Biostatistics (2016), **0**, 0, *pp.* 1–31
doi:10.1093/biostatistics/Ventz˙et˙al˙Adding˙Experimental˙Arms˙to˙Ongoing˙Clinical˙Trials

Adding Experimental Arms to Ongoing Clinical Trials

Steffen Ventz[1,2,4], Matteo Cellamare[1,3,4], Giovanni Parmigiani[1,4] and Lorenzo Trippa[1,4]

[1]*Dana-Farber Cancer Institute, Boston, US*
[2]*University of Rhode Island, Kingston, US*
[3]*Sapienza University of Rome, Rome, Italy*
[4]*Harvard School of Public Health, Boston, US*

steffen@dfci.harvard.edu

Summary

Multi-arm clinical trials use a single control arm to evaluate multiple experimental treatments. In most cases this feature makes multi-arm studies considerably more efficient than two-arm studies that evaluate single experimental treatments. A bottleneck for implementation is the requirement that experimental treatments need to be available at the enrollment of the first patient. New drugs are rarely at the same stage of development. Moreover multi-arm designs may delay the clinical evaluation of new treatments. These limitations motivate our study of statistical methods for adding new experimental arms after a clinical trial started enrolling patients. We consider both balanced and outcome-adaptive randomization for experimental designs that allow investigators to add new arms, discuss their application in a tuberculosis trial, and evaluate the proposed experimental designs using a set of realistic simulation scenarios. Our comparisons include two-arm randomized designs, multi-arm studies and the proposed class of designs with new experimental arms added at different time points to the clinical study.

Key words: Multi-arm trials, platform trial, Bayesian designs, outcome-adaptive randomization.

1. Introduction

Multi-arm studies that test several experimental treatments against a standard of care are substantially more efficient compared to separate two-arm studies, one for each experimental treatment, where patients are randomized to a control arm and a single experimental treatment. This efficiency gain is substantial and has been discussed by various authors (Freidlin *and others*, 2008; Wason *and others*, 2014). Multi-arm studies test experimental treatments against a common control arm, whereas when experimental drugs are evaluated using two-arm studies the control arm is replicated in each study. This difference reduces the overall sample size requirement for multi-arm studies compared to two-arm trials.

The use of response-adaptive assignment algorithms can further strengthen the efficiency gain of multi-arm studies relative to two-arm studies (Berry *and others*, 2010; Trippa *and others*, 2012). As the trial progresses, adaptive algorithms typically increase the randomization probabilities towards the most promising treatments. On average, this translates into higher sample sizes for the arms with positive treatment effects and, in turn into higher power of detecting the best treatments at completion of the study. The randomization probability can also be tailored to the patient profile, which is defined by biomarkers and/or other relevant characteristics, thus reflecting the possibility of treatment-biomarker interactions suggested by the available data (Kim *and others*, 2011; Alexander *and others*, 2013).

Multi-arm studies also tend to reduce fixed costs compared to two-arm trials. The design and planning of a study is a time consuming and costly process that involves a number of clinicians and a variety of investigators from different fields. Multi-arm studies often have the potential to reduce the resources needed to evaluate experimental drugs compared to independent two-arm studies. Based on these arguments, regulatory agencies encourage the use of multi-arm studies (FDA, 2013; Freidlin *and others*, 2008).

Nonetheless, multi-arm studies constitute only a small fraction of the ongoing early stage

clinical studies, both for cancer as well as for several other diseases. A major bottleneck in their implementation is the requirement that all therapies, often drugs from different pharmaceutical companies, have to be available for testing at the time when the clinical trial starts. Experimental drugs are rarely at the same stage of development. During the design period, before the study starts, there are several candidate drugs with promising preclinical or clinical data. But often some of these drugs are not available when the trial starts recruiting patients, due to logistical reasons, investigators' concerns, or because the pharmaceutical company decides to wait for results from other studies, for example from a clinical trial for a different disease. Additionally, it is not uncommon to encounter holdups in the supply chain. Investigators often face a choice between delaying the activation of the trial or starting with a suboptimal subset of the drugs.

Here we consider the design of multi-arm trials wherein new experimental treatments are added at one or multiple time points. Our work is motivated by the endTB trial, a Bayesian response-adaptive Phase III study in tuberculosis that we designed (Cellamare *and others*, 2016). The study originally aimed to evaluate 8 experimental treatments - but while designing the trial, it became clear that 4 drugs would not be available for the initial 12 months of the study or longer. Because there are an increasing number of experimental agents that need to be tested, similar examples exist in several other diseases areas. Recent cancer studies (STAMPEDE, AML15 and AML16), the neurology trial NET-PD, and the schizophrenia study CATIE, to name a few, added or considered adding, experimental drugs to ongoing studies (Hills and Burnett, 2011; Lieberman *and others*, 2005; Burnett *and others*, 2013; Elm *and others*, 2012). Similarly, the pioneering breast cancer trial I-SPY2 (Barker *and others*, 2009) adds and removes arms within a Bayesian randomized trial design, to accelerate the drug development process.

Nonetheless, statistical studies of designs that allows the addition of arms to an ongoing trial are limited. A recent literature review of Cohen *and others* (2015) on trial designs that involved the addition of experimental arms concluded that the statistical approaches remain mostly ad

hoc. Few guidelines are available for controlling and optimizing the operating characteristics of these types of studies, and the criteria for evaluating the designs are unclear. Recent contributions that consider the amendment of one additional arm into an ongoing study and platform designs include Elm *and others* (2012), Hobbs *and others* (2016) and Berry *and others* (2015).

We focus here on randomization procedures and designs for trials, during which new experimental arms are added. We compare three randomization methods. The first one is a version of the standard balanced randomized (BR) design. In this case the arm-specific accrual rate varies over time, with the number of treatments available for randomization. We show that the algorithm yields efficiency gains compared to separate two-arm studies. The other two methods each use data generated by the ongoing study to vary the randomization probabilities adaptively during the trial. One of the algorithms has close similarities with Bayesian adaptive randomization (BAR) (Thall and Wathen, 2007; Lee *and others*, 2010), while the other shares similarities with the doubly adaptive biased coin design (DBCD) (Eisele, 1994; Rosenberger *and others*, 2001). In all three cases the relevant difference between the designs that we consider and BR, BAR and DBCD is the possibility of adding new experimental arms to an ongoing trial.

Moreover, we introduce a bootstrap algorithm to test efficacy for the initial and added treatments under the adaptive sampling design and early stopping for futility and efficacy. The algorithm extends previously introduced bootstrap schemes by Rosenberger and Hu (1999) and Trippa *and others* (2012), and estimates sequentially stopping boundaries corresponding to a discrete pre-specified Type-I error spending function. The randomization algorithms, the bootstrap procedure used in our simulation study are included in an open-source R package.

