Published in final edited form as:
Stat Med. 2022 April 30; 41(9): 1613–1626. doi:10.1002/sim.9314.

An order restricted multi-arm multi-stage clinical trial design

Alessandra Serra[1], **Pavel Mozgunov**[1], **Thomas Jaki**[1,2]

[1]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

[2]Department of Mathematics and Statistics, Lancaster University, Lancaster, UK

Abstract

One family of designs that can noticeably improve efficiency in later stages of drug development are multi-arm multi-stage (MAMS) designs. They allow several arms to be studied concurrently and gain efficiency by dropping poorly performing treatment arms during the trial aswell as by allowing to stop early for benefit. Conventional MAMS designs were developed for the setting, in which treatment arms are independent and hence can be inefficient when an order in the effects of the arms can be assumed (eg,when considering different treatment durations or different doses). In this work, we extend the MAMS framework to incorporate the order of treatment effects when no parametric dose-response or duration-responsemodel is assumed. The design can identify all promising treatments with high probability. We show that the design provides strong control of the family-wise error rate and illustrate the design in a study of symptomatic asthma. Via simulations we show that the inclusion of the ordering information leads to better decision-making compared to a fixed sample and aMAMS design. Specifically, in the considered settings, reductions in sample size of around 15% were achieved in comparison to a conventional MAMS design.

Keywords

adaptive designs; infectious diseases; multi-arm multi-stage; order restriction

1 Introduction

Drug development is costly and time consuming. [1] One family of clinical trial designs that can improve the development process are multi-arm multi-stage designs (MAMS). [2–4] In aMAMS trial, insufficiently promising treatments can be dropped or the trial can be stopped due to overwhelming benefit at a series of interim analyses.

Correspondence to: Alessandra Serra.

Correspondence Alessandra Serra, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK. alessandra.serra@mrc-bsu.cam.ac.uk .

Author Contributions
All authors have directly participated in the planning and execution of the presented work.

Conflict of Interest
The authors declare no potential conflict of interests.

Europe PMC Funders Author Manuscripts

To date these designs have focused on the setting of independent treatment arms and have been argued to be a highly efficient approach to clinical trials. [5–7] They could, however, be suboptimal if an "order" (ie, a monotonic relationship) among the treatment effects can be assumed. Such an order can occur naturally, for example, when multiple doses or administration schedules of the same treatment are tested or when nested combinations of treatments are investigated. Another area where an order can often be assumed is when considering different treatment durations. In infectious diseases such as Tuberculosis (TB) and Hepatitis B (HBV), the treatment duration with current standard regimes is lengthy [8] which results in a large burden on the patients, potentially high costs, increased risk of non-compliance and side effects. [9] In TB and HBV, for example, treatment periods of 6 and 12 months are typical. [10,11] Novel treatments or combinations of treatments in these areas offer the opportunity for both higher efficacy and shorter treatment periods. [12] In the setting of multiple treatment durations Quartagno et al [13] have proposed to model the duration-response curve.While this is an efficient way to understand the duration-effect relationship, it is less clear how to definitively conclude whether a duration is "better" than the current standard.

In this work, we extend the MAMS framework and propose a design that incorporates the order of treatment effects in the decision-making when no parametric dose-response or duration-response model is assumed. The objective of the design is to identify all promising arms (eg, treatment durations, doses, or combination of treatments), including the one associated with the smallest relevant treatment effect.

The rest of the manuscript continues as follows. A case study is introduced in Section 2 before a detailed description of the 3-arm and 2-stage design is provided in Section 3. Section 4 then generalizes the proposed design to an arbitrary number of arms and stages and provides some theoretical results. Section 5 revisits the case study before the design is evaluated via simulations in Section 6. In Section 7, the effect of various critical bounds on the operating characteristics of the proposed design is explored.We conclude with a discussion.

2 Case Study Setting

The *Tiotropium add-on therapy in adolescents with moderate asthma: A 1-year randomized controlled trial* (NCT01257230) [14] is a Phase III study that assessed the efficacy and safety of once-daily tiotropium via Respimat added to inhaled corticosteroid (ICS) with or without a leukotriene receptor antagonist in adolescent patients with moderate symptomatic asthma. Patients were randomized with equal probability to receive 5 μg (2 puffs of 2.5 μg) or 2.5 μg (2 puffs of 1.25 μg) of once-daily tiotropium or placebo (2 puffs). The primary outcome was change from baseline in peak FEV_1 within 3 h after dosing (peak $FEV_{1[0-3h]}$) measured after 24 weeks of treatment. The null hypotheses were tested in a stepwise manner to control the type I error starting from the highest dose suggesting that a monotonic dose-response relationship can be assumed.

3 A 3-Arm 2-Stage Order Restricted Design

In this section, we develop an order restricted design (ORD) for the setting of the case study. We denote the highest dose *(5μg)* by T_1 and the lower dose *(2.5μg)* by T_2. The generalization to an arbitrary number of arms and stages is given in Section 4.

Assume that a patien's response follows a normal distribution with known common variance, σ^2. An alternative approach is outlined in Section 8 for the case of unknown variance. Let $X_i^{(k)} \sim N\left(\mu^{(k)}, \sigma^2\right), k \in \left\{0,1,2\right\}, i = 1{:}n_j^{(k)}$ be the observation of the ith patient on treatment k (the control arm is denoted by 0) and $n_j^{(k)}$ be the number of patients on arm k up to stage j. Let $\theta^{(k)} = \mu^{(k)} - \mu^{(0)}$ be the true treatment effect of active arm $k \in \{1,2\}$ compared to the control. We denote the vector of treatment effects by $\theta = \left(\theta^{(1)}, \theta^{(2)}\right)$. Consider the following order relationship: $\theta^{(1)} \geq \theta^{(2)}$, implying that the treatment effect of the second treatment is at most as large as the treatment effect for the first treatment. Let $r_j^{(k)}$ be the ratio between the number of subjects allocated to treatment $k \in \{0,1,2\}$ and control at each stage j with $r_j^{(0)} = 1$. Let $z_j^{(k)} = \frac{\hat{\mu}_j^{(k)} - \hat{\mu}_j^{(0)}}{\sigma} \sqrt{\frac{r_j^{(0)} n_j^{(k)}}{r_j^{(k)} r_j^{(0)}}}$ be the test statistic [3] at stage j for comparing arm $k \in \{1,2\}$ to control, where $\hat{\mu}_j^{(k)} = \left(n_j^{(k)}\right)^{-1} \sum_{i=1}^{n_j^{(k)}} X_i^{(k)}$ and $n_j^{(k)} = r_j^{(k)} n$, with $k \in \{0,1,2\}$ and n is the sample size in the control group at the first stage. The vector of test statistics follows a multivariate normal distribution $Z \sim N_4(\in, \sum)$ with

$$Z = \left(Z_1^{(1)}, Z_1^{(2)}, Z_2^{(1)}, Z_2^{(2)}\right), \quad \in \; = \left(\frac{\theta^{(1)}}{\sigma} \sqrt{\frac{r_1^{(0)} n_1^{(1)}}{r_1^{(1)} + r_1^{(0)}}}, \frac{\theta^{(2)}}{\sigma} \sqrt{\frac{r_1^{(0)} n_1^{(2)}}{r_1^{(2)} + r_1^{(0)}}}, \frac{\theta^{(1)}}{\sigma} \sqrt{\frac{r_1^{(0)} n_2^{(1)}}{r_2^{(1)} + r_2^{(0)}}}, \frac{\theta^{(2)}}{\sigma} \sqrt{\frac{r_2^{(0)} n_2^{(2)}}{r_2^{(2)} + r_2^{(0)}}} \right)$$

and the covariances between Z-statistics are

$$\operatorname{cov}\left(Z_j^{(k)}, Z_j^{(k)}\right) = 1, \text{ with } k, j \in \{1,2\}, \quad \operatorname{cov}\left(Z_j^{(k)}, Z_j^{(k')}\right) = \sqrt{\frac{r_1^{(k)} n_j^{(k')}}{r_j^{(k)} + r_2^{(0)}} \frac{r_j^{(k')}}{r_j^{(k')} + r_j^{(0)}}}, k \neq k',$$

$$\text{with } k, k', j \in \{1,2\}, \operatorname{Cov}\left(Z_1^{(k)}, Z_2^{(k)}\right) = \sqrt{\frac{r_1^{(k)} r_1^{(0)}}{r_1^{(k)} + r_1^{(0)}} \frac{r_2^{(k)} r_2^{(0)}}{r_2^{(k)} + r_2^{(0)}}}, \text{ with}$$

$$k \in \{1,2\}, \operatorname{Cov}\left(Z_1^{(k)}, Z_1^{(k')}\right) = \sqrt{\frac{r_1^{(k)}}{r_1^{(k)} + r_1^{(0)}} \frac{r_1^{(k')}}{r_1^{(k')} + r_1^{(0)}}} \sqrt{\frac{r_1^{(k)} r_1^{(0)}}{r_1^{(k)} + r_1^{(0)}} \frac{r_2^{(k)} + r_2^{(0)}}{r_1^{(k)} r_2^{(0)}}}, k \neq k', \text{ with}$$

$k, k' \in \{1,2\}$.

