Received: 26 July 2022 | Revised: 22 April 2023 | Accepted: 20 June 2023

DOI: 10.1002/pst.2320

MAIN PAPER

WILEY

Determining the minimum duration of treatment in tuberculosis: An order restricted non-inferiority trial design

Alessandra Serra[1] | **Pavel Mozgunov**[1] | **Geraint Davies**[2] | **Thomas Jaki**[1,3]

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

[2]Department of Clinical Infection, Microbiology and Immunology, Institute of Infection, Veterinary & Ecological Sciences, University of Liverpool, Liverpool, UK

[3]Faculty of 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
Medical Research Council, Grant/Award Numbers: MC_UU_00002/14, MC_UU_00002/19; National Institute for Health and Care Research, Grant/Award Number: NIHR300576; NIHR Cambridge Biomedical Research Centre, Grant/Award Number: BRC-1215-20014

Abstract

Tuberculosis (TB) is one of the biggest killers among infectious diseases worldwide. Together with the identification of drugs that can provide benefits to patients, the challenge in TB is also the optimisation of the duration of these treatments. While conventional duration of treatment in TB is 6 months, there is evidence that shorter durations might be as effective but could be associated with fewer side effects and may be associated with better adherence. Based on a recent proposal of an adaptive order-restricted superiority design that employs the ordering assumptions within various duration of the same drug, we propose a non-inferiority (typically used in TB trials) adaptive design that effectively uses the order assumption. Together with the general construction of the hypothesis testing and expression for type I and type II errors, we focus on how the novel design was proposed for a TB trial concept. We consider a number of practical aspects such as choice of the design parameters, randomisation ratios, and timings of the interim analyses, and how these were discussed with the clinical team.

KEYWORDS

adaptive designs, infectious diseases, multi-arm multi-stage, non-inferiority, order restriction

1 | INTRODUCTION

Tuberculosis remains one the biggest infectious killers worldwide and the global control of the disease is slow.[1] Detection and treatment of infectious pulmonary tuberculosis is a key component of public health strategies in high-burden countries but the current standard of care for first-line therapies requires strong adherence to combination therapy for 6 months to ensure a stable cure. The current first-line regimen evolved over three decades of clinical trial activity and has remained unchanged since the early 1980s.[2] While novel anti-tuberculosis drugs are finally becoming available, new clinical trial evidence and re-examination of historical trials suggest that novel combinations of existing drugs could improve efficacy of the current first-line standard tuberculosis treatment.[3]

TB clinical trials in the past typically evaluated homogenous populations of patients, often those with more severe diseases able to provide repeated bacteriological specimens and modern guidance does not recommend modification of the composition or duration of treatment regimens according to clinical characteristics. Recent data have, however, lent

support to the concept of stratification of treatment on the basis of readily available prognostic factors including rapid microbiological tests and radiology.[4] Strong correlations have been observed in trials and observational studies of measures of baseline bacillary load and long-term outcomes. In a recent pooled re-analysis of three trials of first-line regimens containing fluoroquinolones, the TB-ReFLECT consortium showed that, despite the failure of these trials to show significant non-inferiority, patients on 4-month regimen with low-risk prognostic factors (sputum smear grade <3+ and absence of radiologically apparent cavitation) in fact achieved non-inferior results to the 6-month regimen.[4] These results were also validated externally using a previous clinical trial of prospective simple prognostic stratification,[5] suggesting that the concept of a unitary regimen for all TB patients may result in overlong treatment for some. Furthermore, since two-thirds of patients in these trials fulfilled low-risk prognostic criteria, such a stratified approach clearly has the potential to shorten treatment for a substantial number of patients.

Due to the excellent efficacy of the 6-month first-line regimen, the conventional designs in TB are non-inferiority trials that evaluate whether a novel treatment regimen may shorten treatment of a single reduced duration, estimated on the basis of preclinical and early phase clinical trial data. For instance, Study 31[6] has recently shown that a 4-month rifapentine-based regimen containing moxifloxacin was non-inferior to the standard 6-month regimen in the treatment of tuberculosis. However, the conventional designs are not able to determine the minimum possible duration for any given regimen and their success rests heavily on correctly predicting the target duration from limited data. Such trials thus pose a risk to drug developers and do not provide the information of most interest to policymakers.

Adaptive designs have been proposed as a solution to the question of how to identify the most promising treatment in confirmatory trials rapidly. Multi-arm multi-stage designs (MAMS) have been argued to be a highly efficient approach to clinical trials[7–9] and they have been suggested[10] for improving the treatment of tuberculosis. In this disease area, MAMS have been proposed by Bratton et al.,[11] Cellamare et al.[12] and a completed trial has been adopted by the Pan-African Consortium for the Evaluation of Antituberculosis Antibiotics (PanACEA).[13] In addition, the TRUNCATE-TB trial[14] is a MAMS recently implemented for drug-sensitive tuberculosis. These trials, however, evaluate different treatment regimens, but only a single study duration.

New approaches to estimate the duration of therapy have become available which are capable of evaluating multiple nested durations in an efficient manner.[15,16] The design proposed by Quartagno et al.[15] relies on a parametric model for the duration-response relationship and it does inflate the type I error under certain scenarios where the steepness of the duration-response curve is increasing at the optimal duration. The work proposed by Serra et al.,[16] instead, incorporates the order of treatment effects in the decision-making when no parametric duration-response model is assumed and it guarantees strong control of multiple testing.

Depending on the clinical setting and the disease area, several primary endpoints can be measured and used to address the objectives of the trial. The adaptive design by Serra et al.[16] was proposed considering only normally distributed endpoints in a superiority trial setting. In this work, we extend the design for a non-inferiority trial with a binary endpoint. We describe its application to a clinical trial in TB which aims to compare the efficacy of multiple treatment durations against a standard treatment regimen. We consider several practical aspects such as the choice of design's parameters, randomisation ratios and timing of the interim analyses and how all these aspects were discussed with the clinical team.

The rest of the manuscript continues as follows. The motivating clinical trial is introduced in Section 2 before a detailed description of the novel design is proposed in Section 3. Section 4 describes the choice of the design's parameters for the actual clinical trial before a suggestion for the clinical trial design is provided at the end of the section. A simulation study is described in Section 5 in order to compare the proposed design with a standard MAMS design. Section 6 analyses two different strategies that consider different timings of the interim analyses. We conclude with a discussion.

2 | MOTIVATING TRIAL: RESTRUCTURE TRIAL

The REStrUCTuRe trial is a Randomised Evaluation of Stratified Ultra-Short Combination Tuberculosis Regimens. This is a Phase III trial that aims to determine the shortest possible yet effective duration of treatment for people suffering from pulmonary tuberculosis with low-risk or high-risk prognostic factors using a dose-optimised fluoroquinolone-containing first-line regimen containing Rifampicin (RIF) 40 mg/kg, Pyrazinamide (PZA) 35 mg/kg and Levofloxacin (LFX) 15 mg/kg: $R_{40}Z_{35}L_{15}$. The primary aim is to compare the efficacy of $R_{40}Z_{35}L_{15}$ at durations of treatments from 2 to 4 months, while secondary aims are to compare the safety and tolerability of the investigational drug at different

durations and to estimate the value of measures of culture conversion as predictors of long-term outcome. The schematic of the experimental arms planned to be considered, depending on the patients' prognostic factors, are given in Figure 1. Durations of 2, 3, and 4 months will be tested in the low-risk prognostic sub-population (no cavitation on CXR and/or Smear Grade $< +++$) and durations of 3 and 4 months in the high-risk prognostic sub-population (cavitation on CXR and/or Smear Grade $\geq +++$). Standard of care $2HR_{10}Z_{25}E/4HR_{10}$ is administered daily for 6 months irrespective of participant prognostic score.

The primary outcome of "no durable cure" (NDC) is defined as a confirmed positive mycobacterial culture at the end of treatment with genetically identical isolates (on whole-genome sequencing) to their baseline isolate.

In the next sections, we illustrate and describe the extension of the ordered restricted MAMS[16] to a non-inferiority setting with a binary endpoint and we consider a number of practical aspects.

3 | ORDER RESTRICTED DESIGN

In this section, we first summarise the testing procedure of the original multi-arm multi-stage superiority order restricted design (ORD[16]), and then propose the extension for the non-inferiority setting with a binary endpoint. Non-inferiority trials are designed to test whether new regimens are non-inferior in efficacy compared to the standard regimen currently used. These designs are more appropriate when new regimens may have practical advantages compared to the standard intervention and thus may be preferred in real-life settings even if the new regimen is modestly less efficacious.[3] The level of acceptance is defined by the non-inferiority margin. Given the high efficacy of currently recommended regimens, non-inferiority designs are necessary to be adopted in TB setting.[17]

We start with describing a general design proposal for arbitrary number of arms and stages. Then, given the different treatment durations studied in each sub-population of the motivating trial, we provide the design for each sub-population.

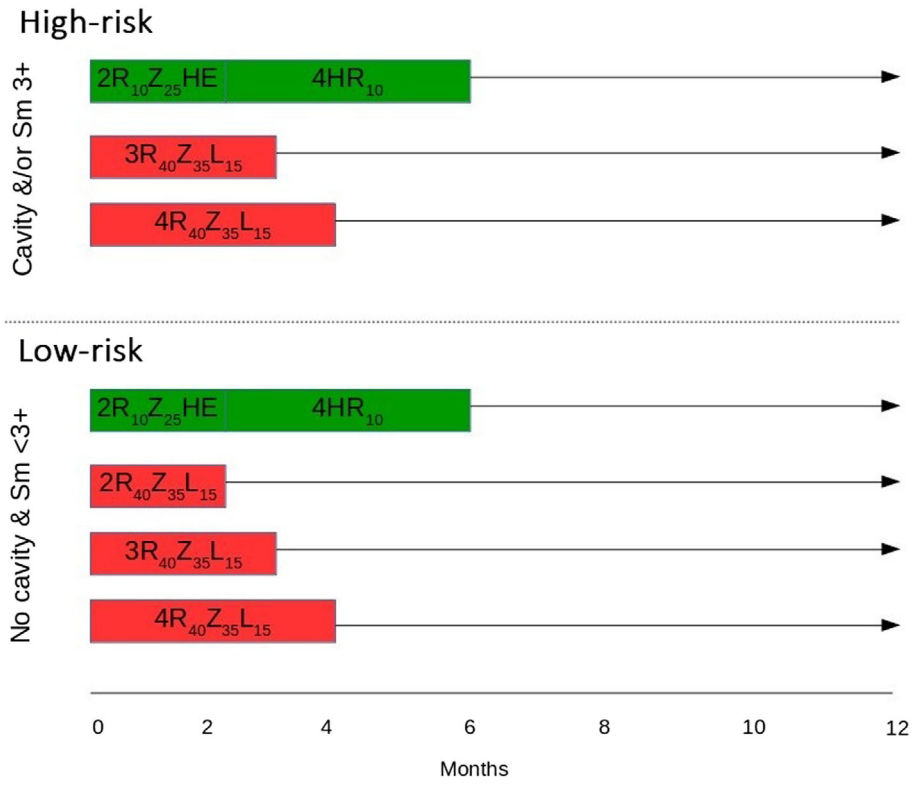

FIGURE 1 Trial design schematic. Coloured blocks represent treatment periods while black arrows represent post-treatment follow-up.

3.1 | Multi-arm multi-stage superiority order-restricted design for normally distributed endpoint

Consider a clinical trial with $K-1$ active treatment arms, $T_1,...,T_{K-1}$, against a control treatment T_0 and J stages at which treatment arms can be dropped or the trial could be stopped for benefit or lack of efficacy. Assume that a patient's response follows a normal distribution with known common variance, σ^2. Let $X_i^{(k)} \sim N\left(\mu^{(k)},\sigma^2\right), k\in\{0,...,K-1\}, i=1:n_j^{(k)}$ be the observation of the i-th 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 $k\in\{1,...,K-1\}$ compared to the control. Denote the vector of treatment effects by $\boldsymbol{\theta}=\left(\theta^{(1)},\theta^{(2)},...,\theta^{(K-1)}\right)$ and assume that the following order relationship holds: $\theta^{(1)}\geq\theta^{(2)}\geq...\geq\theta^{(K-1)}$.