After introducing some notations, we describe in Sections 2.1, 2.2 and 2.3 the three designs for balanced and outcome-adaptive multi-arm trials during which experimental arms can be added. In Section 3 the proposed randomization procedures are combined with early stopping rules and an effective bootstrap algorithm for testing treatment efficacy. We then evaluate the proposed

designs in Section 4 in a simulation study. In Section 5 we compare the performances of the three designs under scenarios tailored to the endTB trial. We conclude the paper in Section 6, with a discussion of the proposed procedures.

2. Adding arms to an ongoing trial

We consider a clinical trial that initially randomizes n_1 patients to either the control arm or A_1 experimental arms. For each patient i, $C_i = a$ indicates the randomization to arm $a = 0, \ldots, A_1$, where $a = 0$ indicates the control arm. In what follows $N'_a(i)$ counts the number of patients randomized to arm a before the i-th arriving patients, while $N_a(i) \leqslant N'_a(i)$ is the number of observed outcomes for arm a before the i-th enrollment. Different values of $N_a(i)$ and $N'_a(i)$ are typically due to a necessary period of time from randomization before the patients' outcome can be measured. We consider binary outcomes, and the random variable $Y_a(i)$ counts the number of observed positive outcomes before the arrival of patient i. It has a binomial distribution with size $N_a(i)$ and response probability θ_a. The available data at the i-th accrual is denoted by $D_i = \left\{\left(N'_a(i), N_a(i), Y_a(i)\right)\right\}_{a \geqslant 0}$. The goal is to test treatment efficacy. The null hypotheses are $H_a : \theta_a \leqslant \theta_0$, one null hypothesis for each arm $a > 0$.

We consider a design where experimental arms are added at K different times points. At the arrival of the M_k-th patient, $k = 2, \cdots, K$, A_k experimental arms are added to the study, and the sample size of the study is increased by an additional n_k patients, so that the final sample size becomes $n = \sum_{k=1}^{K} n_k$. In most cases $K \leqslant 3$ and only one arm will be added $A_k = 1$. But we do not assume that the number of adding times K and $(M_k, A_k), 1 < k \leqslant K$, are known in advance when the study is designed, and we will therefore treat both as random variables.

2.1 *Balanced randomization*

A non-adaptive randomization algorithm assigns patients to the control and the experimental arms with a ratio q_0/q_1. The overall sample size is $n_1 = n_C + A_1 \times n_E$, and the number of patients treated with the control and each experimental arm, n_C and n_E, are selected based on targeted type I/II error probabilities. For the moment we do not consider early stopping.

We describe a randomization scheme for adding new treatments. We first focus on the case of $K = 2$ and define the indicator $I\{N'_a(i) < n_E\}$, which is one if $N'_a(i) < n_E$ and zero otherwise. The first $M_2 - 1$ patients are randomized to the control arm or arms $a = 1, \ldots, A_1$, with probabilities proportional to $q_0 I\{N'_0(i) < n_C\}$ and $q_1 I\{N'_a(i) < n_E\}$. At the M_2-th arrival, the arms $A_1 + 1, \cdots, A_1 + A_2$ are added, and the sample size is extended by $n_2 = n_{C,2} + n_E A_2$ patients, $n_{C,2} \geqslant 0$ patients for the control and n_E for each added arm. Patients $i = M_2, \cdots, n_1 + n_2$ are randomized to the initial arms $a = 0, \ldots, A_1$ or the new arms $a = A_1 + 1, \cdots, A_1 + A_2$, with probability

$$p[C_i = a|D_i] \propto \begin{cases} q_0 \times I\{N'_0(i) < n_C + n_{C,2}\} & \text{if } a = 0, \\ q_1 \times I\{N'_a(i) < n_E\} & \text{if } 0 < a \leqslant A_1, \\ q_2 \times I\{N'_a(i) < n_E\} & \text{if } A_1 < a \leqslant A_1 + A_2. \end{cases} \tag{2.1}$$

At the completion of the study n_E patients have been assigned to the experimental arm $a > 0$ and $n_C + n_{C,2}$ patients to the control. In early phase trials one can potentially set $n_{C,2} = 0$ and use the control data from patients randomized before and after the M_2-th enrollment to evaluate the new experimental arms $a = A_1 + 1, \ldots, A_2$. On the other hand, an additional $n_{C,2} > 0$ patients for the control arm may be necessary for longer trials or trials with slow accrual and potential drifts in the population. The parameter q_2 modulates the enrollment rate to the new arms $a = A_1 + 1, \ldots, A_1 + A_2$ after these have been added to the trial. The choice of q_2 should depend on (q_0, q_1), M_2 and A_2. For example, with q_2 equal to $Q_2 = (q_0 + q_1 A_1)/((n_1 + n_2 - M_2 + 1)/n_E - A_2)$, and $n_{C,2} = 0$, all arms complete accrual at approximately the same time (see Figure 1).

The general case with $K \geqslant 2$ is similar. At the enrollment of the M_k-th patient, A_k new arms are added. The sample size is increased by $n_k = n_{C,k} + A_k n_E$ patients, $n_{C,k} \geqslant 0$ patients for the

control, and $A_k n_E$ for the new arms. Let $\mathcal{A}_k$ be the k-th group of treatments, where $\mathcal{A}_1$ is the set of initial experimental arms. Patient $M_k \leqslant i < M_{k+1}$ is assigned to an active arm a, with probability

$$p[C_i = a|D_i] \propto \begin{cases} q_0 \times I\{N_0'(i) < n_C + \sum_{\ell=1}^{k} n_{C,\ell}\} & \text{if } a = 0, \\ q_1 \times I\{N_a'(i) < n_E\} & \text{if } a \in \mathcal{A}_1, \\ \cdots & \\ q_k \times I\{N_a'(i) < n_E\} & \text{if } a \in \mathcal{A}_k. \end{cases} \tag{2.2}$$

As before, the parameters $q_k, 1 < k \leqslant K$, control how quickly each group of arms $\mathcal{A}_k$ enrolls patients, compared to the previous ones. For example, with $n_{c,k} = 0$ and q_k equal to

$$Q_k = \frac{q_0 + \sum_{j=1}^{k-1} A_j Q_j}{(\sum_{j=1}^{k} n_j - M_k + 1)/n_E - A_k}$$

for $k = 1, \cdots, K$, all arms complete accrual at approximately the same time.

The step function $I\{N_a'(i) \leqslant n_E\}$ leads to a randomization scheme, where the assignment of the last patient(s) enrolled in the trial can be predicted. Alternatively one can replace the indicator by a smooth and monotone function.

