A multi-stage drop-the-losers design for multi-arm clinical trials

James Wason[1], Nigel Stallard[2], Jack Bowden[1], and Christopher Jennison[3]

[1]MRC Biostatistics Unit, Cambridge

[2]Warwick Medical School, University of Warwick, Coventry

[3]Department of Mathematical Sciences, University of Bath, Bath

Abstract

Multi-arm multi-stage (MAMS) trials can improve the efficiency of the drug development process when multiple new treatments are available for testing. A group-sequential approach can be used in order to design MAMS trials, using an extension to the Dunnett multiple testing procedure. The actual sample size used in such a trial is a random variable which has high variability. This can cause problems when applying for funding as the cost will generally also be highly variable. This motivates a type of design which provides the efficiency advantages of a group-sequential MAMS design, but has a fixed sample size. One such design is the two-stage drop-the-losers design, in which a number of experimental treatments, and a control treatment, are assessed at a pre-scheduled interim analysis. The best performing experimental treatment and the control treatment then continue to a second stage. In this paper we discuss extending this design to have more than two stages, which is shown to considerably reduce the sample size required. We also compare the resulting sample size requirements to the sample size distribution of

analogous group-sequential MAMS designs. The sample size required for a multi-stage drop-the-losers design is usually higher than, but close to, the median sample size of a group-sequential MAMS trial. In many practical scenarios, the disadvantage of a slight loss in average efficiency would be overcome by the huge advantage of a fixed sample size. We assess the impact of delay between recruitment and assessment as well as unknown variance on the drop-the-losers designs.

Keywords: clinical trial design; delay; group-sequential designs; interim analysis; multi-arm multi-stage designs; multiple-testing.

1 Introduction

Testing multiple experimental treatments against a control treatment in the same trial provides several advantages over doing so in separate trials. The main advantage is a reduced sample size due to a shared control group being used instead of a separate control group for each treatment. Other advantages include that direct comparisons can be made between experimental treatments and that it is administratively easier to apply for and run one multi-arm clinical trial compared to several traditional trials [1]. Multi-arm multi-stage (MAMS) clinical trials include interim analyses so that experimental treatments can be dropped if they are ineffective; also, if desired, the trial can be designed so that it allows early stopping for efficacy if an effective experimental treatment is found. Two current MAMS trials that are ongoing are the MRC STAMPEDE trial [1], and the TAILoR trial (the design of which is discussed in Magirr, Jaki and Whitehead [2]).

Magirr et al. [2] extend the Dunnett multiple-testing procedure [3] to multiple stages, which we refer to as the group-sequential MAMS design. In this design, futility and efficacy boundaries are pre-specified for each stage of the trial. At each interim analysis, statistics comparing each experimental treatment to the control treatment are calculated and compared to these boundaries. If a statistic is below the futility boundary, then the respective experimental arm is dropped from the trial. If a statistic is above the

efficacy threshold, the trial is stopped with that experimental treatment recommended. Boundaries would generally be required to control the frequentist operating characteristics of the trial. Since there are infinitely many boundaries that do so, a specific boundary can be chosen to minimise the expected number of recruited patients at some treatment effect [4], or by using some boundary function such those of Pocock [5], O'Brien and Fleming [6], or Whitehead and Stratton[7].

The group-sequential MAMS design is efficient in terms of the expected sample size recruited, but has the practical problem that the sample size used is a random variable. This makes planning a trial more difficult than when the sample size is known in advance. An academic investigator applying for funding to conduct a MAMS trial will find that traditional funding mechanisms lack the required flexibility to account for a random sample size [8]. Generally, they would have to apply for the maximum amount that could potentially be used, with the consequence that such trials appear highly expensive to fund. There are also several other logistical issues to consider, such as employing trial staff to work on a trial with a random duration.

An alternative type of MAMS trial is one in which a fixed number of treatments is dropped at each interim analysis. Stallard and Friede [9] propose a group-sequential design where a set number of treatments is dropped at each interim analysis, and the trial stops if the best performing test statistic is above a pre-defined efficacy threshold or below a pre-defined futility threshold. The stopping boundaries are set assuming the maximum test statistic is the sum of the maximum independent increments in the test statistic at each stage, which is generally not true and leads to conservative operating characteristics. A special case of Stallard and Friede's design is the well-studied two-stage drop-the-losers design (see, for example, [10, 11]), in which one interim is conducted, and only the top performing experimental treatment and a control treatment proceed to the second stage. In Thall, Simon and Ellenberg [10], the chosen experimental treatment must be sufficiently effective to continue to the second stage. More flexible two-stage designs have been proposed by several authors, including Bretz et al. [12] and Schmidli

et al. [13]. These designs used closed testing procedures and/or combination tests to control the probability of making a type-I error whilst allowing many modifications to be made at the interim. In the case of multiple experimental arms, there is more scope for improved efficiency by including additional interim analyses, at least for group-sequential MAMS designs [2, 4].

In this paper, we extend the two-stage drop-the-losers design to more than two stages and derive formulae for the frequentist operating characteristics of the design. The resulting design has the advantage of a fixed sample size by maintaining a pre-specified schedule of when treatments are dropped. That is, at each interim analysis, a fixed number of treatments are dropped. Note that this could be thought of as sub-dividing the first stage of a two-stage drop-the-losers trial to allow multiple stages of selection. We show that when there are several treatments, allowing an additional stage of selection noticeably decreases the sample size required for a given power, compared to the two-stage design. We also compare the multi-stage drop-the-losers design to the Dunnett-type MAMS design.

2 Notation

We assume that the trial is to have J stages, i.e. $J-1$ interim analyses and a final analysis, and starts with K experimental treatments and a control treatment. Let $k \in \{0, 1, \ldots, K\}$ index the treatment ($k = 0$ represents the control treatment). Cumulative up to the end of the jth stage of the trial, a total of n_j patients have been recruited to each remaining treatment. The values of n_j are pre-specified, and in particular do not depend on the results of the trial. The ith patient allocated to treatment k has a treatment outcome, X_{ki}, distributed as $N(\mu_k, \sigma_k^2)$. The values of σ_k^2 are assumed to be known.

For $k \in \{1, \ldots, K\}$, define $\delta_k = \mu_k - \mu_0$. The null hypotheses to be tested are $H_0^{(k)} : \delta_k \leq 0$. The global null hypothesis, H_G, is defined as $H_G : \delta_1 = \delta_2 = \ldots = \delta_K = 0$.

