Comparing the MAMS framework with the combination method in multi-arm adaptive trials with binary outcomes

Article

Accepted Version

Comparing the MAMS framework with the combination method in multi-arm adaptive trials with binary outcomes.

Julia E Abery

Department of Mathematics and Statistics, University of Reading, Reading, UK

Susan Todd

Department of Mathematics and Statistics, University of Reading, Reading, UK

Corresponding author:

Julia E Abery

Department of Mathematics and Statistics,

University of Reading, Whiteknights, Reading, RG6 6AH, UK

Email:J.E.Abery@pgr.reading.ac.uk

Comparing the MAMS framework with the combination method in multi-arm adaptive trials with binary outcomes.

Short Title: **Comparing the MAMS framework with the combination method**

Abstract

In multi-arm adaptive trials, several treatments are assessed simultaneously and accumulating data are used to inform decisions about the trial, such as whether treatments are dropped or continued. Different methodological approaches have been developed for such trials and research has compared the performance of different subsets of these. The approach described by Royston et al (2003), for which we use the acronym MAMS(R), has generally not been included in these comparisons because control of the family-wise error rate (FWER) could not be guaranteed. Recently, the MAMS(R) approach has been extended to facilitate the generation of efficient designs which strongly control the FWER. We consider multi-arm two-stage trials with binary outcomes and propose parameterising treatment effects using the log odds ratio. We conduct a simulation study comparing the extended MAMS(R) framework with the well-established combination method both for trials where a different outcome is used for mid-trial analysis and for trials where the same outcome is used throughout. We show how the MAMS(R) framework compares favourably only in scenarios where the same outcome is used. We propose a hybrid selection rule within MAMS(R) methodology and demonstrate that this makes it possible to use the MAMS(R) framework in trials incorporating comparative treatment selection.

Keywords

Multi-arm adaptive trials, MAMS, combination method, log odds ratio, selection rule, family-wise error rate

1 INTRODUCTION

When several treatments are proposed as candidates for a particular medical condition at the same time, the length of time and total sample size required to evaluate each one in a separate conventional clinical trial may be unacceptable. Multi-arm adaptive trials have been developed to offer a more timely and efficient evaluation. In a multi-arm adaptive trial several new therapies may be assessed alongside a single control group; this can speed up the process of evaluation and substantially reduce sample size requirements compared with conducting separate trials. Furthermore, a multi-arm adaptive trial is conducted in stages allowing interim analysis of accumulating data to inform how the trial should progress, for example poorly performing arms may be dropped. A useful application of these methods has been the facility to merge a Phase II with a Phase III trial. These so called 'seamless trials' may substantially reduce sample size requirements and also reduce the potentially lengthy 'white space' between Phase II and Phase III.[1,2]

A key challenge in multi-arm adaptive trial methodology is strong control of the familywise error rate (FWER) so that the probability of recommending an ineffective treatment is not inflated by multiplicity or selection. Several methodological approaches which address this issue have been developed and compared.[3,4] A key feature separating these approaches into two main types concerns the manner in which stage-wise test statistics are obtained. These may either be calculated based on data from each stage separately and then combined at the end of the trial or alternatively may be calculated cumulatively as the trial progresses. The combination method[5,6] is a well-established method of the first type which uses a closed testing procedure[7] to control the FWER strongly. The group sequential method[8,9,10] is of the second type and uses cumulative test statistics, calculated at each stage, which are compared

against boundaries defined by critical values. Different types of boundaries can be specified depending on the requirements of a particular trial, for example to allow early stopping for efficacy or futility. The boundaries are obtained using numerical integration such that the Type 1 error is controlled. A related approach based on cumulative test statistics and stage wise critical values is that proposed by Royston et al,[11] we will refer to this as the MAMS(R) framework. This method allows different outcomes to be used for the intermediate and final analyses, a useful feature in trials where the primary outcome is observable only after a long time period. Critical values are specified which determine the early dropping of poorly performing treatments and the Type 1 error can be calculated for any set of critical values if the correlation between the intermediate and final outcomes is known.

A number of studies have compared the performance of the combination method with the group sequential approach. In the single experimental arm setting, Jennison and Turnbull[12] describe how using the combination method allows greater flexibility regarding stage-wise sample sizes but that unplanned changes reduce efficiency because the final test for the treatment difference is not based on a sufficient statistic. Tsiatis and Mehta[13] show that for trials where such unplanned changes are made, it is always possible to find a group sequential design which has the same sample size and is more powerful. Kelly et al[14] investigated two-stage and five-stage designs in a practical setting and found the group sequential approach yielded similar or slightly greater power compared with the combination method. However, they confirm the greater flexibility of the combination method by showing that changes to sample sizes made on the basis of interim data analysis result in a breach of the Type I error in the group sequential approach, but not in the combination method. Comparisons have also been drawn between different approaches in the multi-arm setting where interim data analysis is used to inform treatment selection.[15] Friede and Stallard[16] compared a number of adaptive

trial designs including the group sequential approach and the combination method. They did not find any method to be consistently more powerful than another. Instead, factors such as the size of the treatment effect and the process chosen for selecting treatments determined which approach performed best. Kunz et al[17] considered multi-arm trials where data regarding an early outcome measure are incorporated in the process of treatment selection. They conducted a comparison study and again found there was no overall advantage for the group sequential approach or the combination method, but that the preferred method depended on treatment effects and correlations between early and final outcomes.

Studies comparing different approaches in multi-arm adaptive design methodology have generally not incorporated the MAMS(R) framework, largely because control of the FWER could not be guaranteed and also because the MAMS(R) framework was developed specifically for trials with survival outcomes. However, MAMS(R) methodology has recently been extended such that binary outcomes may now be accommodated.[18] Strong control of the FWER can also be guaranteed and a systematic search procedure has been developed which can produce efficient designs for trials with any number of stages and treatment arms.[19]

In view of these developments, we consider there is good reason to explore and evaluate the performance of MAMS(R) designs. We propose utilising and further developing the MAMS(R) framework to obtain boundary based trial designs and then use these designs as the basis for a study of the performance of MAMS(R) compared to the combination method in the setting of two-stage trials. Both approaches are relatively easy for clinicians to understand and implement in the multi-arm context and neither method requires the number of treatments selected at an interim analysis to be specified in advance. Furthermore, each of

the approaches can be used in trials where treatment selection is based on the definitive outcome or on an early outcome measure only, without there being any restriction to carry forward only one treatment.

Our motivation is the evaluation of treatments for chronic diseases such as tuberculosis (TB), multiple sclerosis (MS) and osteoporosis. For these conditions, binary outcomes, representing a success or failure recorded for an individual patient, are commonly encountered. A binary outcome in TB could be whether or not a patient converted to a negative sputum culture, in MS whether or not a patient's disability rating has increased by a given number of units and in Osteoporosis whether or not a patient has suffered a fracture during a certain time period.

We consider two types of multi-arm trials, firstly where the intermediate outcome (I) is different from the definitive outcome (D), perhaps because data regarding the definitive outcome would not be available at an early stage in the trial, we refer to these as $I \neq D$ trials, and secondly trials where the same outcome is used throughout, we refer to these as $I = D$ trials. For $I \neq D$ trials, we use as a basis the Phase 2/3 seamless trial described by Bratton[18] in which several treatment regimens for TB were evaluated. The intermediate outcome is whether conversion to negative culture status has occurred after eight weeks treatments and the definitive outcome is whether a patient has relapsed or not during an 18 month period. For $I = D$ trials, we use a two-stage version of the Phase 2 trial, where the outcome related to relapse is used throughout the trial.

In Section 2 we describe how treatment effects for binary outcomes may be parameterised in terms of a probability difference or a log-odds ratio (LOR). We then briefly describe

methodology for the combination method and the MAMS(R) framework. In Section 3 we show how MAMS(R) methodology based on the parameterisation of probability difference can be adapted so that efficient two-stage designs based on the log-odds ratio may be readily obtained. We discuss the selection rules which are currently implemented in the combination method and the MAMS(R) framework and propose a new hybrid selection rule. We then conduct a simulation study comparing the performance of the MAMS(R) and combination methods for both $I \neq D$ and $I = D$ trials with binary outcomes, investigating a number of scenarios and a range of treatment effects.