Example 2.1. Consider 4 experimental treatments and a control arm with response probability of $\theta_0 = 0.3$ at 8 weeks from randomization. A multi-arm trial ($q_0 = q_1 = 1$) with targeted Type I/II error probabilities at 0.1 and 0.2 requires an overall sample size of 265 patients to detect treatment effects of $\theta_a - \theta_0 = 0.2$, with $n_C = n_E = 53$. For an accrual rate of 6 patients per month, the trial duration is approximately 45 months. We can now introduce a departure from this setting. Two treatments $a = 3, 4$ will be available with a delay of approximately 12 and 24 months ($M_2 = 72$, $M_3 = 144$) respectively. We describe three designs. The first one uses all results from the control arm available at the completion of the study to evaluate arms $a = 1, \ldots, 4$. In this case, $n_{C,k} = 0$ and $q_k = Q_k$ for $k = 2, 3$. To avoid bias from possible population trends, the second design estimates the treatment effects of $a = 1, \ldots, 4$ by only using concurrent control outcomes from patients that are randomized during the time window with positive accrual rate for arm a. In this case, to maintain a power of 80% for the added arms, and to keep the accrual

ratios $q_a/q_0 = 1$ constant during the active accrual period of each treatment $a = 1, \ldots, 4$, we set $n_{C,k} = N_0'(M_k)$ at the M_k-th arrival. We also consider a third strategy with three independent trials; one for $a = 1, 2$ and two additional two-arm trials for $a = 3$ and $a = 4$ each study with its own control arm. We assume again an average enrollment of 6 patients per month. The first design requires 265 patients, and the estimates of treatment effects become available 45 months after the first enrollment. The second design requires on average 307 patients, and for $q_k = 1$ 37, 47 and 53 months after the first enrollment. The three independent trials would instead require 371 patients and the treatment effect estimates would be available 46, 60 and 64 months after the first patient is randomized.

2.2 *Bayesian Adaptive Randomization*

Bayesian adaptive randomization (BAR) uses the available data during the trial to assign patients to arms with varying randomization probabilities (Thall and Wathen, 2007; Lee *and others*, 2010). Initially BAR may randomize patients with equal probabilities to each treatment arm. As the trial progresses, information on efficacy becomes available, and randomization favors the most promising treatments. This characteristic can translate into higher power compared to balanced designs (Wason and Trippa, 2014).

We complete the model in the previous paragraphs with a prior $\theta_a \sim p[\theta_a|\nu]$ for θ_a, $a \geqslant 0$. We use a conjugated beta distribution with parameters $\nu = (\nu_1, \nu_2)$. To predict response probabilities of new arms in $\mathcal{A}_k$, even when no outcome data are available for treatments in $\mathcal{A}_k$, we leverage hierarchical modeling with a hyper-prior $\nu \sim p(\nu)$. We use a discrete uniform distribution $p(\nu)$ over a grid of possible ν values.

When we do not add arms, $K = 1$, BAR assigns patient i to arm a with probability

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h(i)} & \text{if } a \in \mathcal{A}_1, \\ c(i) \exp\left\{b \times \left[\max_{a>0} N_a'(i) - N_0'(i)\right]\right\} & \text{if } a = 0, \end{cases} \tag{2.3}$$

where $b > 0$, $c(i) = \sum_{a=1}^{A_1} p[\theta_a > \theta_0|D_i]^{h(i)}$ and $h(\cdot)$ is increasing in the number of enrolled

patients. The function $h(\cdot)$ is used to control the trade-off between exploration and exploitation (Thall and Wathen, 2007). Initially $h(\cdot)$ equals zero and randomization is balanced. Subsequently, as more information becomes available, $h(\cdot)$ increases, and more patients are randomized to the most promising arms. The randomization probability of the control arm in (2.3) is defined to approximately match the sample size of the control and the most promising treatment. This characteristic preserves the power of the adaptive design (Trippa *and others*, 2012).

We modify BAR to allow the addition of new arms. We first consider $K = 2$. At the M_2-th arrival, A_2 new arms are added; and the sample size is increased by n_2 patients. The randomization probabilities are

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h_1(i)} \times q_1(i) & \text{if } a \in \mathcal{A}_1, \\ p[\theta_a > \theta_0|D_i]^{h_2(i)} \times q_2(i) & \text{if } a \in \mathcal{A}_2 \text{ and } i \geqslant M_2, \\ c(i) \exp\left\{b \times \left[\max_{a>0} N'_a(i) - N'_0(i)\right]\right\} & \text{if } a = 0, \end{cases} \tag{2.4}$$

where $c(i) = \sum_{k=1,2;a\in\mathcal{A}_k} I\{M_k \leqslant i\} p[\theta_a > \theta_0|D_i]^{h_k(i)} \times q_k(i)$. We introduce group-specific scaling $q_k(i)$, and power functions $h_k(i)$. The power function $h_k(i)$ control the exploration-exploitation trade-off within each group $\mathcal{A}_k$. The scaling function $q_k(i)$ has two purposes: i) It introduces an initial exploration advantage for newly added experimental treatments, which compete for patient accrual with all open arms. ii) It ensures sufficient exploration of all treatment groups $\mathcal{A}_k$. Several functions serve both purposes. We use a Gompertz function

$$q_k(i) = r_0 + r_1 \exp\left\{ - \exp\left(N'^{(k)}(i) - m_k\right)\right\}, \tag{2.5}$$

where $N'^{(k)}(i)$ is the number of patients randomized to the group of experimental arms $\mathcal{A}_k$ and $m_k, r_1, r_0 > 0$. The function has an initial plateau at $r_0 + r_1$, followed by a subsequent lower plateau at r_0. The initial plateau provides $\mathcal{A}_k$ the necessary exploration advantage when the number of patients randomized to $\mathcal{A}_k$ is small, i.e. $N'^{(k)}(i) < m_k$. During the later stage of the trial once a sufficient number of patients is assigned to treatments in $\mathcal{A}_k$, i.e. $N'^{(k)} > m_k$, the scaling function $q_k(i) \approx r_0$ reaches the lower plateau - and arms in $\mathcal{A}_k$ are assigned following approximately standard BAR (Thall and Wathen, 2007; Lee *and others*, 2010).

In our work, we noted that limiting the maximum number of patients per arm can avoid extremely unbalanced allocations. This may be achieved, for example, by multiplying the Gompertz function in (2.5) by the indicator $I\{N'_a(i) < n'_E\}$, where $n'_E > 0$ represents the desired maximum number of patients in each experimental arm.

We use a function $h_1(\cdot)$ that is monotone in the number of patients randomized to an arm in $\mathcal{A}_1$, and is equal to H for $N'^{(1)} \geqslant n_1$. Similarly, for the added arms in $\mathcal{A}_2$, $h_2(\cdot)$ is monotone in the number of patients randomized to $\mathcal{A}_2$, with a maximum H at n_2. In particular $h_k(i) = H \times [N'^{(k)}(i)/n_k]^\gamma$ if $N'^{(k)}(i) \leqslant n_k$ and H otherwise.

The general case $K \geqslant 2$ is similar. Each patient i is randomized to the available treatments with probabilities

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h_k(i)} \times q_k(i) & \text{if } a \in \mathcal{A}_k \text{ and } M_k \leqslant i, \\ c(i) \exp\left\{b \times \left[\max_a N'_a(i) - N'_0(i)\right]\right\} & \text{if } a = 0. \end{cases} \tag{2.6}$$

where $c(i)$ is defined as in (2.4), and $q_k(i)$ is the Gompertz function defined in (2.5). For $K = 1$ the scheme reduces to standard BAR. The parameter of the scaling function $q_k(i)$ can be selected at the M_k-th arrival such that the expected number of patients assigned to each arm in $a \in \mathcal{A}_k$ under a selected scenario equals a fixed predefined value.