The known variance test-statistic for treatment k at stage j is:

$$Z_j^{(k)} = \left(\frac{\sum_{i=1}^{n_j} X_{ki}}{n_j} - \frac{\sum_{i=1}^{n_j} X_{0i}}{n_j} \right) \sqrt{\frac{n_j}{\sigma_k^2 + \sigma_0^2}}, \tag{1}$$

which has marginal distribution $N(\delta_k \sqrt{\frac{n_j}{\sigma_k^2+\sigma_0^2}}, 1)$.

The covariance between different test statistics can be shown to be:

$$\mathrm{Cov}(Z_j^{(k)}, Z_l^{(m)}) = \begin{cases} \sqrt{\frac{\min(n_j,n_l)}{\max(n_j,n_l)}} & \text{if } k = m; \\ \sqrt{\frac{\sigma_0^4}{(\sigma_k^2+\sigma_0^2)(\sigma_m^2+\sigma_0^2)}} \sqrt{\frac{\min(n_j,n_l)}{\max(n_j,n_l)}} & \text{if } k \neq m. \end{cases} \tag{2}$$

At each stage, a fixed and pre-determined number of experimental treatments are dropped. For J stages, the design is denoted as a '$K : n^{(2)} : \ldots : n^{(J-1)} : n^{(J)}$' design, where $K > n^{(2)} > \ldots > n^{(J-1)} > n^{(J)}$. Thus, at least one experimental treatment is dropped at each analysis. Although $n^{(j)}$ can in principle be more than one, we henceforth only consider designs with $n^{(J)} = 1$, similar to a two-stage drop-the-losers design. The experimental treatments to be dropped are determined by ranking the $Z_j^{(k)}$ statistics of the remaining experimental treatments in order of magnitude, and removing the smallest (least promising) as pre-specified by the design. The control treatment always remains in the trial. At the final stage, one experimental treatment remains, and if its final test statistic is above a threshold, c, that treatment is recommended, and the respective null hypothesis rejected.

It is desirable that the design is chosen in order to control the family-wise type-I error rate (FWER). The FWER is the probability of rejecting at least one true null hypothesis, and strong control of the FWER at level α means that the FWER is $\leq \alpha$ for any configuration of true and false null hypotheses (i.e. for any values of δ_k $k = 1, \ldots, K$). In section 3 we demonstrate how to control the FWER at $\delta_1 = \delta_2 = \ldots = \delta_K = 0$, and show in section 4 that this strongly controls the FWER. As well as the FWER, it is also desirable to control the probability of selecting a genuinely good treatment, were it to exist. To formalise the latter quantity, we use the least favourable configuration (LFC)

of Dunnett [3]. This is the probability of recommending treatment 1 when $\delta_1 = \delta^{(1)}$ and $\delta_2 = \delta_3 = \ldots = \delta_K = \delta^{(0)}$, where $\delta^{(1)}$ is a pre-specified clinically relevant effect, and $\delta^{(0)}$ is some threshold below which a treatment is considered uninteresting. The configuration is called least favourable as it minimises the probability of recommending the effective treatment amongst all configurations such that $\delta_1 \geq \delta^{(1)}$, and none of treatments $2, \ldots, K$ have a treatment effect above $\delta^{(0)}$ [10].

3 Analytic operating characteristics

In this section we provide analytical formulae for the probability of a particular treatment being recommended under a general vector of treatment effects. We also provide formulae for the probability of rejecting any null hypothesis when H_G is true, and the probability to select the best treatment under the LFC. Although the formulae extend naturally to more than 3 stages, the expressions quickly become unwieldy. We thus concentrate on the 3-stage case, where K experimental treatments are included in the first stage, $L < K$ in the second stage, and 1 in the third stage. This is denoted as the $K : L : 1$ design.

3.1 Probability of a specific treatment being recommended

For subsequent development, it is useful to define a ranking of the experimental treatments in terms of how successful they are. We introduce random variables $\psi = (\psi_1, \ldots, \psi_K)$, where ψ_k is the ranking of treatment k. Each of the ψs takes a unique integer value between 1 and K with the following properties:

1. The treatment which reaches the final analysis has rank 1;

2. the treatment which is dropped at the first analysis with the lowest test statistic is given rank, K;

3. if treatment k_1 reaches a later stage than treatment k_2, then $\psi_{k_1} < \psi_{k_2}$, i.e. treatment k_1 has a better ranking;

4. if treatments k_1 and k_2 are dropped at the same stage, and k_1 has a higher test statistic at that stage, then $\psi_{k_1} < \psi_{k_2}$.

For instance, for a three-stage 4:2:1 design where treatment 3 reaches the final stage, treatment 2 is dropped at the second analysis, treatments 1 and 4 are dropped at the first analysis, and treatment 1 has the lowest test statistic at the first analysis, the realised value of ψ is $(4, 2, 1, 3)$.

In terms of $(\psi_1, \ldots, \psi_K)$, the probability of recommending treatment k, i.e. rejecting $H_0^{(k)}$, given mean vector $\delta = (\delta_1, \delta_2, \ldots, \delta_K)$ can be written as (for $J = 3$):

$$\mathbb{P}(\text{Reject } H_0^{(k)}|\delta) = \mathbb{P}(\psi_k = 1, Z_3^{(k)} > c|\delta); \tag{3}$$

that is the kth null hypothesis is rejected if and only if the kth experimental treatment reaches the final stage and its final test statistic is above the critical value c.

WLOG we consider the probability of recommending treatment 1. Let Ψ be the set of all possible realisations of ψ. Then equation (3) becomes:

$$\sum_{\psi \in \{\Psi : \psi_1 = 1\}} \mathbb{P}(Z_3^{(1)} > c, \psi|\delta). \tag{4}$$

We next show how each of the summands in (4) can be written as the tail probability of a multivariate normal distribution. The distribution of $Z = (Z_1^{(1)}, Z_2^{(1)}, \ldots, Z_3^{(K)})$ is multivariate normal with mean $m(\delta)$ and covariance Σ defined in section 2. WLOG, by relabelling treatments and shuffling the entries of δ, we consider the probability $\mathbb{P}(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_3^{(1)} > c|\delta)$. The event $(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_J^{(1)} > c)$ is equivalent to (Treatment $L + 1$ beats treatments $L + 2$, $L + 2$ beats $L + 3$, …, $K - 1$ beats K in the first stage; treatments $1, \ldots, L$ all beat treatment $L + 1$ in the first stage; treatment 1 beats treatment 2, 2 beats 3, …, $L-1$ beats L in the second stage; $Z_3^{(1)} > c$). In total there are $K + L - 1$ conditions. More generally, for a $K : n^{(2)} : n^{(3)} : \ldots : n^{J-1} : 1$ design, the number of conditions required at stage j will be $n^{(j)} - 1$. This is because all treatments that make it through to the next stage must beat the top ranked treatment

that is dropped, and each dropped treatment must beat the next ranked treatment up until the lowest ranked treatment. In addition, there will be 1 final condition representing whether the top ranked treatment beats the critical value in the final stage. Thus, the number of conditions is $K - 1 + \sum_{j=2}^{J-1}(n^{(j)} - 1) + 1$.

