Received: 22 April 2022 | Revised: 27 January 2023 | Accepted: 13 April 2023

DOI: 10.1002/sim.9752

RESEARCH ARTICLE

Statistics in Medicine WILEY

A Bayesian multi-arm multi-stage clinical trial design incorporating information about treatment ordering

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

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

[2]Faculty for Informatics and Data Science, University of Regensburg, Regensburg, Germany

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

Funding information
National Institute for Health Research, Grant/Award Number: NIHR300576; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014; UK Medical Research Council, Grant/Award Numbers: MC_UU_00002/14, MC_UU_00002/19

Multi-Arm Multi-Stage (MAMS) designs can notably improve efficiency in later stages of drug development, but they can be suboptimal when an order in the effects of the arms can be assumed. In this work, we propose a Bayesian multi-arm multi-stage trial design that selects all promising treatments with high probability and can efficiently incorporate information about the order in the treatment effects as well as incorporate prior knowledge on the treatments. A distinguishing feature of the proposed design is that it allows taking into account the uncertainty of the treatment effect order assumption and does not assume any parametric arm-response model. The design can provide control of the family-wise error rate under specific values of the control mean and we illustrate its operating characteristics in a study of symptomatic asthma. Via simulations, we compare the novel Bayesian design with frequentist multi-arm multi-stage designs and a frequentist order restricted design that does not account for the order uncertainty and demonstrate the gains in the sample sizes the proposed design can provide. We also find that the proposed design is robust to violations of the assumptions on the order.

KEYWORDS

adaptive designs, Bayesian inference, infectious diseases, multi-arm multi-stage, order restriction

1 | INTRODUCTION

The development process of clinical trials can be lengthy and costly.[1] Adaptive clinical trial designs, which allow multiple treatment arms to be tested simultaneously,[2-4] can efficiently improve this process by allowing unpromising treatments to be dropped during the trial or by stopping the trial early for overwhelming benefit at a series of interim analyses.

Serra et al[5] have recently proposed an order restricted design that extends the multi-arm multi-stage framework[6-8] when 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, when nested combinations of treatments are investigated or when different treatment durations are simultaneously compared against a standard regimen. In their work, Serra et al[5] assumed that there was no uncertainty about the order of the treatment

Abbreviations: FWER, family-wise error rate; MAMS, multi-arm multi-stage; ORD, order restricted design.

effects. While this assumption is highly plausible in some settings, there could be uncertainty, for example, when an increase in side effects or compliance with the treatment (eg, for longer treatment durations) lead to reduced effects on the treatment efficacy. Thus, methodology that is able to relax the assumption on the order of the treatment effects is needed for such settings.

Although MAMS designs are commonly implemented using frequentist approaches, they can be extended to the Bayesian framework.[9,10] Bayesian sequential designs have been proposed by several researchers for clinical trials.[11-16] These designs are more flexible compared to frequentist designs as they allow incorporation of prior information about the efficacy of the treatment. The frequentist operating characteristics of Bayesian designs are important,[17-19] especially in confirmatory or late phase trials where it is required to control the risk of false positive results, which is the erroneous conclusion that new experimental treatments are efficacious. There are both proposals of Bayesian designs that do not consider control of the type I error[11] and that do (e.g. Ryan et al[15,16]). In the Bayesian setting, the posterior probabilities can be used to derive stopping boundaries, which are used to test the hypotheses and can be constructed to provide control of the frequentist type I error rate at the desirable level α.

In this work, we propose a Bayesian multi-arm multi-stage (B-MAMS) design that incorporates the information of the order among the treatment effects building on the frequentist framework developed by Magirr et al.[3] The proposed design can provide control of the overall type I error under specific values of the control mean and does not assume any parametric arm-response model. 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 design is provided in Section 3. Section 4 revisits the case study through an extensive simulation study before a discussion is provided in Section 5.

2 | CASE STUDY SETTING

The *Tiotropium add-on therapy in adolescents with moderate asthma: A 1-year randomized controlled trial* [NCT01257230][20] is a Phase III study assessing the efficacy and safety of 5 μg or 2.5 μg once-daily tiotropium via Respimat added to inhaled corticosteroid (ICS) with or without a leukotriene receptor antagonist in adolescent patients with moderate symptomatic asthma. The primary efficacy outcome was change from baseline in peak FEV$_1$ within 3 hours after dosing (peak FEV$_{1[0-3h]}$) measured after 24 weeks of treatment. Serra et al[5] revisited this study proposing an Order Restricted Design (ORD) when a monotonic dose-response relationship could be assumed. However, their design effectively disregarded the historical information coming from a previous study called the *Tiotropium in asthmatic adolescents symptomatic despite inhaled corticosteroids: A randomised dose-ranging study* [NCT01122680].[21] This Phase II study evaluated the safety and efficacy of three doses—5 μg, 2.5 μg, 1.25 μg—of once-daily tiotropium versus placebo in three 4-week treatment periods in adolescents with asthma. The primary efficacy outcome was change from baseline in peak FEV$_1$ within 3 hours after dosing (peak FEV$_{1[0-3h]}$).

In the following section, a Bayesian design which allows incorporation of the information from the Phase II NCT01122680 study[21] for the design and analysis of the Phase III trial[20] will be proposed. Moreover, the proposed Bayesian design will be able to account for the uncertainty in the monotonic dose-response relationship.

3 | BAYESIAN ORDER RESTRICTED DESIGN

Consider a clinical trial setting with $K-1$ active treatment arms, $T_1, \dots, T_{K-1}$, a control arm, T_0, and J stages at which treatment arms can be dropped or the trial could be stopped for efficacy or futility. Let t_j be the proportion of the total sample size at which the interim analysis j is done. The patients' responses are assumed to be independent and normally distributed at each stage j with known variances $\sigma_k^2 = \tau^{-1}, k \in \{0 \text{ (control arm)}, 1, \dots, K-1\}$. Let $Y_i^{(k)} \sim N\left(\mu^{(k)}, \sigma_k^2\right), k \in \{0, \dots, K-1\}, i = 1 : n_j^{(k)}, j \in \{1, \dots, J\}$ be the observation of the ith patient on treatment k and $n_j^{(k)}$ be the number of patients on arm k up to stage j. Define $\boldsymbol{n_j} = (n_j^{(1)}, \dots, n_j^{(K-1)})$ as the vector of sample sizes on the active arms at stage j and $n_j^{(0)}$ as the sample size on the control arm at stage j. Let $\overline{Y}_j^{(k)} = (n_j^{(k)})^{-1} \sum_{i=1}^{n_j^{(k)}} Y_i^{(k)},\ k \in \{1, \dots, K-1\}$ be the cumulative outcome mean on treatment k up to stage j and the vector of outcome means up to stage j is $\overline{\boldsymbol{Y}}_{\boldsymbol{j}} = (\overline{Y}_j^{(1)}, \dots, \overline{Y}_j^{(K-1)})$. Denote the covariance matrix of the joint distribution of the active treatment responses by $\boldsymbol{\Sigma}$, which corresponds to a diagonal matrix with the vector $(\sigma_1^2, \dots, \sigma_{K-1}^2)$ on the diagonal and zero on the off-diagonal.

Let $\theta^{(k)} = \mu^{(k)} - \mu^{(0)}$ be the treatment effect of arm $k \in \{1, \dots, K-1\}$ compared to control and $\boldsymbol{\theta} = (\theta^{(1)}, \dots, \theta^{(K-1)})$ be the vector of treatment effects. The null hypotheses of interest are $H_{01} : \{\theta^{(1)} \leq 0\}, \dots, H_{0K-1} : \{\theta^{(K-1)} \leq 0\}$ and the global null hypothesis is defined as $H_0 : \{\theta^{(1)} = \theta^{(2)} = \cdots = \theta^{(K-1)} = 0\}$. Let η_j, ϵ_j be the upper and lower critical values at stage j (with $\eta_J = \epsilon_J$) that are used for the decision-making throughout the trial (as described below).