We test the null hypotheses: $H_{01}{:}\left\{\theta^{(1)} \leq 0\right\}, H_{02}{:}\left\{\theta^{(2)} \leq 0\right\}$ with the global null hypothesis denoted by $H_0{:}\left\{\theta^{(1)} = \theta^{(2)} = 0\right\}$. Let $u_j^{(1)}, l_j^{(1)}$ and $u_j^{(2)}, l_j^{(2)}$ be the critical values at stage j for T_1 and T_2, respectively, used to test the hypotheses, with $u_2^{(k)} = l_2^{(k)}, k \in \{1,2\}$.

The proposed design then takes into account the order among the treatment effects when making the decisions at the first stage and the final analysis and a set of decision rules consistent with this order is given in Table 1. For example, if both Z-statistics cross the upper bounds at the interim analysis, the trial is stopped for efficacy (as in a conventional MAMS design). In contrast to the traditional MAMS design, the trial continues if there is contradicting evidence with respect to the order, for example, if $Z_1^{(2)}$ crosses the upper bound, but there is not enough evidence to claim superiority of T_1 to control, then both arms are continued to the next stage.

The idea behind these decision rules is that at any stage the effectiveness of T_2 can be claimed only if T_1 can be declared superior to the control. Therefore, T_1 can be regarded as a gatekeeper. [15] Following this procedure, depending on the context, alternative decisions could be considered for the cells colored in red in Table 1 (see Section 2 of the Supporting Information for more discussion).

3.1 Family-wise error rate

For confirmatory clinical trials, control of the family wise error rate (FWER) in the strong sense at level α, that is the probability to reject at least one true null hypothesis, is often required.[16] Using the rules described in Table 1, the FWER for the 3-arm 2-stage ORD can be written as

$$P(\text{rejecting at least one true } H_{0k},\, k \in \{1,2\} | H_0) = P\left(Z_1^{(1)} \geq u_1^{(1)} | H_0\right) + P\left(Z_2^{(1)} \geq u_2^{(1)}, l_1^{(1)} < Z_1^{(1)} u_2^{(1)} | H_0\right) + P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | H_0\right). \quad (1)$$

Equation (1) shows that the events used for the computation of the type I error under the global null hypothesis ({Reject H_{01} and H_{02}}, {Reject H_{01} not H_{02}}) are a subset of the events ({Reject $H_{01} \cup H_{02}$}) used in the MAMS design of Magirr et al. [3] Thus, the probability of rejecting at least one hypothesis under the global null will be smaller for the ORD compared to the MAMS design, while the probability of rejecting neither hypothesis will be smaller for MAMS if the same bounds are used. It is worth noting that overall, the critical bounds, if these are the same for all active treatments $u_1^{(1)} = u_1^{(2)} = u_1, u_2^{(1)} = u_2^{(2)} = l_2^{(1)} = l_2^{(2)} = u_2, l_1^{(1)} = l_1^{(2)} = l_1$, for the 3-arm 2-stage ORD are smaller in each stage compared to the MAMS design of Magirr et al [3] (see Section 7 of the Supporting Information). Consequently the ORD design is strictly more powerful than the MAMS design under these assumptions.

The critical bounds for the given treatment arm can be defined as function of a (possibly arm-specific) parameter, that is, $u_j^{(k)} = u_j^{(k)}\left(a^{(k)}\right), l_j^{(k)} = l_j^{(k)}\left(a^{(k)}\right), j \in \{1,2\}, k \in \{1,2\}$, which can be searched over a grid of values for $a^{(k)}$ in order to strongly control the FWER at level α. If $a^{(1)} = a^{(2)} = a$, then a unique solution can be found restricting the search over a such that the expression given in Equation (1) is below α under the global null hypothesis. In this case, the solution is unique either when the search is based on different

boundary shapes $u_j^{(k)}, l_j^{(k)}$ or when the same boundary shapes are used for all experimental arms— $u_j^{(k)} = u_j, l_j^{(k)} = l_j$ for all $k \in \{1,2\}$. If $a^{(k)}$ are not the same for each arm, additional constraints are required for the uniqueness of the solution and to maintain the strong control of the FWER at level α, such as the control under the partial null hypotheses. In the 3-arm setting, for example, this is $(\theta^{(1)}, 0)$. However, different values of $\theta^{(1)}$ can provide different boundaries, and so the solution is unique for the specific value of $\theta^{(1)}$ (see Section 7 for more details).

Theorem 1 below then shows that, if the same bounds, $(u_j, l_i), j \in \{1,2\}$ are used for each arm, and the same allocation ratios (with respect to the control) are used for all active treatments, the FWER is maximized under the global null hypothesis and hence the above ensures strong control of the FWER.

Theorem 1. *Consider a 3-arm 2-stage ORD design and denote the global null hypothesis by* $H_0: \theta^{(1)} = \theta^{(2)} = 0$. *Let* $u_1^{(1)} = u_1^{(2)} = u_1, u_2^{(1)} = u_2^{(2)} = l_2^{(1)} = l_2^{(2)} = u_2, l_1^{(1)} = l_1^{(2)} = 1_1$ *be the critical bounds such that Equation (1) is below under the global null hypothesis. Let us assume that there are equal numbers of patients on each active treatment within each stage:* $r_j^{(k)} = rj, \ \forall k \in \{1,2\}$.

Let $\boldsymbol{\theta}_0$ *be a vector where at least one treatment effect is less or equal to 0. Then,*

$$P(\text{rejecting at least one true } H_{0k}, k \in \{1,2\} | \theta_0) \leq$$
$$P(\text{rejecting at least one true } H_{0k}, k \in \{1,2\} | H_0) \leq \alpha$$

The proofs of all theorems are given in Section 1 of the Supporting Information.

3.2 Power requirement

To power the study, we consider the configuration $\theta = \left(\theta^{(1)}, \theta^{(2)}\right)$, where $\theta^{(1)} \geq \theta^{(2)} \geq \delta_0 > 0$ and δ_0 is the minimum clinically relevant difference. The ORD is then powered at (1 - β) to reject both hypotheses under $\theta = \left(\theta^{(1)}, \theta^{(2)}\right), \theta^{(1)} \geq \theta^{(2)} \geq \delta_0 > 0$ when Equation (2) is satisfied:

$$\begin{aligned}
&P\left(Z_1^{(1)} \geq u_1^{(1)}, z_1^{(2)} \geq u_1^{(2)} | \theta\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \theta\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, z_1^{(1)} \leq l_1^{(1)}, z_1^{(2)} \geq u_1^{(2)} | \theta\right) + \\
&P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \geq u_1^{(1)}, l_1^{(2)} < z_1^{(2)} < u_1^{(2)} | \theta\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < z_1^{(1)} < u_1^{(1)}, l_1^{(2)} < z_1^{(2)} < u_1^{(2)} | \theta\right) \geq 1 - \beta .
\end{aligned} \tag{2}$$

Theoretical considerations and numerical evaluations have shown that, if Pocock boundaries [17] for both treatments are used, $u_1^{(1)} = u_1^{(2)} = u_1, \ l_1^{(1)} = l_1^{(2)} = \ -u_1, u_2^{(1)} = u_2^{(2)} = u_1,$

and critical values found such that Equation (1) is below α, the power of the ORD design is, practically, no smaller than a fixed balanced sample design with the same sample size. Furthermore, for a number of treatment effects, it was found to be strictly positive. Full details of these considerations are given in Section 3 of the Supporting Information.