The probability of this ordering can be written in terms of differences between entries of Z. For example, the probability of treatment $L+1$ beating treatment $L+2$ in the first stage is the same as the probability that $\mathbb{P}(Z_1^{(L+1)} - Z_1^{(L+2)} > 0)$. Note that $Z_1^{(L+1)} - Z_1^{(L+2)}$ is an affine transformation of Z, a multivariate normal random variable, and so is normal. In this way we can write each term in (4) as a multivariate normal tail probability. The mean and covariance of this multivariate normal random variable are $Am(\delta)$ and $A\Sigma A^T$ respectively, where the matrix A represents the affine transformation required.

A general specification of A is difficult to write concisely. For each pairing of one treatment beating a different treatment, there is a row in A with a 1 in the column corresponding to the relevant stage test statistic of the winning treatment, and a -1 in the column corresponding to the relevant stage test statistic of the beaten treatment. There is then an additional row with a 1 in the column corresponding to the last stage test statistic of the last remaining treatment. In total there are $K + L - 1$ rows. For example, in the $4:2:1$ design, A for the event $(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_3^{(1)} > c)$ is the following 5×12 matrix:

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5}$$

For a more general $K : n^{(2)} : n^{(3)} : \ : n^{J-1} : 1$, A will be a matrix with $(K + \sum_{j=2}^{J-1}(n^{(j)} - 1))$ rows and JK columns.

Let the vector AZ be denoted as $(a_1, \ldots, a_{K+L-1})$ which, as stated before, has a multivariate normal distribution with mean $Am(\delta)$ and covariance $A\Sigma A^T$. Then the probability $\mathbb{P}(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_3^{(1)} > c)$ is $\mathbb{P}(a_1 > 0, \ldots, a_{K+L-2} > 0, a_{K+L-1} > c)$. This probability can be efficiently evaluated using the method of Genz and Bretz [14].

Each term in (4) can be calculated efficiently in this way, with the matrix A depending on the term. However it is computationally more efficient to derive A once and permute the entries of δ (which will change $m(\delta)$, but not A or $A\Sigma A^T$) for each separate term, especially as the number of experimental treatments or stages increases.

3.2 Probability of recommending any treatment under the global null hypothesis

We now consider evaluating the probability of recommending any treatment when the global null hypothesis, H_G, is true and all treatment response variances are equal. Under H_G, $m(\delta)$ will be 0. This considerably simplifies the evaluation compared to section 3.1 as now each term in equation (4) is equal. The probability of recommending any treatment under the global null hypothesis will be:

$$K!\, \mathbb{P}(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_J^{(1)} > c | \delta = 0). \tag{6}$$

Equation (6) can be calculated as $K!$ times the tail probability of the multivariate normal random variable, evaluated at $m = 0$, in Section 3.1.

3.3 Probability of recommending a specific treatment under the least favourable configuration

We assume the trial is to be powered to recommend treatment 1 at the LFC, which is when the mean treatment effect of treatment 1 is $\delta^{(1)}$ and all other treatments have mean effect $\delta^{(0)}$. Since all treatments except treatment 1 have the same effect, each term in equation (4) will have the same probability. Since there are $(K-1)!$ possible rankings of

treatments $2, \ldots, K$, the probability of recommending treatment 1 will be:

$$(K-1)!\, \mathbb{P}(\psi_1 = 1, \psi_2 = 2, \ldots, \psi_K = K, Z_J^{(1)} > c | \delta_1 = \delta^{(1)}, \delta_2 = \delta^{(0)}, \ldots, \delta_K = \delta^{(0)}). \quad (7)$$

Equation (7) can be calculated as $(K-1)!$ times the tail probability of the multivariate normal random variable, evaluated at $m(\delta^{(1)}, \delta^{(0)}, \ldots, \delta^{(0)})$.

R code provided online (`https://sites.google.com/site/jmswason`) allows the user to find the values of n and c so that a design has required FWER and power.

3.4 Delay between recruitment and assessment of patients

There is commonly a delay between recruiting a patient and assessment of their response to treatment. This delay can be a fixed period of time, for example if the response is measured at a fixed time after treatment begins. It could instead be variable, for example when the outcome is time-to-event.

A delay between recruitment and assessment means that at the time of an interim analysis, there may be patients who have been recruited but have not yet been assessed, and thus contribute no information to the analysis. This causes the efficiency, in terms of number of patients recruited, of a trial with early stopping to be reduced as patients will have already been recruited to arms that were dropped. It may also mean that an arm may be dropped that would not have been dropped if all patients recruited to it had been fully observed. The potential loss of efficiency depends on not just the extent of delay but also the recruitment rate of the trial. A trial with a faster recruitment rate will be more seriously affected than one with a slower recruitment rate, as more patients will be recruited but not assessed at an interim analysis.

The methodology described in sections 3.1-3.3 can be applied without modification in the case of a fixed delay by simply using the number of patients with outcome information available at the interim analysis instead of the planned number. It cannot be if the delay

is associated with the treatment effect, for example with time-to-event outcomes, where the absence of early events at an interim analysis is informative.

4 Strong control of family-wise error rate

In order to control the probability of recommending an ineffective treatment, we use equation (6) to control the probability of recommending any treatment when the global null hypothesis is true. In the case of a group-sequential MAMS trial, this was shown to control the family-wise error rate (FWER) in a strong sense [2]. In this section, we prove that controlling the FWER at the global null hypothesis strongly controls the FWER for the multi-stage drop-the-losers design also.