Example 2.2. Consider the setting of example 2.1 and a BAR design instead. To simplify comparison, we set the overall sample size to $n = 265$. We can easily verify that if $H = b = 0$, $q_k(i) = 1$, and $n'_E = 53$, the BAR and BR designs with $q_k = 1$ are identical. We now describe the major operating characteristics under three scenarios. In scenarios 1 to 3, only arm $a = 1$, $a = 3$ (added at $M_2 = 72$) or $a = 4$ (added at $M_3 = 144$) has a positive treatment effect $(\theta_a, \theta_0) = (0.5, 0.3)$. In each scenario, the remaining 4 of the 5 arms, including the control, have identical response rates. We tuned the parameters of the adaptive design to maximize power for $K = 1$ and $A_1 = 2$ under the assumption that there is a single effective arm, $(H, \gamma, b) = (3, 1.5, 0.5)$. The tuning parameter for the Gompertz function $(r_0 = 1, r_1 = 3)$ and $(m_1, m_2, m_3) = (20, 30, 45)$

are selected through simulations, to get approximately the same average sample size when all response rates equal 0.3.

In all three scenarios, the trial completes accrual after approximately 45 months, as for BR. In scenario 1, BAR randomizes on average 64 patients to arm 1 and the control across 5000 simulation, while on average (43,46,47) patients are assigned to the ineffective arms 2,3,4 (standard deviations (SDs) of 4.7, 6.4, 7.4, 5.3 and 5.4) (see table 7). The power increases to 85% - compared to 80% for the BR design - with identical overall number of enrolled patients. In scenarios 2 and 3, BAR randomizes on average 64 and 63 patients to arm $a = 3$ and $a = 4$, respectively (SD 5.8 and 5.5). This translates into 86% and 85% power for the added arms 3 and 4, compared to 80% for BR.

2.3 *Doubly Adaptive Coin Design*

The doubly adaptive coin design (DBCD) (Eisele, 1994) is a response adaptive randomization scheme, which assigns patients approximately according to the target proportions $\rho_a = \rho_a(\theta), a \geqslant 1$ that depend on $\theta = \{\theta_a\}_a$. Examples include the Neyman allocation $\rho_a \propto (\theta_a(1-\theta_a))^{1/2}$, and $\rho_a \propto \theta_a^{1/2}$ (Rosenberger *and others*, 2001). Since the response probabilities θ_a are unknown, the target allocation is estimated by $\widehat{\rho}_a(i)$. For $K = 1$, patients are randomized to arm $a = 0, \ldots, A_1$ with probabilities

$$p[C_i = a|D_i] \propto \widehat{\rho}_a(i) \times q_a(i). \tag{2.7}$$

Here $q_a(i) = \left(\widehat{\rho}_a(i) \times i/(N'_a(i)+1)\right)^h$ varies with the ratio of (i) the estimated target allocation proportion and (ii) the current proportion of patients randomized to arm a (Hu and Zhang, 2004). If the current proportion of patients assigned to arm a is smaller than the target, then for the next patient, the randomization probability to arm a will be larger than $\widehat{\rho}_a(i)$ and vice versa. Larger values of h yield stronger corrections towards the target.

We now consider new experimental arms added during the study. Until the M_2-th arriving

patient, the target $\{\rho_a;\ a = 0, \dots A_1\}$ is a function of $\{\theta_a;\ a = 0, \dots A_1\}$, and it is estimated through the hierarchical Bayesian model in subsection 2.2 by $\widehat{\rho}_a(i) = E[\rho_a(\theta)|D_i]$. Patient $i < M_2$ is randomized to the control or experimental arm $a \in \mathcal{A}_1$, with probabilities defined by (2.7). Then, at the enrollment of the M_k-th patient, $k \geqslant 2$, A_k arms are added and the overall sample size is increased by n_k. Before observing any outcome under $a \in \mathcal{A}_k$, the target is re-defined to $\rho_a(\theta)$, with $\theta = \{\theta_a; 0 \leqslant a \leqslant A_1 + \ldots + A_k\}$ and the posterior distribution of the hierarchical model is used to compute $\widehat{\rho}_a(i) = E[\rho_a(\theta)|D_i]$ for all $0 \leqslant a \leqslant A_1 + \ldots + A_k$. Also in this case, the function $q_a(i)$ is used to approximately match the patient allocation to arm a with the estimated target $\hat{\rho}_a(\theta)$. Each patient $i \geqslant 1$ is randomized to the control arm $a = 0$, or to treatments $a \in \mathcal{A}_k$ in groups added before the i-th arrival with probability

$$p(C_i = a|D_i) \propto \widehat{\rho}_a(i) \times q_a(i). \tag{2.8}$$

For treatments in $\mathcal{A}_k$, $1 \leqslant k \leqslant K$, the functions $q_a(i) = \left[\widehat{\rho}_a(i)i/(N'_a(i) + 1)\right]^{h_k(i)}$ correct the current allocation proportions towards the estimated target.

To avoid extremely unbalanced randomization probabilities, we can replace $\widehat{\rho}_a(i) \times q_a(i)$ in expression (2.8) with $\max(\widehat{\rho}_a(i) \times q_a(i), w(i))$, where $w(i)$ is a function of the data D_i. We used $w(i) \propto (1 + \sum_{k;a\in\mathcal{A}_k} I\{M_k \leqslant i, N'_a(i) < n'_E\})^{-1}$, a decreasing function of the number of active arms. Also for the DBCD design, the function $h_k(\cdot)$ increases during time, for $i \geqslant M_k$ the function is $h_k(i) = h_k + H \times (N^{'(k)}(i)/n_k)^\gamma$ if $N^{'(k)}(i) < n_k$ and $h_k + H$ otherwise. The interpretations of the functions $h_k(i)$ in the DBCD and BAR designs are different, and in our simulation studies the parameters are tuned separately for these trial designs. Similar to BAR, we limit the maximum number of patients per arm by multiplying the correction $q_a(i)$ by the indicator $I\{N'_a(i) < n'_E\}$.

Example 2.3. We consider again the setting in examples 2.1 and 2.2, and use a DBCD design for the trial. Following (Rosenberger *and others*, 2001; Tymofyeyev *and others*, 2007) we use the target allocation $\rho_a(\theta) \propto \theta_a^{0.5}$ for $a > 0$. To preserve the power of the design, similarly to example 2.2, we use $\rho_0(\theta) = \max_{a>0} \theta_a^{0.5}$ to approximately match the sample size of the control

and the most promising experimental arm. For comparison to examples 2.1 and 2.2 we use again an overall sample size of 265. If the response probabilities for all arms are 0.3, a DBCD with $(H, \gamma) = (3, 1)$ and $(h_1, h_2, h_3) = (0, 4, 5)$ randomizes on average 50 patients to each experimental arm, and 54 to the control (SD 3.3, 4.7, 4.7, 4.7 and 4.4). We consider again the same three scenarios as in example 2.1 and 2.2. Either arm $a = 1$, the first added arm $a = 3$, or the second added arm $a = 4$, have a positive treatment effect.