Consider the following Bayesian model:

$$\begin{aligned}
&Y_i^{(k)} \sim N\left(\mu^{(k)}, \tau^{-1}\right) \quad k = 0, \dots, K-1 \text{ and } i = 1, \dots, n_j^{(k)}\\
&\mu^{(1)} \sim N\left(\mu_0^{(1)}, \tau_{01}^{-1}\right), \quad \mu^{(0)} \sim N\left(\mu_0^{(0)}, \tau_{00}^{-1}\right)\\
&\delta^{(k)} \sim N\left(\delta_0^{(k)}, \left(\tau_d^{(k)}\right)^{-1}\right), \quad \mu^{(k+1)} = \mu^{(k)} - \delta^{(k)} \quad k = 1, \dots, K-2.
\end{aligned}$$

If $\delta^{(k)}, k \in \{1, \dots K-2\}$ are, for instance, restricted to be positive random variables, then the treatment means are assumed to be ordered: $\mu^{(1)} \geq \dots \geq \mu^{(K-1)}$. However, in this setting, we assume that there is uncertainty in the order of the treatment means effects and we consider a normal prior distribution for $\delta^{(k)}$. This parametrization allows to correlate the treatment means to each other and to capture the uncertainty of the treatment means effects order through the distribution of $\delta^{(k)}$. Alternative parametrization strategies are discussed in Section 5.

The joint distribution of the response means is normally distributed with mean $\boldsymbol{\mu} = (\mu_0^{(1)}, \dots, \mu_0^{(s)}, \dots, \mu_0^{(K-1)})$, where $\mu_0^{(s)} = \mu_0^{(1)} - \sum_{r=1}^{s-1} \delta_0^{(r)}, s \in \{2, \dots, K-1\}$ and covariance matrix $\boldsymbol{\Omega}$. The covariance matrix has diagonal elements defined as $\text{Var}\left(\mu_0^{(s)}\right) = \tau_{01}^{-1} + \sum_{r=1}^{s-1} (\tau_d^{(r)})^{-1}, s \in \{2, \dots, K-1\}$ and the non-diagonal element at row s and column t with $s, t \in \{1, \dots K-1\}$, is $\text{Var}\left(\mu_0^{(\min(s,t))}\right)$. The posterior distribution of the treatment means given the cumulative data up to stage j is

$$\begin{pmatrix} \mu^{(1)} \\ \cdot \\ \cdot \\ \mu^{(K-1)} \end{pmatrix} | \overline{\boldsymbol{Y_j}} \sim N\left(\boldsymbol{\Gamma_j^{-1} a_j}, \boldsymbol{\Gamma_j^{-1}}\right),$$

with $\boldsymbol{\Gamma_j} = \boldsymbol{\Omega^{-1}} + \boldsymbol{n_j \Sigma^{-1}}$, $\boldsymbol{a_j} = \boldsymbol{\Omega^{-1} \mu} + \boldsymbol{\Sigma^{-1} n_j \overline{Y}_j}$.

At each interim analysis, the decision to stop the trial for efficacy or to drop the unpromising arms is based on the evaluation of the posterior probability of the treatment effect given the data up to stage j. At each stage, the posterior probability that the treatment effect relative to arm k is greater than zero, $P(\theta^{(k)} > 0 | \overline{\boldsymbol{Y_j}})$, is computed. If this probability is below a pre-specified threshold, ϵ_j, that is, $P(\theta^{(k)} > 0 | \overline{\boldsymbol{Y_j}}) < \epsilon_j$ then treatment arm k is dropped. Arm k, instead, can be declared as superior to control at stage j if the posterior probability of the treatment effect being greater than zero given the observed data is larger than a specific threshold, η_j, that is, $P(\theta^{(k)} > 0 | \overline{\boldsymbol{Y_j}}) \geq \eta_j$.

If the posterior probability falls inside the two bounds, $\epsilon_j \leq P(\theta^{(k)} > 0 | \overline{\boldsymbol{Y_j}}) < \eta_j$, then the decision to proceed further with arm k and control is made as there is not enough information to decide whether to stop the arm or to claim its efficacy. Differently from the conventional frequentist MAMS design,[3] the trial can still continue when an arm has crossed the upper bound and is stopped for efficacy. In this case, all other treatment arms remaining in the trial can proceed to the next stage. In this way, a decision on each treatment arm can be made at the end of the study, and the power to reject all hypotheses is increased.

3.1 | Family-wise error rate and power requirements

In confirmatory clinical trials, the control of the family-wise error rate (FWER) is often required.[22] For the proposed design, we ensure weak control of the FWER under the global null and it provides the control of the FWER under the partial null hypotheses (see details below). While the design might not guarantee control of the FWER under the entire null space, it can benefit from the Bayesian framework while still providing protection from erroneous conclusions[23] under particular regions of the null space.

The critical bounds η_j, ϵ_j can be chosen in order to maintain such control at level α. The FWER under the global null can be written as

$$\begin{aligned} &P(\text{rejecting at least one true } H_{0k},\ k \in \{1, \dots, K-1\} | H_0) \\ &\quad = P(P(\theta^{(k)} > 0 | \overline{\boldsymbol{Y}_j}) \geq \eta_j \text{ for some } (k,j) \in \{1, \dots, K-1\} \times \{1, \dots, J\} | H_0). \end{aligned} \tag{1}$$

Define $A_{k,j} = \{P(\theta^{(k)} > 0|\overline{\boldsymbol{Y}_j}) < \epsilon_j\}$ and $B_{k,j} = \{\epsilon_j \leq P(\theta^{(k)} > 0|\overline{\boldsymbol{Y}_j}) < \eta_j\}$. If the event $A_{k,j}$ occurs then the treatment arm k is dropped for futility at stage j. If instead $B_{k,j}$ is satisfied, the arm k can proceed to stage $j+1$. The event that all hypotheses $H_{01}, \dots, H_{0K-1}$ fail to be rejected can be defined as

$$R_{K-1} = \bigcap_{k=1}^{K-1} \left(\bigcup_{j=1}^{J} \left[\left\{ \bigcap_{i=1}^{j-1} B_{k,i} \right\} \cup A_{k,j} \right] \right).$$

It follows that

$$P(\text{rejecting at least one true } H_{0k},\ k \in \{1, \dots, K-1\} | H_0) = 1 - P(R_{K-1} | H_0). \tag{2}$$

The objective of the study is to identify all promising treatment arms. Thus, the design is powered to reject all hypotheses for which the true treatment effects are different from zero: $\boldsymbol{\theta} = (\theta^{(1)}, \dots, \theta^{(K-1)})$, with $\theta^{(1)}, \dots, \theta^{(K-1)} > 0$. Using the definition

$$\overline{R}_{K-1} = \bigcap_{k=1}^{K-1} \left(\bigcup_{j=1}^{J} \left[\left\{ \bigcap_{i=1}^{j-1} B_{k,i} \right\} \cup \overline{\{A_{k,j} \cup B_{k,j}\}} \right] \right),$$

the power to reject all hypotheses is equal to $P(\overline{R}_{K-1}|\boldsymbol{\theta})$.

Below, we provide the example on what form the quantities introduced above take under the setting of the motivating 3-arm 2-stage trial.

3.1.1 | 3-arm 2-stage

For the 3-arm 2-stage trial, the event that all hypotheses fail to be rejected is

$$\begin{aligned} R_2 = \Big\{ &\{\{P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_1}) < \epsilon_1\} \cap \{P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_1}) < \epsilon_1\}\} \\ &\cup \{\{P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_1}) < \epsilon_1\} \cap \{\epsilon_1 \leq P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_1}) < \eta_1\} \cap \{P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_2}) < \epsilon_2\}\} \\ &\cup \{\{P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_1}) < \epsilon_1\} \cap \{\epsilon_1 \leq P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_1}) < \eta_1\} \cap \{P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_2}) < \epsilon_2\}\} \\ &\cup \{\{\epsilon_1 \leq P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_1}) < \eta_1\} \cap \{P(\theta^{(2)} > 0|\overline{\boldsymbol{Y}_2}) < \epsilon_2\} \cap \{\epsilon_1 \leq P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_1}) < \eta_1\} \cap \{P(\theta^{(1)} > 0|\overline{\boldsymbol{Y}_2}) < \epsilon_2\}\} \Big\}. \end{aligned}$$

Below, we describe how to search for the critical bounds $\eta_j, \epsilon_j, j \in \{1, \dots, J\}$ and the sample size in order to satisfy the FWER and the power requirements.