We denote by m_j, the fixed number of observations collected in stage j on each surviving treatment and on the control arm. At the end of stage j, the cumulative sample size on each remaining treatment and the control arm is $n_j = m_1 + \ldots + m_j$. Without loss of generality, we assume just one treatment is eliminated in each stage: the reason there is no loss of generality here is that if two or more treatments are to be eliminated, we can suppose that data gathering stages with sample size $m_j = 0$ take place between each elimination.

Initially the set of indices of all treatments is

$$I_0 = \{1, \ldots, K\}$$

and after a treatment has been eliminated at the end of stage j, we denote the set of indices of the $K - j$ remaining treatments by I_j.

Recall for $k = 1, \ldots, K$, we denote the observations on treatment k in stages 1 to j by X_{ki}, $i = 1, \ldots, n_j$, and denote the corresponding observations on the control arm by X_{0i}, $i = 1, \ldots, n_j$. For each $k \in I_{j-1}$, the difference between the sum of responses on

treatment k and the control at the end of stage j is

$$S_{j,k} = \sum_{i=1}^{n_j} (X_{ki} - X_{0i}).$$

We define the terms $S_{j,k}$ for $k \in I_{j-1}$ since these are the statistics observed after gathering new data in stage j. The values $S_{j,k}$, $k \in I_{j-1}$, are used to select the treatment to be eliminated at the end of stage j, and the values $S_{j,k}$, $k \in I_j$, are then carried forward. The set I_{K-1} contains just one treatment index and after data are gathered on this treatment and control in stage K, this $S_{j,K}$ is used to decide whether or not the one treatment in I_{K-1} is superior to the control.

We first consider the general case where treatments 1 to K have treatment effects $\delta_1, \ldots, \delta_K$ relative to the control treatment. For notational convenience, we set

$$S_{0,k} = 0, \quad k = 1, \ldots, K.$$

With normally distributed responses of common variance σ^2, we can describe the data gathering in stage $j \geq 1$ by writing

$$S_{j,k} = S_{j-1,k} + m_j \delta_k + \epsilon_{j,k} \sqrt{m_j \sigma^2} + \xi_j \sqrt{m_j \sigma^2}, \tag{8}$$

where all the $\epsilon_{j,k}$ and ξ_j are independent $N(0,1)$ random variables. Here, $\epsilon_{j,k}$ is associated with the responses on treatment k in stage j; ξ_j is associated with responses on the common control arm in stage j and these terms introduce correlation into the sums $S_{j,k}$, $k \in I_{j-1}$.

After the data gathering part of stage j, the treatment k_j^* with the lowest $S_{j,k}$ for $k \in I_{j-1}$ is eliminated, leaving

$$I_j = I_{j-1} \backslash \{k_j^*\}.$$

After the penultimate stage $K-1$, one treatment, k_{last} say, remains in I_{K-1} and this

treatment and the control are observed in the final stage, K. After stage K, the statistic including the final stage data is $S_{K,k_{last}}$. If

$$S_{K,k_{last}} > c,$$

H_0: $\delta_{k_{last}} \leq 0$ is rejected in favour of $\delta_{k_{last}} > 0$.

The trial is designed to have type I error probability α when $\delta_1 = \ldots = \delta_K = 0$. We wish to show this also implies strong control of the family wise error rate (FWER) for testing the family of hypotheses $H_0^{(k)}$: $\delta_k \leq 0$, $k = 1, \ldots, K$.

Consider two experiments which differ only with respect to values of the treatment effects. In Experiment 1, $\delta_1 = \ldots = \delta_K = 0$ and we use the notation described above. We define a parallel set of notation for Experiment 2. We denote the treatment effects in Experiment 2 by ϕ_l, $l = 1, \ldots, K$, and suppose some of the ϕ_l may be positive, and others negative or equal to zero. Let L_j denote the set of indices of treatments still in the trial after stage j of Experiment 2 and

$$N_j = \{l\text{: } l \in L_j \text{ and } \phi_l \leq 0\},$$

so now a type I error will only occur if one of the hypotheses H_0: $\phi_l \leq 0$ for $l \in N_j$ is eventually rejected. For $j = 1, \ldots, K-1$, let $T_{j,l}$, $l \in L_{j-1}$ be the analogues of Experiment 1's $S_{j,k}$, $k \in I_{j-1}$. For $j = K$, $L_{K-1} = \{l_{last}\}$, $I_{K-1} = \{k_{last}\}$ and $T_{K,l_{last}}$ is the analogue of $S_{K,k_{last}}$.

With

$$T_{0,l} = 0, \quad l = 1, \ldots, K,$$

we can write for each $j \geq 1$

$$T_{j,l} = T_{j-1,l} + m_j\phi_l + \eta_{j,l}\sqrt{m_j\sigma^2} + \xi_j\sqrt{m_j\sigma^2}, \tag{9}$$

where the $\eta_{j,l}$ and ξ_j are independent $N(0,1)$ random variables.

After the data gathering part of stage j, the treatment l_j^* with the lowest $T_{j,l}$ for $l \in L_{j-1}$ is eliminated, leaving

$$L_j = L_{j-1} \backslash \{l_j^*\}.$$

After the penultimate stage $K-1$, only one treatment, l_{last} say, remains. This is observed in stage K and if

$$T_{K,l_{last}} > c,$$

H_0: $\phi_{l_{last}} \leq 0$ is rejected in favour of $\phi_{l_{last}} > 0$.

We shall establish the desired FWER property by a coupling argument which assumes the terms ξ_j in (8) and (9) are equal and which re-uses values $\eta_{j,l}$ in (9) as values for some of the $\epsilon_{j,k}$ in (8). It is straightforward to see that the model for Experiment 1 given by (8) and the model for Experiment 2 given by (9) follow the correct distributional assumptions. The type I error rate for Experiment 1 is α, by construction. Thus, if we can demonstrate that a type I error is made in Experiment 1 whenever a type I error is made in Experiment 2, it follows that Experiment 2 has the smaller type I error probability — and so this must be no greater than α.