In all 3 scenarios, the trial closes after approximately 45 months - as for BR and BAR. In scenario 1, DBCD randomizes on average 59 and 60 patients to arm 1 and the control (the target is 61), and approximately 49 patients to the remaining ineffective arms $a = 2, 3, 4$ (SD 3.9, 4.0, 4.6, 4.4 and 4.2). The power is 82% for arm $a = 1$, while it is 80% and 85% under BR and BAR in examples 2.1 and 2.2, respectively. For scenarios 2 and 3, the DBCD randomizes on average 59 and 58 patients to the effective arms $a = 3$ or $a = 4$ (SD of 3.8 and 3.4). Under scenario 2 the power is 82%, compared to 80% and 86% for BR and BAR, respectively. Similarly in scenario 3, arm 4 DBCD has 82% power compared to BR and BAR with 80% and 85% power. In our simulations, DBCD tends to have a lower variability of patient allocation compared to BAR.

3. Early stopping rules and hypothesis testing

We describe hypothesis testing and early stopping rules. We consider the strategy wherein arm a in $\mathcal{A}_k$ is stopped for futility after the enrollment of the i-th patient if the posterior probability of a treatment effect, under the hierarchical prior in subsection 2.2, falls below the boundary $f_{i,a}$, i.e. $p[\theta_a > \theta_0 | D_i] \leqslant f_{i,a}$. Here $f_{i,a} = f \times \left(N_a(i)/n'_E\right)^g$ increases from 0 to $f \in [0, 1]$ when $N_a(i) = n'_E$, where $n'_E = n_E$ for BR.

If arm $a \in \mathcal{A}_k$ is not stopped for futility, we compute a bootstrap p-value estimate at time τ_a, the time point at which accrual terminates - for example when $N_a(i)$ reaches n'_E, or at the completion of the trial. The bootstrap procedure is similar to algorithms discussed in Rosenberger

and Hu (1999) and in Trippa *and others* (2012). For comparisons, it is useful to use the same testing approach for all 3 randomization schemes described in section 2. We use the statistic T_a, the standardized difference between the estimated response rates of arm $a > 0$ and the control, to test the null $H_a : \theta_a \leqslant \theta_0$ at significance level α. Large values of T_a indicate a large treatment effect. The algorithm estimates the distribution of T_a under the null H_a, and the (possibly adaptive) characteristics of the randomization procedure. If the estimated response probability $\widehat{\theta}_a$ for experimental arm $a > 0$ is smaller than the estimated probability for the control $\widehat{\theta}_0$, we don't reject the null H_a; while if $\widehat{\theta}_a > \widehat{\theta}_0$, we use the following bootstrap procedure (see also Algorithm 1): (i) First, for all arms a' active at or before τ_a, we compute the maximum likelihood estimate (MLE) $\widehat{\theta}_{a'}$. For arm a and the control, we restrict the MLE to $\theta_0 = \theta_a$. (ii) Then the algorithm simulates $c = 1, \cdots, C$ times the study forward until time $\tau_{a,c}$. The definitions of $\tau_{a,c}$ and $T_{a,c}$ are identical to those of τ_a and T_a, and correspond to the c-th simulation. Each simulation starts at the M_k-th enrollment from the available data D_{M_k}, and adds A_j arms at the M_j-th enrollment, for all groups $j \geqslant k$, which were added before $\tau_{a,c}$. Each patient in the bootstrap simulations responds to a' with probability $\widehat{\theta}_{a'}$ and the simulations' accrual rate is identical to the accrual rate of the actual trial. (iii) For each $c = 1, \ldots, C$ we compute $T_{a,c}$, and set $S_{a,c}$ equal to zero if arm a was stopped early, and equal to one otherwise. We then compute an estimated p-value, $\widehat{p}(T_a) = \sum_{c=1}^{C} I\{T_{a,c} \geqslant T_a, S_{a,c} = 1\}/C$.

The algorithm can be modified to include early stopping for efficacy. In this case, there is a connection between Lan and DeMets (1983) α-spending method and our algorithm explained next. We consider J interim analyses (IA), conducted after a pre-specified set of observed outcomes. The Type I error probability α is partitioned into $\sum_1^J \alpha^{(j)} = \alpha$. The algorithm estimates the thresholds t_j, such that, under the null H_a, the probability $p\big(T_a^{(j)} \geqslant t^{(j)}, S_a^{(j)} = 1\big) \approx \alpha^{(j)}$. Here $S_a^{(j)} = 0$ if arm a is stopped before the j-th IA and equals 1 otherwise, while $T_a^{(j)}$ is a summary statistics at the j-th IA identical to T_a. We first describe the procedure assuming that a is the

only experimental arm that involves early stopping for efficacy, and then relax this assumption.

At IA $j = 1, \cdots, J$, we compute $T_a^{(j)}$ and the MLEs of the response probabilities restricted to the null H_a using the available data unless arm a has been previously stopped. The algorithm generates $c = 1, \cdots, C$ simulations that cover the time window from the M_k-th enrollment until the j-th IA. In these simulations, patients respond to treatments accordingly to the MLEs restricted to H_a. We compute $\left(T_{a,c}^{(\ell)}, S_{a,c}^{(\ell)}\right)$ for each simulation, for $\ell \leqslant j$, with definitions identical to those of $\left(S_a^{(\ell)}, T_a^{(\ell)}\right)$ $\ell \leqslant j$ for the actual trial. Then iteratively, we compute $\hat{t}^{(\ell)} = \min\left\{t : \sum_c I\{T_{a,c}^{(\ell)} \geqslant t, S_{a,c}^{(\ell)} = 1\}/C \leqslant \alpha^{(\ell)}\right\}$ for each $\ell \leqslant j$. If the observed statistics $T_a^{(j)}$ is larger than $\hat{t}^{(j)}$, we reject the null H_a at the j-th IA.

Next, we relax the assumption that early stopping for efficacy involves only a single arm. At IA $j = 1, \ldots, J$, the arms evaluated for efficacy vary, and the pre-specified $\alpha_a^{(j)} \in [0, 1]$ that partitions $\alpha = \sum_j \alpha_a^{(j)}$ can vary across arms. The algorithm estimates the thresholds $t_a^{(j)}$ defined by the following target: under the (unknown) combination of response rates $(\theta_0, \ldots, \theta_{a-1}, \theta_0, \theta_{a+1}, \ldots)$, where we replace θ_a with θ_0, the probability of stopping arm a for efficacy at the j-th IA is $\alpha_a^{(j)}$. In what follows, simulations under H_a are generated using the estimates $(\hat{\theta}_0, \ldots, \hat{\theta}_{a-1}, \hat{\theta}_0, \hat{\theta}_{a+1}, \ldots)$. We tested the algorithm for up to 6 IA and 8 arms:

[`1st IA`] We compute the estimates $\hat{\theta}_a$ and the statistics $T_a^{(1)}$ for all arms that enrolled patients before the 1st IA. Then, separately for each of these arms a, we use $c = 1, \cdots C$ simulations under H_a, from the 1st patient until the 1st IA, to approximate $t_a^{(1)}$ by $\hat{t}_a^{(1)} = \min_t \{t : \sum_c I\{T_{a,c}^{(1)} \geqslant t, S_{a,c}^{(1)} = 1\}/C \leqslant \alpha_a^{(1)}\}$. If $T_a^{(1)} \geqslant \hat{t}_a^{(1)}$ and $S_a^{(1)} = 1$, we reject H_a. When new arms are added before the 1st IA, say $\mathcal{A}_2$, then all simulations will include them, starting from the M_2-th randomization, and will generate thresholds $\hat{t}_a^{(1)}$ for $a \in \mathcal{A}_2$.