3.2 | Search of the critical bounds and the sample size

To control the FWER at level α under the global null hypothesis and to power the design at $1-\beta$ under the configuration $\boldsymbol{\theta}$, the following inequalities need to be satisfied

$$1 - P(R_{K-1}|H_0) \leq \alpha, \text{ and } P(\overline{R}_{K-1}|\boldsymbol{\theta}) \geq (1-\beta). \tag{3}$$

Firstly, consider how to compute the single probability events $P(A_{k,j})$ and $P(B_{k,j})$. It can be seen that the inequalities $P(\theta^{(k)} > 0|\overline{\mathbf{Y}_j}) < \epsilon_j$ and $\epsilon_j \leq P(\theta^{(k)} > 0|\overline{\mathbf{Y}_j}) < \eta_j$ are satisfied if

$$\frac{\mathbb{E}\left[\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right] - \mathbb{E}\left[\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right]}{\sqrt{\mathrm{Var}\left(\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right) + \mathrm{Var}\left(\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right)}} < z_{\epsilon_j} \text{ and } z_{\epsilon_j} \leq \frac{\mathbb{E}\left[\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right] - \mathbb{E}\left[\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right]}{\sqrt{\mathrm{Var}\left(\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right) + \mathrm{Var}\left(\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right)}} < z_{\eta_j},$$

respectively, where $z_{\eta_j}, z_{\epsilon_j}$ are the η_j, ϵ_j-percentiles of a standard normal. Given the Bayesian model described in Section 3 it follows that

$$\mathbb{E}\left[\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right] = \frac{\mu_0^{(0)}\tau_{00} + \overline{Y}_j^{(0)} I_j^{(0)}}{\tau_{00} + I_j^{(0)}}, \quad I_j^{(0)} = \frac{n_j^{(0)}}{\sigma_0^2}, \quad \mathrm{Var}\left(\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right) = \frac{1}{\tau_{00} + I_j^{(0)}},$$

$$\mathbb{E}\left[\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right] = \sum_{i=1}^{K-1} c_{ij}^{(k)} \overline{Y}_j^{(i)} + \gamma_j^{(k)}, \quad c_{ij}^{(k)} = \left(\mathbf{\Gamma_j^{-1}\Sigma^{-1}n_j}\right)[k,i], \quad \gamma_j^{(k)} = \left(\mathbf{\Gamma_j^{-1}\Omega^{-1}}\boldsymbol{\mu}\right)[k], \quad \mathrm{Var}\left(\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right) = (\mathbf{\Gamma_j^{-1}})[k,k],$$

where the notation $[k, i]$ indicates the element of the matrix at row k and column i, while $[k]$ refers to the kth position in the vector. Thus, the inequality $P(\theta^{(k)} > 0|\overline{\mathbf{Y}_j}) < \epsilon_j$ translates into solving

$$\sum_{i=1}^{K-1} c_{ij}^{(k)} \overline{Y}_j^{(i)} - \frac{\overline{Y}_j^{(0)} I_j^{(0)}}{\tau_{00} + I_j^{(0)}} < -\gamma_j^{(k)} + \frac{\mu_0^{(0)}\tau_{00}}{\tau_{00} + I_j^{(0)}} + z_{\epsilon_j}\sqrt{\mathrm{Var}\left(\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right) + \mathrm{Var}\left(\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right)},$$

and the distribution of the linear combination of the outcome variable means is

$$\sum_{i=1}^{K-1} c_{ij}^{(k)} \overline{Y}_j^{(i)} - c_j^{(0)} \overline{Y}_j^{(0)} \sim N\left(\sum_{i=1}^{K-1} c_{ij}^{(k)} \mu^{(i)} - c_j^{(0)} \mu^{(0)}, \sum_{i=1}^{K-1} \frac{(c_{ij}^{(k)}\sigma_i)^2}{n_j^{(i)}} + \frac{(c_j^{(0)}\sigma_0)^2}{n_j^{(0)}}\right), \tag{4}$$

where $c_j^{(0)} = \frac{I_j^{(0)}}{\tau_{00}+I_j^{(0)}}$.

Thus, the probability of the event $A_{k,j}$ becomes $P(A_{k,j}) = F\left(-\gamma_j^{(k)} + \frac{\mu_0^{(0)}\tau_{00}}{\tau_{00}+I_j^{(0)}} + z_{\epsilon_j} v_j^k\right)$, where $v_j^k = \sqrt{\mathrm{Var}\left(\mu^{(k)}_{|\overline{\mathbf{Y}_j}}\right) + \mathrm{Var}\left(\mu^{(0)}_{|\overline{Y}^{(0)}_j}\right)}$ and $F(x)$ is the cumulative density function of the normal distribution in (4). Similarly, the probability of the event $B_{k,j}$ can be computed.

When multiple arms are concurrently studied, the joint distribution of the probability events is a multivariate normal distribution and the covariance matrix changes depending on the number of arms that have reached a particular stage. For instance, the covariance between arms s, t at stage j and $j^\star$, $j \leq j^\star$ is

$$\mathrm{Cov}\left(\sum_{i=1}^{K-1} c_{ij}^{(s)} \overline{Y}_j^{(i)} - c_j^{(0)} \overline{Y}_j^{(0)}, \sum_{i=1}^{K-1} c_{ij^\star}^{(t)} \overline{Y}_{j^\star}^{(i)} - c_{j^\star}^{(0)} \overline{Y}_{j^\star}^{(0)}\right) = \sum_{i=1}^{K-1} c_{ij}^{(s)} c_{ij^\star}^{(t)} \mathrm{Cov}(\overline{Y}_j^{(i)}, \overline{Y}_{j^\star}^{(i)}) + (c_j^{(0)} c_{j^\star}^{(0)})\mathrm{Var}(\overline{Y}_{j^\star}^{(0)}),$$

where $\mathrm{Cov}(\overline{Y}_j^{(i)}, \overline{Y}_{j^\star}^{(i)}) = t_j \mathrm{Var}(\overline{Y}_j^{(i)})$ if treatment i has reached stage $j^\star \neq j$ and the interim analysis is done at t_j% of the total population, otherwise $\mathrm{Cov}(\overline{Y}_j^{(i)}, \overline{Y}_{j^\star}^{(i)}) = \mathrm{Var}(\overline{Y}_j^{(i)})$, with $\mathrm{Var}(\overline{Y}_j^{(i)}) = \frac{\sigma_i^2}{n_j^{(i)}}$.

In order to compute $P(R_{K-1})$, for each event in R_{K-1}, one needs to find $c_{ij}^{(k)}, \gamma^{(k)}$ and the covariance matrix of the multivariate normal distribution. The critical values can be expressed as a function of a parameter, p, $\epsilon_j = \epsilon_j(p), \eta_j = \eta_j(p)$, so that the search is restricted to a grid of values for p with $\epsilon_j(p), \eta_j(p) \in [0, 1]$ and the sample size $\boldsymbol{n_j}$ in order to satisfy the inequalities in (3). Additional constraints can be considered in order to maintain control of the FWER under the partial null hypotheses. In the 3-arm setting, we use these partial null configurations: $(\theta^{(1)}, 0)$ and $(0, \theta^{(2)})$. However, it can be observed that the FWER computed under the global null hypothesis depends on the value of $\mu^{(0)}$. As outlined in Stallard et al,[24] control of the type I error is not possible for every value of $\mu^{(0)}$. Indeed, under the hypothesis that $\mu^{(i)} = \mu^{(0)}, i \in \{1, \dots, K-1\}$, if for example $\mu^{(0)} \to \infty$ and $\sum_{i=1}^{K-1} c_{ij}^{(k)} - c_j^{(0)} > 0$ then the expected value of the normal distribution in Equation (4) goes to infinity and thus in this case it is not possible to control the FWER at level α. Thus, the critical bounds can be found for specific values of the treatment means under which the FWER is controlled. Exploratory analyses can then be conducted in order to explore the operating characteristics of the design for different true values of the control mean.

The description of the probability events to compute the FWER for the special case of a 3-arm 2-stage design is described in detail in Section 1 of Data S1.

In the next section, an extensive simulation study will be performed in order to evaluate the operating characteristics of the proposed design in the setting of the motivating trial.[20]

4 | CASE STUDY REVISITED

4.1 | 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][20] using prior information from the study *Tiotropium in asthmatic adolescents symptomatic despite inhaled corticosteroids: A randomised dose-ranging study* [NCT01122680][21] introduced in Section 2.

In the phase III study (NCT01257230), patients were randomized in a 1:1:1 ratio to receive 5 μg or 2.5 μg of once-daily tiotropium or placebo. The null hypotheses were tested in a stepwise manner to control the type I error at level $\alpha = 0.025$. The study was powered at 80% to detect a difference of 120 mL between treatments in the change from baseline of peak $\text{FEV}_{1[0-3h]}$ assuming a common standard deviation 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 proposed Bayesian design and the prior distributions are constructed using information from the previous study (NCT01122680).