Define the ratio of the sample size on treatment $k\in\{0,...,K-1\}$ at time point $j\in\{1,...,J\}$ over the sample size on the control in the first stage as $r_j^{(k)}$ with $r_1^{(0)}=1$. Let $Z_j^{(k)}=\frac{\widehat{\mu}_j^{(k)}-\widehat{\mu}_j^{(0)}}{\sigma}\sqrt{\frac{r_j^{(0)}n_j^{(k)}}{r_j^{(k)}+r_j^{(0)}}}$ be the test statistic[18] at stage j for comparing arm k to control, where $\widehat{\mu}_j^{(k)}=\left(n_j^{(k)}\right)^{-1}\sum_{i=1}^{n_j^{(k)}}X_i^{(k)}$ and $n_j^{(k)}=r_j^{(k)}n$, with $k\in\{0,...,K-1\}$ and n is the sample size in the control group at the first stage.

The null hypotheses of interest are $H_{01}:\left\{\theta^{(1)}\leq 0\right\},...,H_{0K-1}:\left\{\theta^{(K-1)}\leq 0\right\}$. Let $u_j^{(k)},l_j^{(k)},k\in\{1,...,K-1\},j\in\{1,...,J\}$ be the critical values at stage j used to test the hypotheses with $u_J^{(k)}=l_J^{(k)},k\in\{1,...,K-1\}$.

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 $T_{\mathscr{L}_j},...,T_k$ is stopped, where $\mathscr{L}_j$ is the lowest index on the 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 continues for all arms between k and k'. If there is sufficient evidence to drop arm k, that is when $Z_j^{(k)}\leq l_j^{(k)}$, and if there is not any contradicting evidence for $k'>k$ then the recruitment to arms $T_k,...,T_{\mathscr{H}_j}$ is stopped, where $\mathscr{H}_j$ is the highest index on the treatment arms remaining in the trial at stage j.

The description of the decision rules for the special case of a 3-arm 2-stage design is provided in the Appendix in Table A1, while we refer to Serra et al.[16] for a general algorithm for the decision-making in this setting.

3.2 | Multi-arm multi-stage non-inferiority order-restricted design for binary endpoint

Assume that patients' responses follow a Bernoulli distribution. Let $X_i^{(k)}\sim Bern\left(p^{(k)}\right), k=\{0,1,...,K-1\}, i=1:n_j^{(k)}$ be the observation of the i-th 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 $\theta^{(k)}=p^{(k)}-p^{(0)}$ be the true treatment effect of arm k compared to the control. Denote the number of subjects on the control in the first stage by n and the total sample size on treatment k up to and including stage j is defined as $n_j^{(k)}=r_j^{(k)}n$. Let $\delta>0$ be the non-inferiority margin and $Z_j^{(k)}=\frac{\widehat{p}_j^{(k)}-\widehat{p}_j^{(0)}+\delta}{\sqrt{\frac{p^{(k)}\left(1-p^{(k)}\right)}{n_j^{(k)}}+\frac{p^{(0)}\left(1-p^{(0)}\right)}{n_j^{(0)}}}}$ be the test statistic[19] at stage j for comparing arm k to control and $\widehat{p}_j^{(k)}=\left(n_j^{(k)}\right)^{-1}\sum_{i=1}^{n_j^{(k)}}X_i^{(k)}$, with $k=\{0,1,...,K-1\}$. The vector $\boldsymbol{Z}=\left(Z_1^{(1)},Z_1^{(2)},...,Z_1^{(K-1)},Z_2^{(1)},...,Z_2^{(K-1)},...,Z_J^{(1)},...,Z_J^{(K-1)}\right)$ of test statistics is approximately multivariate normally distributed with $\boldsymbol{Z}\sim N_{(K-1)J}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. Note that if the estimated response rates are used in the denominator of the test statistic—instead of the true values this would result in a multivariate t distribution. In this case, one could transform the single test statistics to the normal scale using a quantile transformation to obtain approximately normal test statistics.

The mean value of the test statistic is $\mathbb{E}\left[Z_j^{(k)}\right]=\frac{\theta^{(k)}+\delta}{\sqrt{\frac{p^{(k)}\left(1-p^{(k)}\right)}{n_j^{(k)}}+\frac{p^{(0)}\left(1-p^{(0)}\right)}{n_j^{(0)}}}}$ and the covariances between Z-statistics are:

$$\mathrm{Cov}\left(Z_j^{(k)}, Z_j^{(k)}\right) = 1,$$

$$\mathrm{Cov}\left(Z_j^{(k)}, Z_j^{(k')}\right) = \frac{\sqrt{r_j^{(k)} r_j^{(k')}}\, p^{(0)}\left(1-p^{(0)}\right)}{\sqrt{r_j^{(0)} p^{(k)}\left(1-p^{(k)}\right) + r_j^{(k)} p^{(0)}\left(1-p^{(0)}\right)}\sqrt{r_j^{(0)} p^{(k')}\left(1-p^{(k')}\right) + r_j^{(k')} p^{(0)}\left(1-p^{(0)}\right)}},$$

$$\mathrm{Cov}\left(Z_j^{(k)}, Z_{j+1}^{(k)}\right) = \frac{\sqrt{r_j^{(k)} r_j^{(0)} r_{j+1}^{(k)} r_{j+1}^{(0)}}}{\sqrt{r_j^{(0)} p^{(k)}\left(1-p^{(k)}\right) + r_j^{(k)} p^{(0)}\left(1-p^{(0)}\right)}\sqrt{r_{j+1}^{(0)} p^{(k)}\left(1-p^{(k)}\right) + r_{j+1}^{(k)} p^{(0)}\left(1-p^{(0)}\right)}}\left[\frac{p^{(k)}\left(1-p^{(k)}\right)}{r_{j+1}^{(k)}} + \frac{p^{(0)}\left(1-p^{(0)}\right)}{r_{j+1}^{(0)}}\right],$$

$$\mathrm{Cov}\left(Z_j^{(k)}, Z_{j+1}^{(k')}\right) = \frac{\sqrt{r_j^{(k)} r_{j+1}^{(k')} r_j^{(0)}}\, p^{(0)}\left(1-p^{(0)}\right)}{\sqrt{r_{j+1}^{(0)}}\sqrt{r_j^{(0)} p^{(k)}\left(1-p^{(k)}\right) + r_j^{(k)} p^{(0)}\left(1-p^{(0)}\right)}\sqrt{r_{j+1}^{(0)} p^{(k')}\left(1-p^{(k')}\right) + r_{j+1}^{(k')} p^{(0)}\left(1-p^{(0)}\right)}},$$

with $k, k' \in \{1,...,K-1\}, j \in \{1,...,J\}$. We test the null hypotheses: $H_{01}: \{\theta^{(1)} \le -\delta\}, ..., H_{0K-1}: \{\theta^{(K-1)} \le -\delta\}$ with the global null hypothesis denoted by $H_0: \{\theta^{(1)} = ... = \theta^{(K-1)} = -\delta\}$.

3.2.1 | Number of arms and stages for the REStrUCTuRe trial

Given the different treatment durations studied in each sub-population, for the high-risk sub-population we denote the vector of treatment effects by $\boldsymbol{\theta_{12}} = \left(\theta_h^{(1)}, \theta_h^{(2)}\right)$ and the vector of response rates by $\boldsymbol{p_{12}} = \left(p_h^{(0)}, p_h^{(1)}, p_h^{(2)}\right)$. Consider the following order relationship: $\theta_h^{(1)} \ge \theta_h^{(2)}$. Thus, T_1 and T_2 are the treatments at the longest and shortest durations, respectively, in this subgroup. We denote the global null hypothesis as $H_{012}: \left\{\theta_h^{(1)} = \theta_h^{(2)} = -\delta\right\}$. We define the set of the indices for the active treatment arms in this subgroup as $K_h = \{1,2\}$.

For the low-risk sub-population, the vector of treatment effects is denoted by $\boldsymbol{\theta_{123}} = \left(\theta_l^{(1)}, \theta_l^{(2)}, \theta_l^{(3)}\right)$ and the vector of response rates by $\boldsymbol{p_{123}} = \left(p_l^{(0)}, p_l^{(1)}, p_l^{(2)}, p_l^{(3)}\right)$. We consider the following order relationship: $\theta_l^{(1)} \ge \theta_l^{(2)} \ge \theta_l^{(3)}$. Thus, we denote the longest treatment duration by T_1, the medium duration by T_2 and the shortest duration by T_3. We denote the global null hypothesis as $H_{0123}: \left\{\theta_l^{(1)} = \theta_l^{(2)} = \theta_l^{(3)} = -\delta\right\}$. We define the set of the indices for the active treatment arms in this subgroup as $K_l = \{1,2,3\}$.

For both sub-populations, a single interim analysis is planned after half of the total planned maximum number of participants have completed their treatment. From the theory of group sequential design,[20] the design performs well in terms of operating characteristics when the interim analysis is planned after having observed between 30% and 70% of the total information. Thus, it was decided to consider a middle value among those and plan the interim analysis after having observed 50% of the total information. Thus in the next sections, we consider a 3-arm (for the high-risk sub-population) and a 4-arm (for the low-risk sub-population) 2-stage design.

3.3 | Family-wise error rate

For confirmatory clinical trials, control of the family-wise error rate (FWER) in the strong sense at level α (one-sided), that is the probability to reject at least one true null hypothesis, is typically required.[21] Using the rules described in Table A2, the FWER for the 3-arm 2-stage ORD can be written as

$$P(\text{rejecting at least one true } H_{0k}, k \in K_h | H_{012}, \boldsymbol{p_{12}}) = P\left(Z_1^{(1)} \geq u_1^{(1)} | H_{012}, \boldsymbol{p_{12}}\right) + P\left(Z_2^{(1)} \geq u_2^{(1)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)} | H_{012}, \boldsymbol{p_{12}}\right) + P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | H_{012}, \boldsymbol{p_{12}}\right). \quad (1)$$

while the FWER for the 4-arm 2-stage design is

$$P(\text{rejecting at least one true } H_{0k}, k \in K_l | H_{0123}, \boldsymbol{p_{123}}) = P(\text{rejecting at least one true } H_{0k}, k \in \{1,2\} \mid H_{012}, \boldsymbol{p_{12}}) + P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} < u_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | H_{0123}, \boldsymbol{p_{123}}\right). \quad (2)$$

As proved in Serra et al.,[16] the design maintains strong control of the FWER if the same bounds, $u_j^{(k)} = u_j, l_j^{(k)} = l_j, j \in \{1,2\}$ are used for each arm with $k \in K_h$ for the high-risk sub-population and $k \in K_l$ for the low-risk sub-population, and let us assume that there are equal numbers of patients on each active treatment within each stage: $r_j^{(k)} = r_j, \forall k \in K_h$ for the high-risk sub-population and $k \in K_l$ for the low-risk sub-population. We use the same allocation ratios within each sub-population and we provide the design for each sub-population separately. In this study, we do not combine the test statistics. In addition, the futility boundaries considered in this study are binding and this implies that valid inference is possible only if the decision rules at the interim analyses are strictly followed.

3.4 | Power requirement

As outlined in Fleming,[22] non-inferiority clinical settings present several challenges that should be evaluated during the design, the conduct and the analysis of the trial. One of those concerns the determination of the sample size, which can be very large for example when the non-inferiority margins are too rigorous. The sample size for a non-inferiority clinical trial should be planned so that the trial will have adequate statistical power to conclude that the non-inferiority margin is ruled out if the active treatment is truly non-inferior compared to the control.[23] If the active arm is discovered to be more effective than the control, then it will be easier to rule out any non-inferiority margin than if the active arm is equivalent or slightly inferior to the control, and a smaller sample size could be used. A less effective active arm will require a larger sample size.

For this study, we decided to power the study under the configuration where there is no difference between the control and intervention arms, that is $\boldsymbol{\theta_{12}} = \left(\theta_h^{(1)}, \theta_h^{(2)}\right)$ for the 3-arm trial with $\theta_h^{(1)} = \theta_h^{(2)} = 0$ and $\boldsymbol{\theta_{123}} = \left(\theta_l^{(1)}, \theta_l^{(2)}, \theta_l^{(3)}\right)$ for the 4-arm trial with $\theta_l^{(1)} = \theta_l^{(2)} = \theta_l^{(3)} = 0$.