4 Generalization Of The Ord To *K*-Arm *J*-Stage

Consider a clinical trial with $K-1$ active treatment arms, $T_1, \ldots, T_{K-1}$, against a control treatment and J stages and denote the treatment effect comparing treatment k against control by $\theta^{(k)}$. We denote the vector of treatment effects by $\theta = \left(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k-1)}\right)$. The null hypotheses of interest are $H_{01}: \left\{\theta^{(1)} \leq 0\right\}, \ldots, H_{0k-1}: \left\{\theta^{(k-1)} \leq 0\right\}$. Let $Z_j^{(k)}$ denote the test statistic based on all data up to stage j for comparison $k \in \{1, \ldots, K-1\}$ as before and assume that the following order relationship holds: $\theta^{(1)} \geq \theta^{(2)} \geq \cdots \geq \theta^{(k-1)}$. Let $u_j^{(k)}, l_j^{(k)}, k \in \{1, \ldots, K-1\}, j \in \{1, \ldots, J\}$ be the critical values at stage j with $u_J^{(k)} = l_J^{(k)}, k \in \{1, \ldots, K-1\}$.

The decision rules at the interim analyses follow the same principle as for the 3-arm 2-stage design defined above. The decisions are made in order to be able to select all promising treatment arms at the end of the trial and H_{0k} can only be rejected if all $H_{0k'}k' < k$ have been rejected. Once H_{0k} has been rejected, the recruitment to arms $T_{\mathscr{L}_j}, \ldots, T_k$ is stopped, where $\mathscr{L}_j$ is the lowest index of a treatment arms remaining in the trial at stage *j*. If there is contradicting evidence with respect to the order at stage *j*, that is when $Z_j^{(k)} \geq u_j^{(k)}$ and there is at least one $k' < k$ such that $Z_j^{(k')} < u_j^{(k')}$, then recruitment to these arms continue. As for the 3-arm 2-stage design, if there is sufficient evidence to drop arm *k*, that is when $Z_j^{(k)} \leq u_j^{(k)}$, and if there is any contradicting evidence for $k' > k$ then the recruitment to arms $T_k, \ldots T_{\mathscr{H}_j}$ is stopped, where $\mathscr{H}_j$ is the highest index on the treatment arms remaining in the trial at stage j. A general algorithm for the decision-making in this setting is given in Algorithm 1.

Algorithm 1 Rules for *K*-arm *J*-stage ORD when $\theta^{(1)} \geq \theta^{(2)} \geq \cdots \geq \theta^{(K-1)}$

1. Let $\mathscr{L}_j$ and $\mathscr{H}_j$ be the lowest and highest indices on the active treatment arms remaining in the trial at stage *j*, respectively. At stage *j*, compute $Z_j^{(k)}$ for $k \in \kappa = \left\{\mathscr{L}_j, \ldots \mathscr{H}_j\right\}$.
2. Stop recruitment to arm k at stage *j* for *efficacy* if for all $k' \in \kappa, k' \leq k, z_j^{(k')} \geq u_j^{(k')}$.
3. Stop recruitment to arm k at stage *j* for *futility* if
   - $Z_j^{(k)} < l_j^{(k)}$ and for all $k' \in \kappa, k'' > k, Z_j^{(k'')} < u_j^{(k'')}$, or

- $l_j^{(k)} < Z_j^{(k)} < u_j^{(k)}$ and for at least one $k^\star \in \mathcal{K}^\star, k^\star < k, Z_j^{(k^\star)} < l_j^{(k^\star)}$ and for no $k'' \in \kappa, k'' > k^\star, Z_j^{(k'')} > u_j^{(k'')}$.

4. Stop the trial when the recruitment to all arms is stopped.

Let M_j be a random variable representing the number of arms (including the control) at stage j when H_{01} failed to be rejected at stage $j-1$. Because of the hierarchy in testing the hypotheses, the FWER for an K-arm J-stage design can be written as

$$\begin{aligned} &P(\text{rejecting at least one true} H_{0k},\ k \in \{1, \ldots, K-1\} | H_0) = \\ &\sum_{j=1}^{J} P(\text{rejecting } H_{01} \text{ at } j\text{thstage}, H_{0k} \text{not rejected at stages}, \forall s < j | H_0 = \\ &P\left(Z_1^{(1)} \geq Z_1^{(1)} | H_0\right) + \sum_{j=2}^{J} \sum_{m=2}^{K} P\left(Z_1^{(1)} \geq u_1^{(1)} | M_j = m, H_0\right) \times P\left(M_j = m\right), \end{aligned} \tag{3}$$

where $P(M_j = m)$ is the probability that at the previous stage H_{01} failed to be rejected and the number of arms were at least m. One can show that the following iterative equality holds

$$P\left(M_j = m\right) = \sum_{c=m}^{K} P\left(A_{j,c-m+1}^{(c-1)} | M_{j-1} = c, H_0\right) \times P\left(M_{j-1} = c\right),$$

with at the first stage $P(M_1 = K) = 1$ and 0 otherwise. The set $A_{j,c-m+1}^{(c-1)}$ defines the event that H_{01} failed to be rejected at stage $j-1$ when the number of treatment arms (including the control arm) in the trial were c. This set is formally defined in Table 2. In the definition of $A_{j,c-m+1}^{(c-1)}$, the superscript $(c-1)$ indicates the number of active treatment arms that are still in the trial at stage $j-1$, while the subscript $c-m+1$ refers to the number of active treatment arms (that is equal to $c-m$) that have been dropped before reaching the stage j.

While the expression for the FWER in the general case is cumbersome, for a fixed number of stages (arms), the FWER for K-arm (J-stage) can be found iteratively—an example for 2-stage trials is given in Section 4 of the Supporting Information.

The critical bounds for a K-arm J-stage ORD can be, again, defined as functions of parameters $a^{(k)}$ such that $u_j^{(k)} = u_j^{(k)}\left(a^{(k)}\right), l_j^{(k)} = l_j^{(k)}\left(a^{(k)}\right), j \in \{1, \ldots, J\}, k \in \{1, \ldots, K-1\}$. Thus, as for the 3-arm 2-stage setting, under the constraint of $a^{(1)} = a^{(2)} = \cdots = a^{(k-1)} = a$, a unique solution can be found to control the FWER at level α in the strong sense. Theorem 2 below then shows that, if the same bounds, $(u_j, l_j), j \in \{1, \ldots, J\}$ are used for each arm and the same allocation ratios (with respect to the control) are used for all active treatments, the FWER is maximized under the global null hypothesis $H_0: \left\{\theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(k-1)} = 0\right\}$ and hence the above ensures strong control of the FWER.

Europe PMC Funders Author Manuscripts

Theorem 2. *Consider a K-arm J-stage ORD design and denote the global null hypothesis by* $H_0: \left\{\theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(k-1)} = 0\right\}$. *Let* $u_j^{(k)} = u_j, l_j^{(k)}, u_J^{(k)} = l_J^{(k)} = u_J, k \in \{1, \ldots, K-1\}$ *be the critical values such that Equation (3) is below* α *under the global null hypothesis. Assume that there are equal numbers of patients on each active treatment within each stage:* $r_j^{(k)} = r_j, \ \forall k \in \{1, \ldots, K-1\}$. *Let* $\boldsymbol{\theta}_0$ *be the vector where at least one treatment effect is less or equal to 0.*

Then,

$$P(\text{rejecting at least one true } H_{0k}, k \in \{1, \ldots, K-1\} | \theta_0) \leq$$
$$P(\text{rejecting at least one true } H_{0k}, , k \in \{1, \ldots, K-1\} | H_0) \leq \alpha$$

In the next session, a simulation study will be described in order to apply the proposed design in the context of the asthma trial. [14]

5 Case Study

5.1 Setting

We revisit the results of the clinical trial of *Tiotropium add-on therapy in adolescents with moderate asthma: A 1-year randomized controlled trial* (NCT01257230). [14] Patients were randomized in a 1:1:1 ratio to receive 5μg or 2.5μg of once-daily tiotropium or placebo. The null hypotheses were tested in a stepwise manner to control the type I error at level α = 0.025. The study was powered at 80% to detect a difference of 120 mL between treatments in the change from baseline of peak $FEV_{1[0\text{-}3h]}$ assuming a common SD of 340 mL. It was found that 127 patients per group were needed, resulting in a maximum sample size of 381 patients. The trial is revisited using the ORD, which can be applied assuming a monotonic dose-response relationship.