In line with the Phase III trial, we assume that the change from baseline of peak $\text{FEV}_{1[0-3h]}$ is normally distributed with standard deviation $\sigma_k = 340, k \in \{0, 1, 2\}$. For the prior distributions, we set $\mu_0^{(0)} = 489$ and $\mu_0^{(1)} = 602$, as these are the estimated mean responses obtained at the end of the Phase II NCT01122680 study. We center the prior distribution of $\delta^{(1)}$ at 0, so that a-priori we do not assume a difference between the treatment means. Given the lack of information regarding the between-patient variability, we firstly consider uninformative priors with $\tau_{00} = \tau_{01} = \tau_d^{(1)} = 10^{-6}$. Additional analyses are then conducted in order to explore how the design performs with more informative prior distributions.

For the simulation scenarios, we set the true $\mu^{(0)}$ to be equal to the estimated mean response obtained at the end of Phase II, that is $\mu^{(0)} = 489$, while for the treatment arms $\mu^{(k)} = \mu^{(0)} + \theta^{(k)}, k \in \{1, 2\}$. As in the original study, we consider an improvement of FEV_1 of 120 of interest. The trial is then designed to achieve 80% power under the configuration $\boldsymbol{\theta} = (120{,}120)$.

We evaluate the performance of the design under various treatment effect configurations when a monotonic order is assumed to be true: $\theta^{(1)} = 120$ and $\theta^{(2)} \in \{0, 20, 40, 60, 80{,}120\}$ and the configurations where the order is violated: $\theta^{(2)} = 120$ and $\theta^{(1)} \in \{0, 20, 40, 60, 80{,}120\}$.

The sample size and the critical bounds are found in order to reject all hypotheses at 80% when $\boldsymbol{\theta} = (120{,}120)$ with $\mu^{(1)} = 609, \mu^{(2)} = 609, \mu^{(0)} = 489$ and to control the FWER at level $\alpha = 0.025$ under $H_0 = \{\mu^{(1)} = \mu^{(2)} = \mu^{(0)} = 489\}$ and two partial nulls $H_{01} = \{\mu^{(1)} = \mu^{(0)} = 489, \mu^{(2)} = 609\}$ and $H_{02} = \{\mu^{(1)} = 609, \mu^{(2)} = \mu^{(0)} = 489\}$. A single interim analysis after half of the total study population is observed and triangular[25] critical bounds are used, that is $\eta_2 = \epsilon_2 = \Phi\left(\frac{\Phi^{-1}(\eta_1)}{(1+0.5)}\frac{2}{\sqrt{2}}\right), \epsilon_1 = \Phi\left(\frac{\Phi^{-1}(\eta_1)}{(1+0.5)}\frac{1}{2}\right)$, where $\Phi(\cdot)$ is the cumulative density function of a standard normal distribution. 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.

The proposed Bayesian design is compared to the frequentist ORD,[5] developed under an order assumption among the treatment effects, the modified MAMS design, MAMS(m),[3] whose modification allows all treatment arms that have not crossed any bounds to continue to the next stage even when one arm crossed the upper boundary, and the frequentist design proposed by Urach and Posch[26] (Urach & Posch) with separate stopping rules and futility critical boundaries which apply the closure principle to define a sequentially rejective group sequential test.

TABLE 1 Probability to reject at least one hypothesis (FWER), maximum sample size (Max SS) and expected sample size (ESS) for each design under the global null hypothesis H_0 with $\mu^{(1)} = \mu^{(2)} = \mu^{(0)} = 489$ when all designs are powered at 80% to reject all hypotheses under $\boldsymbol{\theta} = (120{,}120)$.

Design	u_1, u_2, l_1, v_1, v_2	$\eta_1, \eta_2, \epsilon_1$	Max SS	ESS	FWER
MAMS(m)	2.482, 2.34, 0.827, -, -		612	380.85	0.025
B-MAMS(U)		0.9934, 0.9906, 0.7959	612	378.43	0.025
ORD	2.223, 2.095, 0.741, -, -		534	316.91	0.024
B-MAMS(D)		0.9927, 0.9893, 0.7921	576	355.19	0.024
Urach & Posch	2.482, 2.34, 0.827, 2.222, 2.095		552	343.12	0.025
B-MAMS(C)		0.9906, 0.9866, 0.7832	492	303.75	0.024

Notes: Results are provided using 10^4 simulations. Upper—$u_j, j \in \{1, 2\}$—and lower—l_1—triangular critical bounds are provided for the ORD and MAMS(m) frequentist designs. For the Urach & Posch design $u_j, j \in \{1, 2\}$ and l_1 are the upper and lower global boundaries, while $v_j, j \in \{1, 2\}$ are the elementary boundaries.

All designs use triangular critical bounds[25] and the numerical results are found using R[27] and 10^4 replicate simulations. In the simulations, the posterior distributions are estimated using MCMC samples through the package `rjags`.[28] The critical bounds and the sample size for the Bayesian design are found computing all probabilities to reject at least one null hypothesis under the global null and the probability to reject both hypotheses under the partial null configurations as described in Sections 3.1 and 3.2, without the use of MCMC sampling for the posterior distributions. The package `mvtnorm`[29] is used to compute the integrals in R.[27] Results using Pocock[30] critical bounds are provided in Section 2 of Data S1.

4.2 | Numerical results

We start by considering the Bayesian design with no prior information for the treatment means and the treatment mean difference, that is $\tau_{00} = \tau_{01} = \tau_d^{(1)} = 10^{-6}$. We refer to this version of the design as to B-MAMS(U). The obtained design parameters and results of the simulations under the global null hypothesis for all considered approaches are given in Table 1. One can note that all considered designs control the FWER at level $\alpha = 0.025$ as required.

The B-MAMS(U) design requires a total maximum sample size of 612 patients, which is the same maximum total sample size required by the MAMS(m) to achieve 80% of power to reject both hypotheses. At the same time, the ORD and the Urach & Posch designs require a total of 534 and 552 patients, respectively.

One of the advantages of using a Bayesian design is that it allows incorporating historical information. One possibility is, for example, to include information on the treatment mean difference between the active treatment arms. Specifically, we use $\tau_d^{(1)} = 5.9 \times 10^{-5}$ (and $\tau_{00} = \tau_{01} = 10^{-6}$ as before) and refer to this design option of the Bayesian design as to B-MAMS(D). The Bayesian design B-MAMS(D) can control the FWER at level α under the global and partial null hypotheses and the power to reject both hypotheses is 80%. This design requires a total maximum sample size of 576 patients, which is higher compared to the ORD and Urach & Posch designs—for these 42 and 24 fewer patients are required respectively—but lower compared to the MAMS(m) design.

Another possibility is to include prior information on the control arm. We refer to this design as to B-MAMS(C) and we consider a prior on control that is worth an extra 20% of data compared to the frequentist MAMS(m) design. If $\tau_{00} = 0.00039$, $\tau_{01} = \tau_d^{(1)} = 10^{-6}$, then the Bayesian design would require 82 patients per arm per stage that leads to a total maximum sample size of 492 patients compared to 612 patients in the frequentist MAMS(m) design and 534 for the ORD design.

The designs' performances under the non-global null configurations are given in Figures 1 and 2 showing the probability to reject both hypotheses and the second hypothesis, respectively. Note that, when the first or the second treatment are no different from control, $\theta^{(1)} = 0$ or $\theta^{(2)} = 0$, the probability to reject all hypotheses is controlled at level α for all considered designs.