2 BACKGROUND

2.1 Choice of treatment difference for trials with binary outcomes

For a clinical trial with a binary outcome, let the proportion of patients who have a positive response regarding a chosen outcome be denoted p_E under the experimental treatment and p_C under the control treatment. To compare the new therapy with the control, the difference in proportions, $p_E - p_C$, may be used as a measure of the treatment effect. This option has the merit of simplicity. However, an alternative measure of treatment difference for binary outcomes is the log odds ratio (LOR) defined as $\theta = \log\{p_E(1-p_C)/p_C(1-p_E)\}$. Unlike the measure 'difference in proportions' the LOR is asymptotically normally distributed and this may be an advantage when significance tests are based on assumptions of normality. Also, the LOR is closely linked to the logit, the natural parameter used in logistic modelling. There may be times when it is desirable to express a clinical outcome using a modelling approach, perhaps to allow inclusion of relevant covariates. Using the LOR makes this transition more straightforward. In this paper we have chosen to consider the LOR

parameterisation and describe its implementation in the MAMS(R) framework and the combination method.

2.2 MAMS(R) framework

MAMS(R) trials were initially proposed by Royston et al[11] to address the need for increased efficiency in evaluating treatments for diseases where the main outcome of interest is a survival time response. Their approach was developed for the $I \neq D$ case, but can easily be applied in the $I = D$ case. Recently, MAMS(R) methodology has been extended to accommodate binary outcomes[18] with the difference in success rate between the control and the experimental treatments being used to parameterise treatment effects. Briefly, the fundamental elements of the MAMS(R) framework, irrespective of outcome type, are as follows. To obtain a design for a trial with K experimental treatments and a single control, we assume the same null and alternative hypotheses for all treatment arms. We denote the treatment effect, comparing experimental treatment T_i with the control treatment, by θ_{ij}, where i denotes the treatment arm and j denotes the intermediate or definitive outcome. The hypotheses of interest are then

$$H_{0(i)}: \theta_{ij} \leq \theta_j^0 \quad i = 1, \ldots K, j = I, D$$

$$H_{A(i)}: \theta_{ij} > \theta_j^0 \quad i = 1, \ldots K, j = I, D$$

For the $I \neq D$ case, θ_I^0 and θ_D^0 represent the null hypotheses for the early and definitive outcome respectively whereas for the $I = D$ case, $\theta_I^0 = \theta_D^0$. For a superiority trial, θ_j^0 is usually equal to 0 whereas for a non-inferiority trial θ_j^0 generally takes a small but negative value.

In a two stage MAMS(R) trial, a test of $H_{0(i)}$ against $H_{A(i)}$ is conducted for each treatment arm at the end of each stage. Cumulative test statistics calculated for each treatment are compared against predetermined critical values. At the end of stage one, a treatment is dropped if the test statistic falls below the stage one critical value (C_1). At the end of the second stage, any remaining treatment is declared beneficial if the stage two critical value (C_2) is exceeded. A key issue in MAMS(R) methodology is how to determine C_1 and C_2 so that the Type I error is controlled at some specified value. Originally, although designs included several experimental treatment arms, the pair-wise error rate (PWER) was controlled rather than the FWER. Assuming the null hypothesis is true, let standardised test statistics obtained for a given treatment arm at stage one and stage two be denoted S_{ij}^1 and S_{ij}^2 respectively. Then,

$$\begin{pmatrix} S_{ij}^1 \\ S_{ij}^2 \end{pmatrix} \sim BVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where BVN denotes the bivariate normal distribution and ρ is the correlation between the test statistics obtained at the interim and final analyses. For $I = D$, ρ is a function of the stage-wise sample sizes whereas for $I \neq D$, ρ is a function of the stage-wise sample sizes and also the correlation between the intermediate and definitive outcomes. The probability of a given treatment passing both stages and thereby being declared effective is given by $pr\big((S_{ij}^1 \geq C_1, S_{ij}^2 \geq C_2)|H_{0(i)}\big) = PWER$. The PWER is calculated by integration of the tail areas of the joint distribution. Similar expressions for pair-wise power can be obtained by considering the probability of a treatment passing both stages when the alternative hypothesis is true. The critical values C_1 and C_2 can be chosen on a trial and error basis such that the PWER is no greater than some α and the pairwise power no less than some ω. This approach has been used for designing both $I = D$ and $I \neq D$ trials. However, Bratton[19] suggests that although

this method is appropriate when $I = D$, it may not be suitable when $I \neq D$ since in this case the maximum PWER in fact occurs when a treatment is ineffective on the definitive outcome, but is fully effective on the intermediate outcome so that C_2 should be determined solely by the target α as in a single stage trial.

Bratton[19] proposes a method for obtaining a set of critical values for a MAMS(R) trial which ensures that the FWER is controlled at a specified level. Following the approach to trial design first suggested by Simon[20] and then developed by Jung et al[21] and Mander et al[22], Bratton developed software in which a systematic search procedure is used to generate a set of designs which achieve a specified FWER and pair-wise power; such designs are termed 'feasible'. The overall expected sample size of each feasible trial, denoted N, is then calculated under two scenarios, firstly under the global null hypothesis and secondly under the situation where all arms have treatment effects on I and D equal to some reference values, denoted θ_I^R and θ_D^R, which are specified by the user. We denote these two scenarios using $H_{0(G)}$ and $H_{R(G)}$ respectively. Designs which minimise a weighted sum of these two measures are identified as 'admissible'. For a trial with K experimental treatment arms with a target FWER of α, the PWER for each treatment arm is first set to α^*, where α^* satisfies the Dunnett probability $\alpha = \phi_K(z_{\alpha^*}, \ldots, z_{\alpha^*}; C)$, where ϕ_K is the K-dimensional multivariate normal distribution function and C is the between-arm correlation matrix. A search is then carried out over many possible combinations of values for C_1 and C_2 and designs which are feasible retained. Identifying admissible designs requires calculation of $E\left(N\middle|H_{0(G)}\right)$ and $E(N|H_{R(G)})$. Obtaining these measures requires calculation of the per-treatment stage-wise sample sizes and also the numerical evaluation of the probability that k out of K treatments will reach stage two of the trial under each hypothesis. This probability may be obtained using a simulation approach somewhat similar to the method described by Wason and Jaki[23].

Test statistics (S_i^s) with the appropriate correlation structure are generated for each treatment at each stage (s)of a MAMS(R) trial in accordance with Equation 5.1 described by Bratton,[19] which in the notation of this paper and for equal allocation to experimental and control treatment arms can be expressed as

$$S_i^s = \sqrt{0.5}\, x_0^s + \sqrt{0.5} x_i^s + \frac{\theta_i^s - \theta^{0(s)}}{\sigma_i^s},$$

where x_i^s are standard normally distributed random variables generated for $i = 0,1..K$ with the appropriate between stage correlation of test statistics, θ_i^s is the true treatment effect for treatment i on the outcome of interest at stage s, $\theta^{0(s)}$ is the treatment effect at stage s under the null hypothesis and σ_i^s is the standard deviation of the observed treatment effects under θ_i^s. By simulating test statistics for a large number of trials and observing the proportion of trials where k out of K treatments pass stage one, the required probabilities can be estimated and then used to determine $E\left(N\middle|H_{0(G)}\right)$ and $E(N|H_{R(G)})$. A loss function denoted L, similar to that proposed by Mander et al[23] is then specified. L is a weighted sum of $E\left(N\middle|H_{0(G)}\right)$ and $E\left(N\middle|H_{R(G)}\right)$ and admissible designs are defined as those which minimise the loss function for a chosen weight (q), such that $L_{(q)} = qE(N|H_{0(G)}) + (1 - q)E\left(N\ |H_{R(G)}\right)$, where $0 < q < 1$. Using this extended methodology, designs which strongly control the FWER can be readily produced for both $I \neq D$ and $I = D$ trials with any number of treatment arms and any number of stages. In principle, the methods could be extended to accommodate any outcome measure which has an asymptotically normally distributed test statistic provided the between stage correlation structure is known.

2.3 Combination method

Combination test methodology can accommodate a variety of outcome types and the test statistics used for treatment selection at stage one can relate either to the definitive outcome $(I = D)$ or to a suitable early outcome $(I \neq D)$. The fundamental elements of the combination test are as follows. Consider again a two-stage trial where there are K experimental treatment arms and a single control arm. Taking first the case when $I = D$, the treatment effect calculated at the end of each stage is denoted θ_i and the hypotheses of interest are then

$$H_{0(i)}: \theta_i \leq \theta^0 \quad i = 1, \dots K$$

$$H_{A(i)}: \theta_i > \theta^0 \quad i = 1, \dots K$$

At the end of the first stage, test statistics are calculated to test $H_{0(i)}$ against $H_{A(i)}$ for each treatment arm. These test statistics are used initially to make a decision concerning which treatments should be continued into the second stage of the trial, for example the treatment arm associated with the largest test statistic may be selected. At the end of the second stage, test statistics relating to each treatment arm still in the trial are calculated as before, using data from the second stage only.