In line with the original trial we assume that the change from baseline of peak $FEV_{1[0\text{-}3h]}$ is normally distributed with SD σ = 340, $k \in \{0,1,2\}$, and common baseline mean FEV_1 of $\mu^{(0)}$= 2747. As in the original study, we consider an improvement of FEV_1 of 120 of interest and hence consider the following values for $\theta^{(k)}, k \in \{1,2\}: \theta = (0,0), \theta = (120,0)$ and $\boldsymbol{\theta}$ = (120 120). The design is powered at 80% to reject all hypotheses or at least one hypothesis (in order to compare the sample sizes between the ORD and the original trial design) when all doses have the same effect compared to the placebo. Additionally to the achieved power, the efficiency of the proposed design is measured by its expected sample size (ESS), that is the mean number of patients recruited to the trial before it is terminated.

We consider one- and two-stage ORD designs and note that the one-stage ORD design corresponds to the hierarchical testing strategy used in the original design. For the two-stage design the interim analysis takes place after half of the total sample size has been observed and triangular critical bounds [18] are used. The numerical results found using R[19] and 10^6 replicate simulations.

5.2 Numerical results

Consider the 3-arm 1-stage ORD using the maximum total sample size of 381 patients that corresponds to the same maximum total sample size originally planned for the study. In this setting, the critical bound at the final analysis is $u_1 = z_{1-\alpha} = z_{0.975} = 1.96$ Table 3 describes the results of the simulation.

It can be seen that the FWER is controlled at level $\alpha = 0.025$ under all considered null scenarios. For the scenarios where there is at least one dose that is superior to control, the probability to reject at least one hypothesis is 80% as required in the original study. Note that 80% is the probability of rejecting at least one dose considering any rejections and not only correct rejections. Therefore, when no interim analyses are planned, the ORD requires the same maximum sample size as in the original study if it is powered to reject at least one hypothesis. It is worth noting that the probability of rejecting all hypotheses, if all of them are true, is around 69%.

The distinguishing feature of the proposed ORD is that it allows to include interim analyses during the trial. Table 3 shows the operating characteristics of the design when an interim analysis takes place after observing half of the maximum sample sizes. If the study is powered to reject at least one hypothesis at 80%, a maximum sample size of 426 patients is needed—45 more patients than for the 1-stage design. The gain from using a two-stage design arises in terms of the expected sample size—on average the number of patients is expected to be below 332 under each scenario. At the same time, under this power configuration, the probability to identify the smallest promising dose is at 68% similar to the single-stage design.

Furthermore, it is argued by the construction of the ORD proposed in Section 3 that if it is of interest to identify the lowest dose with the promising treatment effect, the trial should be powered to reject all hypotheses. In order to identify the lowest effective dose with the desirable 80% probability, a single-stage trial would require 474 patients and the two-stage design 534 patients. As before, the inclusion of an interim analysis and the use of triangular bounds lead to the reduction of the expected number of patients. Specifically, on average the number of patients is expected to be below 400 under each scenario.

Overall, the ORD was found to reproduce the sample size calculation of the original study when no interim analyses were planned and is powered to reject at least one hypothesis. At the same time, the ORD allows the inclusion of interim analyses during the trial. When an interim analysis is planned during the trial, the ORD has shown to be an efficient design as it allows to stop the trial earlier or to drop unpromising doses with high probability before the end of the study.

The results discussed in this section use triangular bounds. Qualitatively similar results using Pocock [17] and O'Brien and Fleming [20] boundaries are provided in Section 6 of the Supporting Information. To further investigate the design characteristics of the proposed ORD, an extensive simulation study is conducted in Section 6.

6 Numerical Evaluations

6.1 Setting

As in the motivating example, consider a clinical trial setting with 3 treatments arms, 2 experimental and a control with 1:1:1 allocation ratio. Consider a single-stage design (that is a fixed sample design (FSD) with hierarchical testing, FSD(h)) and a two-stage design with one pre-planned interim analysis at the middle of the trial. The objective of the trial is to find all promising treatment arms. The FWER is to be controlled below α = 0.05 and the power of trial is to be at least 80% to reject both hypotheses when both treatment arms have the same effect compared to the control. The patients' responses on treatment arm k are assumed to have normal distribution with mean $\mu^{(k)}$ and SD σ = 1. For the control group $\mu^{(0)} = 1$, while for the treatment arms $\mu^{(k)} = \mu^{(0)} + \theta^{(k)}$. We fix the clinically relevant difference to be 0.5. Therefore, the scenario under which the ORD is powered is $\boldsymbol{\theta}$ = (0.5, 0.5). We evaluate the performance of the design under various treatment effects configurations when the treatment effect on the first arm is fixed to be 0.5 and the treatment effect on the second arm is varied: $\theta = \left(0.5, \theta^{(2)}\right), \theta^{(1)} \geq \theta^{(1)} \geq \theta^{(2)}$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For the 3-arm 2-stage ORD, the bounds are found under the conditions $a^{(k)} = a$ (ie, the same for each treatment arm) and they are found using a grid search—based on the triangular boundary shape [18] for the all experimental treatments—over one single parameter. Thus, the solution found is unique.

6.2 Competing approaches

The proposed ORD is compared to the FSD, [21] in which the total sample size is specified at the design stage of the trial and it is not subject to adaptations during the process of the trial. The hypotheses are tested at the end of the trial only and any hypothesis can be rejected independently of the other.

The second comparator is the MAMS design by Magirr et al. [3] However, the conventional MAMS design is not an appropriate comparator as it does stop as soon as at least one hypothesis is rejected. Therefore, the following modification of the MAMS design is considered for the comparison. At the interim analysis, if a Z-statistic corresponding to one arm crosses the upper or lower bound while another does not, the trial will still continue with the treatment that did not cross a bound. The design is referred to as the MAMS(m).

The FWER expression for MAMS(m) is the same as derived by Magirr et al[3] but the power expression changes. To power a 3-arm 2-stage design, for a given configuration of $\boldsymbol{\theta}$ and a pre-specified level β, we search for the sample size that satisfies

$$\begin{aligned} &P\left(Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \theta\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)} | \theta\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \theta\right) + \\ &P\left(Z_2^{(2)} \geq u_2^{(2)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}, Z_1^{(1)} \geq u_1^{(1)} | \theta\right) \geq 1 - \beta \end{aligned} \tag{2}$$

6.3 Numerical results

Two main simulation studies are conducted to compare the three competing approaches. In the first one, each design is constructed such that it yields 80% power to reject both null hypotheses under the alternative hypothesis. The second one compares the power between approaches based on the common sample size.

6.3.1 Same power requirement for all designs—In this subsection, the results when all design are powered at 80% to reject both null hypotheses are provided. The resulting design specifications and operating characteristics under the global null hypothesis are provided in Table 4.

Under $\boldsymbol{\theta} = (0,0)$, all designs control the type I error at level $\alpha = 0.05$ as expected. The total maximum sample sizes necessary to reach a power of 80% is 231 for the FSD, 192 for the FSD with a hierarchical test (FSD(h)) which is akin to a single-stage ORD design, 222 for the 3-arm 2-stage ORD and 264 for the MAMS(m) designs. The maximum sample size (to achieve the same power to reject both hypotheses) for the FSD is greater compared to FSD(h) and the two-stage ORD as it does not account for the hierarchy in testing.