It can be observed that the B-MAMS(U) design matches the operating characteristics of the frequentist MAMS(m) design. Indeed, for all considered scenarios the two designs have the same probability to reject both hypotheses and the same probability to reject the second hypothesis H_{02}. The two designs provide also the same ESS under all considered scenarios. Compared to the other two frequentist approaches, when the order is satisfied, the B-MAMS(U) design shows

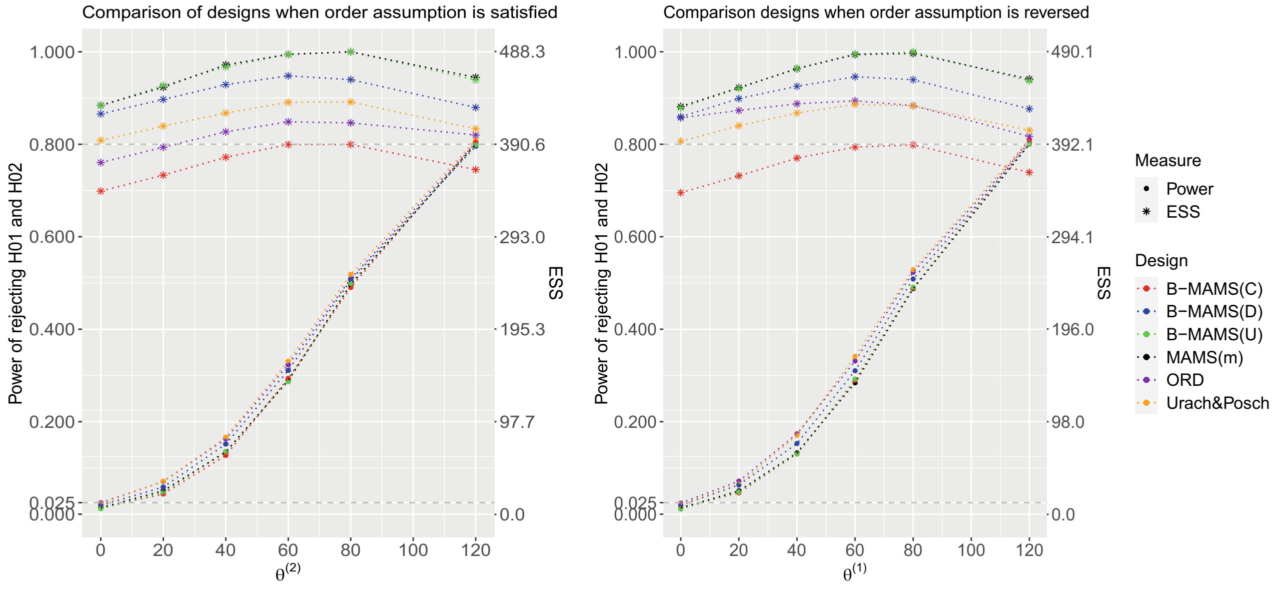

FIGURE 1 Power to reject all hypotheses and expected sample sizes (ESS) under $\boldsymbol{\theta} = (120, \theta^{(2)})$ and $\theta^{(2)} \in \{0, 20, 40, 60, 80, 120\}$ (left) and under $\boldsymbol{\theta} = (\theta^{(1)}, 120)$ and $\theta^{(1)} \in \{0, 20, 40, 60, 80, 120\}$ (right) for the 3-arm 2-stage MAMS(m), ORD, Urach & Posch and Bayesian designs when all designs are powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (120,120)$. All designs use triangular bounds. Results are provided using 10^4 replications.

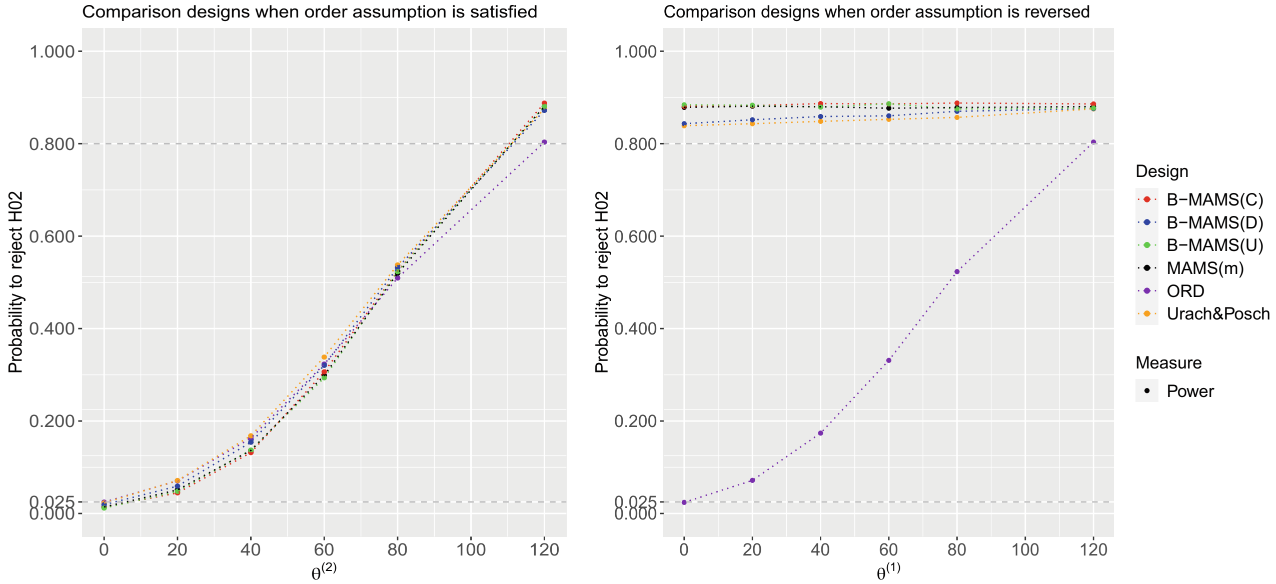

FIGURE 2 Probability to reject H_{02} under $\boldsymbol{\theta} = (120, \theta^{(2)})$ and $\theta^{(2)} \in \{0, 20, 40, 60, 80, 120\}$ (left) and under $\boldsymbol{\theta} = (\theta^{(1)}, 120)$ and $\theta^{(1)} \in \{0, 20, 40, 60, 80, 120\}$ (right) for the 3-arm 2-stage MAMS(m), ORD, Urach & Posch and Bayesian designs when all designs are powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (120,120)$. All designs use triangular bounds. Results are provided using 10^4 replications.

lower power of rejecting both hypotheses—a difference up to 4.4% and 3.6% compared to the Urach & Posch and ORD designs respectively (see the left panel in Figure 1). As shown on the left panel in Figure 2, the probability to reject the second hypothesis is smaller for the B-MAMS(U) design when $\theta^{(2)} \leq 80$ compared to the Urach & Posch design—a difference up to 4.4%—and smaller compared to the ORD when $\theta^{(2)} \leq 60$—a difference up to around 3%. When $\theta^{(2)} = 120$, the B-MAMS(U) shows an increase of around 7% in the probability to reject the second hypothesis compared to the ORD. In terms of ESS, the B-MAMS(U) shows an increase up to 12% and 18% depending on the scenario compared to the Urach & Posch and ORD designs, respectively, when the assumption on the treatment effect order is satisfied. When this assumption is violated, the B-MAMS(U) has lower power—a difference up to 4.9% and 4.3% compared to the Urach &

Posch and ORD designs—to reject both hypotheses. However, the probability to reject the second hypothesis is higher for the B-MAMS(U) design compared to the Urach & Posch and ORD designs—an increase up to 4.5% and 86%, respectively depending on the simulation scenario (right panel in Figure 2). In terms of ESS, the B-MAMS(U) shows an increase up to 13% and 14% depending on the scenario compared to the Urach & Posch and ORD designs, respectively.

When the order is satisfied, the probability to reject both hypotheses for the B-MAMS(D) is slightly lower—a difference up to around 1.2% and 2%—compared to the frequentist ORD and Urach & Posch designs respectively, while it is slightly higher compared to the MAMS(m)—an increase up to 2%. The B-MAMS(D) design shows higher probability to reject the second hypothesis compared to the ORD when $\theta^{(2)} \geq 80$—an increase of around 3% and 7% when $\theta^{(2)} = 80$ and $\theta^{(2)} = 120$, respectively. However, the probability to reject the second hypothesis is slightly lower compared to the Urach & Posch design (see plot on the left in Figure 2). In addition, the B-MAMS(D) design has lower ESS compared to the MAMS(m)—a reduction up to 6.8% depending on the simulation scenario, while it has higher ESS compared to the ORD and Urach & Posch designs—up to 13% and 7%, respectively, depending on the scenario when the order assumption is satisfied. When the order is reversed, the probability to reject the second hypothesis is higher for the B-MAMS(D) design: it is above 80% and slightly higher compared to the Urach & Posch design, while it is noticeably smaller for the ORD. However, the B-MAMS(D) design shows a decrease up to around 2% and 3% compared to the frequentist ORD and Urach & Posch designs respectively in the probability to reject both hypotheses. The increase in power for the frequentist designs is due to the hierarchical testing procedure that is not used in the proposed Bayesian designs.