The primary objective of the trial is to identify the shortest possible treatment duration in each sub-population. Thus, the trial is to be powered to reject all correct hypotheses. However, alternative power strategies can be also considered, for example to reject at least one hypothesis or to reject the first and the second hypotheses only in the 4-arm trial. The equations that need to be satisfied to power the design are provided in Appendix A.

3.5 | No early stopping for declaring non-inferiority

As explained in Section 2, we look for the culture status at the end of the treatment. However, early success claims might be misleading, if patients transition to a failed state after longer follow-up, given the slow-growing nature of Tuberculosis. To accommodate this, an alternative approach could be to consider an interim analysis with only futility bounds and then evaluate efficacy outcomes at the final analysis—the decision rules used in this case are summarised in Tables A5 and A6 for 3-arm and 4-arm designs, respectively. In this way, the interim analysis considers the primary endpoint evaluated at the end of the treatment, while the final analysis would be at some months—for example 6 or 12—post-randomization. All other design features considered above remain the same.

In the following section, we will explore the operating characteristics of the proposed design when applied to the TB trial.

4 | DESIGN'S PARAMETERS FOR THE RESTRUCTURE TRIAL

4.1 | Setting

In the setting of the motivating trial we consider a study where patients are randomised with equal probability to receive either the current first-line regimen $2HR_{10}Z_{25}E/4HR_{10}$ at a fixed duration of 6 months or the experimental regimen $R_{40}Z_{35}L_{15}$ at duration determined in two sub-populations defined by their baseline factors. Within each sub-population, the allocation between the arms will be balanced (low-risk 1:1:1:1, high-risk 1:1:1). Patients are enrolled and divided into sub-groups depending on their prognostic factors. The distribution in the screened population is expected to be approximately 2:1 in favour of the low-risk group and we consider around 10% losses to follow-up. This is to ensure that the total sample size is feasible for the trial. Note that losses to follow-up in the intervention arms in the low-risk stratum could be lower due to shorter treatment duration, but also could be higher due to safety or lack of efficacy in some of the lower duration groups.

The assumed true response rates in the control arm for the low-risk and high-risk sub-populations are 92% and 86%, respectively. One interim analysis is planned after half of the total population has their primary endpoints evaluated. The design is constructed to control the FWER at level $\alpha = 5\%$ and to reach 80% of power under various power strategies (see details in Section 4.3).

Below, we specify the required parameters to design the study, such as the non-inferiority margin, the shape of the critical bounds, and the appropriate power configurations together with the rationale that was used in the discussion with the clinical team to justify them.

4.2 | Non-inferiority margin

In a non-inferiority setting, the determination of the non-inferiority margin is a critical step and it is often challenging.[23] In recent Phase III trials in TB, non-inferiority margins of between 6% and 12% have been accepted by regulators in different contexts—see Table 1. With the exception of STAND, trials evaluating relative reductions in the duration of treatment of a third have all used non-inferiority margins of 6%–8% whereas trials with more ambitious goals have used 10% for a halving of duration (STREAM) and 12% for a reduction of two-thirds (TRUNCATE-TB). According to FDA guidance on setting non-inferiority margins,[23] and assuming cure rates of 30% and 90% in those receiving no treatment and standard of care, respectively, a margin of 6% ($M_2 =$ the largest clinically acceptable difference (degree of inferiority) of the test drug compared to the active control) preserves 90% of the treatment effect of Standard of Care (SOC) while 12% ($M_1 =$ the entire effect of the active control assumed to be present in the NI study) preserves 80%. The relative reduction in duration to be evaluated in REStrUCTuRE is between one third (in either prognostic sub-population) and two-thirds (in the low-risk prognostic sub-population). Selecting a non-inferiority margin of 10% for the trial was thought therefore to be consistent with the range used in previous recent clinical trials and preserves 83% of the estimated treatment effect of SOC. Thus, the choice of 10% was made to be within M_2 and M_1 and risk is well balanced against the expected benefit in adherence and reduction in resistance. For this study, the NI margin is chosen to be equal in the two sub-populations to have consistent design parameters for the two separate trials. However, different non-inferiority margins could be also assumed for each trial.

4.3 | Type of power configuration and proposal for the design

The next step consists of evaluating various power strategies for the objective of the trial taking into account the recruitment feasibility and the distribution of patients across sub-population. We analyse the differences among the strategies for both sub-populations and we describe how the final decision on the design was reached with the clinical team. We consider the setting as described at the beginning of this section and a non-inferiority margin of 10% for both sub-populations. In order to ensure strong control of the FWER,[16] we consider the case where the same bounds, $u_j^{(k)} = u_j, l_j^{(k)} = l_j, u_2 = l_2, j \in \{1,2\}$, are used for each arm with $k \in K_h$ for the high-risk sub-population and $k \in K_l$ for the low-risk sub-population.

We derive the critical bounds separately for each sub-population as these are considered as separate trials and no combined hypotheses testing is planned. First, we determine analytically—relying on the asymptotical normal

approximation of the vector of test statistics—the sample sizes and the critical values for each sub-population and power configuration. The normal approximation of the test statistics and the design's performance are evaluated using simulations. The critical bounds and the sample size are then tuned in order to get closer to 5% in the probability to reject at least one hypothesis under the global null and to reach 80% under the alternative hypothesis. Specifically, if after the first round of simulations, the pre-specified α-level is found to be inflated under the global null and/or the desired power level is not reached under the power configuration, then further simulations are run considering a grid of values for the critical bounds and the sample sizes—the grid of values is constructed around the analytical values previously found. The pair of values—critical bounds and sample size—that satisfy the FWER and power requirements is finally selected.

The first six rows in Table 2 provide the results for each sub-population and power configuration under the global and alternative hypotheses for the ORD with triangular critical bounds (given in Table 1 in the Supplementary Materials). The remaining rows of this Table report the results for the competing approach that is described in Section 6.

One possibility for the TB trial is to use a configuration of "reject at least one" for the high-risk sub-population and "reject all hypotheses" in the low-risk sub-population. This strategy would require 624 in the low-risk sub-population and 522 in the high-risk sub-population for a maximum sample size of 1146 and an expected sample of 826 under the alternative hypothesis. Inflating the sample size for 10%—that is simply a standard inflation of the required sample size—loss to follow-up gives a maximum sample size of 1261 and expected sample sizes of 909. Under the null hypothesis, the sample sizes are expected to be 375 and 316 for the low-risk and high-risk sub-populations, leading to a total of 691 patients.

Alternatively, ensuring the rejection of at least one hypothesis in each sub-population with the same design parameters, the maximum sample sizes for the low-risk and high-risk sub-populations were 432 and 522, respectively, leading to a total maximum sample size of 954. However, under the alternative hypothesis, the expected sample sizes were 321 and 377, a total expected sample size of 698. Adjusting these for 10% loss to follow-up gives maximum and expected sample sizes of 1050 and 768, respectively. Under the null hypothesis, the sample sizes are expected to be 260 and 316 for the low-risk and high-risk sub-populations, leading to a total expected sample size of 576 patients.

Despite the maximum and expected sample size under the alternative hypothesis being 20% and 18% larger in the suggested strategy compared to the global "reject at least one" strategy, the first approach is more 'efficient' in the sense that if we power the design to reject at least one hypothesis for the high-risk sub-population and to reject all for the low-risk sub-population we can get closer (the ratio of low-risk to high-risk participants is 1.2 instead of 0.83) to the 2:1 ratio—in favour of the low-risk sub-population—of the expected screened population. Instead, if the designs are both powered to reject at least one hypothesis then we are slightly far from the expected ratio in the screened population. Thus, our suggestion for the design is to use a configuration of "reject at least one" for the high-risk sub-population and "reject all hypotheses" in the low-risk sub-population at a β level of 0.20 and a one-sided $\alpha = 0.05$.

4.4 | Critical boundaries' shape

One of the features of the proposed design is the shape of the critical boundaries used to test the hypotheses. In order to decide which shape of the bounds suits best the objective of the trial, we explore the operating characteristics of the

TABLE 1 Design parameters of recent non-inferiority trials in TB.

Trial	Arms	δ	Test	$\alpha-$level	Treatment arms
RIFAQUIN[31]	3	6%	One-sided	5%	First-line, 6 versus 4 months
ReMOX[32]	3	6%	One-sided	1.25%	First-line, 6 versus 4 months
OFLOTUB[33]	2	6%	One-sided	2.5%	First-line, 6 versus 4 months
TBTC Study 31[34]	3	6.6%	Two-sided	5%	First-line, 6 versus 4 months
RIFASHORT[35]	3	8%	One-sided	5%	First-line, 6 versus 4 months
STREAM[36]	2	10%	Two-sided	5%	MDR-TB, 9 versus 18–21 months
STAND[37]	4	12%	One-sided	2.5%	First-line, 6 versus 4 months
TRUNCATE[14]	5	12%	One-sided	1.25%	First-line, 6 versus 2 months

TABLE 2 Maximum sample size (MaxSS), expected sample size (ESS) under global null and alternative hypotheses, FWER and probability to reject at least one hypothesis, all hypotheses and the longest and medium durations (LM) for each sub-population under the alternative for the ORD and MAMS(m) designs when triangular boundaries are used.

ORD							
sub-populations	**MaxSS**	**ESS$_{H_0}$**	**ESS$_{H_1}$**	**FWER**	**Reject at least one H_{0k}**	**Reject all**	**Reject LM**
high-risk	672	405	475	0.051	0.88	**0.80**	–
	522	316	377	0.047	**0.80**	0.67	–
low-risk	624	375	449	0.049	0.93	**0.82**	0.88
	432	260	321	0.050	**0.82**	0.61	0.70
	536	322	408	0.049	0.88	0.72	**0.80**
MAMS(m)							
sub-populations	**MaxSS**	**ESS$_{H_0}$**	**ESS$_{H_1}$**	**FWER**	**Reject at least one H_{0k}**	**Reject all**	**Reject LM**
high-risk	444	273	349	0.049	**0.82**	0.50	–
low-risk	832	521	578	0.049	0.99	**0.83**	0.88

Note: The global null hypothesis for the high-risk sub-population is: $H_{012} = \{p^{(0)} = 0.86, p^{(1)} = p^{(2)} = 0.76\}$, while for the low-risk $H_{0123} = \{p^{(0)} = 0.92, p^{(1)} = p^{(2)} = p^{(3)} = 0.82\}$. The alternative hypothesis for the high-risk sub-population is: $H_1 = \{p^{(0)} = p^{(1)} = p^{(2)} = 0.86\}$, while for the low-risk $H_1 = \{p^{(0)} = p^{(1)} = p^{(2)} = p^{(3)} = 0.92\}$. Cells coloured in blue correspond to the chosen power configurations for the TB trial. Results are provided using 10^6 replications. Values in bold refer to the target probability (around 80%) for each power configuration.

design under the global and alternative hypotheses considering Pocock,[24] O'Brien and Fleming[25] and Triangular[26] boundaries. We consider the case where the same bounds, $u_j^{(k)} = u_j, l_j^{(k)} = l_j, u_2 = l_2, j \in \{1,2\}$, are used for each arm with $k \in K_h$ for the high-risk sub-population and $k \in K_l$ for the low-risk sub-population.

In this section, results are provided without tuning of the parameters. The expected sample sizes (ESS), that are the mean number of patients recruited to the trial before it is terminated, are also measured under the global and alternative hypotheses. The numerical results are found using R[27] and 10^6 replicate simulations.

Table 3 provides the maximum sample sizes, the expected sample sizes and the probabilities to reject at least one hypothesis, all hypotheses or at least the first two for each sub-population and using Pocock, O'Brien and Fleming and Triangular boundaries. For each power configuration and for each sub-population, the design with O'Brien and Fleming critical bounds requires the least total maximum sample size. The efficiency of the proposed design, however, is measured by the ESS. Indeed, the two-stage design allows to reduce the maximum total sample size by a fraction that depends on the different critical bounds used at the interim and final analyses. For the O'Brien and Fleming bounds, the reduction in sample size (RSS), which is RSS $= 1 - \frac{\text{ESS}}{\text{MaxSS}}$, under the global and alternative hypotheses is quite small—a reduction up to 1% under the global null for each design and up to 14% under the alternative hypothesis. Despite triangular bounds require for each design and sub-population almost the highest total maximum sample size—which is still feasible to recruit given previous TB studies—compared to the other critical bounds, this shape of bounds provides the largest reduction in sample size under the global null and alternative hypotheses—a reduction of around 40% under the global null for each design and up to 29% under the alternative hypothesis. Thus, in order to minimize the expected sample size under the global null and alternative hypotheses, we decided to consider triangular boundaries for the REStrUCTuRe trial.