The designs' performances under the configuration $\theta^{(1)} \geq \theta^{(2)}$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ are presented in Figure 1. When the second treatment arm is no different to control, $\theta^{(2)} = 0$, the probability of rejecting the null hypothesis for this arm is controlled at $\alpha = 0.05$ for all designs and when $\boldsymbol{\theta}$=(0.5,0.5) all the designs satisfy the power requirement at 80%. However, under all other considered non-zero values of $\theta^{(2)}$, the probability of rejecting both hypotheses is higher for the approaches accounting for the hierarchy, FSD(h) and ORD, than for other competing designs. The gain from using a two-stage ORD design compared to FSD(h) arises in terms of expected sample sizes which is strictly lower for the two-stage design. The 3-arm 2-stage ORD has noticeably lower expected sample size (ESS) compared to all other designs ranging from 13% to 35% depending on the simulation scenario. The largest difference in power (an increase of around 5%) between the 3-arm 2-stage ORD and the MAMS(m) design is achieved under $\theta^{(2)} = 0.2$.

6.3.2 Common sample size for all designs—In this subsection, the three designs are compared with a common maximum sample size of 222 patients, which is the maximum sample size necessary for the 2-stage ORD in order to reject both hypotheses at 80% under $\boldsymbol{\theta} = (0.5, 0.5)$.

The design specifications and operating characteristics of the designs under the global null hypothesis are provided in Table 5, while the designs' performances under the configuration $\theta^{(1)} \geq \theta^{(2)}$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ are presented in Figure 2.

As expected all designs control the FWER under the global null. Under the scenario when the second treatment arm is no different to control, $\theta^{(2)} = 0$, the probability of rejecting the null hypothesis for this arm is controlled at $\alpha = 0.05$. Under all other considered non-zero values of $\theta^{(2)}$, the approaches accounting for the hierarchy, FSD(h) and ORD, result in

higher power compared to the other competing designs. The gain from using a two-stage ORD design compared to FSD(h) can once more be seen in terms of expected sample sizes which is strictly lower for the two-stage design. The 2-stage ORD has lower ESS compared to the other designs with reductions between 3% and 32%. The largest difference in power—around 9.8%—between the ORD and the MAMS(m) design is achieved under $\theta^{(2)} = 0.3$.

Overall, the ORD results in noticeable gains across all considered scenarios both in terms of power and expected sample size. Therefore, the inclusion of the order restriction into the decision rules for the decision-making can provide advantages in power and/or expected sample size compared to standard approach to multi-arm trials, specifically, the FSD and the MAMS(m).

7 Different Bounds For Each Treatment Arm

In the results above the same bounds are used for both treatments. However, the proposed design allows for different boundary shapes to be used for different treatments which could lead to potential benefit in terms of power. In this section, we explore the effect of various boundary shapes on the operating characteristics of the designs.

We consider the setting as in Section 6.1 and let $u_j^{(1)} = u_j^{(1)}\left(a^{(1)}\right), l_j^{(1)} = l_j^{(1)}\left(a^{(1)}\right)$ be the upper and lower bounds for T_1 at the stage *j*, and $u_j^{(2)} = u_j^{(2)}\left(a^{(2)}\right), l_j^{(2)} = l_j^{(2)}\left(a^{(2)}\right)$ be the boundaries for T_2 at the stage *j*, being functions of $a^{(1)}$ and $a^{(2)}$, respectively. The critical bounds and the sample size could be searched over a grid of values for $a^{(1)}$ and $a^{(2)}$ in order to strongly control the FWER at level α and to satisfy the power requirements in Equation (2). To maintain strong control of the FWER at level α it is necessary to control the type I error when $\boldsymbol{\theta}_0 = (\theta^{(1)}, 0)$. Indeed, as shown in Section 1 of the Supporting Information, under $\boldsymbol{\theta}_0 = (\theta^{(1)},0)$ it holds

$$P(\text{rejecting at least one true } H_{0k}, k \in \{1,2\}|\theta_0) =$$
$$P(\text{reject } H_{02}|\text{reject } H_{01}, \theta_0) \times P(\text{reject} H_{01}|\theta_0) \leq P(\text{reject} H_{02}|\text{reject} H_{01}, \tilde{\theta}_0 = (\infty, 0),$$

where $P\left(\text{reject} H_{02}\middle|\text{reject } H_{01},\ \tilde{\theta}_0 = (\infty, 0)\right)$ tends to

$$P\left(Z_1^{(2)} > u_1^{(2)}|H_{02}\right) + P\left(Z_2^{(2)} > u_2^{(2)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}|H_{02}\right).$$

Thus, the bounds for the second arm are searched over a grid of values of $a^{(2)}$ as for a 2-arm design with 2 stages [22] and then the bounds for T_1 are searched in order to satisfy Equation (1) under the null hypothesis. Finally, the sample size is searched to satisfy the power requirements, assuming equal allocation to all arms.

Several combinations of boundary shapes are compared considering all possibilities with constant POC, [17] O'Brien and Fleming OBF [20] and triangular TRIAN [18] bounds. Among all these nine combinations of bounds, a subset of six is selected. For each shape of the bounds for T_1, two combinations are selected. These are those that provide the smallest ESS and

the highest power compared to the other ones. The combinations that are excluded are the ones that use constant bounds for T_1 and T_2, the combination with O'Brien and Fleming and constant bounds for T_1 and T_2 respectively and the combination with triangular and constant bounds for T^1 and T_2 respectively. The complete set of results is provided in Section 5 of the Supporting Information.

The summary of operating characteristics, probability to reject both and probability to reject only one, and the expected sample size, for the proposed design using the six remaining combinations of boundary shapes is provided in Figure 3.

The design resulting in the highest power in rejecting both hypotheses among the subset of selected bounds is the one that uses triangular bounds for the treatment associated to the highest effect and O'Brien and Fleming bounds for the arm associated to the lowest treatment effect. Indeed, in the way that O'Brien and Fleming bounds are constructed, if $u_1^{(2)} > u_1^{(1)}$, then the trial tends to stop later and the final decision on T_2 is based on more data. Therefore, the test becomes more powerful compared to the test that tends to make a final decision on T_2 earlier. This selection of bounds also corresponds to the smallest probability of rejecting the first hypothesis and not the second one (that is the probability of making an error when $\theta^{(2)} < 0$) compared to the other combinations.

The combination that uses O'Brien and Fleming bounds for both treatment arms, the combination with Pocock bound for T_1 and O'Brien and Fleming bounds for T_2, and the combination of triangular and O'Brien and Fleming bounds result in similar power to reject both hypotheses but the first two combinations require lower maximum sample size compared to the latter—198 and 204 patients, respectively against 210 patients. At the same time, the three combinations differ in terms of ESS and in terms of probability of rejecting only the first hypothesis. Indeed, among these three combinations, the one with triangular bound for T_1 and O'Brien and Fleming bound for T_2 is the one with the smallest ESS and presents the smallest probability of rejecting only the first hypothesis.

The combination that uses O'Brien and Fleming bound for T_1 and triangular bound for T_2 is the one with smallest power and highest probability of rejecting only the first hypothesis. While the combination that uses triangular bounds for both treatment arms is the one with the smallest ESS (indeed triangular bounds are constructed to minimize the ESS [18]) for each configuration of $\boldsymbol{\theta}$ compared to the other ones, even though it is one of the combinations with the highest maximum sample size—222 patients. Nevertheless, this combination has slightly smaller power of rejecting both hypotheses and slightly smaller probability of rejecting only the first hypothesis compared to the combination with Pocock bounds for T_1 and triangular bound for T_2.

Overall, the results suggest that there is a benefit, in terms of power and ESS, in using different bounds for each treatment arm. However, the final choice of the bounds will depend on the specific objectives of the trial. For example, in order to minimize the ESS it is recommended the use of triangular bounds for both treatment arms, whereas in order to maximize the power of rejecting all hypotheses it is recommended the use of triangular bounds for T_1 and O'Brien and Fleming bounds for T_2.

8 Discussion

The aim of the current study was to explore MAMS designs that could select the most promising arm associated to the minimum treatment effect. An approach that takes the order relationship among the treatment effects (when no parametric dose-response or duration-response model is assumed) into account has been proposed. In the proposed approach we claim the effectiveness of the arm associated to the minimum treatment effect compared to the control only if we claim that the treatment associated to the maximum effect is efficacious. Through theoretical arguments and extensive numerical evaluation we show that the proposed design can provide noticeable advantages in power and/or expected sample sizes required in the trial compared to the alternatives.