At the end of the trial, the test statistics arising from each stage are used in a closed testing procedure[7] (CTP) to produce a set of stage one and stage two p-values. In a CTP, p-values must be obtained for all possible composite or intersection null hypotheses as well as each individual null hypothesis. For example, in a trial with three experimental treatment arms $(T_1, T_2 \; and \; T_3)$, stage one and stage two p-values are obtained for individual null hypotheses $H_{0(1)}$ $H_{0(2)}$ and $H_{0(3)}$ and for the intersection null hypotheses $H_{0(1,2)}$ $H_{0(1,3)}$ $H_{0(2,3)}$ and $H_{0(1,2,3)}$. For the intersection hypotheses, the methods of Dunnett[24] can be applied such that

for $H_{0(1,2,3)}$, the p-value will equate to the Dunnett-adjusted p-value relating to the largest of the three observed test statistics. For the final analysis of treatment effectiveness, the stage-wise p-values for each individual and intersection null hypothesis are combined, producing an overall p-value for each one, the only requirement being that the distribution of p_2 conditional on p_1 should be stochastically no larger than the uniform distribution.[25] One approach is to use the weighted inverse normal method proposed by Lehmacher and Wassmer[26] which calculates the final p-value using $C(p_1, p_2) = 1 - \Phi[w_1\Phi^{-1}(1-p_1) + w_2\Phi^{-1}(1-p_2)]$, where Φ denotes the normal distribution function and w_1 and w_2 are predetermined weights specified for each stage, s, of the trial such that $w_s > 0$ and $w_1^2 + w_2^2 = 1$, the weights being determined by the stage-wise sample sizes. An intersection hypothesis is rejected at level α if $C(p_1, p_2) \leq \alpha$. An experimental treatment is declared superior to the control at level α only if the individual null hypothesis and all relevant intersection hypotheses are rejected. For example, at the end of the trial T_2 is declared beneficial only if $H_{0(2)}$, $H_{0(1,2)}$, $H_{0(2,3)}$ and $H_{0(1,2,3)}$ are all rejected at level α. Using a CTP in this way ensures strong control of the FWER when multiple hypotheses are being tested.

In the second stage, a subset defining an intersection hypothesis may contain a dropped treatment. In this instance, following the methods adopted by Posch et al[27] and Friede et al,[28] the second stage p-value for this intersection hypothesis is set as the p-value for the group of treatments contained in the original subset *and* selected for the second stage. If the set is empty then the second stage p value is set to 1. For the case where $I \neq D$, the same procedure is used except that the test statistics initially obtained at the end of stage one relate to an early outcome. These test statistics are used to inform treatment selection but are *not* used in the final analysis of treatment efficacy. Once data regarding the definitive outcome

becomes available, these are used to obtain the test statistics for the stage one group of patients, which are then used exactly as for the $I = D$ case.

3 METHODS

For the investigations detailed in this paper we modified routines available in Stata as described in Section 3.1. We used these modified routines to obtain designs for trials when $I \neq D$ and when $I = D$. An integral part of any multi-arm adaptive trial is the selection rule and in Section 3.2 we consider this in detail and also suggest a hybrid rule for MAMS(R) trials when $I = D$. Based on the trial designs obtained, we conducted a simulation study comparing the MAMS(R) framework with the combination method, as described in Section 3.3.

3.1 Adapting trial designs in the MAMS(R) framework for the LOR

Feasible and admissible designs for trials with binary outcomes, where the treatment difference is parameterised as 'difference in proportions', can be readily generated according to the approach described in Section 2.2 by using the **nstagebin** and **nstagebinopt** MAMS(R) programs for Stata.[19] We adapted these programmes to produce designs for two-stage MAMS(R) trials with a binary outcome and the treatment effects parameterised as a LOR. Formulae used in the routines for calculating suggested sample sizes and the variance of the treatment effect were modified to reflect the LOR parameterisation. Sample sizes suggested by the LOR formulae are approximate and may be over-estimated under the LOR,[29] so we incorporated a new routine to refine stage-wise sample sizes so that the Type 1 error is as close to the target value as possible. Further details are given in the Appendix.

Bratton[19] derived expressions based on the parameterisation 'difference in proportions' for the between stage correlation of test statistics. For trials when $I = D$, these expressions remain the same under the LOR. However, for trials when $I \neq D$, we were unable to obtain an analytical expression based on the LOR. Therefore, we adapted for the binary context a simulation based approach for approximating between-stage correlations of early and definitive test statistics which was proposed by Bratton et al[30] for trials with a survival outcome. Again, further details are given in the Appendix.

3.2 Selection Rules

There are a number of different types of selection rule which may be used in a multi-arm adaptive trial, for example a rule may specify that the k best performing treatments are continued in the trial or alternatively that all treatments meeting a certain standard are continued. A particularly flexible selection rule which encompasses many different selection options is the 'epsilon' rule[31] whereby the treatment associated with the largest test statistic is selected to continue along with all others whose test statistic is within a specified range (ε) of the largest. Note that when $\varepsilon = 0$, only the best treatment is selected and when $\varepsilon = \infty$ all treatments are selected to continue.

The MAMS(R) framework uses thresholds for treatment selection as well as in the final analysis of treatment efficacy. When $I \neq D$, the threshold for the early outcome is not binding (see Section 2.2) and therefore an epsilon rule may be used in place of the threshold without inflating the Type 1 error rate. However, when $I = D$, all thresholds, including those which determine the treatments selected to continue, are binding and therefore control of the FWER is not guaranteed if an epsilon rule is used. For $I = D$ trials where a more

comparative selection rule is required, we propose implementing a 'hybrid' rule in the MAMS(R) framework, where the selection process occurs in two steps. Firstly, the interim test statistics associated with each treatment group are compared to the threshold and only those meeting this standard retained. Then an epsilon selection rule is implemented, so that the best performing of the retained treatments is selected along with any other treatment where the test statistic is within epsilon of the largest.

The combination method can accommodate a variety of selection rules and the user may choose a rule which facilitates the aims of the particular trial. For example, if the objective is for the early dropping of poorly performing arms then a threshold rule may be chosen. Alternatively, if the aim is for a more comparative approach such that only the best performing treatments are selected, then an epsilon rule may be implemented.

3.3 Simulation Study

We conducted a simulation study to compare the performance of the MAMS(R) framework and the combination method for conducting two-stage trials with a binary outcome, using the LOR parameterisation and a variety of selection rules. We considered first the case when $I \neq D$ and then the case when $I = D$. As highlighted above, modified STATA routines were used to obtain MAMS(R) trial designs. When implementing the combination method we used a number of routines from the R package 'asd' by Parsons et al.[32]

3.3.1 Trials when $I \neq D$

The trial which motivates the simulation study is a Phase 2/3 trial described by Bratton[18] in which a Phase 2 superiority trial and a Phase 3 non-inferiority trial were combined to create a seamless trial. We specify a one sided FWER of 0.025 (to match a conventional two-sided error rate of 0.05), a pair-wise power of 0.9 and a 1:1 allocation ratio. Control arm event rates for the I and D outcomes are 0.75 and 0.90 respectively. Treatment effects for the I outcome are set at $\theta_I^0 = 0$ and $\theta_I^R = 0.894$ and for the D outcome at $\theta_D^0 = -0.539$ and $\theta_D^R = 0$. We used our revised routines based on the LOR to produce feasible and admissible MAMS(R) designs for two-stage three-arm ($K = 2$) and six-arm ($K = 5$) trials where $I \neq D$, choosing the design which is admissible across the widest range of q (see Section 2.2). Details of the chosen MAMS(R) designs are given in Table 1.

Table 1. Summary of two stage $I \neq D$ designs used in simulation study

Two experimental treatment arms ($K = 2$)			
	α (critical value)	ω	Cumulative per-arm sample size
Stage 1	0.0700 (1.476)	0.97	207
Stage 2	0.0135 (2.212)	0.82	743
Five experimental treatment arms ($K = 5$)			
	α (critical value)	ω	Cumulative per-arm sample size
Stage 1	0.0400 (1.751)	0.97	244
Stage 2	0.0060 (2.511)	0.82	895

Using the designs, we evaluated performance across a range of values for the underlying treatment effect of T_1 on the definitive outcome, denoted θ_{1D}. The effect of T_1 on the early outcome was held constant at θ_I^R. For each value of θ_{1D}, we calculated the percentage of

trials where any non-null treatment was declared beneficial at the end of the trial. Based on the sample sizes specified for the chosen three-arm ($K = 2$) and six-arm ($K = 5$) designs, we simulated individual patient data for 100 000 trials for each value of θ_{1D} under two different scenarios. In the first scenario, all other experimental treatments other than T_1 were ineffective on both early and definitive outcomes. In the second scenario, other experimental treatments were partially effective, with treatment effects equal to $\theta_{1D}/4$ for the definitive outcome and held constant at $\theta_I^R/4$ for the early outcome. Using a threshold rule initially, we compared the performance of the MAMS(R) framework and the combination method. We then implemented an epsilon rule. We set $\varepsilon = 1$ to emulate a moderately stringent rule, partway between selecting one treatment and selecting all treatments. Again we compared the performance of the MAMS(R) framework and the combination method.