A key step in the coupling argument is to define the relationship between treatments $k \in I_{j-1}$ and $l \in L_{j-1}$ which specifies how values $\eta_{j,l}$ in (9) are to be used as values for the $\epsilon_{j,k}$ in (8). Define

$$N_0 = \{l\colon\ \phi_l \leq 0\},$$

and, as noted previously,

$$N_j = \{l\colon\ l \in L_j \text{ and } \phi_l \leq 0\}, \quad \text{for } j = 1, \ldots, K-1.$$

For $j = 0$, define

$$\pi_0(l) = l, \quad \text{for each } l \in N_0.$$

In applying (9) for $j = 1$, generate independent random variables $\xi_1 \sim N(0,1)$ and $\eta_{1,l} \sim N(0,1)$, $l \in L_0$. Then, in applying (8) for $j = 1$, use the same value ξ_1 as in (9), set

$$\epsilon_{1,\pi_0(l)} = \eta_{1,l} \quad \text{for each } l \in N_0,$$

and generate the remaining $\epsilon_{1,k}$ values as additional independent $N(0,1)$ variates. It follows that

$$T_{1,l} \leq S_{1,\pi_0(l)} \quad \text{for each } l \in N_0. \tag{10}$$

Our aim is to define injective functions π_j from N_j to I_j at the end of each stage $j = 1, \ldots, K-1$, such that

$$T_{j,l} \leq S_{j,\pi_j(l)} \quad \text{for each } l \in N_j. \tag{11}$$

Intuitively, this means that for each treatment arm in Experiment 2 which has a treatment effect less than or equal to zero, and so would produce a type I error if the associated null hypothesis were rejected, there is a treatment arm in Experiment 1 which has a treatment effect of zero and more positive current data — and so this should be more inclined to lead to a type I error. Finally, after stage K, we have the control and just one treatment, k_{last} in Experiment 1 and l_{last} in Experiment 2 and final statistics $S_{K,k_{last}}$ and $T_{K,l_{last}}$.

Assuming we can define the desired functions π_j, there are two possibilities at the end of the trial when stage $j = K$ is completed. The first possibility is that, on entering stage K, the set N_{K-1} is empty and a type I error cannot be made in Experiment 2. The second is that N_{K-1} is non-empty and contains a single element, so $\phi_{l_{last}} \leq 0$ and $\pi_{K-1}(l_{last}) = k_{last}$ (the only element of I_{K-1}): before the final stage data are seen,

$$T_{K-1,l_{last}} \leq S_{K-1,k_{last}},$$

then with the (coupled) final stage data,

$$T_{K,l_{last}} \leq S_{K,k_{last}}.$$

A type I error in Experiment 2 requires $T_{K,l_{last}} > c$ and this can only occur if

$$S_{K,k_{last}} > c,$$

in which case a type I error is also made in Experiment 1. This establishes the desired property that a type I error is made in Experiment 1 whenever a type I error is made in Experiment 2 and the FWER result follows.

It remains to show that injective functions π_j from N_j to I_j, $j = 1, \ldots, K-1$, can be defined with the required property (11). For the case $j = 1$, we know that property (10) holds before a treatment is eliminated at the end of stage 1 and we need to define a function π_1 from N_1 to I_1 satisfying (11) with $j = 1$, after the first treatment has been eliminated. The eliminated treatments are k_1^* in Experiment 1 and l_1^* in Experiment 2, where

$$S_{1,k_1^*} \leq S_{1,k} \quad \text{for } k \in I_0,\, k \neq k_1^* \tag{12}$$

and

$$T_{1,l_1^*} \leq T_{1,l} \quad \text{for } l \in L_0,\, l \neq l_1^*.$$

In defining π_1 from N_1 to I_1, we need to consider values $l \in N_1 = N_0 \backslash \{l_1^*\}$. For each value $l \in N_1$ with $\pi_0(l) \neq k_1^*$, we set

$$\pi_1(l) = \pi_0(l) \ \in \ I_1 = I_0 \backslash \{k_1^*\}.$$

It follows from (10) that $T_{1,l} \leq S_{1,\pi_1(l)}$ for these values of l. Now suppose there is a value $\tilde{l} \in N_1$ for which $\pi_0(\tilde{l}) = k_1^*$ and thus $\pi_0(\tilde{l}) \notin I_1 = I_0 \backslash \{k_1^*\}$. In this case, we can set $\pi_1(\tilde{l})$ to be any index in I_1 which is not already defined as $\pi_1(l)$ for some other $l \in N_1$ (since

I_1 has at least as many elements as N_1, there will be at least one option to choose here). The resulting π_1 has the injective property. Now, by (10) and (12),

$$T_{1,\tilde{l}} \leq S_{1,\pi_0(\tilde{l})} = S_{1,k_1^*} \leq S_{1,\pi_1(\tilde{l})}$$

so (11) is satisfied for $j = 1$ and $l = \tilde{l}$. This completes the definition of π_1.

The construction of functions π_j for $j = 2, \ldots, K-1$ and proof of their properties continues by induction. For a general j, we apply (8) and (9) using the same ξ_j in both cases and with

$$\epsilon_{j,\pi_{j-1}(l)} = \eta_{j,l} \quad \text{for each } l \in N_{j-1}.$$

With property (11) for $j-1$, we have

$$T_{j-1,l} \leq S_{j-1,\pi_{j-1}(l)} \quad \text{for each } l \in N_{j-1}$$

and, because of the common values of $\epsilon_{j,\pi_{j-1}(l)}$ and $\eta_{j,l}$ and the common ξ_j arising in (8) and (9), this ensures that

$$T_{j,l} \leq S_{j,\pi_{j-1}(l)} \quad \text{for each } l \in N_{j-1}.$$

Thus, we can define π_j by setting

$$\pi_j(l) = \pi_{j-1}(l) \in I_j$$

for each value $l \in N_j$ with $\pi_{j-1}(l) \neq k_j^*$. If there is a value $\tilde{l} \in N_j$ for which $\pi_{j-1}(\tilde{l}) = k_j^*$, we can set $\pi_j(\tilde{l})$ to be any element of in I_j which is not already defined as $\pi_j(l)$ for some other $l \in N_j$. The same reasoning as in the case $j = 1$ shows that the resulting π_j from N_j to I_j has the injective property and satisfies (11), which proves the inductive step.

As noted earlier, the inductive properties at stage K imply that if $\phi_{l_{last}} \leq 0$: before

collecting the final stage data, we have $\pi_{K-1}(l_{last}) = k_{last}$ and

$$T_{K-1,l_{last}} \leq S_{K-1,k_{last}},$$

then with the (coupled) final stage data,

$$T_{K,l_{last}} \leq S_{K,k_{last}};$$

a type I error in Experiment 2 requires $T_{K,l_{last}} > c$ and this can only occur if

$$S_{K,k_{last}} > c,$$

in which case a type I error is also made in Experiment 1, as required.