The proposed design can be applied to a wide range of clinical trial settings where it could be assumed an order among the treatment effects. For example, in clinical trial designs applied to infectious diseases, such as TB and HBV, where it can be assumed that longer treatment durations correspond to a higher efficacy. In this case, the focus translates into the problem of selecting the shortest promising treatment duration. The proposed design can also be applied to clinical trial settings where nested combinations of treatments are tested against a common control arm.

The proposed design is closely linked to the hierarchical procedures described in the literature for example, by Glimm et al, [23] Tamhane et al. [24,25] In particular, in the literature, the hierarchical procedure is applied for testing multiple end-points in a specific order. In some applications, the interest is to test the secondary endpoint just if the primary endpoint has been rejected. Therefore, the endpoints could be tested on hierarchically order and various strategies can be adopted to test the hypotheses [23] depending on the study objectives. It can be noted that when non-binding futility boundaries $\left(l_1^{(1)} = l_1^{(2)} = -\infty\right)$ are used in the 3-arm 2-stage ordered restricted design, the overall testing procedure [23] coincides with the proposed method.

In this study, it has been assumed that the common variance is known. However, the effect of this assumption is not negligible especially with small sample sizes. In this case, a possible approach would be to transform the individual test statistics using the function $f(x) = \Phi^{-1}\{T_d(x)\}$, where T denotes the t distribution function with d degrees of freedom and Φ is the standard cumulative density function. More details of this approach are given in Jennison and Turnbull. [26]

Several avenues of future research present itself. First, focus has been given to superiority tests in this work. In certain diseases, such as in TB, non-inferiority designs are the norm and hence further research on non-inferiority hypothesis tests is of interest. Second, we assume that information time is the same for all treatments. When considering different durations of treatment, however, this information accumulates a different times and hence further work will consider optimal designs in this setting. Finally, we assume in this work that there is no uncertainty about the order of the treatment effects. Using a Bayesian

framework, however, would naturally allow for uncertainty in that assumption to be considered.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This report is independent research supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr Pavel Mozgunov, NIHR300576; and Prof Jaki's Senior Research Fellowship, NIHR-SRF-2015-08-001) and by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research or the Department of Health and Social Care (DHSC). T Jaki also received funding from UK Medical Research Council (MC_UU_00002/14).

Funding information

National Institute for Health Research, Grant/Award Numbers: NIHR-SRF-2015-08-001, NIHR300576; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014; UK Medical Research Council, Grant/Award Number: MC_UU_00002/14

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Abbreviations

FSD	fixed sample design
MAMS	multi-arm multi-stage
ORD	ordered restricted design

References

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ. 2016; 47: 20–33. DOI: 10.1016/j.jhealeco.2016.01.012 [PubMed: 26928437]
2. Stallard N, Todd S. Sequential designs for phase III clinical trials incorporating treatment selection. Stat Med. 2003; 22: 689–703. DOI: 10.1002/sim.1362 [PubMed: 12587100]
3. Magirr D, Jaki T, Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. Biometrika. 2012; 99: 494–501. DOI: 10.1093/biomet/ass002
4. Bauer P, Kieser M. Combining different phases in the development of medical treatments within a single trial. Stat Med. 1999; 18: 1833–1848. DOI: 10.1002/(SICI)1097-0258(19990730)18:14<1833::AID-SIM221>3.0.CO;2-3 [PubMed: 10407255]
5. Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? Clin Investig. 2015; 5: 393–399. DOI: 10.4155/cli.15.13
6. Pallmann P, Bedding AW, Choodari-Oskooei B, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. BMC Med. 2018; 16: 1–15. DOI: 10.1186/s12916-018-1017-7
7. Burnett T, Mozgunov P, Pallmann P, Villar S, Wheeler GM, Jaki T. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. BMC Med. 2020; 18 (1) 1–21. [PubMed: 31898501]

8. Ginsberg AM, Spigelman M. Commentary challenges in tuberculosis drug research and development. Nat Med. 2007; 13: 290–294. [PubMed: 17342142]
9. Davies GR, Phillips PPJ, Jaki T. Adaptive clinical trials in tuberculosis: applications, challenges and solutions. Int J Tuberc Lung Dis. 2015; 19: 626–634. DOI: 10.5588/ijtld.14.0988 [PubMed: 25946350]
10. Papatheodoridis G, Buti M, Cornberg MS, et al. EASL clinical practice guidelines: management of chronic hepatitis B virus infection. J Hepatol. 2012; 57: 167–185. DOI: 10.1016/j.jhep.2012.02.010 [PubMed: 22436845]
11. WHO. Guidelines for treatment of drug-susceptible tuberculosis and patient care. 2017.
12. Lienhardt C, Nunn A, Chaisson R, et al. Advances in clinical trial design: weaving tomorrow's TB treatments. PLoS Med. 2020; 17 e1003059 doi: 10.1371/journal.pmed.1003.059 [PubMed: 32106220]
13. Quartagno M, Walker AS, Carpenter JR, Phillips PPJ, Parmar MKB. Rethinking non-inferiority: a practical trial design for optimising treatment duration. Clin Trials. 2018; 15: 477–488. DOI: 10.1177/1740774518778027 [PubMed: 29871495]
14. Hamelmann E, Bateman ED, Vogelberg C, et al. Tiotropium add-on therapy in adolescents with moderate asthma: a 1-year randomized controlled trial. J Allergy Clin Immunol. 2016; 138: 441–450. e8 doi: 10.1016/j.jaci.2016.01.011 [PubMed: 26960245]
15. Dmitrienko A, Tamhane AC. Gatekeeping procedures with clinical trial applications. Pharm Stat. 2007; 6: 171–180. DOI: 10.1002/pst.291 [PubMed: 17583553]
16. European Agency for the Evaluation of Medicinal Products. Committee for proprietary medicinal products: points to consider on multiplicity issues in clinical trials. 2002.
17. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. Biometrika. 1977; 64: 191. doi: 10.2307/2335684
18. Whitehead, J. The Design and Analysis of Sequential Clinical Trials. Hoboken, NJ: John Wiley & Sons; 1997.
19. Core Team R. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.
20. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. Biometrics. 1979; 35: 549. doi: 10.2307/2530245 [PubMed: 497341]
21. Dunnett CW. A multiple comparison procedure for comparing several treatments with a control. J Am Stat Assoc. 1955; 50: 1096–1121.
22. Jaki T, Pallmann P, Magirr D. The R package MAMS for designing multi-arm multi-stage clinical trials. J Stat Softw. 2019; 88: 1–25. DOI: 10.18637/jss.v088.i04
23. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. Stat Med. 2010; 29: 219–228. DOI: 10.1002/sim.3748 [PubMed: 19827011]
24. Tamhane AC, Mehta CR, Liu L. Testing a primary and a secondary endpoint in a group sequential design. Biometrics. 2010; 66: 1174–1184. DOI: 10.1111/j.1541-0420.2010.01402.x [PubMed: 20337631]
25. Tamhane AC, Gou J, Jennison C, Mehta CR, Curto T. A gatekeeping procedure to test a primary and a secondary endpoint in a group sequential design with multiple interim looks. Biometrics. 2018; 74: 40–48. DOI: 10.1111/biom.12732 [PubMed: 28589692]
26. Jennison, C, Turnbull, BW. Group Sequential Methods with Applications to Clinical Trials. Boca Raton FL: Chapman & Hall / CRC Press; 2000.