4.5 | Proposal for the design with only futility bounds at the interim analysis

In this section, results for the alternative approach described in Section 3.5 are provided. Table 4 provides the maximum sample sizes, the actual maximum sample size—that is the mean number of the patients recruited to the trial when all treatment arms proceed to the final analysis—, the expected sample sizes and the probabilities to reject all, at least one or the first two hypotheses for each sub-population and using triangular boundaries when for both sub-populations we are expected to recruit around 30 patients per month.

In this setting, we would require 450 and 544—in the worst case scenarios 470 and 560, respectively—patients in order to ensure 80% power to reject at least one and all hypotheses for the high-risk and the low-risk groups,

TABLE 3 Maximum sample size (MaxSS), expected sample size (ESS) under global null and alternative hypotheses, FWER, probability to reject at least one hypothesis, all hypotheses and the longest and medium durations (LM) for each sub-population under the alternative and reduction in sample size (RSS) $\left(=1-\frac{\text{ESS}}{\text{MaxSS}}\right)$ for each sub-population under global null and alternative hypotheses for the ORD.

Pocock bounds—$u_1=u_2=1.876, l_1=-1.876$									
Sub-populations	**MaxSS**	**ESS_{H_0}**	**ESS_{H_1}**	**FWER**	**Reject at least one H_{0k}**	**Reject all**	**Reject LM**	**RSS_{H_0}**	**RSS_{H_1}**
High-risk	642	625	479	0.056	0.87	**0.79**	–	0.03	0.25
	504	490	417	0.047	**0.78**	0.66	–	0.03	0.17
Low-risk	592	570	439	0.049	0.93	**0.83**	0.87	0.02	0.25
	416	403	332	0.044	**0.84**	0.66	0.74	0.03	0.20
	528	510	424	0.053	0.86	0.70	**0.77**	0.03	0.20
O'Brien and Fleming bounds—$u_1=2.373, u_2=1.68, l_1=-2.373$									
sub-populations	**MaxSS**	**ESS_{H_0}**	**ESS_{H_1}**	**FWER**	**Reject at least one H_{0k}**	**Reject all**	**Reject LM**	**RSS_{H_0}**	**RSS_{H_1}**
High-risk	582	578	502	0.049	0.87	**0.78**	–	0.01	0.14
	456	452	413	0.046	**0.82**	0.71	–	0.01	0.09
Low-risk	536	530	462	0.052	0.89	**0.76**	0.82	0.01	0.14
	376	373	346	0.045	**0.83**	0.65	0.72	0.01	0.08
	480	475	438	0.051	0.86	0.69	**0.77**	0.01	0.09
Triangular bounds—$u_1=1.899, u_2=1.79, l_1=0.633$									
Sub-populations	**MaxSS**	**ESS_{H_0}**	**ESS_{H_1}**	**FWER**	**Reject at least one H_{0k}**	**Reject all**	**Reject LM**	**RSS_{H_0}**	**RSS_{H_1}**
High-risk	672	405	475	0.051	0.88	**0.80**	–	0.40	0.29
	516	314	394	0.048	**0.79**	0.66	–	0.39	0.24
Low-risk	608	369	438	0.044	0.92	**0.79**	0.85	0.39	0.28
	416	253	334	0.042	**0.80**	0.59	0.68	0.39	0.20
	536	322	408	0.049	0.88	0.72	**0.79**	0.40	0.24

Note: The global null hypothesis for the high-risk sub-population is: $H_{012}=\{p^{(0)}=0.86, p^{(1)}=p^{(2)}=0.76\}$, while for the low-risk $H_{0123}=\{p^{(0)}=0.92, p^{(1)}=p^{(2)}=p^{(3)}=0.82\}$. The alternative hypothesis for the high-risk sub-population is: $H_1=\{p^{(0)}=p^{(1)}=p^{(2)}=0.86\}$, while for the low-risk $H_1=\{p^{(0)}=p^{(1)}=p^{(2)}=p^{(3)}=0.92\}$. Results are provided using 10^6 replications. Values in bold refer to the target probability (around 80%) for each power configuration.

respectively. The trial is expected to end after 15.0 and 17.6 months for the high-risk group and after 18.4 and 21.9 months for the low-risk group under the null and alternative hypotheses, respectively.

5 | NUMERICAL EVALUATION

5.1 | Competing approaches

In this section, we evaluate the operating characteristics of the proposed design under various treatment configurations. The proposed non-inferiority design is compared to the MAMS(m) design proposed by Magirr et al.[18] but that allows all treatment arms that have not crossed any critical bounds to continue to the next stage. This design has been adapted for the non-inferiority setting. The FWER expression for MAMS(m) is the same as derived by Magirr et al[18] but the power expression changes. Tables A2–A4 in the Appendix, highlight the differences in the decision rules between the ORD and MAMS(m) designs for 3-arm 2-stage and 4-arm 2-stage settings.

5.2 | Setting

We consider the setting as described in Section 4 and we evaluate the performance of the design under intermediate treatment effects configurations when the treatment effect on the first arm is fixed to be zero, $\theta^{(1)}=0$, and the treatment effects on the other arms are varied. For the 3-arm 2-stage design we consider

TABLE 4 Maximum sample size (MaxSS), actual maximum sample size (AMS) under global null and alternative hypotheses, expected sample size (ESS) under global null and alternative hypotheses, FWER and probability to reject at least one hypothesis, all hypotheses and the longest and medium durations (LM) for each sub-population under the alternative for the ORD designs when triangular boundaries are used and only futility bounds are used at the interim analysis. D_{H_0} and D_{H_1} are the expected durations (in months) of the trial under the null and the alternative hypotheses, respectively.

Sub-populations	MaxSS	AMS_{H_0}	AMS_{H_1}	ESS_{H_0}	ESS_{H_1}	FWER	Reject at least one H_{0k}	Reject all	Reject LM	l_1, u_2	D_{H_0}	D_{H_1}
High-risk	450	470	469	452	468	0.049	**0.82**	0.69	-	0.565, 1.600	15.0	17.6
Low-risk	544	560	560	526	559	0.049	0.92	**0.81**	0.87	0.570, 1.612	18.4	21.9

Note: The global null hypothesis for the high-risk sub-population is: $H_{012} = \{p^{(0)} = 0.86, p^{(1)} = p^{(2)} = 0.76\}$, while for the low-risk $H_{0123} = \{p^{(0)} = 0.92, p^{(1)} = p^{(2)} = p^{(3)} = 0.82\}$. The alternative hypothesis for the high-risk sub-population is: $H_1 = \{p^{(0)} = p^{(1)} = p^{(2)} = 0.86\}$, while for the low-risk $H_1 = \{p^{(0)} = p^{(1)} = p^{(2)} = p^{(3)} = 0.92\}$. A recruitment rate of around 30 patients per month is assumed for both risk groups. Results are provided using 10^5 replications. Values in bold refer to the target probability (around 80%) for each power configuration.

$$\theta^{(2)} \in \{-0.10, -0.07, -0.04, -0.01, 0.02, 0.05, 0.08, 0\},$$

while for the 4-arm 2-stage design we consider

$$\left(\theta^{(2)}, \theta^{(3)}\right) \in \{(-0.1, -0.11), (-0.07, -0.08), (-0.04, -0.06), (-0.01, -0.03), (0.02, 0), (0.05, 0.02), (0.08, 0.05), (0, 0)\}.$$

Note that the considered values were just chosen to cover values in the interval $[-\delta, \delta]$, where δ is the selected non-inferiority margin. As one does not know what the treatment effect will actually be in the trial, there is the need to explore different scenarios. In particular, for the 3-arm 2-stage ORD a sequence of values between $-\delta$ and $\delta - 0.02$ was chosen with a gap of 0.03 units. For the 4-arm 2-stage design instead the same values for $\theta^{(2)}$ as for the 3-arm 2-stage design were used, while for $\theta^{(3)}$ other values in the interval $[-0.11, 0.05]$ where considered.

For the high-risk sub-population, both designs are powered at 80% to reject at least one hypothesis, while for the low-risk sub-population designs are powered at 80% to reject all hypotheses when all durations have the same response rates as the standard of care. Additionally to the achieved power, the efficiency of the designs is measured by their expected sample size (ESS). However, the ESS depends on the recruitment rate and the time of the interim analysis and this might be close or equal to the maximum sample size of the selected design. Another metric to look at is the time to first positive claim. This is a metric, where a difference could be observed, but it would require assumptions on recruitment speed in order to gauge if a sequential design has the potential to decrease the study cost (or not). However, it is worth noting that the time to first positive claim is a metric that might not necessarily take into account the correct decision at the end of the trial (e.g., to reject two or more hypotheses when all arms are non-inferior to the control).

5.3 | Numerical results

In this section, we provide the results of the simulation studies for the two separate sub-populations.

Table 2 provides the results for each sub-population and power configuration under the global and alternative hypotheses for the chosen power configurations of the ORD—blue rows in Table 2—and the MAMS(m). The design's performances under the considered non-null treatment configurations are provided in Figures 2 and 3, describing for each scenario the probability to reject the alternative hypotheses, the expected sample size, the duration of the trial and the time to first positive claim for the 3-arm and 4-arm design, respectively. We consider a recruitment rate of around 30 patients per month for each sub-group and for the case where we do not have any rejection, the time to first positive claim is set to the end of the trial.