[`2nd IA`] We recompute $\hat{\theta}$. Then, separately for each arm, we re-estimate $\hat{t}_a^{(1)}$ by using a new set of $c = 1, \cdots C$ simulations under H_a that cover the time window between the 1st patient and the 1st IA. After the $\hat{t}_a^{(1)}$'s have been re-computed, we extend the simulations in time to

cover the window between the 1st and the 2nd IAs. In simulation c, if $T_{a,c}^{(1)} > \hat{t}_a^{(1)}$, then arm a is stopped for efficacy. This part of the algorithm creates, for each arm a evaluated at the 2nd IA, $c = 1, \cdots, C$ simulations under H_a, from the 1st patient until the 2nd IA; importantly, these simulations include early stopping at the 1st IA. We can therefore compute $\hat{t}_a^{(2)} = \min_t \left\{ t : \sum_c I\{T_{a,c}^{(2)} \geqslant t, S_{a,c}^{(2)} = 1\}/C \leqslant \alpha_a^{(2)} \right\}$. Analogously to the 1st IA, if new arms (for example $\mathcal{A}_2$) are added between the 1st and 2nd IAs, then all simulations, starting from the M_2-th randomization, will include the added arms.

The same procedure is iterated, similarly to $j = 2$, for $j = 3, \cdots, J$. In some simulations of the multi-arm study under H_a, where $a \in \mathcal{A}_k$, the arm a might not appear because the experimental arms have been dropped and the trial stopped before M_k enrollments. To account for this simulations under H_a, when $a \in \mathcal{A}_k$, are generated conditional on the event that the multi-arm study enrolls more than M_k patients .

Example 3.1. We continue example 2.1 and apply the bootstrap algorithm to the BR design. The application to BAR and DBCD is similar. The trial starts with 2 initial experimental arms and $n_1 = 159$ patients. Two arms are added, one after 12 months and one after 24 months, $(M_2, M_3) = (72, 144)$, and the overall sample size is 265. An interim efficacy analysis is performed after 100 and 200 observed outcomes, and a final analysis is conducted after all outcomes are observed, so $J = 3$. For the initial set of arms $a = 1, 2$ the Type I error probabilities are $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0.025, 0.025, 0.05)$, whereas $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0, 0.05, 0.05)$ for arms $a = 3, 4$. We consider 4 scenarios: in scenario 1, all arms have identical response rates of 0.3. Scenarios 2 to 4 are identical to those in examples 2.1 and 2.3. In scenario 2, arm 1 has a treatment effect, and in scenarios 3 and 4 the first added arm $a = 3$, or the second added arm $a = 4$ has a treatment effect of 0.5. All ineffective arms have a response rate equal to the control of 0.3. We applied the bootstrap procedure with $C = 10,000$.

For scenario 1, the Type I error across 5000 simulated trials was $0.10, 0.09, 0.11$ and 0.09,

for arms 1 to 4. Arms 1 and 2 were stopped early for futility in 50% of all simulations, whereas ineffective arms 3 and 4 were stopped early for futility in 43% of all simulations. In scenario 2, the initial arm $a = 1$ has 79% power. The probability of rejecting the null H_1 in stages one to three are $0.31, 0.25$ and 0.23, respectively, and the empirical Type I error rates for arms $a = 2, 3, 4$ are $0.11, 0.10$ and 0.10. Similarly in scenarios 3 and 4, the first added arm $a = 3$ and the second added arm $a = 4$ have 79% and 78% power, respectively, with estimated type I error rates of (0.11, 0.10, 0.10) in scenario 3, and (0.10, 0.11, 0.11) in scenario 4, for the remaining 3 ineffective arms. In scenario 3 the probability of rejecting H_3 at stage 2 and 3 are 0.59 and 0.20, whereas in scenario 4 H_4 is rejected at stage 2 or 3 with probability 0.54 and 0.24, respectively. Effective arms in scenarios 2 to 4 were stopped early for futility in less than 2% of all simulations.

4. Simulation Study

Continuing examples 2.1, 2.2 and 2.3, we consider the same four scenarios as in example 3.1. In scenario 1 no experimental arm has an effect, and in scenarios 1 to 3 either the first initial arm, the first added arm $a = 3$, or the second added arm $a = 4$ have an effect of 0.5, and all other ineffective arms have response rates equal to the control rate of 0.3. The initial and overall sample sizes are 159 and 265 patients, respectively, and the Type I error is controlled at 10%. For BAR and DBCD, the maximum number per arm is $n'_E = 69 \approx 1.3 \times n_E$, namely BAR and DBCD can assign at most 69 patients to each experimental arm. As explained above, for DBCD, the minimum randomization probability to each active arm was restricted to values larger than one over three times the number of active arms.

We first summarize the performance of the three designs without early stopping to illustrate the characteristics of the randomization schemes and compare the designs to the current practice where the investigators conduct three independent trials; one trial for the initial two experimental arms, and two independent two-arm studies for arm $a = 3$ and $a = 4$ with their own control.

The overall rate of accrual of the three concurrent *balanced randomized* and *independent* (BRI) trials is set to 6 patients per month, and is assumed to be identical for the competing designs in Section 2.

For arms $1, 3$ and 4 Figure 7 shows the median number of patients randomized to each arm as a function of the overall number of patients enrolled in the trial. For each scenario and design, the plotted graph represents for a fixed arm a the median number of patients assigned to arm a over 5000 simulated trials (y-axis), after a total of 1 to 265 (371 for BRI) patients have been enrolled to the trial (x-axis). Under BRI, 2×53 additional patients are necessary for the two additional control arms, this prolongs the trials and slows down the accrual to the experimental arms.