5 Results

5.1 Motivating trial

As a case-study for the results in this paper, we consider the currently ongoing TAILoR (TelmisArtan and InsuLin Resistance in HIV) trial (the design of this trial is discussed in Magirr et al. [2]). This trial was originally designed to test four different doses of Telmisartan. Telmisartan is thought to reduce insulin resistance in HIV-positive individuals on combination antiretroviral therapy (cART). The primary endpoint was reduction in insulin resistance in the telmisartan-treated groups in comparison with the control group as measured by HOMA-IR at 24 weeks. A group-sequential MAMS design was used to avoid assumptions regarding monotonicity of dose-response relationship, which were thought to be invalid based on a previous trial of the treatment in a different indication.

The trial design controls the FWER at 0.05 with 90% power under the LFC with $\delta^{(1)} = 0.545$ and $\delta^{(0)} = 0.178, \sigma_0^2 = \sigma_1^2 = \ldots = \sigma_K^2 = 1$. The value of $\delta^{(1)}$ was chosen so that the probability of a patient allocated to a treatment with treatment effect $\delta^{(1)}$

having a better treatment response than if they were given the control treatment was 0.65. The value of $\delta^{(0)}$ was chosen to make this probability 0.55.

5.2 Comparison of two- and three-stage drop-the-losers designs

We first show that extending the drop-the-losers design beyond two stages is worthwhile. For $(\alpha, 1-\beta, \delta_1, \delta_0) = (0.05, 0.9, 0.545, 0.178)$, we found the required sample size of the one-stage (i.e. no interim analyses), two-stage drop-the-losers design and the most efficient (i.e. the one that gives the lowest total required sample size) three-stage drop-the-losers design for different numbers of experimental arms, K, using equations (6) and (7). Table 1 shows the required total sample size for one-stage (i.e. all K experimental arms continue throughout the trial), two-stage and three-stage drop-the-losers designs, and the percentage reduction that results from going from two to three stages. The percentage reduction is low for $K = 3$, but increases in the number of experimental arms. It is likely that at least $K = 4$ experimental arms would be required before the additional administrative burden of a third stage would be worthwhile.

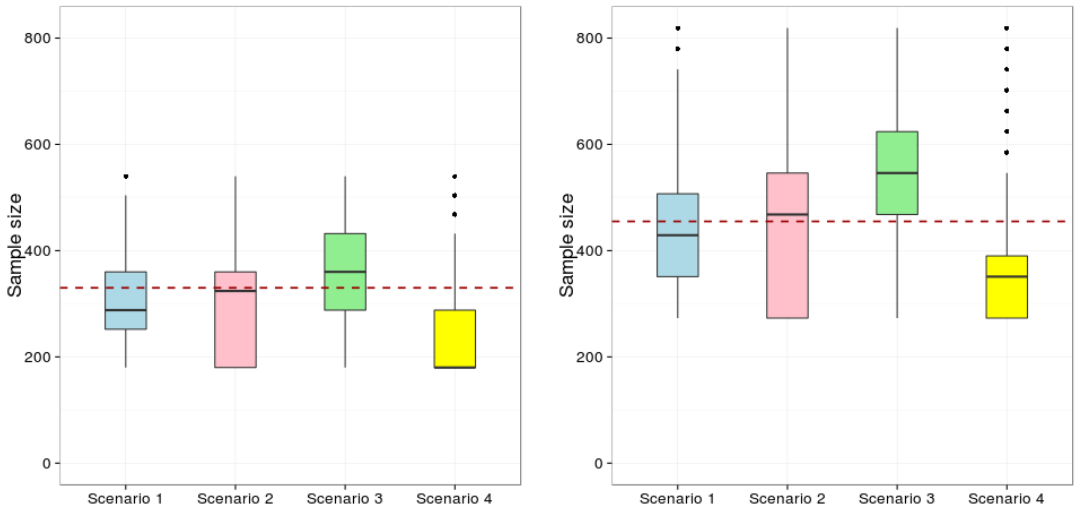
	Total sample size required for 90% Power			Percentage reduction in sample size	
K	$J = 1$	$J = 2$	$J = 3$	$J = 1$ to $J = 2$	$J = 2$ to $J = 3$
3	312	282	270	9.6	4.2
4	420	364	330	13.3	9.3
6	637	531	455	16.6	14.3
8	864	715	585	17.2	18.2

Table 1: Sample sizes required for one-stage design, together with two-stage and three-stage drop-the-losers designs for different numbers of experimental arms (K). The three-stage designs shown are the most efficient of all possible three-stage designs (i.e. the one with the lowest sample size required for 90% power). Parameters used are $\alpha = 0.05, \beta = 0.1, \delta^{(1)} = 0.545, \delta^{(0)} = 0.178$.

It should be pointed out that for $K \geq 4$, the percentage drop in sample size when going from two stages to three stages is similar to the percentage drop when going from one stage to two stages. Thus if one were to think it was worthwhile to include a single interim analysis, it should also be worthwhile to include two.

5.3 Comparison of three-stage group-sequential MAMS and drop-the-losers designs

We next compare the sample size properties of the group-sequential MAMS design to the drop-the-losers design. Both designs have have the same design parameters as in the previous section. For the group-sequential MAMS designs we use the triangular test boundaries of Whitehead and Stratton [7]. These generally give good expected sample size properties [2]. Figure 1 shows boxplots of the sample size distribution (using 250000 replicates) for the three-stage group-sequential MAMS design under four scenarios: 1) under H_G; 2) under the LFC; 3) when $\delta_1 = \delta_2 = \ldots = \delta_K = \delta^{(0)}$ and 4) when $\delta_1 = \delta_2 = \ldots = \delta_K = -\delta^{(0)}$. Both $K = 4$ and $K = 6$ are considered. The solid black line represents the median sample size. The dashed lines represent the (fixed) sample sizes of the respective optimal three-stage drop-the-losers designs ($4:2:1$ for $K = 4$ and $6:3:1$ for $K = 6$).



(a) $K = 4$ (b) $K = 6$

Figure 1: Sample size distribution for three-stage group-sequential MAMS designs with $K = 4$ and $K = 6$ different scenarios of treatment effects. Scenario 1 - global null hypothesis (H_G); scenario 2 - the least-favourable configuration (LFC); Scenario 3 - all experimental treatments have uninteresting treatment effect ($\delta^{(0)}$); Scenario 4 - all experimental treatments have effect $-\delta^{(0)}$. Dashed red line gives the required sample size for the equivalent three-stage drop-the-losers design. Design parameters used are $\alpha = 0.05, \beta = 0.1, \delta^{(1)} = 0.545, \delta^{(0)} = 0.178$.