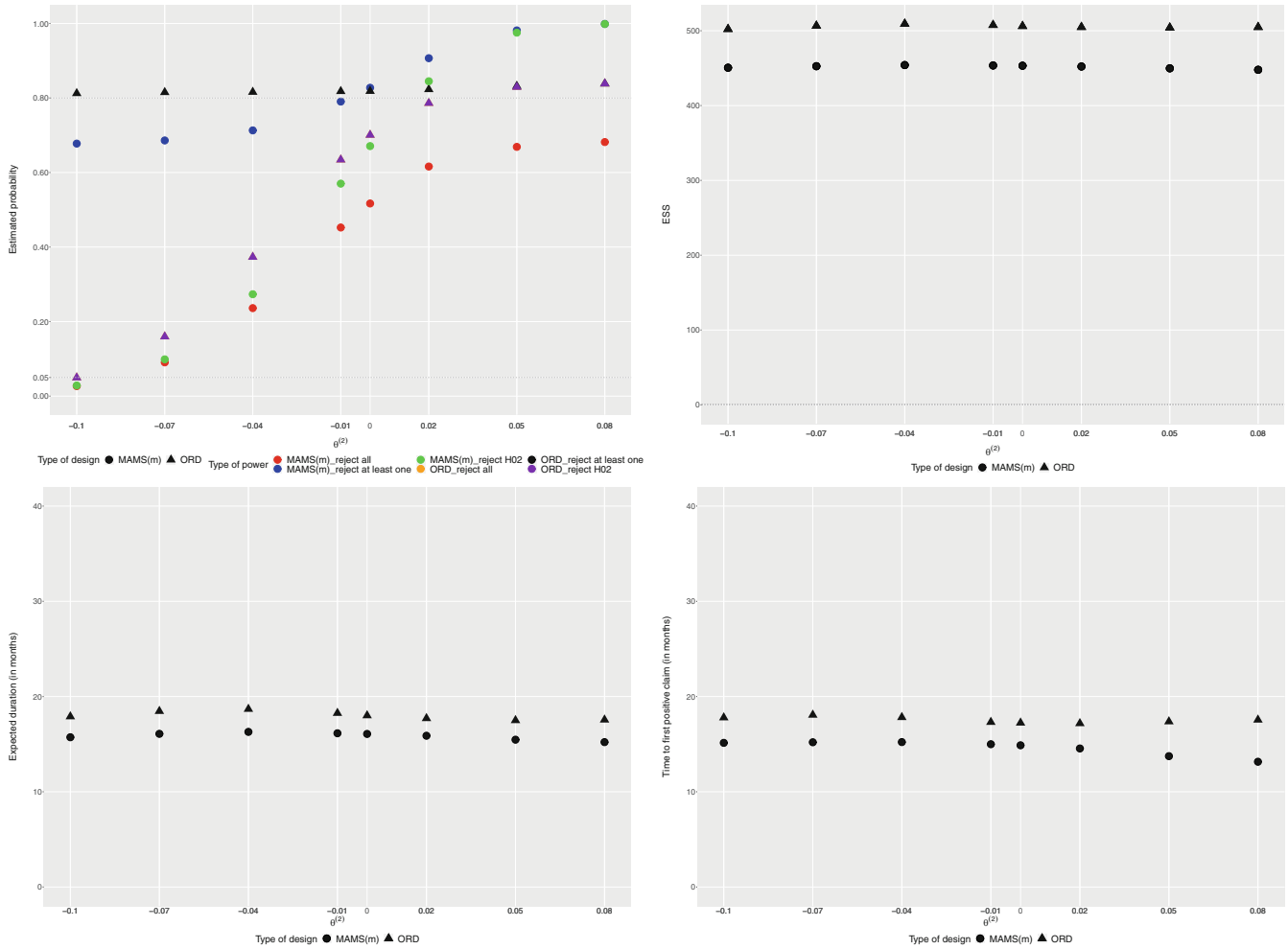

FIGURE 2 Probability to reject at least one hypothesis, to reject all hypotheses, to reject the second hypothesis—top left panel—, expected sample sizes (ESS)—top right panel—, duration of the trial—bottom left panel—and time to first positive claim—bottom right panel—for the 3-arm 2-stage MAMS(m) and ORD under $\theta^{(1)} = 0, \theta^{(2)} \in \{-0.10, -0.07, -0.04, -0.01, 0, 0.02, 0.05, 0.08\}$ when both designs are powered to reject at least one hypothesis at 80% under $\theta = (0,0)$. 3-arm 2-stage MAMS(m) and ORD designs use triangular critical bounds. Note that the ORD_reject all and the ORD_reject H02 points overlap in the Figure. A recruitment rate of around 30 patients per month is assumed for the high-risk group. Results are provided using 5×10^4 replications.

5.3.1 | High-risk sub-population

Total maximum sample sizes of 522 and 444 patients are required to reach a power of 80% to reject at least one null hypothesis under $\theta = (0,0)$ for the ORD and the MAMS(m), respectively. The sample size is lower for the MAMS design because we are powering to reject at least one hypothesis (rather than all hypotheses). The MAMS design does not take into account the order in rejecting the hypotheses. Thus, we need a smaller sample size in order to reject at least one hypothesis compared to the ORD. Both designs control the FWER under the global null at level α. Despite the differences in total maximum sample size, it can be observed that while the ORD has a power of around 70% to reject all true hypotheses under $\theta = (0,0)$, the MAMS(m) design reaches only 52%.

Figure 2 shows that the ORD has higher power to reject all hypotheses—that coincides with the probability to reject H_{02}—for all considered scenarios compared to the MAMS(m)—a difference up to 18%. For all scenarios, the ORD has around 80% of power to reject at least one hypothesis, while the MAMS(m) design has lower power when $\theta \leq -0.01$—a difference up to 13%. When $\theta > 0$, the MAMS(m) has higher power to reject at least one hypothesis compare to the ORD—a difference at most of 16%. The ORD has also higher power to reject the second hypothesis when $\theta < 0.02$—up to 10%—compared to the MAMS(m) but it has lower power when $\theta \geq 0.02$—a decrease up to 16%.

In terms of ESS, the average number of patients is expected to be at most 10% higher for the ORD compared to the MAMS(m) depending on the considered scenario. In terms of duration of the trial, the ORD is expected to be slightly

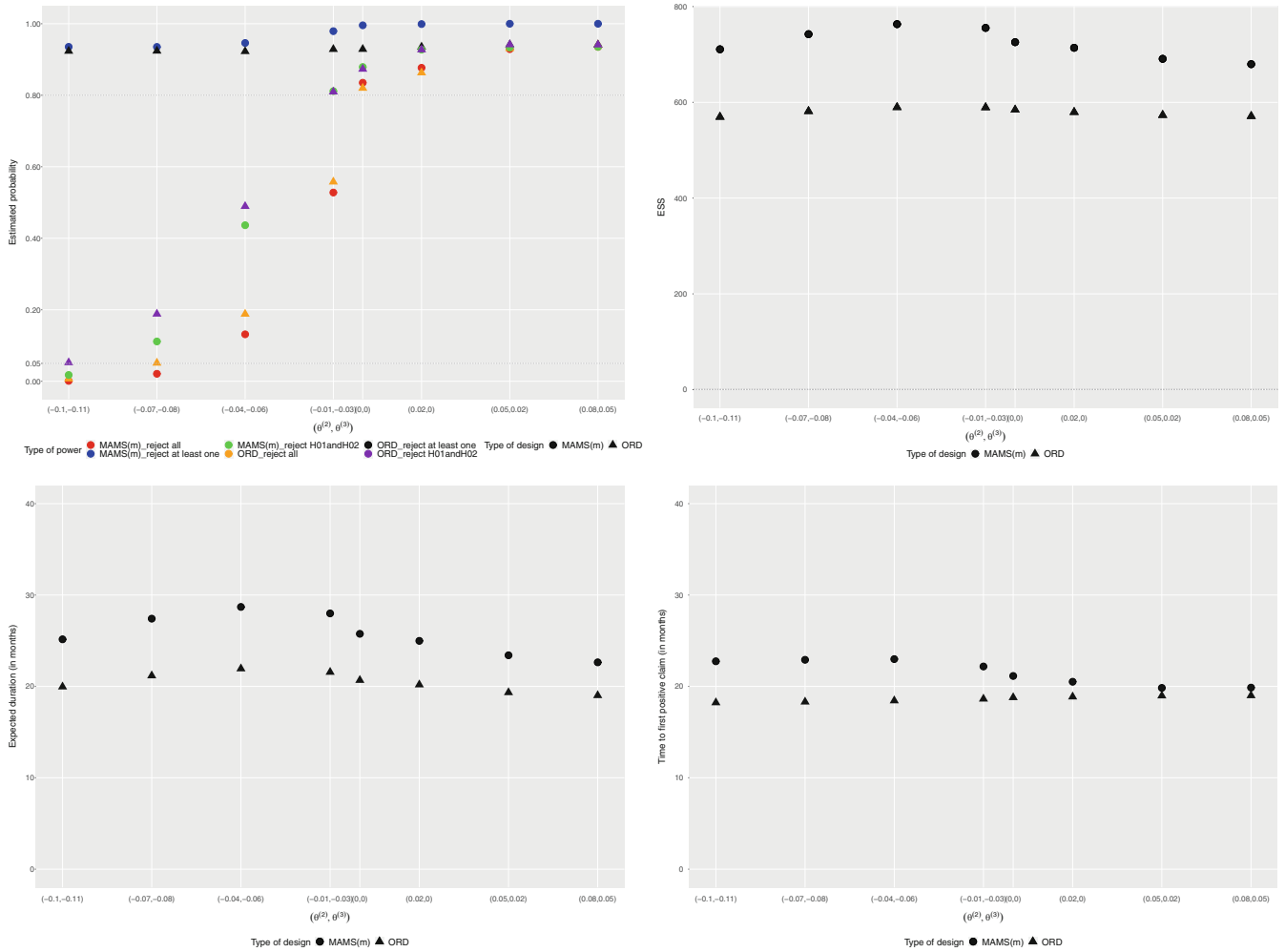

FIGURE 3 Probability to reject at least one hypothesis, to reject all hypotheses, to reject the first and the second hypotheses—top left panel—, expected sample sizes (ESS)—top right panel—, duration of the trial—bottom left panel—and time to first positive claim—bottom right panel—for the 4-arm 2-stage MAMS(m) and ORD under $\theta^{(1)} = 0, \left(\theta^{(2)}, \theta^{(3)}\right) \in \{(-0.1, -0.11), (-0.07, -0.08), (-0.04, -0.06), (-0.01, , -0.03), (0,0), (0.02,0), (0.05,0.02), (0.08,0.05)\}$ when both designs are powered to reject all hypotheses at 80% under $\theta = (0,0,0)$. 4-arm 2-stage MAMS(m) and ORD designs use triangular critical bounds. A recruitment rate of around 30 patients per month is assumed for the low-risk group. Results are provided using 5×10^4 replications.

longer compared to the MAMS for each scenario and the time to first positive claim is smaller—up to 4 months—for the MAMS design compared to the ORD in each scenario.

5.3.2 | Low-risk sub-population

Total maximum sample sizes of 624 and 832 patients are required to reach a power of around 80% to reject all hypotheses under $\boldsymbol{\theta} = (0,0,0)$ for the ORD and the MAMS(m), respectively. Here, we power to reject all hypotheses and thus we require more patients in MAMS compared to the ORD. Both designs control the FWER under the global null at level α.

Figure 3 shows that the ORD has higher power to reject all hypotheses for almost all considered scenarios compared to the MAMS(m)—a difference up to 5%. For all scenarios, the ORD has around 92% of power to reject at least one hypothesis, while the MAMS(m) design has higher power for all scenarios—a difference at most of 6.7%. The ORD has also higher power to reject the first two hypotheses for almost all considered scenarios compared to the MAMS(m)—up to 7.7% of difference.

In terms of ESS, the ORD can provide a reduction up to 23%, depending on the scenario, compared to the MAMS(m). In terms of duration of the trial, the ORD is expected to be shorter—up to 6 months—compared to

the MAMS design for each scenario with larger differences for scenarios with two negative treatment effects and the time to first positive claim is smaller—up to 5 months—for the MAMS design compared to the ORD for scenarios where all treatment effects are above zero.

Overall, the non-inferiority ORD can provide higher power to reject all hypotheses and smaller expected sample size compared to the non-inferiority MAMS(m) design when it is powered to reject all hypotheses and the order assumption among the treatment effects is satisfied. The MAMS design outperforms the ORD when the order assumption is not satisfied, that is when $\theta^{(1)} \geq \theta^{(2)} \geq \theta^{(3)}$ is not satisfied. This is expected because the MAMS design does not take into account the order in rejecting the hypotheses.

6 | TIMING OF THE INTERIM ANALYSES

In the original ORD[16] and in the results obtained above, the information time was assumed to be the same for all treatments. When considering different durations of treatment, however, the information for each arm accumulates at different times. For example, consider the high-risk sub-population where the standard regimen is at fixed duration of 6 months, while the experimental regimens are at 4 and 3 months, respectively. As represented in Figure 4, assume that we start to recruit patients on each arm at the same time. At time t_1 the first block of patients is recruited and allocated to the three treatment arms. Similarly, at time t_2 and t_3 the second and third blocks of patients are recruited and allocated to the treatment arms, respectively. The recruitment process continues in this way, but, while the first block of patients that were recruited at time t_1 have their efficacy outcomes evaluated in the next 6 months if they were allocated to the control arm, the other patients will have their efficacy outcomes evaluated after 4 and 3 months if they were allocated to the longest or shortest duration arms, respectively. Thus, information accumulates at different times and at the time of the interim analysis a different number of patients in each arm has completed their treatment. Hence, for this clinical trial setting, different strategies about the timing of the interim analysis can be evaluated. Below, we propose and examine two possible strategies.