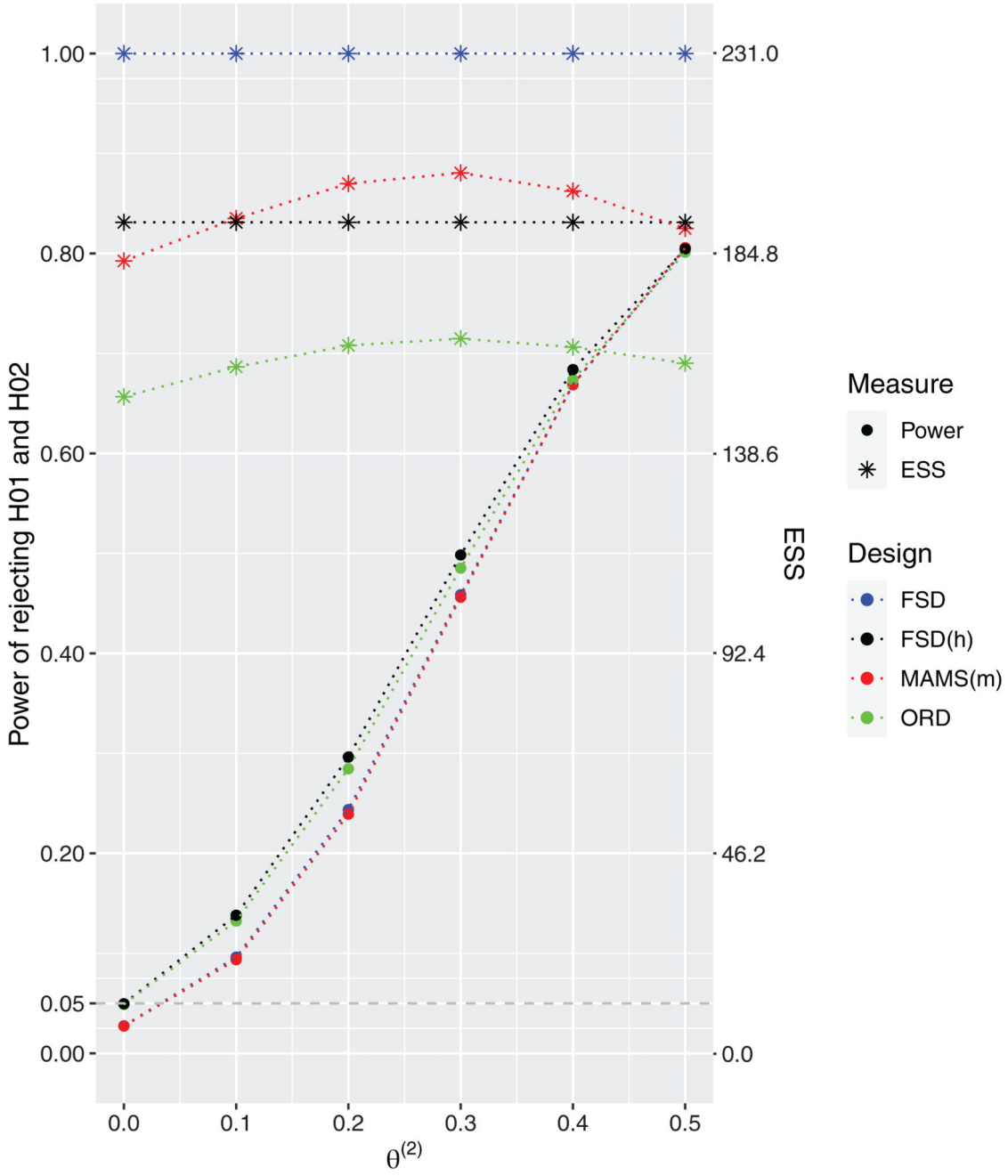

Figure 1.

Power and expected sample sizes (ESS) under $\theta = (0.5, \theta^{(2)})$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for the FSD, 3-arm 1-stage ORD (FSD(h)), 3-arm 2-stage ORD, and MAMS(m) designs when all designs are powered at 80% to reject both hypotheses under $\theta = (0.5, 0.5)$. 3-arm 2-stage ORD and MAMS(m) use triangular bounds. Results are provided using 10^6 replications

Europe PMC Funders Author Manuscripts

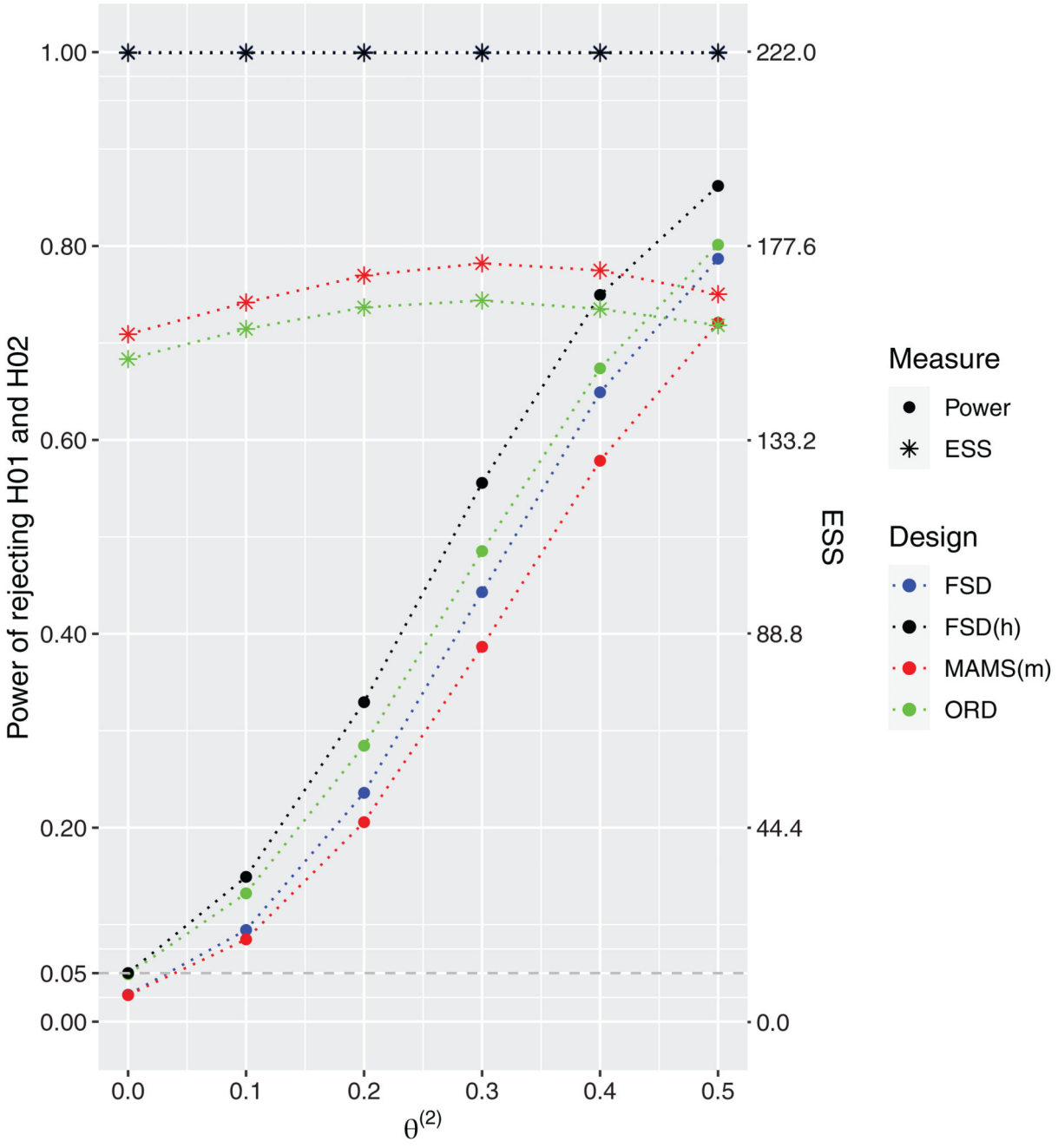

Figure 2.