Figure 1 demonstrates that the sample size used by a group-sequential MAMS design is highly variable, and strongly depends on the configuration of treatment effects asssumed. For scenario 1 (H_G), the drop-the-losers sample size is higher than the median sample size used by the group-sequential MAMS design. Under the LFC (scenario 2), the drop-the-losers design sample size is very close. Scenarios 3 and 4 are more distinct - under scenario 3, when all treatments are slightly effective, the drop-the-losers sample size is lower than the median group-sequential sample size. Under scenario 4, however, it is considerably higher. There are differences between $K = 4$ and $K = 6$ - generally the drop-the-losers design is comparatively more efficient for $K = 6$. These results are generally encouraging for the drop-the-losers design. The indications are that some efficiency (in terms of average numbers of patients recruited) may be lost, with the loss being higher when there are fewer treatments and when they are ineffective. In some cases, the average efficiency of the drop-the-losers design is higher. In the situation where all experimental treatments are ineffective, the considerable loss in efficiency from using the drop-the-losers design can be mitigated by adding a futility rule, which we consider further in the discussion.

6 Spacing of interim analyses when there is delay between recruitment and assessment of patients

In previous sections we have assumed there is no delay between recruitment and assessment of patients. Of course in reality there will always be some delay, and often it will be considerable. For example, in the TAILoR trial the final endpoint is measured 24 weeks after treatment.

In this section we explore the optimal spacing of analyses when there is delay. As assessing design characteristics for a drop-the-loser design is very efficient, we can explore a wide variety of spacings in a short time. We assume that the primary endpoint is measured 6 months after recruitment, comparable to the delay in the TAILoR trial. We

consider the $4:2:1$ and $4:1$ designs with $\delta^{(1)} = 0.545$, $\delta^{(0)} = 0.178$, $\alpha = 0.05$, $1-\beta = 0.9$. For each design, a grid of possible spacings are explored. For the $4:2:1$ design, spacings are expressed in terms of $(1, \omega_2, \omega_3)$. If the group size of a design is n, and the spacing is $(1, \omega_2, \omega_3)$, then the first interim analysis will take place after n patients have been *recruited*, the second interim analysis will take place after a further $\omega_2 n$ patients have been recruited to each remaining arm, and the last analysis will take place after a further $\omega_3 n$ patients have been recruited and *assessed*. That is, the last interim analysis takes place after all patients recruited have been assessed. We assume that once the decision has been made to drop an experimental arm, that decision cannot be reversed after seeing data from patients who were previously recruited but not assessed. For the $4:1$ design, spacings are expressed in terms of $(1, \omega_2)$. For each design, the optimal spacing, i.e. the spacing that gives the lowest total required sample size is found by searching over the ω_2 and ω_3 parameters.

Table 2 shows the optimal spacing parameters and total sample size for both designs when the mean number of patients recruited per week, m, varies. Note that the design which tests four experimental treatments without any interim analyses requires 420 patients in total.

	Optimal spacing		Max SS		Percentage reduction in SS	
m	J=2	J=3	J=2	J=3	J=2	J=3
0*	(1,0.9)	(1,0.9,0.8)	361	326	14.0	9.7
1	(1,0.8)	(1,0.9,0.45)	377	344	10.2	8.8
2	(1,0.5)	(1,0.95,0.2)	390	363	7.1	6.9
4	(1,0.35)	(1,0.75,0.05)	422	405	-0.5	3.6

Table 2: Properties of $4:2:1$ and $4:1$ designs when there is a 6 month delay between recruitment and assessment of patients for different mean recruitment rates (m). Uniform recruitment is assumed. SS = sample size. *m=0 corresponds to the no-delay case.

Table 2 shows that as the mean recruitment rate increases, there is a lower efficiency gain from including interim analyses. The reduction in sample size, compared to the design with one fewer stage, decreases significantly from 14% to an increase of 0.5% for $J = 2$ and from 9.7% to 3.6% for $J = 3$. The optimal spacing of interim analyses impacts

the sample size considerably. For example, for a mean recruitment rate of 2 patients per week, a $4:2:1$ design with equally spaced interim analyses (i.e. $(\omega_2, \omega_3) = (1, 1)$) would use a total of 390 patients, compared to the 363 patients used with the optimal spacing.

These results caution that the impact of delay on the efficiency of an adaptive design is large. If there is significant delay and quick recruitment, a non-adaptive design will sometimes be more efficient. However, for fairly reasonable recruitment rates, these results show that including both one and two interim analyses will reduce the sample size requirements compared to a design without interim analyses.

7 Discussion

Multi-arm multi-stage (MAMS) designs are of great interest in practice, as their use means more new treatments can be tested with the same limited pool of patients. Much of the methodology about designing MAMS trials has focused on designs in which treatments are dropped early if their test statistics are below some pre-specified futility boundary. This leads to uncertainty in the number of treatments that will be in the trial at each stage, and therefore uncertainty in the total sample size required. This leads to issues in applying for funding to conduct a MAMS trial, as well as other logistical issues such as staff employment. A design that does have a fixed sample size is the two-stage drop-the-losers design, where multiple experimental treatments are evaluated at an interim analysis, then the best performing experimental treatment goes through to the second stage. We have investigated design issues in extending the drop-the-losers design to have more than two stages. If there are four or more treatments, we find that a third stage results in a considerable reduction in sample size required. In addition, the fixed sample size compares well to the median sample size used in a group-sequential MAMS design. The design therefore retains many of the efficiency benefits of a MAMS design whilst also having a fixed sample size, which is very useful in practice. We have mainly considered the utility of adding a third stage, as each additional interim analysis increases

the administrative burden of the trial. Adding a fourth stage provides a substantially lower additional efficiency advantage unless there are a lot of treatments being tested.

In this paper we assumed a known variance of the normally distributed outcome. However the method of quantile substitution, described in Jennison and Turnbull [15], can be used to change the final critical value so that the type-I error rate is controlled when the variance is estimated from the data. We carried out simulations that showed this method performs very well in practice (results not shown), similarly to the group-sequential [16] and group-sequential MAMS cases [17].