3.3.2 Trials when $I = D$

The trial motivating the simulation study is a two-stage Phase 2 superiority trial as described by Bratton.[18] A one sided FWER of 0.025, a pair-wise power of 0.8 and a 1:1 allocation ratio are specified. Control arm event rates and treatment effects are the same for both stages of the trial and are as described for the D outcome in Section 3.3.1. Using the approach described for $I \neq D$ we obtained MAMS(R) designs for two-stage three-arm ($K = 2$) and six-arm ($K = 5$) trials where $I = D$. The chosen MAMS(R) designs are described in Table 2.

Table 2. Summary of two stage $I = D$ designs used in simulation study

Two experimental treatment arms ($K = 2$)			
	α (critical value)	ω	Cumulative per-arm sample size
Stage 1	0.2300 (0.739)	0.94	92
Stage 2	0.0160 (2.144)	0.94	250
Five experimental treatment arms ($K = 5$)			
	α (critical value)	ω	Cumulative per-arm sample size
Stage 1	0.1900 (0.878)	0.95	113
Stage 2	0.0070 (2.457)	0.93	286

We compared the performance of the MAMS(R) framework and the combination method in the same manner as for $I \neq D$. Since for $I = D$ the intermediate and definitive outcome are the same, we do not use the subscript D for θ, the underlying treatment effect for T_1 being simply denoted θ_1. As before, we simulated individual patient data for 100 000 trials for each value of θ_1 under two different scenarios such that in the first, all other experimental treatments other than T_1 were ineffective and in the second, other experimental treatments were partially effective, with treatment effects equal to $\theta_1/4$. As for $I \neq D$, the performance of MAMS(R) framework and the combination method were compared when a threshold rule is used. We then implemented an epsilon rule $(\varepsilon = 1)$ for the combination method and used the new hybrid rule for the MAMS(R) framework.

4 RESULTS

4.1 Trials when I ≠ D

In this section, two sets of results are presented relating to the case where $I \neq D$. The first gives a direct comparison of performance between the MAMS(R) framework and the combination method when both implement a threshold selection rule, this reflects the usual mode of operation for the MAMS(R) framework. The second set gives a further comparison of performance to show the effect of implementing a different selection rule.

4.1.1 Comparison of the MAMS framework and the combination method using a threshold selection rule

Table 3 [insert Table 3] presents estimated probabilities to declare effectiveness on the final outcome across a range of values for θ_{1D}, firstly for any non-null treatments and secondly for null or partially effective treatments only. Results for the three-arm design ($K = 2$) are presented in the upper section of the table and for the six-arm design ($K = 5$) in the lower section. On the left-hand side of the table results are presented for scenarios where treatments other than T_1 are ineffective on both the early and the final outcome ($\theta_{iI} = \theta_I^0$, $\theta_{iD} = \theta_D^0 \; for \; all \; i \neq 1$) while results for scenarios where treatments other than T_1 are partially effective on both the early and final outcome ($\theta_{iI} = \theta_I^R/4$, $\theta_{iD} = \theta_{1D}/4 \; for \; all \; i \neq 1$) are given on the right-hand side. The rows of the table refer to the different values of θ_{1D} investigated. Results in bold show the percentage of trials in which any non-null treatment is declared beneficial, for different values of θ_{1D} (the effect of T_1 on the early outcome being

held constant at θ_I^R). The results in parentheses give the percentage of trials in which at least one of the null or partially effective treatments is declared beneficial.

In Table 3, the results in bold show that under a threshold selection rule the combination method results in marginally greater power than the MAMS(R) framework. This general finding is observed for the three-arm ($K = 2$) and the six-arm design ($K = 5$) and across all scenarios and treatment effects investigated. The slight power advantage of the combination method over the MAMS(R) framework is larger for the six-arm design ($K = 5$) than for the three-arm design ($K = 2$). However, the advantage is somewhat less for scenarios where partially effective treatments are present compared with scenarios where treatments other than T_1 are ineffective. The results in parentheses on the left-hand side of Table 3 show that when treatments other than T_1 are ineffective, the percentage of trials in which null treatments are declared effective is very low for both methods, as expected. As θ_{1D} increases, this percentage increases slightly for the combination method because for any given trial, the presence of the more effective treatment makes rejection of any intersection hypothesis which encompasses the null hypothesis for this treatment more likely. This increase does not occur for the MAMS(R) framework where the progress of individual treatment arms is not affected by the performance of other treatments. The reason why the percentages increase substantially for $\theta_{1D} = \theta_D^0$ is because when T_1 is ineffective on the final outcome, it will be more likely than other treatments to progress to the second stage and be declared effective on the final outcome due to the early outcome effect being held constant at θ_I^R for T_1 across all values of θ_{1D}. The percentage is much lower than 2.5% because the trials are designed such that the target FWER is 2.5% when all treatments are fully effective on the early outcome but ineffective on the final outcome (see Section 2.2). As θ_{1D} increases, there is a sharp increase in the percentage of trials in which partially effective treatments are declared effective, shown

by the results in parentheses on the right-hand side of the Table 1. This is an expected finding when selection is determined by a threshold. The rate tends to be slightly lower for MAMS(R) than for the combination method.

4.1.2 Performance of the MAMS(R) framework and the combination method under the epsilon selection rule

In Figure 1 [insert Figure 1], power curves are presented showing the performance of the MAMS(R) framework and the combination method under both the threshold and the epsilon selection rule. The upper sets of four lines are obtained by plotting the percentage of trials where any non-null treatment is declared effective on the final outcome, for different values of θ_{1D}. The lower sets of four lines show the percentage of trials where at least one null or partially-effective treatment is declared beneficial on the final outcome. Panels i) and ii) show results for the three-arm ($K = 2$) design and panels iii) and iv) for the six-arm ($K = 5$) design. In panels i) and iii), results are presented for scenarios where treatments other than T_1 are ineffective on both the early and the final outcome ($\theta_{iI} = \theta_I^0\ ,\ \theta_{iD} = \theta_D^0\ \ for\ all\ i \neq 1$). Results for scenarios where treatments other than T_1 are partially effective on both the early and final outcome ($\theta_{iI} = \theta_I^R/4\ ,\ \theta_{iD} = \theta_{1D}/4\ \ for\ all\ i \neq 1$) are shown in panels ii) and iv).

Considering the upper sets of lines in Figure 1, the percentage of trials where a non-null treatment is declared effective is consistently greater when an epsilon rule is used in place of the threshold rule. This is true for both the MAMS(R) framework and the combination method and reflects the operation of the epsilon selection rule at the interim analysis, allowing the most effective treatment through to the second stage even when the threshold

required by the other methods has not been met. The separation resulting from the change in selection rules is larger in the context of the combination method than in the MAMS(R) framework, this is most obvious at the higher values of θ_{1D} investigated and for the scenarios where partially effective treatments are present (panels ii) and iv)). As discussed in Section 4.1.1, under a threshold rule the combination method is marginally more powerful than the MAMS(R) framework across all the scenarios investigated, although there is less difference between the two methods when partially effective treatments are present. Under an epsilon rule the combination method is again more powerful than the MAMS(R) framework, but the advantage tends to be larger and is not reduced when partially effective treatments are present. For the six-arm design where partially effective treatments are present (panel iv)) the combination method with the epsilon rule clearly provides the greatest power across all treatment effects.

Considering the lower sets of lines in Figure 1, it is clear that, compared with the threshold rule, implementing an epsilon selection rule substantially reduces the rate at which partially effective treatments are declared effective at the final analysis. In some settings this may be viewed as desirable. In the MAMS(R) framework the usual use of a threshold rule facilitates the objective of declaring any non-null treatment(s) effective whereas moving away from the threshold towards an epsilon selection rule results in a more directed result, with greater power to select the best treatment and a reduced probability of declaring inferior treatments beneficial

.