For all the considered scenarios, the probability to reject both hypotheses is almost the same for the B-MAMS(C) and the MAMS(m) designs independently of the order assumption. However, the total maximum sample size for the Bayesian design is 20% smaller compared to the frequentist design. The B-MAMS(C) provides advantages to reject the second hypothesis compared to all other designs when the order assumption is reversed. Indeed, the B-MAMS(C) has the highest probability to reject the second hypothesis and the smallest ESS compared to all other designs for all considered scenarios—a reduction in ESS up to 14%, 19%, and 20% compared to the Urach & Posch, ORD and MAMS(m) designs respectively.

Overall, the results suggest that the proposed Bayesian design provides benefits compared to the considered frequentist approaches in terms of ESS—the B-MAMS(C) provides a reduction up to 14%, 19%, and 20% compared to the Urach & Posch, ORD, and MAMS(m) designs, respectively, while the B-MAMS(D) provides a reduction up to 6.8% compared to the MAMS(m) design—and probability to reject the second hypothesis—an increase up to around 1.3%, 4.2%, and 4.6% for the B-MAMS(D), B-MAMS(C), and B-MAMS(U) compared to the Urach & Posch and an increase up to 86% compared to the ORD—when the order assumption is not satisfied. Nevertheless, the proposed design shows a reduction in power to reject both hypotheses—the B-MAMS(D) shows a difference up to around 2% and 3% compared to the ORD and Urach & Posch designs, while the B-MAMS(C) and B-MAMS(U) provide both a difference of around 5% and 4.3% compared to the Urach & Posch and ORD designs respectively. In addition, when historical information is disregarded, the proposed design can match the operating characteristics of the MAMS(m) frequentist design.

4.3 | Explorations of the FWER control

In the simulation studies described above, $\mu^{(0)}$ was assumed to be equal to the true value mean on control. In this section, we investigate the robustness of the results when the true value and the prior mean on control differ.

In Figure 3, we explore the probabilities to reject at least one null hypothesis under the global null (ie, $\boldsymbol{\theta} = (0, 0)$) and the probability to reject both hypotheses under the partial null configurations (ie, $\boldsymbol{\theta} = (120, 0)$ and $\boldsymbol{\theta} = (0, 120)$) for several values of the mean of the control arm ($\mu^{(0)}$). The numerical results are provided using an analytical expression—computing all probabilities as described in Sections 3.1 and 3.2, without the use of MCMC sampling for the posterior distributions.

It can be observed that, for the B-MAMS(U) and the B-MAMS(D), the top and the middle panel of Figure 3, respectively, the FWER under the global null hypothesis is at most around 2.6% (an inflation of around 0.1%) when the true mean is off the prior mean by a factor of 3, $\mu^{(0)} = 1489$, whereas the probability to reject both hypotheses under $\boldsymbol{\theta} = (0,120)$ and $\boldsymbol{\theta} = (120, 0)$ is controlled for both designs—it is below 2% for all values of $\mu^{(0)}$. The bottom panel of Figure 3 shows how the probability to reject at least one or all hypotheses changes when $\mu^{(0)}$ is shifted from the true prior mean for the B-MAMS(C) design. It can be observed that the FWER under the global null hypothesis is sharply increasing for the considered values of $\mu^{(0)}$, reaching around 27% when the true mean is off the prior mean by a factor of 1.3, $\mu^{(0)} = 656$, whereas the probability to reject both hypotheses under the partial null configurations is around 16% for $\mu^{(0)} = 656$. This design

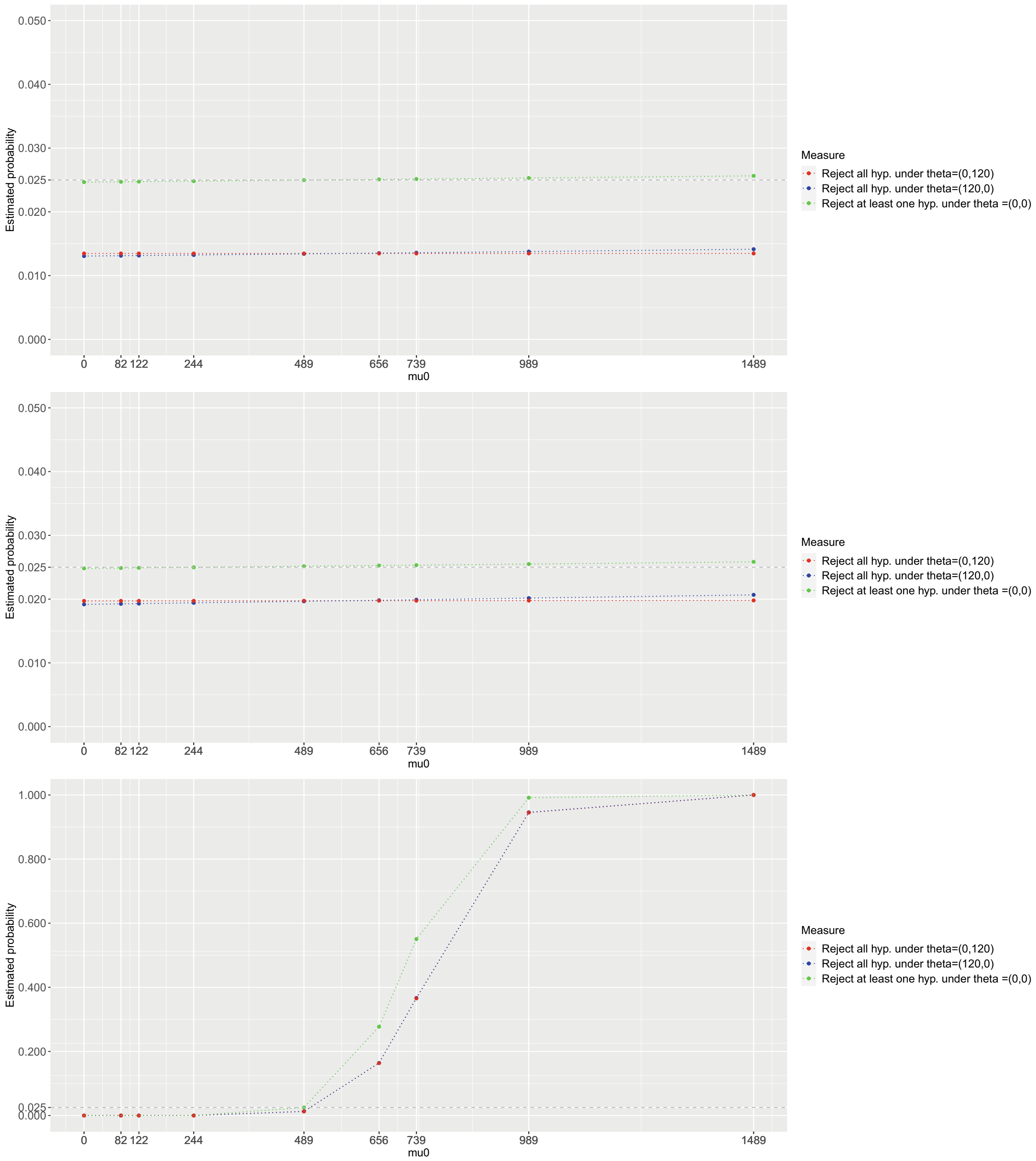

FIGURE 3 Probability to reject at least one H_{0k} under $\boldsymbol{\theta} = (0, 0)$ and probability to reject all hypotheses under $\boldsymbol{\theta} = (120, 0)$ and $\boldsymbol{\theta} = (0{,}120)$ for the 3-arm 2-stage Bayesian design when it is powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (120{,}120)$. Figure at the top refers to the B-MAMS(U) design which incorporates no prior information on the treatment means—prior of $\mu^{(0)}$ is centered in $\mu_0^{(0)} = 489$ and has standard deviation $\sigma = \sqrt{\frac{1}{\tau_{00}}} = 1000$. Figure in the middle refers to the B-MAMS(D) design. Figure at the bottom refers to the B-MAMS(C) design which incorporates prior information on the control mean—prior of $\mu^{(0)}$ is centered in $\mu_0^{(0)} = 489$ and has standard deviation $\sigma = \sqrt{\frac{1}{\tau_{00}}} = 50.63$. 3-arm 2-stage designs use triangular bounds.