Power and expected sample sizes (ESS) under $\theta = (0.5, \theta^{(2)})$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for the FSD, 3-arm 1-stage ORD (FSD(h)), 3-arm 2-stage ORD, and MAMS(m) designs when all designs have the same common total sample size equal to 222 patients. 3-arm 2-stage ORD and MAMS(m) use triangular bounds. Results are provided using 10^6 replications

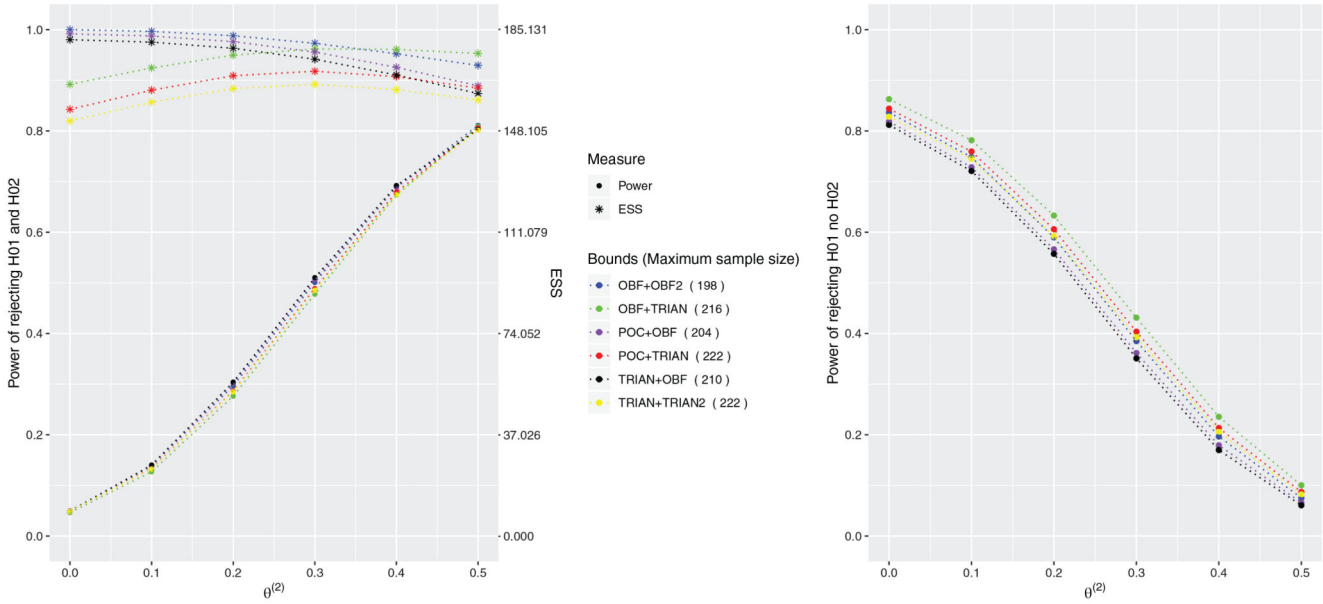

Figure 3.

Probability of rejecting both hypotheses (left) and probability of rejecting the first but not the second hypothesis (right) under $\theta = (0.5, \theta^{(2)})$ and $\theta^{(2)} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for the 3-arm 2-stage ORD design when it is powered at 80% to reject both hypotheses under $\theta = (0.5, 0.5)$. ORD uses the selected combination of bounds which control the type I error under $\theta = (\infty, 0)$. Results are provided using 10^6 replications

Europe PMC Funders Author Manuscripts

Table 1

Combination of the decision rules in the 3-arm 2-stage trial with $\theta^{(1)} \geq \theta^{(2)}$

	$Z_j^{(1)} \geq u_j^{(1)}$	$l_j^{(1)} < Z_j^{(1)} < u_j^{(1)}$	$Z_j^{(1)} \leq l_j^{(1)}$
$Z_j^{(2)} \geq u_j^{(2)}$	Stop: select T_1, T_2	Proceed with T_1, T_2	Proceed with T_1, T_2
$l_j^{(2)} < Z_j^{(2)} < u_j^{(2)}$	Proceed with T_2	Proceed with T_1, T_2	Drop both arms
$Z_j^{(2)} \leq l_j^{(2)}$	Stop: select T_1	Proceed with T_1	Drop both arms

Note: Cells colored in red correspond to contradicting evidence.

Table 2

Sets of events for K-arm J-stage design with $k \in \{1, \ldots, K-1\}$

Set	Definition
$C_j^{(k)}$	$\left\{ l_{j-1}^{(k)} < Z_{j-1}^{(k)} < u_{j-1}^{(k)} \right\}$
$S_j^{(k)}$	$\left\{ Z_{j-1}^{(k)} \leq l_{j-1}^{(k)} \right\}$
$E_j^{(k)}$	$\left\{ \left(\left(C_j^{(1)} \cup S_j^{(1)} \right) \cap \left(C_j^{(k)} \cup S_j^{(k)} \right) \right) \cup \left(E_j^{(k-1)} \cap C_j^{(k)} \right) \right\}$
$E_j^{(0)}$	Ω
$A_{j,1}^{(k)}, k > 0$	$E_j^{(k)}$
$A_{j,2}^{(k)}, k > 1$	$E_j^{(k-1)} \cap S_j^{(k)}$
$A_{j,k+1-s}^{(k)}, k > 2, s = 1:k-2$	$E_j^{(s)} \cap S_j^{(s+1)} \cap_{t=s+2}^{k} \left(C_j^{(t)} \cup S_j^{(t)} \right)$

Note: Ω is the whole sample space.

Europe PMC Funders Author Manuscripts

Table 3
Results of the simulations that revisit the NCT01257230 trial using the ORD

Design powered to reject all hypotheses							
$\theta_{(1)}$	$\theta_{(2)}$	**Max. SS**	**Stages**	**Reject all**	**Reject H_{01} not H_{02}**	**Reject at least one H_{0k}**	**ESS**
0	0	474	1	0.005	0.020	0.025	474.00
		534	2	0.004	0.021	0.025	316.39
120	0	474	1	0.025	0.856	0.881	474.00
		534	2	0.025	0.854	0.879	371.83
120	120	474	1	**0.803**	0.078	0.881	474.00
		534	2	**0.802**	0.081	0.883	399.81
Design powered to reject at least one hypothesis							
$\theta_{(1)}$	$\theta_{(2)}$	**Max. SS**	**Stages**	**Reject all**	**Reject H_{01} not H_{02}**	**Reject at least one H_{0k}**	**ESS**
0	0	381	1	0.005	0.021	0.025	381.00
		426	2	0.004	0.021	0.025	252.43
120	0	381	1	0.025	**0.779**	**0.803**	381.00
		426	2	0.024	0.774	0.798	304.67
120	120	381	1	**0.691**	0.112	**0.803**	381.00
		426	2	0.684	0.117	**0.802**	331.89

Note: For the two-stage design, the triangular bounds are used. Proportions refer to 10^6 replications and values of interest are in bold.
Abbreviations: ESS, expected sample size; Max. SS, maximum sample size.

Table 4

FWER, maximum sample sizes (Max. SS), expected sample sizes (ESS), and critical bounds under $\boldsymbol{\theta} = (0, 0)$ for the 3-arm 2-stage ORD, 3-arm 1-stage ORD (FSD(h)), FSD, and MAMS(m) designs when all designs are powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (0.5,0.5)$

Design	u_1, u_2, l_1	Max. SS	ESS	Reject at least one H_{0k}
FSD	1.917, -, 1.917	231	231.0	0.05
FSD(h)	1.644, -, 1.644	192	192.0	0.05
ORD	1.898, 1.789, 0.633	222	134.4	0.05
MAMS(m)	2.179, 2.055, 0.726	264	166.6	0.05

Note: 3-arm 2-stage ORD and MAMS(m) use triangular bounds. Results are provided using 10^6 replications.

Europe PMC Funders Author Manuscripts

Table 5

FWER, maximum sample sizes (Max. SS), expected sample sizes (ESS), and critical bounds under $\boldsymbol{\theta} = (0,0)$ for the 3-arm 2-stage ORD, 3-arm 1-stage ORD (FSD(h)), FSD, and MAMS(m) designs when all designs have the same common total sample size equal to 222 patients

Design	u_1, u_2, l_1	Max. SS	ESS	Reject at least one H_{0k}
FSD	1.917, -, 1.917	222	222.0	0.05
FSD(h)	1.644, -, 1.644	222	222.0	0.05
ORD	1.898, 1.789, 0.633	222	134.4	0.05
MAMS(m)	2.179, 2.055, 0.726	222	140.1	0.05

Note: 3-arm 2-stage ORD, and MAMS(m) use triangular bounds. Results are provided using 10^6 replications.