4.2 Trials when I = D

In this section, results for the case where $I = D$ are considered. As before, two sets of results are presented, the first set relating to a direct comparison under a threshold selection rule and the second set showing the effect of implementing different selection rules; results are given for the combination method under the threshold and the epsilon rule and for the MAMS(R) framework under the threshold and the hybrid rule (see Section 3.2).

4.2.1 Comparison of the MAMS(R) framework and the combination method using a threshold selection rule

Table 4 [insert Table 4] presents estimated probabilities to declare effectiveness, firstly for any non-null treatment and secondly for any null or partially effective treatment(s). The structure of the table is as for Table 3. Note that on the left-hand side of the table results are presented for scenarios where treatments other than T_1 are ineffective $(\theta_i = \theta^0 = 0 \; for \; all \; i \neq 1)$ while results for scenarios where treatments other than T_1 are partially effective $(\theta_i = \theta_1/4 \; for \; all \; i \neq 1)$ are given on the right-hand side.

In contrast to the $I = D$ case, the results in Table 4 show that under a threshold rule the MAMS(R) framework results in slightly greater power, compared with the combination method. This opposite finding may be due to the fact that when $I = D$, there is a binding threshold at stage one and this allows for a more liberal critical value at stage two compared with the $I \neq D$ case. This general finding is observed for both the three-arm $(K = 2)$ and the six-arm design $(K = 5)$ and across all scenarios and treatment effects investigated. It was also verified for an alternative trial scenario which had different treatment effects and stage-

wise sample sizes (results not shown). The power advantage of the MAMS(R) framework over the combination method is marginal, but is greater for the scenarios where a large number of partially effective treatments are present. The results in parentheses on the left-hand side of Table 4 show the percentage of trials in which null treatments are declared effective. Under the global null hypothesis ($\theta_i = \theta^0 = 0 \; for \; all \; i$) the estimated FWER is larger for the MAMS(R) framework than for the combination method. However, at most of the other treatment effects investigated, null treatments are declared beneficial at a similar or lower rate for the MAMS(R) framework compared with the combination method. For the reasons described in the context of Table 3, as θ_1 increases this rate rises slightly for the combination method, but not for the MAMS(R) framework. As θ_1 increases, there is a substantial increase in the percentage of trials in which partially effective treatments are declared effective, shown by the results in parentheses on the right-hand side of Table 4. For the three-arm design ($K = 2$) the rate tends to be lower for MAMS(R) than for the combination method whereas for the six-arm design ($K = 5$) it is slightly greater for MAMS(R) across all values of θ_1.

4.2.2 Performance of the MAMS(R) framework and the combination method under different selection rules

In Figure 2 [insert Figure 2], power curves are presented for four different schemes: the MAMS(R) framework and the combination method under the threshold rule, the combination method under the epsilon rule and the MAMS(R) framework under the hybrid rule. The layout of the figure is as described for Figure 1. Note that in panels i) and iii) results are presented for scenarios where treatments other than T_1 are ineffective ($\theta_i = \theta^0 = 0 \; for \; i \neq$

1) while results for scenarios where treatments other than T_1 are partially effective ($\theta_i = \theta_1/4 \; for \; i \neq 1$) are shown in panels ii) and iv).

Looking at the upper sets of lines, for the combination method power is consistently greater when an epsilon rule rather than a threshold rule is implemented. The differences become larger as θ_1 increases, reflecting the operation of the epsilon selection rule as discussed in Section 4.1.2. The separation resulting from the change in selection rule is most obvious for higher values of θ_1, because at lower values of θ_1 even if T_1 is selected at an interim it would be unlikely to be declared effective on the final outcome at the end of stage two. However, in the MAMS(R) framework, when the hybrid selection rule replaces the threshold rule the percentage of trials where T_1 is declared effective is slightly reduced because the hybrid rule is a more stringent selection rule than the threshold. As discussed in Section 4.2.1, under the threshold rule the MAMS(R) framework is more powerful than the combination method across all the scenarios investigated, particularly when a large number of partially effective treatments are present. Moving away from using a threshold rule to implementing the epsilon rule for the combination method or the hybrid rule for MAMS(R), this advantage reverses, at least for the majority of scenarios. For the three-arm trial ($K = 2$) the combination method under the epsilon rule gives greater power than the other schemes, particularly at larger treatment effects. However, for the six-arm trial when partially effective treatments are present, there is no clear advantage. The MAMS(R) framework under the threshold or hybrid rule results in similar power at higher treatment effects and better power at lower treatment effects compared with the combination method under the epsilon rule (see panel iv).

Looking at the lower sets of lines, implementing the epsilon or hybrid rule substantially reduces the rate at which null and partially effective treatments are declared beneficial at the

final analysis. It can be clearly seen in Figure 2 that as θ_1 increases, there is no steep rise in the proportion of partially effective treatments which are declared beneficial, such as is observed under the threshold rule, (see panels ii) and iv)). This is because as θ_1 increases the numerical distance between θ_1 and the treatment effect of the partially effective treatments increases and this will tend to reduce the number of trials where these arms progress even though the absolute value of the effect in these arms is increasing. Across all the scenarios we investigated, the MAMS(R) framework under the hybrid selection rule achieved consistently lower rates for recommending null or partially effective treatments compared to any other scheme. This result can be seen clearly by noting the relative position of the lines in the lower section of each panel in Figure 2. The black dashed line showing the results for the MAMS(R) framework under the hybrid rule consistently occupies a lower position than the other lines.

5. DISCUSSION

By adapting and implementing recent developments in methodology, we have used the MAMS(R) framework to obtain efficient boundary based trial designs for multi-stage adaptive trials where the outcomes are binary and where treatment effects are parameterised as the LOR. Since methodology now allows the FWER to be controlled in MAMS(R) trials, we were able to carry out a simulation study to make an in-depth comparison of MAMS(R) trials with the well-established combination method in multi-arm multi-stage trials incorporating treatment selection, both for trials when $I \neq D$ and for trials when $I = D$.

For trials when $I \neq D$, we found that the combination method achieves greater power than the MAMS(R) framework across all scenarios investigated. This was the case both under a

threshold selection rule and an epsilon rule. The advantage of the combination method over MAMS(R) is most clearly seen for the six-arm ($K = 5$) design and when an epsilon rule is implemented. The reason why the combination method is more powerful may be that MAMS(R) designs for trials where $I \neq D$ tend to be inherently conservative. The conservatism occurs because, to ensure the FWER is strongly controlled, the critical value for the final stage is determined assuming that treatments are fully effective on the I outcome, as explained in Section 2.2. For both the MAMS(R) framework and the combination method, power is greater if an epsilon rule rather than a threshold rule is used.

In contrast, however, we found that for $I = D$ trials, where this conservative approach is not required, the MAMS(R) framework achieves slightly greater power than the combination method when a threshold selection rule is used. This finding is observed across all scenarios, irrespective of the size of the treatment effect or whether partially effective treatments are present. Generally, the differences are slightly greater for the six-arm ($K = 5$) design and when partially effective treatments are present. One possible reason for the combination method having less power is that the combining of evidence from the two stages of the trial means that final comparisons of treatments may not be based on a sufficient statistic for the treatment difference; this has been suggested for the single arm setting by authors such as Jennison and Turnbull[12] and Kelly et al.[14] We also showed that a hybrid selection rule can be implemented in the MAMS(R) framework to facilitate a more comparative selection procedure. However, when comparing the combination method under the epsilon rule with the MAMS(R) framework under the hybrid rule, we found that MAMS(R) no longer has a consistent advantage, the combination method achieving similar or greater power in some scenarios. We found that the rate at which partially effective treatments are recommended is

lower for MAMS(R) under the hybrid rule than for any other scheme we investigated including the combination method under the epsilon rule.

In this paper we have explored the use of the MAMS(R) framework to obtain boundary based trial designs. This approach has the advantage of being relatively simple to understand and implement and of accommodating treatment selection based either on the definitive outcome or purely on an early outcome measure. We acknowledge that the MAMS(R) framework is mainly appropriate for trials where no early stopping for efficacy is envisaged. In contrast, the multi-arm group sequential designs developed by Magirr et al[10] specify both efficacy and futility boundaries so that trial designs which incorporate early stopping for efficacy may be obtained.