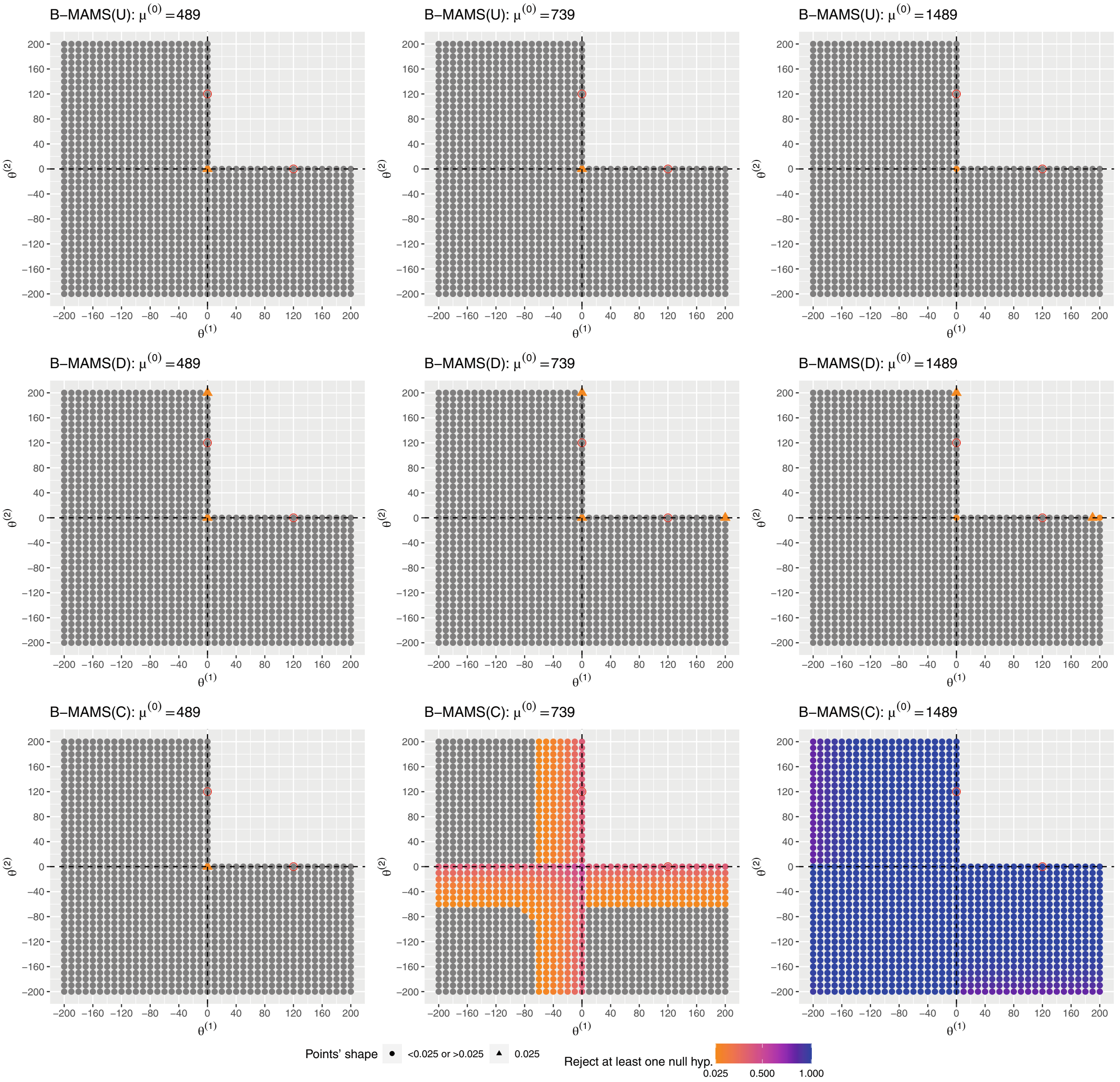

FIGURE 4 Probability to reject at least one null hypothesis when $\mu^{(0)} = 489{,}739, 1489$ for different values of the treatment effects for the 3-arm 2-stage B-MAMS(C), B-MAMS(D), B-MAMS(U) designs when they are powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (120{,}120)$. Points in grey colour refer to probability values below 0.025. Partial null configurations—$\boldsymbol{\theta} = (120, 0)$ and $\boldsymbol{\theta} = (0{,}120)$—are marked with red circles.

shows major inflations on the FWER because it is the one that considers an informative prior distribution on the control mean and thus requires the least sample size compared to all other designs.

Figure 4 provides further explorations of the probability to reject at least one null hypothesis for three values of the control mean, $\mu^{(0)} = 489{,}739, 1489$, but for varying values of the treatment effects for the 3-arm 2-stage B-MAMS(C), B-MAMS(D), B-MAMS(U) designs when they are powered at 80% to reject both hypotheses under $\boldsymbol{\theta} = (120{,}120)$. Specifically, we investigate how the FWER behaves for several values of the treatment effects, $\theta^{(1)}$ and $\theta^{(2)} \in (-200{,}200)$.

It can be observed that the B-MAMS(U) and B-MAMS(D) designs (top and middle panels of Figure 4) are both robust in controlling the FWER around the global and partial null hypotheses for all the considered values of the control mean. Under the global null hypothesis the FWER is equal to 0.025 when the control mean is $\mu^{(0)} = 489$ or 739, while it is slightly inflated—around 2.6%—when $\mu^{(0)} = 1489$. The B-MAMS(D) design shows also some minor inflations—around 0.1%—when the treatment effects are close to 200.

The B-MAMS(C) shows major inflations for all considered values of the treatment effects around the global and partial nulls when the control mean is off by a factor between half and one standard deviation to the prior mean. These inflations are more pronounced for all considered treatment effects when the control mean is off by a factor of 3.

Overall, these exploratory analyses have shown minor inflations on the FWER when the true control mean is different from the prior assumption and the precision on the control mean is small. Thus, the proposed design can provide robust results when the assumption on the value of the control mean is violated and it provides benefits in terms of power compared to other frequentist approaches. However, there can be substantial inflation of the FWER when using an informative prior distribution on the control mean.

Nevertheless, the analyses presented were restricted to the setting of the case study. To evaluate the operating characteristics of the proposed design under different scenarios and possible misspecification of the order assumption compared to the motivating trial, we have conducted further analyses considering two more settings: one with more treatment arms, a 4-arm 2-stage trial, and one with more stages, a 3-arm 3-stage trial. For these settings, we have explored how the FWER behaves under different values of the control mean.

For the 4-arm 2-stage trial, we have considered the same clinical setting as for the case study, but with one more additional arm. We investigate how the FWER behaves for several values of the treatment effects, with $\theta^{(1)}$ and $\theta^{(2)}$ varying from -200 to 200 by steps of 40, while $\theta^{(3)}$ is varying from -200 to 200 by steps of 80. In this way, all possible misspecifications of the order effects can be explored. As for the 3-arm 2-stage setting, minor inflations (up to 0.7%) of the FWER are shown when the true control mean is different from the prior assumption and the precision on the control mean is small for the B-MAMS(U) and B-MAMS(D) designs. However, there can be substantial inflation of the FWER when using an informative prior distribution on the control mean for the B-MAMS(C) design. For example, for this chosen design, it can be observed (Figure 8 in Section 3 of Data S1) that the probability to reject at least one hypothesis is almost 1 when $\theta^{(1)}$ or $\theta^{(2)}$ are equal to zero and $\theta^{(3)} = -200$. In addition, for the 4-arm 2-stage design, we have also included some results to compare how the design performs compared to the frequentist ordered and multi-arm multi-stage designs. The corresponding results are presented in Section 3 of Data S1.

For the 3-arm 3-stage design, we considered other scenarios and normalized treatment effects. For this design we have explored the probability to reject at least one null hypothesis when $\mu^{(0)} = 0, 0.5, 1$ and with the following parameters for the Bayesian model: $\tau = 1, \tau_{01} = 10^{-6}, \mu_0^{(0)} = 0, \mu_0^{(1)} = 0, \delta_0^{(1)} = 0$. In Table 3 in Data S1 we provide the sample sizes and the critical bounds that are found to control the FWER under the global and partial nulls at level $\alpha = 0.025$ when $\mu^{(0)} = 0$ and the designs are powered at 80% to reject all hypotheses under $\boldsymbol{\theta} = (0.5, 0.5)$. However, similar conclusions to the 3-arm 2-stage and 4-arm 2-stage designs can be drawn for all the considered variants of the proposed design. The complete results are provided in Section 4 of Data S1.

5 DISCUSSION

The aim of this study was to develop a MAMS design that can account for historical information and incorporate an order relationship among the treatment arms. In the proposed approach, we model the correlation among the treatment means through the use of prior distributions for the treatment mean difference among two consecutive treatment arms. Through simulations and theoretical results we show that the proposed approach can provide benefits compared to the frequentist ORD[5] when the order assumption is not satisfied. For specific values of the model parameters, the Bayesian design can also match the operating characteristics of the frequentist approaches.