Figure 3 shows the variability of treatment assignment at the end of the trial. Under scenario 1, BRI and BR randomize at the end of the study 53 patients to each arm. BRI requires 106 additional patients for the two additional control arms. In scenario 1, DBCD has a median accrual of 52 patients for all the experimental arms with interquartiles (IQ) $(49, 56)$ for arms 1 and 2, and an IQ of $(49, 55)$ for arms 3 and 4. In comparison, using BAR, the median accrual for the first two experimental arms is 49 (IQ: 42, 58), 50 (IQ: 45, 56) for arm 3 and 52 (IQ: 48, 57) for arm 4. In scenario 2, where the first initial arm has a positive effect, BAR and DBCD have a median accrual of 66 (IQ: 61,70) and 59 (IQ: 57,62) patients for this arm, with 85% and 82% power, compared to 80% using BR (Table 7). In scenarios 3, BAR and DBCD have 86% and 82% power of detecting the effect of the first added arm, respectively, compared to 80% under BR (Table 7). The median accrual for the first added arm is 65 (IQ: 60,69) patients for BAR and 59 (IQ: 57, 62) for DBCD. Lastly, in scenario 4 the second added arm has a positive effect. BAR and DBCD assigns a median number of 63 (IQ: 56, 66) and 58 (IQ: 56, 61) patients to this arm, which translates into 85% and 83% power, respectively.

We now compare BR, BAR and DBCD, when early stopping for efficacy and futility are

included as described in Section 3. The tuning parameters of the futility stopping boundaries (f, g) were selected such that the probability of stopping an effective initial arm early for futility is around 1%, $(f, g) = (0.25, 1.5)$ for BR and $(f, g) = (0.2, 1.5)$ for BAR and DBCD. Larger values of g (1 to 2.5) decrease the probability of dropping an arm for futility during the study. As before, the overall Type I error bound α was set to 10%, with error rates of $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0.025, 0.025, 0.05)$ for the initial arms after 100, 200 and 265 observed outcomes, and $\left(\alpha_a^{(1)}, \alpha_a^{(2)}, \alpha_a^{(3)}\right) = (0, 0.05, 0.05)$ for the first and second added arm $a = 3, 4$.

Table 7 shows the average sample size, standard deviation and power for experimental arms $a = 1, 3$ and 4 across 5000 simulated trials. Under scenario 1, BAR and DBCD have a higher average overall sample size than BR, with 260 and 261 patients for BAR and DBCD, compared to 245 for BR. This is expected; once an arm a that enrolled $N'_a(i)$ patients is stopped, the final overall sample size in a BR trial is reduced by $53 - N'_a(i)$, while BAR and DBCD assign these patients to the remaining active arms. The Type I error probabilities across simulations are close to the target of 10%. In scenario 2, BR randomizes on average 52 patients (SD 3) to the superior arm 1, compared to 54 (SD 13.2) for BAR and 60 (SD 4.5) for DBCD. The power under the three designs is 79%, 84% and 81%, with probabilities of rejecting H_1 at interim analyses 1, 2 and 3 equal to $(0.31, 0.25, 0.23)$ for BR, $(0.33, 0.27, 0.24)$ for BAR and $(0.32, 0.25, 0.24)$ for DBCD. In scenario 3, BAR and DBCD have 84% and 81% power, respectively, compared to 79% for BR, with mean accrual of 52 (SD 3), 54 (SD 13) and 61 (SD 4.5) patients for BR, BAR and DBCD. The probability of stopping the effective arm incorrectly for futility is 1.2% for BR compared to $< 1\%$ for BAR and DBCD. BAR and DBCD randomize on average less patients to ineffective experimental arms compared to BR. Lastly in scenario 4, where the second added arm has a positive effect, BR , BAR and DBCD assign on average 52, 59 and 60 patients to this arm (SD 3.9, 6.9 and 4.2) with power of 78%, 84% and 81%, respectively. For BR, the probabilities of rejecting H_4 at the second and third interim analyses are $(0.55, 0.23)$, compared to $(0.58, 0.26)$ for

BAR and $(0.57, 0.24)$ for DBCD. The probability of dropping the second added arm incorrectly for futility was 1.5% for BR and $< 1\%$ for BAR and DBCD.

5. THE endTB TRIAL

Our motivation for adding arms to an ongoing study is the endTB trial in multi-drug resistant Tuberculosis (MD-TB) (Cellamare *and others*, 2016). The trial tests five experimental treatments under a response-adaptive BAR design similar to the one described in section 2.2. We initially designed the trial with 8 experimental arms, but we were later informed that four of the experimental treatments would not be available at the activation of the trial. The investigators asked if the treatments could be added in one or two groups at a later point. Previous trials showed response probabilities of approximately 0.55 after 6 months of treatment with the control therapy. We consider a response probability of 0.7 as clinically relevant increase for experimental treatment. The study expects an accrual rate of 10 patients per month.

We present a simulation with $A_1 = 4$ initial experimental arms, and an initial sample size of $n_1 = 500$ patients. Two groups of $A_2 = A_3 = 2$ arms are added after $M_2 = 200$ and $M_3 = 300$ patients have been enrolled. The overall sample size is increased to 700 and 900 patients at the enrollment of patient $M_2 = 200$ and $M_3 = 300$. The Type I error is controlled at the $\alpha = 5\%$ level.

We consider the four scenarios summarized in Table 3. In scenario 1, all 8 experimental arms are ineffective, with response rates identical to the control. In scenarios 2 and 3, the initial arm $a = 1$ and added arm $a = 5 \in \mathcal{A}_2$ (scenario 2) or $a = 7 \in \mathcal{A}_3$ (scenario 3) are effective, with response rates of 0.7 and 0.75. Lastly, in scenario 4, arm 1 and the added arms 5 and 7 are effective, with response probabilities $0.7, 0.75$ and 0.7.

Table 4 shows the mean number of patients randomized to the control and arms $1, 5$ and 7 across 5000 simulations, together with the standard deviation and the power. Under scenario 1,

BR randomizes on average 98 and 79 patients to the control and the initial arms, and 80 and 82 patients to arms in the 2nd and 3rd group (SD $5.1, 25.4, 25.4$ and 24.4), respectively, compared to (134, 93, 95, 97) for BAR (SD $10.9, 26.3, 23.9$ and 21.4) and (106, 99, 98,98) for DBCD (SD $5, 8, 6.4, 6.3$ and 6.2), respectively. Under scenario 2, BR has 70% and 90% power of detecting the arms with response rates 0.7 and 0.75. BAR and DBCD have 10% and 3% higher power for $a = 1$ (80% and 73%), and 7% and 3% higher power for $a = 5$ (97% and 93%) associated to an increase in the average allocation of 28 (BAR) and 7 (DBCD) patients for $a = 1$, and 32 (BAR) and 8 (DBCD) patients for $a = 5$. In scenario 3, BR randomizes on average 99 (SD 7.4) patients to arm 1, compared to 127 (SD 15.6) for BAR and 106 (SD 5.0) for DBCD, respectively. This translates into a power of 70% for BR, 80% for BAR and 74% for DBCD. For the added arm $a = 7$, BR has 92% power compared to 97% and 93% for BAR and DBCD with mean accrual of 100, 130 and 108 under BR, BAR and DBCD, respectively. Lastly, in scenario 4, where arms 1,5 and 7 are effective with response rates $0.7, 0.75$ and 0.7, BR randomizes an average $(99, 100, 99)$ patients to these arms (SD 8.7, 2.7 and 5.2) with power $70\%, 90\%$ and 70%. In comparison BAR and DBCD randomize on average $(121, 128, 118)$ patients and $(105, 107, 103)$ patients to arms $a = 1, 5, 7$. These gains in mean sample sizes translate into 79%, 96% and 79% power under BAR, and 72%, 93% and 73% under DBCD, respectively.