Based on our findings, we suggest that for multi-arm two-stage trials with binary outcomes where $I \neq D$, the combination method may be a more suitable choice than MAMS(R), particularly for trials with many treatment arms. For either method, the selection rule which best meets the objectives of the trial can be chosen. Since the stage one critical value is not binding, an epsilon rule may be implemented in the MAMS(R) context without inflating the FWER. This rule was shown to increase power compared with the threshold rule. By contrast, for trials where $I = D$, if the objectives of the trial are best met by using a threshold selection rule, the MAMS(R) framework may be a more suitable option than the combination method, particularly for trials with a substantial number of experimental arms and where partially effective treatments are likely to be present. Our results suggest that by implementing the hybrid rule, the MAMS(R) framework may also be successfully used for trials where the aim is to recommend the best treatments and that this may provide an effective way to minimise the probability of inferior but partially effective treatments being

declared effective at the end of the trial. However, the more stringent hybrid rule does mean that some of the power advantage of MAMS(R) over the combination method seen under the threshold rule is lost. Where the main treatment effect is likely to be large and other treatments likely to be ineffective, the combination method under the epsilon rule may be a better choice since we found it achieves greater power in these scenarios. However, for a proposed trial with many treatment arms where some are likely to be partially effective and it is desirable to minimise the rate at which these are recommended, we suggest that MAMS(R) under the hybrid rule should be considered since it provides comparable power to the combination method whilst keeping the rate for inferior treatments substantially lower. Since no method consistently out-performs the others, the choice of which method to adopt for a given trial is best considered on an individual trial basis. We recommend that simulations based on the specific context and objectives of a particular trial should be conducted at the outset and the results used to determine which approach is the most suitable.

Finally, in this study only two-stage trials were considered. Both the MAMS(R) and combination methodologies described in this paper can readily extend to include more than two stages[19,33], this is a possible area for future work. Similarly, now that methodology exists for calculating FWER in the context[30] of trials with survival outcomes, it would be useful to develop methodology for feasible and admissible designs for this context such that further comparisons between MAMS(R) and the combination method may be conducted.

Acknowledgements

This work was supported by the EPSRC [PhD studentship to JEA]. We thank the two referees for their comments which greatly improved this manuscript. The authors declare that there are no conflicts of interest.

References

1. Bretz F et al. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: general concepts. *Biometrical Journal* 2006; 48: 623-634

2. Schmidli H et al. Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal* 2006; 48: 635-643

3. Friede T and Stallard N, A comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; 50: 767-781

4. Koenig F et al. Adaptive Dunnett tests for treatment selection. *Statistics in Medicine* 2008; 27:1612–1625

5. Bauer P and Kohne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994; 50:1029-1041

6. Bauer P and Kieser M. Combining different phases in the development of medical treatments within a single trial. *Statistics in Medicine* 1999; 18: 1833-1848

7. Marcus R, Peritz E and Gabriel K. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* 1976; 63: 655-660

8. Stallard N and Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. *Statistics in Medicine* 2003; 22: 689-703

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

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

11. Royston et al. Novel designs for multi-arm clinical trials with survival outcomes with an application in ovarian cancer. *Statistics in Medicine* 2003; 22: 2239-2256

12. Jennison C and Turnbull B. Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine* 2003; 22: 971-973

13. Tsiatis A and Mehta C. On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika* 2003; 90: 367-378

14. Kelly P, Sooriyarachchi R et al. A practical comparison of group-sequential and adaptive methods. *Journal of Biopharmaceutical Statistics* 2005; 15: 719-738

15. Stallard N and Todd S. Seamless phase II/III designs. *Statistical Methods in Medical Research* 2010; 20(6): 623-634

16. Friede T and Stallard N. a comparison of methods for adaptive treatment selection. *Biometrical Journal* 2008; 50: 767-781

17. Kunz C et al. A comparison of methods for treatment selection in seamless phase II/III clinical trials incorporating information on short-term endpoints. *Journal of Biopharmaceutical Statistics* 2015; 25:170-189

18. Bratton D, Phillips P, Parmer M. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Medical Research Methodology* 2013; 13: 139-153

19. Bratton D. *Design issues and extensions of multi-arm multi-stage clinical trials*. PhD Thesis. University College London, UK, 2015

20. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; 10:1–10.

21. Jung SH et al. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 2004; 23:561-569

22. Mander A et al. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics* 2012; 11(2):91-96

23. Wason J and Jaki T. Optimal design of multi-arm multi-stage trials. *Statistics in Medicine* 2012;31: 4269-4279

24. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical association* 1955; 50: 1096-1121

25. Brannath W, Posch M, Bauer P. Recursive combination tests. *Journal of the American Statistical association* 2002; 97:236-244

26. Lehmacher W and Wassmer G. Adaptive sample size calculations in group sequential trials. *Biometrics* 1999; 55:1286-1290

27. Posch M et al. Testing and estimation in flexible group sequential designs with adaptive treatment selection. *Statistics in medicine* 2005; 24: 3697-3714

28. Friede T et al. Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine* 2011; 30(13): 1528-1540

29. Siqueira A, Todd S and Whitehead. Sample size considerations in active-control non-inferiority trials with binary data based on the odds ratio. A. *Statistical Methods in Medical Research* 2015; 24(4): 453-461)

30. Bratton D, Choodari-Oskooei B, Royston P. A menu-driven facility for sample-size calculation in mutiarm, multistage randomised controlled trials with time-to-event outcomes:Update. *The Stata Journal* 2015;15(2):350-368

31. Kelly P, Stallard N, Todd S.An adaptive group sequential design for phase II/III clinical trials that select a single treatment from several. *Journal of Pharmaceutical Statistics* 2005; 15(4): 641-658

32. Parsons et al. An R package for implementing simulations for seamless phase II/III clinical trials using early outcomes for treatment selection. *Computational Statistics and Data Analysis* 2012; 56: 1150-1160

33. Wassmer G, Eisebitt R, Coburger S. Flexible interim analyses in clinical trials using multistage adaptive test designs.*Drugs Information Journal* 2010; 35(4): 1131-1146

34. Royston P et al. Designs for clinical trials with time-to-event outcomes based on stopping guidelines for lack of benefit. *Trials* 2011;12: 81

Table 3. Comparison of power for MAMS(R) framework and the combination method under a threshold selection rule for trials where $I \neq D$

	% trials Treatment 1 declared beneficial *(% trials where one or more null treatment(s) declared beneficial)*				**% trials any non-null treatment declared beneficial** *(% trials where one or more partially effective treatment(s) declared beneficial)*			
$\boldsymbol{\theta_{1D}}$	$\boldsymbol{K = 2\ (\theta_{2D} = \theta_D^0)}$				$\boldsymbol{K = 2\ (\theta_{2D} = \theta_{1D}/4)}$			
	Combination		MAMS(R)		Combination		MAMS(R)	
0.077	**88.84**	*(0.40)*	**87.97**	*(0.25)*	**88.15**	*(6.42)*	**87.93**	*(4.5)*
0	**80.95**	*(0.40)*	**79.59**	*(0.25)*	**80.12**	*(5.45)*	**79.74**	*(3.73)*
-0.077	**69.61**	*(0.37)*	**67.50**	*(0.21)*	**68.52**	*(4.59)*	**67.71**	*(3.18)*
-0.154	**54.66**	*(0.36)*	**52.11**	*(0.22)*	**53.34**	*(3.67)*	**52.25**	*(2.58)*
-0.231	**38.21**	*(0.39)*	**35.57**	*(0.24)*	**37.54**	*(2.88)*	**35.95**	*(1.99)*
-0.308	**23.24**	*(0.34)*	**21.01**	*(0.23)*	**22.82**	*(2.18)*	**21.61**	*(1.57)*
-0.385	**12.07**	*(0.28)*	**10.37**	*(0.23)*	**11.93**	*(1.49)*	**11.1**	*(1.20)*
-0.462	**5.08**	*(0.22)*	**4.19**	*(0.23)*	**5.37**	*(1.04)*	**5.01**	*(0.97)*
-0.539	**1.82**	*(1.97)*	**1.43**	*(1.63)*	**2.08**	---	**1.97**	---
	$\boldsymbol{K = 5\ (\theta_{2D} = \theta_{3D} = \theta_{4D} = \theta_{5D} = \theta_D^0)}$				$\boldsymbol{K = 5\ (\theta_{2D} = \theta_{3D} = \theta_{4D} = \theta_{5D} = \theta_{1D}/4)}$			
	Combination		MAMS(R)		Combination		MAMS(R)	
0.077	**90.71**	*(0.36)*	**88.87**	*(0.25)*	**89.24**	*(9.06)*	**88.88**	*(7.52)*
0	**83.13**	*(0.33)*	**79.99**	*(0.22)*	**80.89**	*(7.48)*	**80.19**	*(6.21)*
-0.077	**70.85**	*(0.35)*	**66.46**	*(0.24)*	**68.04**	*(5.94)*	**66.71**	*(4.94)*
-0.154	**54.57**	*(0.36)*	**49.22**	*(0.24)*	**51.55**	*(4.79)*	**49.8**	*(3.95)*
-0.231	**36.91**	*(0.34)*	**31.56**	*(0.23)*	**33.98**	*(3.52)*	**31.93**	*(3.00)*
-0.308	**20.97**	*(0.33)*	**16.73**	*(0.25)*	**19.33**	*(2.57)*	**17.47**	*(2.31)*
-0.385	**9.92**	*(0.31)*	**7.26**	*(0.24)*	**9.47**	*(1.77)*	**8.21**	*(1.69)*
-0.462	**3.88**	*(0.26)*	**2.51**	*(0.24)*	**3.82**	*(1.19)*	**3.36**	*(1.29)*
-0.539	**1.11**	*(1.25)*	**0.65**	*(0.81)*	**1.44**	---	**1.38**	---

(--- denotes scenarios where no treatments which are partially effective on the final outcome are present)

Figure 1. Comparison of the MAMS(R) framework and combination method under threshold and epsilon selection rules for trials where $I \neq D$.. Upper lines are estimated power to declare any non-null treatment beneficial and lower lines show the percentage of trials where at least one null or partially-effective treatment is declared beneficial.

