ellenberg SS. Two-stage selection and testing designs for comparitive clinical trials. *Biometrika* 1988; **75**:303–310.
3. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; **22**:689–703.
4. Stallard N, Friede T. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine* 2008; **27**:6209–6227.
5. Follmann DA, Proschan MA, Geller NL. Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics* 1994; **50**:325–336.
6. Chen YHJ, DeMets DL, Lan KKG. Some drop-the-loser designs for monitoring multiple doses. *Statistics in Medicine* 2010; **29**:1793–1807.
7. Magirr D, Jaki T, Whitehead J. A generalised dunnett test for multi-arm, multi-stage clinical studies with treatment selection. *Biometrika* 2012; **99**:494–501.
8. Brannath W, Koenig F, Bauer P. Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics* 2007; **6**:205–216.

9. Bauer P. Multistage testing with adaptive designs (with discussion). *Biometrie und Informatik in Medizin und Biologie* 1989; **20**:130–148.
10. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; **50**:1029–1041. Correction: *Biometrics* 1996; 52:380.
11. Hommel G. Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal* 2001; **43**(5):581–589.
12. Müller HH, Schäfer H. A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine* 2004; **23**:2497–2508.
13. Tsiatis AA, Mehta C. On the inefficiency of adaptive design for monitoring clinical trials. *Biometrika* 2003; **90**:367–378.
14. Jennison C, Turnbull BW. Adaptive and nonadaptive group sequential tests. *Biometrika* 2006; **93**:1–21.
15. Koenig F, Brannath W, Bretz F, Posch M. Adaptive dunnett tests for treatment selection. *Statistics in Medicine* 2008; **27**:1612–1625.
16. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association* 1955; **50**:1096–1121.
17. Friede T, Stallard N. A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; **50**:767–781.
18. Di Scala L, Glimm E. Time-to-event analysis with treatment arm selection at interim. *Statistics in Medicine* 2011; **30**(26):3067–3081. DOI: 10.1002/sim.4342. http://dx.doi.org/10.1002/sim.4342 [Accessed on 29 April 2014].
19. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Statistics in Medicine* 2012; **31**(30):4309–4320. DOI: 10.1002/sim.5541. http://dx.doi.org/10.1002/sim.5541 [Accessed on 29 April 2014].
20. Slud E, Wei LJ. Two-sample repeated significance tests based on the modified wilcoxon statistic. *Journal of the American Statistical Association* 1982; **77**:862–868.
21. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
22. Jennison C, Turnbull BW. Confirmatory seamless phase II/III clinical trials with hypothesis selection at interim: opportunities and limitations. *Biometrical Journal* 2006; **48**:650–655.
23. Dunnett CW. Selection of the best treatment in comparison to a control with an application to a medical trial. In *Design of experiments: Ranking and selection*, Santer T, Tamhane A (eds). Marcel Dekker: New York, 1984; 47–66.
24. Marcus R, Peritz E, Gabriel KR. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; **63**:655–660.
25. Holm S. A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics* 1979; **6**:65–70.
26. Lehmacher W, Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; **55**(4):1286–1290 (English). http://www.jstor.org/stable/2533757 [Accessed on 29 April 2014].
27. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical Association* 2002; **97**:236–244.
28. Kelly PJ, Stallard N, Todd S. An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Biopharmaceutical Statistics* 2005; **15**:641–658.
29. Bauer P, Posch M. Letter to the editor. *Statistics in Medicine* 2004; **23**:1333–1334.
30. Stallard N. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine* 2010; **29**(9):959–971.
31. Jenkins M, Stone A, Jennison C. An adaptive seamless phase II/III design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical Statistics* 2011; **10**:347–356.
32. Irle S, Schäfer H. Interim design modifications in time-to-event studies. *Journal of the American Statistical Association* 2012; **107**:341–348.
33. Magirr D, Jaki T, Posch M, Klinglmueller F. Simultaneous confidence intervals that are compatible with closed testing in adaptive designs. *Biometrika* 2013; **100**(4):985–996.
34. Hayter AJ, Hsu JC. On the relationship between stepwise decision procedures and confidence sets. *Journal of the American Statistical Association* 1994; **89**(425):128–136. http://www.jstor.org/stable/2291208 [Accessed on 29 April 2014].
35. Strassburger K, Bretz F. Compatible simultaneous lower confidence bounds for the Holm procedure and other Bonferroni-based closed tests. *Statistics in Medicine* 2008; **27**:4914–4927.
36. Whitehead J, Jaki T. One- and two-stage design proposals for a phase ii trial comparing three active treatments with a control using an ordered categorical endpoint. *Statistics in Medicine* 2009; **28**:828–847.
37. Jaki T, Magirr D. MAMS: designing multi-arm multi-stage studies, 2012. http://CRAN.R-project.org/package=MAMS, R package version 0.2, [Accessed on 29 April 2014].
38. Anderson TW. *An Introduction to Multivariate Statistical Analysis*. Wiley: New York, 1984.
39. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: multivariate normal and t distributions, 2010. http://CRAN.R-project.org/package=mvtnorm, R package version 0.9–99991, [Accessed on 29 April 2014].