In the considered setting, we are interested in selecting all arms (eg, doses) that are efficacious compared to the standard one. We allow termination of one arm first (eg, for efficacy) and then continuation of the trial with the remaining arms. Indeed, in some multiple dose settings, it is not guaranteed that higher doses lead to increased efficacy compared to lower doses—see for example, the TAILoR study.[31] Likewise, in the setting where multiple durations of the same treatment arm are compared against a standard regimen, one could assume that longer treatment durations lead to higher efficacy compared to the shortest ones. But there could be uncertainty about this assumption as the increase in side effects or compliance with the treatment can lead to reduced effects on the treatment efficacy. So, in this case, stopping shorter durations at interim and continuing with longer ones could be a reasonable choice. Conversely, if a monotonic order assumption among the treatment effects is certain—that is, higher doses (or longer treatment durations) lead to higher efficacy compared to lower doses (or shorter durations), then the order restricted multi-arm multi-stage design[5] can be implemented.

The proposed design ensures weak control of the FWER under specific values of the control mean. While strong control of the FWER is desirable and recommended by the Food and Drug Administration,[32] this is not a must—see for example, recent COVID-19 trials such as the Coverage study.[33] Moreover, strong control in a Bayesian framework is challenging in general as any prior distribution that give notable weight outside of the null space leads to inflation of the FWER unless very conservative thresholds under the global null are used. We believe that ensuring weak control is still important. It is however essential to understand how the FWER behaves under reasonable misspecifications of the design's assumptions through exploratory analyses. The proposed design has been shown to be robust in controlling the FWER for several plausible values of the control mean.

Alternative approaches can be considered for modelling the ordering among the treatment arms. For example, one could consider prior distributions on the difference between each treatment mean and the largest treatment mean, that is, $\mu^{(1)}$. This alternative parametrization could be easily implemented and linked to the one proposed in this paper if normal distributions are used for modelling the treatment mean differences. In addition, one could consider prior distributions on the treatment effects and their differences. However, also in this case one needs to incorporate some prior distribution on the control mean. The Bayesian model can also be adapted when a strict order can be assumed among the treatment means. One could, for instance, define the prior distributions of $\delta^{(k)}, k \in \{1, \dots, K-2\}$ to be lognormal distributions. In this case, the treatment means will result ordered as $\mu^{(1)} \geq \dots \geq \mu^{(K-1)}$.

In this study, we assumed that the variance of patients' responses is known. However, in practice this might not be true. Further work needs to be done in order to incorporate prior information about $\sigma_k, k \in \{0, \dots, K-1\}$ into the model. In addition, in this work, we have just considered normally distributed outcomes. Further work is needed in order to extend this framework to non-normally distributed endpoints.

AUTHOR CONTRIBUTIONS

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

ACKNOWLEDGMENTS

This report is independent research supported by the National Institute for Health Research (NIHR Advanced Fellowship, Dr Pavel Mozgunov, NIHR300576) 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 and P Mozgunov also received funding from UK Medical Research Council (MC_UU_00002/14 and MC_UU_00002/19, respectively). For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

No datasets were generated or analyzed during the current study. Programming code for reproducing the numerical results is available at GitHub: https://github.com/OrderedRestrictedDesign/3arm2stageDesigns.

ORCID

Alessandra Serra https://orcid.org/0000-0001-8431-5154
Pavel Mozgunov https://orcid.org/0000-0001-6810-0284
Thomas Jaki https://orcid.org/0000-0002-1096-188X

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
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
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
5. Serra A, Mozgunov P, Jaki T. An order restricted multi-arm multi-stage clinical trial design. *Stat Med*. 2022; 41(9):1613-1626. doi:10.1002/sim.9314
6. 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
7. 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
8. 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.
9. Jacob L, Uvarova M, Boulet S, Begaj I, Chevret S. Evaluation of a multi-arm multi-stage Bayesian design for phase II drug selection trials–an example in hemato-oncology. *BMC Med Res Methodol*. 2016;16(1):1-15.
10. Whitehead J, Cleary F, Turner A. Bayesian sample sizes for exploratory clinical trials comparing multiple experimental treatments with a control. *Stat Med*. 2015;34:2048-2061. doi:10.1002/sim.6469
11. Spiegelhaltert DJ, Parmar MKB, Freedman LS. Bayesian approaches to randomized trials. *J R Stat Soc A*. 1994;157:357-416.
12. Berry SM, Carlin BP, Lee JJ, Muller P. *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton: CRC Press; 2010.
13. Zhu H, Yu Q. A Bayesian sequential design using alpha spending function to control type I error. *Stat Methods Med Res*. 2017;26:2184-2196. doi:10.1177/0962280215595058
14. Zhu L, Yu Q, Mercante DE. A Bayesian sequential design for clinical trials with time-to-event outcomes. *Stat Biopharm Res*. 2019;11(4):387-397.
15. Ryan EG, Bruce J, Metcalfe AJ, et al. Using Bayesian adaptive designs to improve phase III trials: a respiratory care example. *BMC Med Res Methodol*. 2019;19(1):1-10.
16. Ryan EG, Lamb SE, Williamson E, Gates S. Bayesian adaptive designs for multi-arm trials: an orthopaedic case study. *Trials*. 2020;21(1):1-16. doi:10.1186/s13063-019-4021-0
17. US FDA. Guidance for the use of Bayesian statistics in medical device clinical trials. Guidance for industry and FDA staff US FDA Docket 2010(2006D-0191); 2010:50.
18. Ventz S, Trippa L. Bayesian designs and the control of frequentist characteristics: A practical solution. *Biometrics*. 2015;71:218-226. doi:10.1111/biom.12226
19. Emerson SS, Kittelson JM, Gillen DL. Frequentist evaluation of group sequential clinical trial designs. *Stat Med*. 2007;26(28):5047-5080.
20. 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
21. Vogelberg C, Engel M, Moroni-Zentgraf P, et al. Tiotropium in asthmatic adolescents symptomatic despite inhaled corticosteroids: a randomised dose-ranging study. *Respir Med*. 2014;108(9):1268-1276. doi:10.1016/j.rmed.2014.06.011
22. European Agency for the Evaluation of Medicinal Products. Committee for Proprietary Medicinal Products: points to consider on multiplicity issues in clinical trials. 2002.
23. Grieve AP. Idle thoughts of a 'well-calibrated' Bayesian in clinical drug development. *Pharm Stat*. 2016;15(2):96-108.
24. Stallard N, Todd S, Ryan EG, Gates S. Comparison of Bayesian and frequentist group-sequential clinical trial designs. *BMC Med Res Methodol*. 2020;20:1-14. doi:10.1186/s12874-019-0892-8
25. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. Chichester, New York: John Wiley & Sons; 1997.
26. Urach S, Posch M. Multi-arm group sequential designs with a simultaneous stopping rule. *Stat Med*. 2016;35:5536-5550. doi:10.1002/sim.7077
27. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2019.
28. Plummer M. rjags: Bayesian Graphical Models using MCMC. R package version 4-10; 2019.
29. Genz A, Bretz F, Miwa T, et al. mvtnorm: Multivariate normal and t distributions. R package version 1.1-3; 2021.
30. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64:191. doi:10.2307/2335684
31. Pushpakom S, Kolamunnage-Dona R, Taylor C, et al. TAILoR (TelmisArtan and InsuLin Resistance in Human Immunodeficiency Virus [HIV]): an adaptive-design, dose-ranging phase IIb randomized trial of telmisartan for the reduction of insulin resistance in HIV-positive individuals on combination antiretroviral therapy. *Clin Infect Dis*. 2020;70(10):2062-2072.
32. US FDA. Multiple endpoints in clinical trials. Guidance for industry. FDA Draft Guidance; 2017.
33. Duvignaud A, Lhomme E, Onaisi R, et al. Inhaled ciclesonide for outpatient treatment of COVID-19 in adults at risk of adverse outcomes: a randomised controlled trial (COVERAGE). *Clin Microbiol Infect*. 2022; 28(7):1010-1016.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Serra A, Mozgunov P, Jaki T. A Bayesian multi-arm multi-stage clinical trial design incorporating information about treatment ordering. *Statistics in Medicine*. 2023;42(16):2841-2854. doi: 10.1002/sim.9752