The first one consists on having the same amount of information—that is, the same number of patients with their primary endpoints evaluated—on each treatment duration at the time of the planned interim analysis. We refer to this strategy as to SI (Same information at the Interim analysis). Thus, the interim analysis is performed when the same number of patients $n_1 = n_1^{(k)}$ have completed their treatment k. The recruitment of patients ends when the total number of patients in the control arm is equal to $2n_1$. This means that at the end of the trial, more patients will be recruited in the longer duration arms that are still in the trial compared to the shorter durations. Indeed, in order to have the same

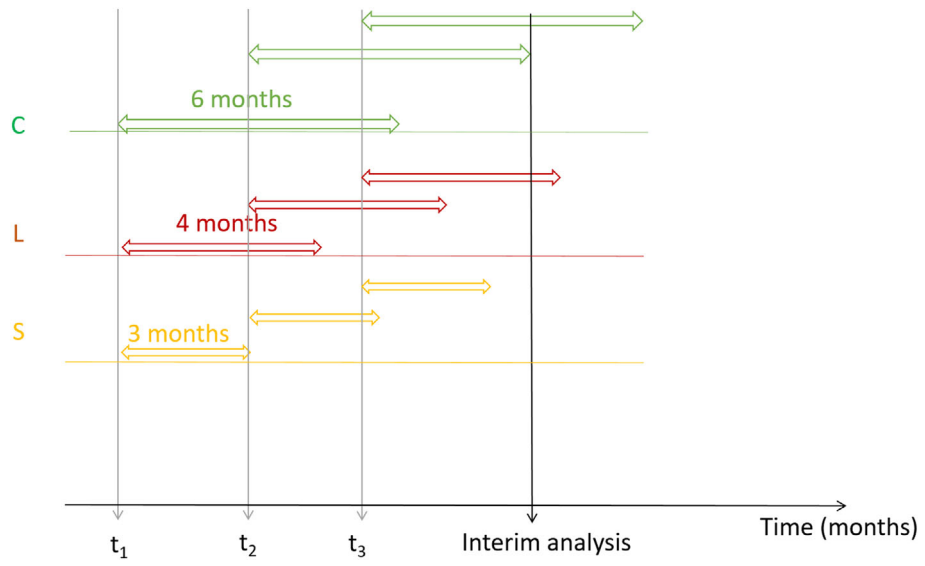

FIGURE 4 Schematic of the accumulation of information for the high-risk sub-population. C is the control arm, L is the longest active duration and S is the shortest active duration. Vectors indicate the length of the treatment. Patients are recruited and allocated to the treatment arms at times $t_i, i \in \{1,2,...,T\}$, with T being the end of the trial.

number of patients in each arm at the interim analysis, each month we need to recruit more patients on the longer durations and fewer in the shorter ones. Thus, for this strategy, we want to modify the allocation ratio in order to recruit more patients under longer treatment durations—as they require more time in order to have their primary outcomes evaluated—and fewer patients on the shorter arms. Figure 5 shows a schematic of the SI strategy.

In the second strategy, instead, the same amount of information is observed in the final analysis—the number of patients in each arm is equal to $n_2 = n_2^{(k)}$. We refer to this strategy as to SF (Same information at the Final analysis). In this case, more information on the shortest durations will be accumulated—as represented in Figure 4—at the interim analysis, which is done when half of the population in the control group has been observed. Note that at the end of the trial all accumulated data is used and included into the analysis. Thus, no data is thrown away and patients will still be followed up if shorter treatment arms have been dropped for futility at the interim analysis.

6.1 | Strategy SI

Let us denote with x_k the number of patients recruited in each arm and for each month with $k \in \{\{0\} \cup K_h\}$ for the high-risk and $k \in \{\{0\} \cup K_l\}$ for the low-risk sub-population. Let d_k be the expected number of months that are needed to recruit x_k patients on treatment arm k and $\boldsymbol{D} = (D_0, D_1, D_2, D_3), \boldsymbol{D} = (D_0, D_1, D_2)$ be the vectors of durations (in months) of treatments that are tested for the low-risk and high-risk sub-populations, respectively. Let R be the number of patients that are recruited per month in each sub-population and $n_1 = n_1^{(k)}$ be the number of patients in each arm up to stage 1. Thus, for the low-risk sub-population we satisfy the following

$$\begin{cases} \min \ \ x_3 d_3 - x_2 d_2 + x_1 d_1 - x_0 d_0 \\ \text{s.t.} \ \ x_3 d_3 - x_2 d_2 + x_1 d_1 - x_0 d_0 \geq 0 \\ x_0 d_0 = x_1 d_1 = x_2 d_2 = x_3 d_3 = n_1 \\ x_0 + x_1 + x_2 + x_3 = R \\ d_1 + D_1 = d_2 + D_2 = d_3 + D_3 = d_0 + D_0 \end{cases}$$

and for the high-risk sub-population

$$\begin{cases} \min \ \ x_2 d_2 - x_1 d_1 + x_0 d_0 \\ \text{s.t.} \ \ x_2 d_2 - x_1 d_1 + x_0 d_0 \geq 0 \\ x_0 d_0 = x_1 d_1 = x_2 d_2 = n_1 \\ x_0 + x_1 + x_2 = R \\ d_1 + D_1 = d_2 + D_2 = d_0 + D_0 \end{cases}$$

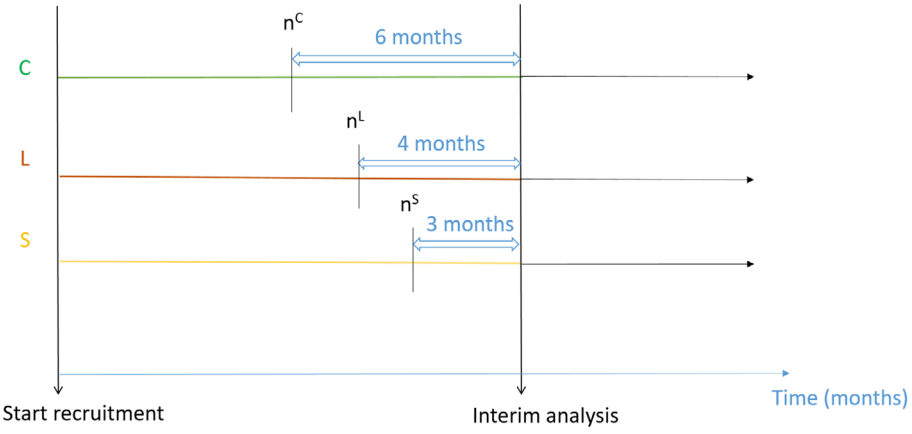

FIGURE 5 Schematic of the SI strategy for the high-risk sub-population. C is the control arm, L is the longest active duration and S is the shortest active duration.

Note that given the equality constraints, the solutions of the systems are the roots of polynomial equations of degree 3 for the high-risk and 4 for the low-risk sub-populations. To solve these systems we use the package nloptr[28] in R.[27] For the low-risk sub-population we initialize the variables as

$$\begin{cases} x_k = \dfrac{R}{4}, k \in \{\{0\} \cup K_l\} \\ d_3 = \dfrac{n_1}{x_3}, d_0 = d_3 + (D_3 - D_0), d_1 = d_0 - (D_1 - D_0), d_2 = d_0 - (D_2 - D_0) \end{cases}$$

and for the high-risk as

$$\begin{cases} x_k = \dfrac{R}{3}, k \in \{\{0\} \cup K_h\} \\ d_2 = \dfrac{n_1}{x_2}, d_0 = d_2 + (D_2 - D_0), d_1 = d_0 - (D_1 - D_0) \end{cases}$$

where the initial n_1 is found analytically in order to satisfy the power requirements and considering the ratios of sample sizes $r_j^{(k)} = j$ for $j \in \{1,2\}$ and $k \in \{\{0\} \cup K_h\}$ for the high-risk and $k \in \{\{0\} \cup K_l\}$ for the low-risk sub-populations .

We solve the system and find $r_2^{(k)} = 2\frac{x_k}{x_0}$ with $k \in K_l$ for the low-risk and $k \in K_h$ for the high-risk sub-populations. We then find analytically the sample size per arm per stage and the critical bounds that are necessary to reach the desired power and to control the FWER at level α with $r_2^{(k)}$ and $r_1^{(k)} = 1$ and $k \in K_l$ for the low-risk and $k \in K_h$ for the high-risk sub-populations. We repeat this procedure until we get a stable solution. We finally run simulations in order to explore the operating characteristics of the design. The interim analysis will be done at month $d_k + D_k$.

6.2 | Strategy SF

For this strategy, patients are recruited until the same number of patients have been observed at the second stage in each arm. That is $n_2 = n_2^{(k)} = 2n$, and thus $r_2^{(k)} = 2$, with $k \in \{\{0\} \cup K_l\}$ for the low-risk sub-population and $k \in \{\{0\} \cup K_h\}$ for the high-risk. At the interim we have $r_1^{(k)} > 1$ for $k \in K_h$ and $k \in K_l$ in the high-risk and low-risk sub-populations, respectively.

As described in the previous section, let us denote by x_k the number of patients per arm per each month and R the total number of patients recruited per month in low-risk and high-risk sub-populations, respectively. It follows that $x_k = \frac{R}{4}, k \in \{\{0\} \cup K_l\}$ in the low-risk, whereas $x_k = \frac{R}{3}, k \in \{\{0\} \cup K_h\}$ in the high-risk sub-population. The interim analysis is done when half of the population in the control group has been observed. So at the first stage we have $n_1^{(0)} = n, n_1^{(1)} = n + x_1(D_0 - D_1), n_1^{(2)} = n + x_2(D_0 - D_2), n_1^{(3)} = n + x_3(D_0 - D_3)$. It follows that $r_1^{(1)} = \frac{n + x_1(D_0 - D_1)}{n}, r_1^{(2)} = \frac{n + x_2(D_0 - D_2)}{n}, r_1^{(3)} = \frac{n + x_3(D_0 - D_3)}{n}, r_1^{(0)} = 1$. We then find analytically the sample size per arm per stage and the critical bounds that are necessary to reach the desired power and to control the FWER at level α with $r_1^{(k)}$ and $r_2^{(k)} = 2$. We run simulations in order to explore the operating characteristics of the design. The final analysis will be done at $\frac{2n}{x_0} + D_0$ months.

6.3 | Numerical evaluation of the two strategies

We consider the same design settings as described in Section 4 and the suggested design proposed in Section 4.3. Based on the clinical team experience, we provide the results considering a recruitment rate of around 30 patients per month, $R = 30$. Thus, we expect to recruit around one patient per day and this is randomly allocated to a treatment arm k with probability $p_k = x_k/R$. We consider the following treatment durations (in months) $\boldsymbol{D} = (D_0, D_1, D_2, D_3) = (6,4,3,2)$ and $\boldsymbol{D} = (D_0, D_1, D_2) = (6,4,3)$ for the low-risk and high-risk sub-population, respectively. The numerical results are found using 10^5 replicate simulations.

Table 5 provides the operating characteristics of the two strategies under the global and alternative hypotheses for each sub-population together with the theoretical maximum sample size, the actual total sample size, that is the mean number of the patients recruited to the trial when all treatment arms proceed to the final analysis and the expected sample size.

For the high-risk sub-population, a total of 476 and 504 patients—502 and 519 actual maximum sample sizes, respectively—are required for the SI and SF strategies, respectively, if the design's parameters provided in Table 6 are used in the trial. It can be observed that the SF requires a larger total sample size compared to the SI strategy in order to reach 80% power to reject at least one hypothesis. However, it can be observed that the ESS at the interim analysis for both strategies is still below the actual maximum sample size. Thus, on average, not all patients are recruited at the time of the interim analysis. The expected duration of the trial under the alternative hypothesis is estimated to be around 15.6 and 19.2 months for the SI and the SF strategies, respectively. This small difference in durations is due to the difference in the number of patients that are required to be observed in each arm for the first interim analysis—the interim analysis in SI is done when the same amount of patients has completed their treatment, while the interim analysis in SF is done as soon as the last patient in the control arm has ended the treatment.

For the low-risk sub-population, a total of 566 and 584 patients—586 and 600 actual maximum sample sizes, respectively—are required for the SI and SF strategies, respectively, in order to reach 80% of power to reject all hypotheses, if the design's parameters provided in Table 6 are used in the trial. As for the high-risk sub-population, the ESS at the interim analysis for both strategies is still below the actual maximum sample size. Under the alternative hypothesis, the total duration of the trial is expected to be around 18.5 and 20.8 months for the SI and SF strategies, respectively. As for the high-risk sub-population, the difference in durations is reflected by the differences in the number of patients that are required to be observed in each arm for the first interim analysis.

Overall, the results suggest that the strategy that matches the sample size at the interim analysis is one that minimizes the total maximum sample sizes under the considered simulation scenarios, while the SF strategy minimizes the expected sample size under the global null and alternative hypotheses. The two strategies have almost the same duration under the two hypotheses—small differences are due to the different sample sizes and different probabilities to stop at the interim and final analyses. Thus, in order to minimize the expected sample size, the SF strategy is preferred.