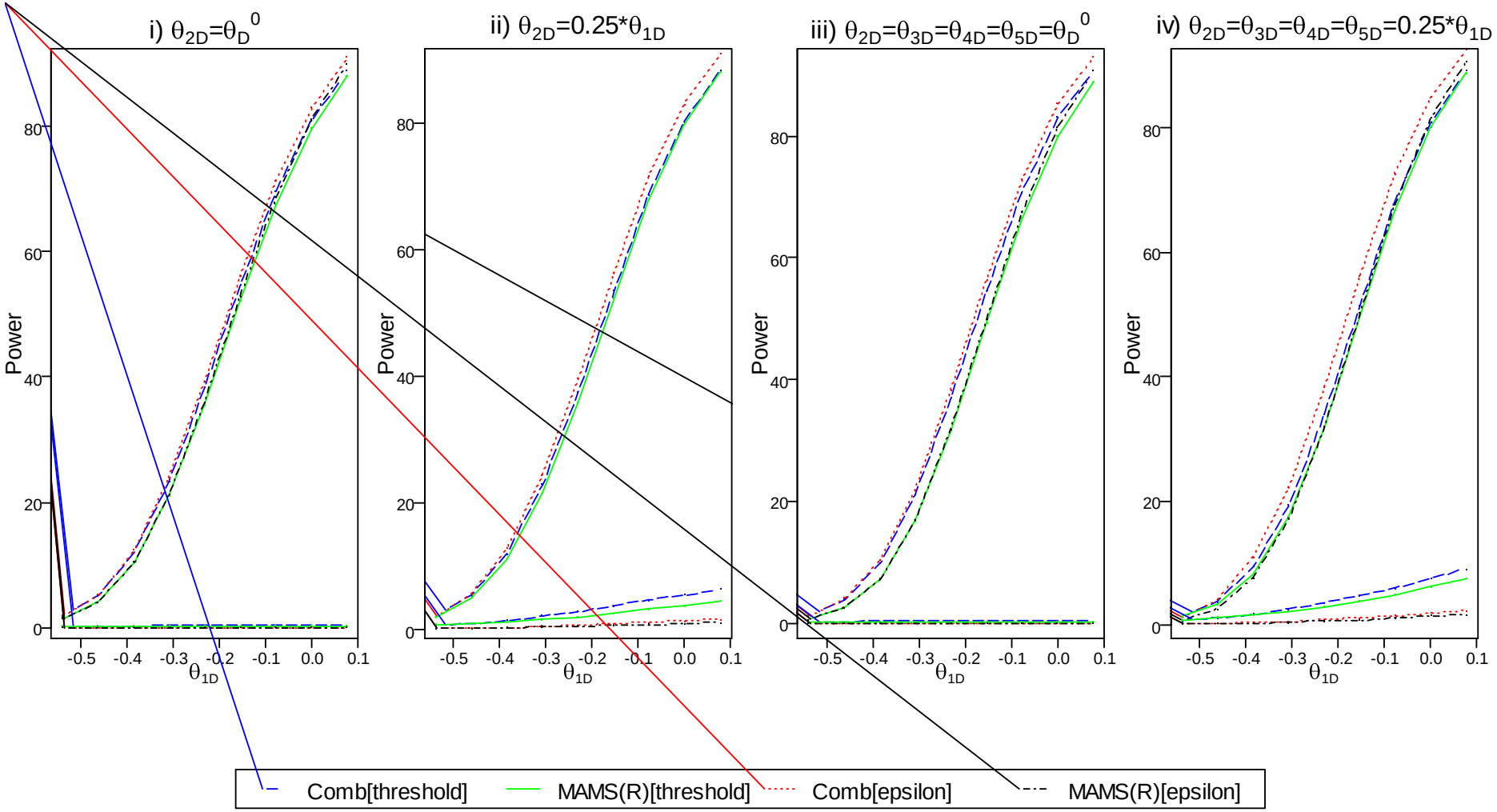
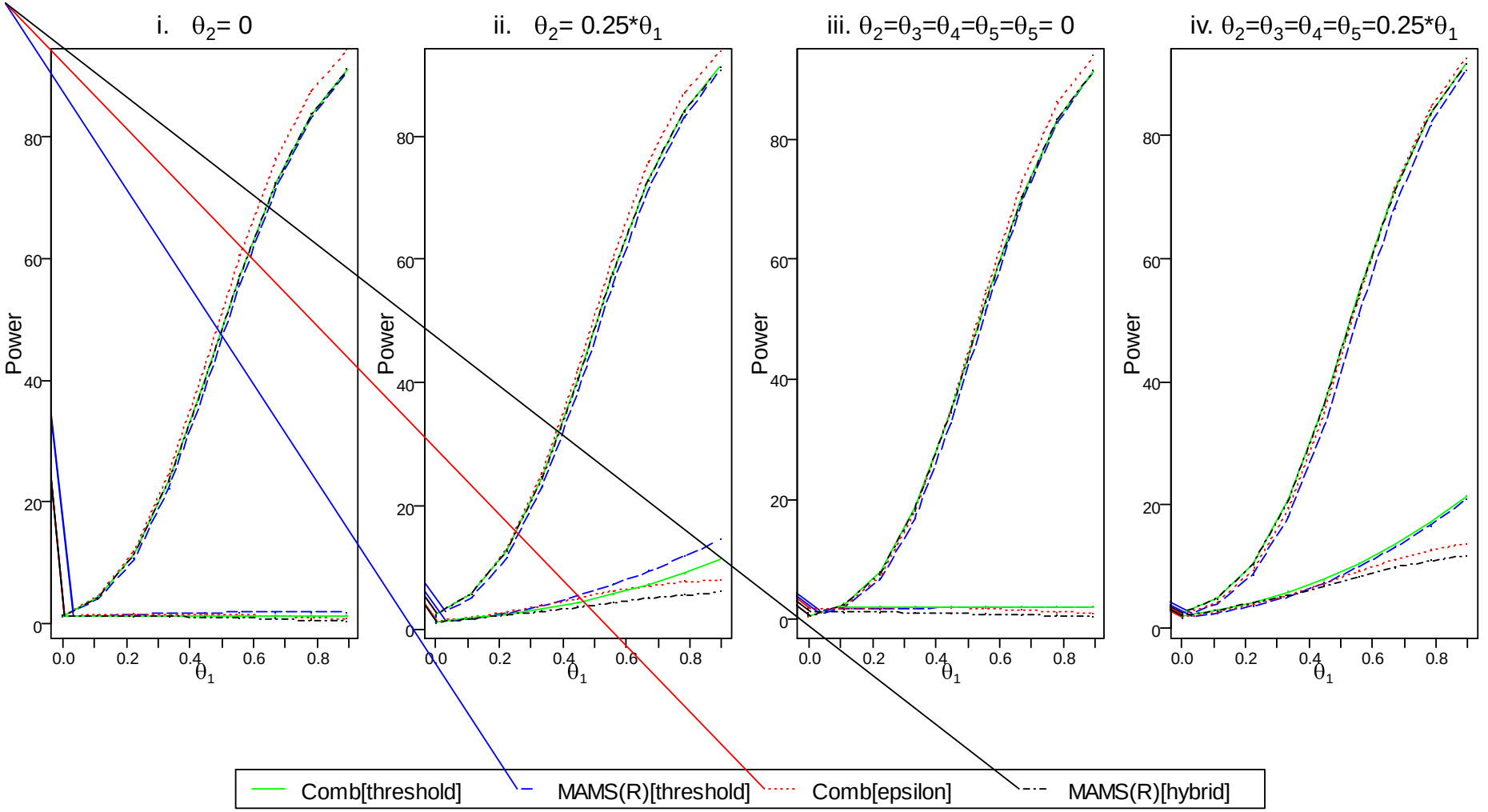