In practice the requirement to drop a fixed number of treatments at each stage may be difficult to keep to. For example, if all treatments are performing poorly in comparison to control, then it may be unethical to continue with even the best performing. Any changes to the design during the trial will affect the operating characteristics of the trial. However, dropping more treatments than planned will lead to a lower than nominal FWER rather than an inflation. If one wishes to keep more treatments in the trial than originally planned, then this will lead to an inflation in FWER. However, by modifying the final critical value suitably, this inflation can be reduced. The analytical formulae in this paper can be modified in order to calculate the required critical value if more sophisticated stopping rules are used.

An alternative design that controls the number of treatments passing each analysis but also allows early stopping of the trial for futility or efficacy is the design of Stallard and Friede [9]. The multi-stage drop-the-losers design is somewhat less flexible than the Stallard and Friede design, but does have the advantage of having analytical formulae that provide exact operating characteristics of the design. The formulae for the Stallard and Friede design are conservative, especially when there are more than two stages. Of course simulation could be used to evaluate the operating characteristics exactly, but this makes it difficult to evaluate a large number of potential designs. We have shown that this is important in the case of delay between recruitment and assessment, where the spacing of the interim analyses becomes very important. The multi-stage drop-the-losers design

can be evalauted extremely quickly, which allows the optimal interim analysis spacing to be found quickly.

One worrying factor for the efficiency of adaptive trials in general, and drop-the-loser design specifically, is delay between recruiting a patient and assessing their outcome. Such delay means that at a given interim analysis there will be patients who are recruited but not yet assessed. These patients will not contribute to that interim analysis, nor to any subsequent analysis if the treatment they are on is dropped. We have investigated the effect of delay and show that drop-the-loser designs can still provide efficiency gains over a multi-arm design without interim analyses if the recruitment rate is below some level. This level will depend on the extent of delay and the total sample size of the trial. There are two factors that may go someway towards mitigating the impact of delay. Firstly, there may well be early outcomes that correlate well with the final outcome [18]. For example, in the TAILoR trial, the final outcome is HOMA-IR at 24 weeks, but if earlier measurements could be made, these may well be highly informative for the 24 week endpoint. In that case, more patients could be included in the interim analysis. A second factor is that trial recruitment tends to start slowly and increase over time, perhaps as more centres are added to the trial. This means that a greater proportion of patients may be available for assessment at earlier interim analyses compared to the uniform recruitment case we considered here. Research into the effect of delay on group-sequential MAMS trials and strategies to account for it (extending the work of Hampson and Jennison [19] to multi-arm trials) would be very useful.

This paper has considered design issues in multi-stage drop-the-losers trials. A drawback of adaptive designs in general is that estimation of relevant quantities, such as the mean treatment effect, after the trial is more complicated than in a traditional trial. For example, using the maximum likelihood estimate in two-stage trials will result in bias [20, 21, 22]. The issue of estimation for multi-stage drop-the-losers trials is considered in Bowden and Glimm [23].

Acknowledgements

This work was funded by the Medical Research Council (grant numbers G0800860 and MR/J004979/1). We thank Dr Ekkehard Glimm and two anonymous referees for their useful comments.

References

[1] M.R. Sydes, M.K.B. Parmar, N.D. James, et al. Issues in applying multi-arm multi-stage methodology to a clinical trial in prostate cancer: the MRC STAMPEDE trial. *Trials*, 10, 2009.

[2] D. Magirr, T. Jaki, and J. Whitehead. A generalized Dunnett test for multiarm-multistage clinical studies with treatment selection. *Biometrika*, 99:494–501, 2012.

[3] C.W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50:1096–1121, 1955.

[4] J.M.S. Wason and T. Jaki. Optimal design of multi-arm multi-stage clinical trials. *Statistics in Medicine*, Epub.

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

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

[7] J. Whitehead and I. Stratton. Group sequential clinical trials with triangular continuation regions. *Biometrics*, 39:227–236, 1983.

[8] J. Kairalla, C. Coffey, M. Thomann, and K. Muller. Adaptive trial designs: a review of barriers and opportunities. *Trials*, 13, 2012.

[9] N. Stallard and T. Friede. A group-sequential design for clinical trials with treatment selection. *Statistics in Medicine*, 27:6209–6227, 2008.

[10] P.F. Thall and S.S. Simon, R.and Ellenberg. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics*, 45:537–547, 1989.

[11] A. Sampson and M. Sill. Drop-the-losers design: normal case. *Biometrical Journal*, 47:257–268, 2005.

[12] F. Bretz, H. Schmidli, F. Konig, A. Racine, and W. Maurer. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, 48:623–634, 2006.

[13] H. Schmidli, F. Bretz, A. Racine, and W. Maurer. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: Applications and practical considerations. *Biometrical Journal*, 48:635–643, 2006.

[14] A. Genz and F. Bretz. Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics*, 11:950–971, 2002.

[15] C. Jennison and B.W. Turnbull. *Group sequential methods with applications to clinical trials*. Chapman and Hall, Boca Raton FL, 2000.

[16] J.M.S. Wason, A.P. Mander, and S.G. Thompson. Optimal multi-stage designs for randomised clinical trials with continuous outcomes. *Statistics in Medicine*, 31:301–312, 2012.

[17] J.M.S. Wason and T. Jaki. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine*, 31:4269–4279, 2012.

[18] N. Stallard. A confirmatory seamless phase II/III clinical trial design incorporating short-term endpoint information. *Statistics in Medicine*, 29:959–971.

[19] L.V. Hampson and C. Jennison. Group sequential tests for delayed responses. *Journal of the Royal Statistical Society B*, 75:1–37, 2013.

[20] A. Cohen and H.B. Sackrowitz. Two stage conditionally unbiased estimators of the selected mean. *Statistics and Probability Letters*, 8:273–278, 1989.

[21] J. Bowden and Ekkehard Glimm. Unbiased estimation of selected treatment means in two-stage trials. *Biometrical Journal*, 50:515–527, 2008.

[22] PK. Kimani, S. Todd, and N. Stallard. Conditionally unbiased estimation in phase II/III clinical trials with early stopping for futility. *Statistics in Medicine*, 32:2893–2901, 2013.

[23] J. Bowden and E. Glimm. Conditionally unbiased and near unbiased estimation of the selected treatment mean for multi-stage drop-the-losers trials. *Submitted to Biometrical Journal*.