7 | DISCUSSION

The aim of this work was to describe the application of the order restricted design proposed by Serra et al.[16] in the context of a tuberculosis trial. In this clinical trial setting, non-inferiority trials are the norm and hence an extension of the original design has been proposed. Practical considerations were provided regarding how to choose some design's parameters such as the non-inferiority margin and the shape of the critical bounds for the considered TB trial. Theoretical and practical considerations regarding several types of power configurations were provided and two different strategies were proposed in order to take into account the fact that the information is accumulating at different times when multiple treatment durations are simultaneously tested in the same trial.

The primary objective of the trial was to identify the shortest possible treatment duration in each sub-population. Thus, the trial is to be powered to reject all correct hypotheses. However, alternative power strategies can be more feasible. For example, when even some reduction in the duration of the treatment is of interest or the resources are limited. In these cases, one could consider to power the design in order to reject the correct hypotheses relating to the particular number of experimental arms (rather than all of them).

The ORD has been shown to be an efficient design that can be applied when multiple treatment durations are simultaneously tested in the same trial. Nevertheless, all the considerations provided in this work were specific to the REStrUCTuRe trial. The choice of the primary endpoint was driven by the clinical investigator in the team, and the design in the manuscript is proposed under the assumption that this endpoint is valid. This manuscript is focused on the questions on how one could design such a study upon the agreement on the particular choice of the endpoint. The methodology proposed in this work can be applied to different definitions of primary endpoint, for example endpoints that consider a minimum follow-up for patients after they have ended their treatment. Indeed, even though culture might be negative, bacteria may be left and it takes long for TB to grow, such that recurrence could only be observed months after treatment completion (while most appear to happen relatively shortly after completion). Culture conversion at treatment completion is not a perfect surrogate for long-term success—that's why the STEP design[29] has been proposed to de-risk subsequent Phase III studies. The efficiency in using the proposed adaptive design with another definition of primary endpoint would be determined by the specific setting, specifically, by the expected recruitment rate as well as the time of follow-up for every patient. This, however, holds for any adaptive design.[30]

In this trial, we have assumed to consider an interim analysis when half of the total maximum sample size has completed their treatment. However, depending on the recruitment speed, it might happen that all patients are already

TABLE 5 Sample size at the first stage on the control arm $\left(n_1^{(0)}\right)$, maximum sample size (MaxSS), actual maximum sample size (AMS) under global null and alternative hypotheses, expected sample size (ESS) under global null and alternative hypotheses, expected sample size $\left(\mathrm{ESS}^{IA}\right)$ at the interim analysis (IA) under global null and alternative hypotheses, FWER, the probability to reject at least one hypothesis, all hypotheses and the longest and medium durations (LM) for each strategy under the alternative for the high-risk and low-risk sub-populations and the expected total duration of the trial, expected time of the interim analysis (IA) in months under the global null $\left(\mathrm{D}_{H_0}, \mathrm{D}_{H_0}^{IA}\right)$ and alternative hypotheses $\left(\mathrm{D}_{H_1}, \mathrm{D}_{H_1}^{IA}\right)$.

High-risk															
Strategy	$n_1^{(0)}$	**MaxSS**	$\mathbf{AMS}_{H_0}$	$\mathbf{AMS}_{H_1}$	$\mathbf{ESS}_{H_0}$	$\mathbf{ESS}_{H_1}$	$\mathbf{ESS}_{H_0}^{IA}$	$\mathbf{ESS}_{H_1}^{IA}$	**FWER**	**Rej. at least one**	**Rej. all**	$\mathbf{D}_{H_0}^{IA}$	$\mathbf{D}_{H_1}^{IA}$	$\mathbf{D}_{H_0}$	$\mathbf{D}_{H_1}$
SI	94	476	501	502	491	494	488	488	0.050	**0.81**	0.68	14.2	14.2	14.9	15.6
SF	84	504	519	518	462	481	447	447	0.050	**0.81**	0.70	14.9	14.9	16.9	19.2
Low-risk															
Strategy	$n_1^{(0)}$	**MaxSS**	$\mathbf{AMS}_{H_0}$	$\mathbf{AMS}_{H_1}$	$\mathbf{ESS}_{H_0}$	$\mathbf{ESS}_{H_1}$	$\mathbf{ESS}_{H_0}^{IA}$	$\mathbf{ESS}_{H_1}^{IA}$	**FWER**	**Rej. at least one**	**Rej. all**	**Rej. LM**	$\mathbf{D}_{H_0}^{IA}$	$\mathbf{D}_{H_1}^{IA}$	$\mathbf{D}_{H_0}$	$\mathbf{D}_{H_1}$
SI	86	566	585	586	559	567	552	552	0.048	0.93	**0.81**	0.86	16.2	16.2	17.5	18.5
SF	73	584	600	597	509	535	487	487	0.049	0.92	**0.81**	0.85	16.2	16.2	18.5	20.8

Note: A recruitment rate of around 30 patients per month is assumed for both risk groups. Results are provided using 10^5 replications. Values in bold refer to the target probability (around 80%) for each power configuration.

TABLE 6 Triangular critical bounds and ratios of sample sizes for each strategy and sub-population.

High-risk							
Strategy	**MaxSS**	u_1	u_2	l_1	$r_1^{(1)}, r_2^{(1)}$		$r_1^{(2)}, r_2^{(2)}$
SI	476	1.895	1.787	0.632	1, 1.595		1, 1.448
SF	504	1.896	1.788	0.632	1.238, 2		1.357, 2
Low-risk							
Strategy	**MaxSS**	u_1	u_2	l_1	$r_1^{(1)}, r_2^{(1)}$	$r_1^{(2)}, r_2^{(2)}$	$r_1^{(3)}, r_2^{(3)}$
SI	566	1.896	1.788	0.632	1, 1.649	1, 1.516	1, 1.403
SF	584	1.896	1.788	0.632	1.205, 2	1.308, 2	1.410, 2

enrolled at the time of the interim analysis. Thus, other practical considerations and modifications of the design should be considered depending on the specific trial setting.

In addition, one of the limitations of this work is that it relies on the asymptotic normal distribution of the test statistics. Especially when the sample size is small, the normal distribution could be a poor approximation for the test statistic. Moreover, alternative strategies for the different timings of the analyses can be explored, that is, staggering the opening of the treatment durations in order to get the same amount of information at the interim or final analyses.

AUTHOR CONTRIBUTIONS

All authors contributed equally to 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 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.

FUNDING INFORMATION

None reported.

CONFLICT OF INTEREST STATEMENT

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

Programming code for reproducing the numerical results is available at GitHub:https://github.com/OrderedRestrictedDesign/Non-inferiority.

ORCID

Alessandra Serra https://orcid.org/0000-0001-8431-5154
Pavel Mozgunov https://orcid.org/0000-0001-6810-0284

REFERENCES

1. World Health Organization and others. Global tuberculosis report 2020: executive summary. 2020.
2. Murray JF, Schraufnagel DE, Hopewell PC. Treatment of tuberculosis. A historical perspective. *Ann Am Thorac soc*. 2015;12(12):1749-1759.
3. 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.1003059

4. Imperial MZ, Nahid P, Phillips PP, et al. A patient-level pooled analysis of treatment-shortening regimens for drug-susceptible pulmonary tuberculosis. *Nat Med*. 2018;24(11):1708-1715.
5. Johnson JL, Hadad DJ, Dietze R, et al. Shortening treatment in adults with noncavitary tuberculosis and 2-month culture conversion. *Am J Respir Crit Care Med*. 2009;180(6):558-563.
6. Dorman SE, Nahid P, Kurbatova EV, et al. Four-month rifapentine regimens with or without moxifloxacin for tuberculosis. *N Engl J Med*. 2021;384(18):1705-1718.
7. Jaki T. Multi-arm clinical trials with treatment selection: what can be gained and at what price? *J Clin Invest*. 2015;5:393-399. doi:10.4155/cli.15.13
8. 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
9. Burnett T, Mozgunov P, Pallmann P, Villar SS, 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.
10. Phillips PP, Gillespie SH, Boeree M, et al. Innovative trial designs are practical solutions for improving the treatment of tuberculosis. *J Infect Dis*. 2012;205(suppl_2):S250-S257.
11. Bratton DJ, Phillips PP, Parmar MK. A multi-arm multi-stage clinical trial design for binary outcomes with application to tuberculosis. *BMC Med Res Methodol*. 2013;13(1):1-14.
12. Cellamare M, Ventz S, Baudin E, Mitnick CD, Trippa L. A Bayesian response-adaptive trial in tuberculosis: the end TB trial. *Clin Trials*. 2017;14(1):17-28.
13. Boeree MJ, Heinrich N, Aarnoutse R, et al. High-dose rifampicin, moxifloxacin, and SQ109 for treating tuberculosis: a multi-arm, multi-stage randomised controlled trial. *Lancet Infect Dis*. 2017;17(1):39-49.
14. Papineni P, Phillips P, Lu Q, Cheung Y, Nunn A, Paton N. TRUNCATE-TB: an innovative trial design for drug-sensitive tuberculosis. *Int J Infect Dis*. 2016;45:404.
15. Quartagno M, Carpenter JR, Walker AS, Clements M, Parmar MKB. The DURATIONS randomised trial design: estimation targets, analysis methods and operating characteristics. *Clin Trials*. 2020;17:644-653.
16. Serra A, Mozgunov P, Jaki T. An order restricted multi-arm multi-stage clinical trial design. *Stat Med*. 2022;41:1613-1626. doi:10.1002/sim.9314
17. Nunn AJ, Phillips PP, Gillespie SH. Design issues in pivotal drug trials for drug sensitive tuberculosis (TB). *Tuberculosis*. 2008;88:S85-S92.
18. Magirr D, Jaki T, Whitehead J. Ageneralized Dunnett test for multi-armmulti-stage clinical studies with treatment selection. *Biometrika*. 2012;99:494-501. doi:10.1093/biomet/ass002
19. Zhang Z. Non-inferiority testing with a variable margin. *Biom J*. 2006;48(6):948-965.
20. Wassmer G, Brannath W. *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials. 301*. Springer; 2016.
21. European Agency for the Evaluation of Medicinal Products. Committee for Proprietary Medicinal Products: points to consider on multiplicity issues in clinical trials. 2002.
22. Fleming TR. Current issues in non-inferiority trials. *Stat Med*. 2008;27(3):317-332.
23. U.S. Department of Health and Human Services Food and Drug Administration. Non-inferiority clinical trials to establish effectiveness guidance for industry; 2016.
24. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika*. 1977;64:191. doi:10.2307/2335684
25. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics*. 1979;35:549. doi:10.2307/2530245
26. Whitehead J. *The Design and Analysis of Sequential Clinical Trials*. John Wiley & Sons; 1997.
27. Team RC. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2019. http://www.R-project.org/.
28. Ypma J, Borchers HW, Eddelbuettel D, Ypma MJ. Package 'nloptr'. 2018.
29. Phillips PP, Dooley KE, Gillespie SH, et al. A new trial design to accelerate tuberculosis drug development: the phase IIC selection trial with extended post-treatment follow-up (STEP). *BMC Med*. 2016;14(1):1-11.
30. Wason J, Brocklehurst P, Yap C. When to keep it simple–adaptive designs are not always useful. *BMC Med*. 2019;17(1):1-7.
31. Nunn A. An international multi-Centre controlled clinical trial to evaluate high dose rifapentine and a quinolone in the treatment of pulmonary tuberculosis (RIFAQUIN). *Trop Med Int Health*. 2010;15(8):S22-S23.
32. Gillespie SH, Crook AM, McHugh TD, et al. Four-month moxifloxacin-based regimens for drug-sensitive tuberculosis. *N Engl J Med*. 2014;371(17):1577-1587.
33. Merle CS, Sismanidis C, Sow OB, et al. A pivotal registration phase III, multicenter, randomized tuberculosis controlled trial: design issues and lessons learnt from the Gatifloxacin for TB (OFLOTUB) project. *Trials*. 2012;13(1):1-10.
34. Dorman SE, Nahid P, Kurbatova EV, et al. High-dose rifapentine with or without moxifloxacin for shortening treatment of pulmonary tuberculosis: study protocol for TBTC study 31/ACTG A5349 phase 3 clinical trial. *Contemp Clin Trials*. 2020;90:105938.
35. Gov C. A randomised trial to evaluate toxicity and efficacy of 1200mg and 1800mg rifampicin for pulmonary tuberculosis (RIFASHORT).
36. Nunn AJ, Rusen I, Van Deun A, et al. Evaluation of a standardized treatment regimen of anti-tuberculosis drugs for patients with multi-drug-resistant tuberculosis (STREAM): study protocol for a randomized controlled trial. *Trials*. 2014;15(1):1-10.