Table 4. Comparison of power for MAMS(R) framework and the combination method under a threshold selection rule for trials where $I = D$

	% trials Treatment 1 declared beneficial *(% trials where one or more null treatment(s) declared beneficial)*				**% trials any non-null treatment declared beneficial** *(% trials where one or more partially effective treatment(s) declared beneficial)*			
$\boldsymbol{\theta_1}$	$\boldsymbol{K = 2\ (\theta_2 = 0)}$				$\boldsymbol{K = 2\ (\theta_2 = \theta_1/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.894	**90.83**	*(1.94)*	**91.10**	*(1.29)*	**90.58**	*(14.58)*	**91.20**	*(11.43)*
0.782	**83.18**	*(1.93)*	**83.77**	*(1.29)*	**82.98**	*(11.87)*	**84.10**	*(9.16)*
0.67	**71.46**	*(1.94)*	**72.48**	*(1.31)*	**71.27**	*(9.48)*	**72.92**	*(7.31)*
0.558	**55.82**	*(1.86)*	**57.23**	*(1.29)*	**55.7**	*(7.31)*	**57.75**	*(5.70)*
0.447	**38.26**	*(1.80)*	**39.85**	*(1.31)*	**38.63**	*(5.47)*	**40.74**	*(4.39)*
0.335	**22.18**	*(1.65)*	**23.57**	*(1.31)*	**23.08**	*(3.93)*	**24.83**	*(3.36)*
0.224	**10.64**	*(1.48)*	**11.51**	*(1.30)*	**11.68**	*(2.67)*	**12.84**	*(2.47)*
0.112	**4.06**	*(1.32)*	**4.43**	*(1.30)*	**5.11**	*(1.78)*	**5.73**	*(1.81)*
0	**1.20**	*(2.13)*	**1.304**	*(2.42)*	**2.13**	---	**2.43**	---
	$\boldsymbol{K = 5\ (\theta_2 = \theta_3 = \theta_4 = \theta_5 = 0)}$				$\boldsymbol{K = 5\ (\theta_2 = \theta_3 = \theta_4 = \theta_5 = \theta_1/4)}$			
	Combination		**MAMS(R)**		Combination		**MAMS(R)**	
0.894	**91.36**	*(2.08)*	**91.43**	*(2.09)*	**90.55**	*(20.93)*	**91.66**	*(21.43)*
0.782	**82.99**	*(2.06)*	**83.35**	*(2.06)*	**81.75**	*(16.71)*	**83.38**	*(17.21)*
0.67	**69.44**	*(2.05)*	**70.40**	*(2.06)*	**68.23**	*(13.17)*	**71.21**	*(13.65)*
0.558	**51.54**	*(2.06)*	**53.25**	*(2.07)*	**50.8**	*(10.00)*	**54.48**	*(10.51)*
0.447	**32.56**	*(1.99)*	**34.63**	*(2.10)*	**32.89**	*(7.31)*	**36.52**	*(7.94)*
0.335	**16.73**	*(1.89)*	**18.54**	*(2.08)*	**18.16**	*(5.19)*	**20.97**	*(5.87)*
0.224	**6.81**	*(1.73)*	**7.90**	*(2.06)*	**8.7**	*(3.58)*	**10.50**	*(4.26)*
0.112	**2.08**	*(1.66)*	**2.53**	*(2.07)*	**3.92**	*(2.41)*	**4.95**	*(3.01)*
0	**0.49**	*(1.94)*	**0.59**	*(2.52)*	**1.93**	---	**2.47**	---

(--- denotes scenarios where no treatments which are partially effective on the final outcome are present)

Figure 2. Comparison of the MAMS(R) framework and combination method under threshold and epsilon selection rules for trials where $I = D$. Upper lines are estimated power to declare any non-null treatment beneficial and lower lines show the percentage of trials where at least one null or partially-effective treatment is declared beneficial.



Appendix

Adapting trial designs in the MAMS(R) framework for the LOR

In order to compute the designs in this paper we modified MAMS(R) programs for Stata as follows. The expression for calculating the control arm sample size, denoted n_C, was changed to the formula based on the LOR, $n_C = \{p_C(1-p_C) + p_E(1-p_E)/(p_E - p_C)^2\}(z_{1-\alpha} + z_\omega)^2$. Sample sizes are calculated at each stage with p_C and p_E being determined by θ_j^0 and θ_j^R, and α and ω relating to the stage-wise alpha and power of the given design. For a 1:1 allocation ration, the suggested sample size for each experimental arm, denoted n_E, is equal to n_C.

The formula provides an approximate sample size but due to the discrete nature of binary data, target Type I error and power may not be achieved exactly. Siqueira et al[27] investigated the accuracy of the Wald-Type formula and reported that sample sizes obtained under the LOR may deviate from the true requirement, with sample sizes tending to be overestimated. We found that under the LOR, there was some deviation from the target stage-wise Type I error and that this occurred rather more than under the original parameterisation. We incorporated a simulation based routine to check over sample sizes in the near neighbourhood of the value suggested by the formula, selecting the size which achieved a Type I error rate closest to the target for that stage. We explored searching over sample sizes within three, five and ten units of the value suggested by the formula. We chose to use +/- 5 as we found that a search of this size provided improvement in Type 1 error accuracy and could be conducted in a reasonable timeframe.

The numerical calculation of FWER based on simulation of normally distributed test statistics under $H_{0(G)}$ remains the same under the LOR although the correlation matrix will reflect different stage-wise sample sizes under the LOR. However, some changes to the routine where admissible designs are identified are required, where estimates for both $E(N|H_{0(G)})$ and $E(N|H_{R(G)})$ are obtained, since these estimates require that the probabilities that k out of K treatments pass to the next stage under each stated hypothesis are known (see Section 2.2). The simulated test statistics used to obtain these probabilities are generated using Equation 5.1 in Bratton.[19] The final term of this expression disappears under $H_{0(G)}$ but not under $H_{R(G)}$ and hence, in our routines the variance of the treatment difference which is included in this final term is changed to reflect the LOR, using $Var(\theta) = 1/n_C p_C + 1/n_C(1-p_C) + 1/n_E p_E + 1/n_E(1-p_E)$, with p_C ,p_E , n_C and n_E being defined for each stage.

We verified that when $I = D,$ under the LOR the between stage correlation of test statistics remains unchanged at $\sqrt{n_C^1/n_C^2}$ where n_C^1 and n_C^2 are the control-arm sample sizes at stage one and stage two respectively. For the case when $I \neq D$, Bratton[19] derived an expression based on the parameterisation 'difference in proportions' for the between stage correlation between early and definitive test statistics. This expression requires an estimate of the positive predictive value (PPV), which is the probability that an individual will have a positive outcome on the definitive outcome given that the outcome for the early outcome was positive, usually obtained by reference to previous trials. We were unable to obtain a similar analytical expression based on the LOR. In the context of survival outcomes, if $I = D$ between-stage correlations for log hazard ratios (LHR) can be expressed analytically as $\sqrt{(e_C^1/e_C^2)}$, where e_C^1 and e_C^2 are the number of control arm events observed on the outcome of interest at stage one and stage two respectively.[34] However, the correlations appear to be intractable when $I \neq D$.[34] Bratton et al[30] suggest that the correlation may be approximated

using $c\sqrt{(e_C^1/e_C^2)}$, where c is an attenuating constant which is an estimate of the correlation between LHRs for I and D, obtained when full data on both outcomes are known; this constant may be obtained from expert opinion based on previous similar trials. Alternatively he proposes a simulation approach where the between-stage correlations are obtained using information about survival outcome correlations from individual patients. We adapted these ideas for the binary context. We approximated the between stage correlations of the early and definitive test statistics using $c\sqrt{n_C^1/n_C^2}$ where c is the estimated correlation between LORs for I and D. We obtained an estimate for c by generating individual patient data using estimates of the PPV, simulating 100 000 trials and obtaining the correlation between the LORs for I and D. We tested the validity of this approach using the 'difference in success probabilities' parameterisation where we could compare the correlations obtained by simulation with those obtained using the analytical expression and found very good agreement with stage-wise correlations agreeing to at least two decimal places.