37. Global Alliance for TB Drug Development. Shortening Treatment by Advancing Novel Drugs (STAND). https://clinicaltrials.gov/ct2/show/results/NCT02342886?view=results

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, Davies G, Jaki T. Determining the minimum duration of treatment in tuberculosis: An order restricted non-inferiority trial design. *Pharmaceutical Statistics*. 2023;22(5): 938-962. doi:10.1002/pst.2320

APPENDIX A: Equations to power the design

To power the 3-arm trial design at level $1-\beta$ to reject at least one hypothesis, the following Equation (A1) should be satisfied:

$$\begin{aligned} &P\left(Z_1^{(1)} \geq u_1^{(1)} | \theta_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + P\left(Z_2^{(1)} \geq u_2^{(1)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)} | \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) \geq 1-\beta. \end{aligned} \tag{A1}$$

while for the 4-arm trial we have

$$\begin{aligned} &P(\text{rejecting at least one true } H_{0k}, k \in K_l | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}) = \\ &P(\text{rejecting at least one true } H_{0k}, k \in \{1,2\} \,|\, \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} < u_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) \geq 1-\beta. \end{aligned} \tag{A2}$$

To power the design in order to reject the first and second treatments in a 4-arm 2-stage design at level $1-\beta$, the following Equation (A3) needs to be satisfied:

$$\begin{aligned} &P\left(Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, Z_1^{(2)} > l_1^{(2)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} < u_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) + \\ &P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \geq u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) + \\ &P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \leq l_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta}_{\mathbf{123}}, \boldsymbol{p}_{\mathbf{123}}\right) \geq 1-\beta. \end{aligned} \tag{A3}$$

In order to power the designs to reject all hypotheses, Equations (A4) needs to be satisfied for the 3-arm 2-stage trial

$$\begin{aligned} &P\left(Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \leq l_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)} | \theta_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + \\ &P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_1^{(1)} \geq u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)} | \theta_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) + \\ &P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)} | \boldsymbol{\theta}_{\mathbf{12}}, \boldsymbol{p}_{\mathbf{12}}\right) \geq 1-\beta. \end{aligned} \tag{A4}$$

while Equation (A5) needs to be satisfied for the 4-arm trial.

$$
\begin{aligned}
&P\left(Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(3)} \geq u_2^{(3)}, Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)}, l_1^{(3)} < Z_1^{(3)} < u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_2^{(3)} \geq u_2^{(3)}, Z_1^{(1)} \geq u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}, l_1^{(3)} < Z_1^{(3)} < u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(2)} \geq u_2^{(2)}, Z_2^{(3)} \geq u_2^{(3)}, Z_1^{(1)} \geq u_1^{(1)}, Z_1^{(2)} < u_1^{(2)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_2^{(3)} \geq u_2^{(3)}, l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}, l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}, l_1^{(3)} < Z_1^{(3)} < u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_2^{(3)} \geq u_2^{(3)}, Z_1^{(1)} < u_1^{(1)}, Z_1^{(3)} \geq u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) + \\
&P\left(Z_2^{(1)} \geq u_2^{(1)}, Z_2^{(2)} \geq u_2^{(2)}, Z_2^{(3)} \geq u_2^{(3)}, Z_1^{(1)} < u_1^{(1)}, Z_1^{(2)} \geq u_1^{(2)}, l_1^{(3)} < Z_1^{(3)} < u_1^{(3)} | \boldsymbol{\theta_{123}}, \boldsymbol{p_{123}}\right) \geq 1 - \beta.
\end{aligned}
\tag{A5}
$$

TABLE A1 Combination of the decision rules at the interim (top panel) and at the final (bottom panel) analyses for the 3-arm 2-stage trial with $\theta^{(1)} \geq \theta^{(2)}$.

	$\boldsymbol{Z_1^{(1)} \geq u_1^{(1)}}$	$\boldsymbol{l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}}$	$\boldsymbol{Z_1^{(1)} \leq l_1^{(1)}}$
$Z_1^{(2)} \geq u_1^{(2)}$	Stop: select T_1, T_2	Proceed with T_1, T_2	Proceed with T_1, T_2
$l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}$	Proceed with T_2	Proceed with T_1, T_2	Drop both arms
$Z_1^{(2)} \leq l_1^{(2)}$	Stop: select T_1	Proceed with T_1	Drop both arms
	$\boldsymbol{Z_2^{(1)} \geq u_2^{(1)}}$		$\boldsymbol{Z_2^{(1)} < u_2^{(1)}}$
$Z_2^{(2)} \geq u_2^{(2)}$	Select T_1, T_2		Select none
$Z_2^{(2)} < u_2^{(2)}$	Select T_1		Select none

Note: Cells coloured in red correspond to contradicting evidence.

TABLE A2 Combination of the decision rules at the interim (top panel) and at the final (bottom panel) analyses for the 3-arm 2-stage ORD—black colour—and MAMS(m)—blue colour—trial.

	$\boldsymbol{Z_1^{(1)} \geq u_1^{(1)}}$	$\boldsymbol{l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}}$	$\boldsymbol{Z_1^{(1)} \leq l_1^{(1)}}$
$Z_1^{(2)} \geq u_1^{(2)}$	Stop: select T_1, T_2	Proceed with T_1, T_2	Proceed with T_1, T_2
	Stop: select T_1, T_2	Proceed with T_1	Stop: select T_2
$l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}$	Proceed with T_2	Proceed with T_1, T_2	Drop both arms
	Proceed with T_2	Proceed with T_1, T_2	Proceed with T_2
$Z_1^{(2)} \leq l_1^{(2)}$	Stop: select T_1	Proceed with T_1	Drop both arms
	Stop: select T_1	Proceed with T_1	Drop both arms
	$\boldsymbol{Z_2^{(1)} \geq u_2^{(1)}}$		$\boldsymbol{Z_2^{(1)} < u_2^{(1)}}$
$Z_2^{(2)} \geq u_2^{(2)}$	Select T_1, T_2		Select none
	Select T_1, T_2		Select T_2
$Z_2^{(2)} < u_2^{(2)}$	Select T_1		Select none
	Select T_1		Select none

TABLE A3 Combination of the decision rules at the interim analysis for the 4-arm 2-stage ORD—black colour—and MAMS(m)—blue colour—trial.

		$Z_1^{(1)} \geq u_1^{(1)}$	$l_1^{(1)} < Z_1^{(1)} < u_1^{(1)}$	$Z_1^{(1)} \leq l_1^{(1)}$
$Z_1^{(2)} \geq u_1^{(2)}$	$Z_1^{(3)} \geq u_1^{(3)}$	Stop: select all—select all	Proceed with T_1, T_2, T_3—proceed with T_1	Proceed with T_1, T_2, T_3—stop: select T_2,T_3
	$l_1^{(3)} < Z_1^{(3)} < u_1^{(3)}$	Proceed with T_3—proceed with T_3	Proceed with T_1, T_2, T_3—proceed with T_1,T_3	Proceed with T_1, T_2, T_3—proceed with T_3
	$Z_1^{(3)} \leq l_1^{(3)}$	Stop: select T_1, T_2—stop: select T_1,T_2	Proceed with T_1, T_2—proceed with T_1	Proceed with T_1, T_2—stop: select T_2
$l_1^{(2)} < Z_1^{(2)} < u_1^{(2)}$	$Z_1^{(3)} \geq u_1^{(3)}$	Proceed with T_2, T_3—proceed with T_2	Proceed with T_1, T_2, T_3—proceed with T_1,T_2	Proceed with T_1, T_2, T_3—proceed with T_2
	$l_1^{(3)} < Z_1^{(3)} < u_1^{(3)}$	Proceed with T_2, T_3—proceed with T_2,T_3	Proceed with T_1, T_2, T_3—proceed with T_1,T_2,T_3	Drop all arms—proceed with T_2,T_3
	$Z_1^{(3)} \leq l_1^{(3)}$	Proceed with T_2—proceed with T_2	Proceed with T_1, T_2—proceed with T_1,T_2	Drop all arms—proceed with T_2
$Z_1^{(2)} \leq l_1^{(2)}$	$Z_1^{(3)} \geq u_1^{(3)}$	Proceed with T_1, T_2, T_3—stop: select T_1,T_3	Proceed with T_1, T_2, T_3—proceed with T_1	Proceed with T_1, T_2, T_3—stop: select T_3
	$l_1^{(3)} < Z_1^{(3)} < u_1^{(3)}$	Stop: select T_1—proceed with T_3	Proceed with T_1—proceed with T_1,T_3	Drop all arms—proceed with T_3
	$Z_1^{(3)} \leq l_1^{(3)}$	Stop: select T_1—stop: select T_1	Proceed with T_1—proceed with T_1	Drop all arms—drop all arms

TABLE A4 Combination of the decision rules at the final analysis for the 4-arm 2-stage ORD—black colour—and MAMS(m)—blue colour—trial.

		$Z_2^{(1)} \geq u_2^{(1)}$	$Z_2^{(1)} < u_2^{(1)}$
$Z_2^{(2)} \geq u_2^{(2)}$	$Z_2^{(3)} \geq u_2^{(3)}$	Select all—select all	Select none—select T_2,T_3
	$Z_2^{(3)} < u_2^{(3)}$	Select T_1, T_2—select T_1,T_2	Select none—select T_2
$Z_2^{(2)} < u_2^{(2)}$	$Z_2^{(3)} \geq u_2^{(3)}$	Select T_1—select T_1,T_3	Select none—select T_3
	$Z_2^{(3)} < u_2^{(3)}$	Select T_1—select T_1	Select none—select none

TABLE A5 Combination of the decision rules at the interim (top panel) and at the final (bottom panel) analyses for the 3-arm 2-stage ORD only considering futility bounds.

	$Z_1^{(1)} \geq l_1^{(1)}$	$Z_1^{(1)} < l_1^{(1)}$
$Z_1^{(2)} \geq l_1^{(2)}$	Proceed with T_1, T_2	Proceed with T_1, T_2
$Z_1^{(2)} < l_1^{(2)}$	Proceed with T_1	Stop trial
	$Z_2^{(1)} \geq u_2^{(1)}$	$Z_2^{(1)} < u_2^{(1)}$
$Z_2^{(2)} \geq u_2^{(2)}$	Select T_1, T_2	Select none
$Z_2^{(2)} < u_2^{(2)}$	Select T_1	Select none

TABLE A6 Combination of the decision rules at the interim (top panel) and at the final (bottom panel) analyses for the 4-arm 2-stage ORD with only futility bounds.

		$Z_1^{(1)} \geq l_1^{(1)}$	$Z_1^{(1)} < l_1^{(1)}$
$Z_1^{(2)} \geq l_1^{(2)}$	$Z_1^{(3)} \geq l_1^{(3)}$	Proceed with all	Proceed with all
	$Z_1^{(3)} < l_1^{(3)}$	Proceed with T_1, T_2	Proceed with T_1, T_2
$Z_1^{(2)} < l_1^{(2)}$	$Z_1^{(3)} \geq l_1^{(3)}$	Proceed with all	Proceed with all
	$Z_1^{(3)} < l_1^{(3)}$	Proceed with T_1	Stop trial
		$Z_2^{(1)} \geq u_2^{(1)}$	$Z_2^{(1)} < u_2^{(1)}$
$Z_2^{(2)} \geq u_2^{(2)}$	$Z_2^{(3)} \geq u_2^{(3)}$	Select all	Select none
	$Z_2^{(3)} < u_2^{(3)}$	Select T_1, T_2	Select none
$Z_2^{(2)} < u_2^{(2)}$	$Z_2^{(3)} \geq u_2^{(3)}$	Select T_1	Select none
	$Z_2^{(3)} < u_2^{(3)}$	Select T_1	Select none