6. Discussion

Drug development in oncology, infectious diseases and other disease areas focuses increasingly on targeted patient populations defined by biological pathways. Drugs targeting biological pathways are usually at different stages of development, and low accrual rates for rare subpopulations require efficient allocation of patients in clinical studies. Multi-arms studies are strongly encouraged by regulatory institutions, to promote comparisons to the standard of care without redundant replicates of control arms. For example, given that in metastatic breast cancer, hormone re-

ceptor positive patients eventually become resistant to the standard endocrine therapy, several trials with overlapping accrual windows recently explored mTOR and CDK4/6 inhibitors in combination with endocrine therapy (NCT00721409, NCT02246621, NCT02107703, NCT01958021, NCT01958021 and NCT00863655). Adding arms to clinical trials could save resources, and a higher proportion of patients could be treated with promising novel therapies. Sharing an active control arm among multiple experimental treatments reduces the proportion of patients allocated to the control and can simplify the inclusion of an active control arm.

Here we explore three randomization schemes for adding experimental arms to an ongoing study. The designs vary in their level of complexity and in the resources required for their implementation. Adding treatments to a trial under BR can be implemented without a substantial increase in the complexity of the design, and can improve efficiency substantially. BAR and DBCD require simulations for parameter tuning, but can potentially increase the power of the multi-arm study. Sequential stopping rules for BR which target a predefined Type I error can be implemented using a standard error spending function approach. For outcome-adaptive BAR and DBCD designs the Type I error probabilities can be controlled with the proposed bootstrap procedure in section 3.

7. Software

An R package which implements the proposed designs is available at http://bcb.dfci.harvard.edu/~steffen/software.html

Acknowledgments

The authors would like to thank Sonal Jhaveri for valuable suggestions. We also like to thank Carole D. Mitnick, principle investigator for the endTB trial, who initiated the project.

References

Alexander, Brian M, Wen, Patrick Y, Trippa, Lorenzo, Reardon, David A, Yung, Wai-Kwan Alfred, Parmigiani, Giovanni and Berry, Donald A. (2013). Biomarker-based adaptive trials for patients with glioblastoma – lessons from i-spy 2. *Neuro-oncology*, not088.

Barker, AD, Sigman, CC, Kelloff, GJ, Hylton, NM, Berry, DA and Esserman, LJ. (2009). I-spy 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* **86**(1), 97–100.

Berry, Scott M, Carlin, Bradley P, Lee, J Jack and Muller, Peter. (2010). *Bayesian adaptive methods for clinical trials*. CRC press.

Berry, Scott M, Connor, Jason T and Lewis, Roger J. (2015). The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA* **313**(16), 1619–1620.

Burnett, Alan K, Russell, Nigel H, Hills, Robert K, Hunter, Ann E, Kjeldsen, Lars, Yin, John, Gibson, Brenda ES, Wheatley, Keith and Milligan, Donald. (2013). Optimization of chemotherapy for younger patients with acute myeloid leukemia: results of the medical research council aml15 trial. *Journal of Clinical Oncology*, JCO–2012.

Cellamare, Matteo, Milstein, Meredith, Ventz, Steffen, Baudin, Elisabeth, Trippa, Lorenzo and Mitnick, Carole. (2016). Bayesian adaptive randomization in a clinical trial to identify new regimens for multidrug-resistant tuberculosis: the endtb trial. *International Journal of Tuberculosis and Lung Disease* **(in press)**.

Cohen, Dena R, Todd, Susan, Gregory, Walter M and Brown, Julia M. (2015). Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials* **16**, 179.

Eisele, Jeffrey R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference* **38**, 249–261.

Elm, Jordan J, Palesch, Yuko Y, Koch, Gary G, Hinson, Vanessa, Ravina, Bernard and Zhao, Wenle. (2012). Flexible analytical methods for adding a treatment arm mid-study to an ongoing clinical trial. *Journal of biopharmaceutical statistics* **22**(4), 758–772.

FDA. (2013). (US Food and Drug Administration): guidance for industry: codevelopment of two or more new investigational drugs for use in combination.

Freidlin, Boris, Korn, Edward L, Gray, Robert and Martin, Alison. (2008). Multi-arm clinical trials of new agents: some design considerations. *Clinical Cancer Research* **14**, 4368–4371.

Hills, Robert K and Burnett, Alan K. (2011). Applicability of a pick a winner trial design to acute myeloid leukemia. *Blood* **118**(9), 2389–2394.

Hobbs, Brian P, Chen, Nan and Lee, J Jack. (2016). Controlled multi-arm platform design using predictive probability. *Statistical methods in medical research*, 0962280215620696.

Hu, Feifang and Zhang, Li-Xin. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Annals of Statistics*, 268–301.

Kim, Edward S, Herbst, Roy S, Wistuba, Ignacio I, Lee, J Jack, Blumenschein, George R, Tsao, Anne, Stewart, David J, Hicks, Marshall E, Erasmus, Jeremy, Gupta, Sanjay *and others*. (2011). The battle trial: personalizing therapy for lung cancer. *Cancer discovery* **1**(1), 44–53.

Lan, KK Gordon and DeMets, David L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**(3), 659–663.

Lee, J Jack, Gu, Xuemin and Liu, Suyu. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* **7**, 584–596.

Lieberman, Jeffrey A, Stroup, T Scott, McEvoy, Joseph P, Swartz, Marvin S, Rosenheck, Robert A, Perkins, Diana O, Keefe, Richard SE, Davis, Sonia M, Davis, Clarence E, Lebowitz, Barry D *and others*. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *New Engl. Journal of Med* **353**(12), 1209–1223.

Rosenberger, W. F. and Hu, F. (1999). Bootstrap methods for adaptive designs. *Stat Med* **18**, 1757–1767.

Rosenberger, William F, Stallard, Nigel, Ivanova, Anastasia, Harper, Cherice N and Ricks, Michelle L. (2001). Optimal adaptive designs for binary response trials. *Biometrics* **57**, 909–913.

Thall, Peter F and Wathen, J Kyle. (2007). Practical bayesian adaptive randomisation in clinical trials. *European Journal of Cancer* **43**, 859–866.

Trippa, Lorenzo, Lee, Eudocia Q, Wen, Patrick Y, Batchelor, Tracy T, Cloughesy, Timothy, Parmigiani, Giovanni and Alexander, Brian M. (2012). Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *JCO* **30**, 3258–3263.

Tymofyeyev, Yevgen, Rosenberger, William F and Hu, Feifang. (2007). Implementing optimal allocation in sequential binary response experiments. *Journal of the American Statistical Association* **102**(477).

Wason, JM, Stecher, Lynne and Mander, Adrian P. (2014). Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* **15**(1), 364.

Wason, James and Trippa, Lorenzo. (2014). A comparison of bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat. Med* **33**(13), 2206–2221.

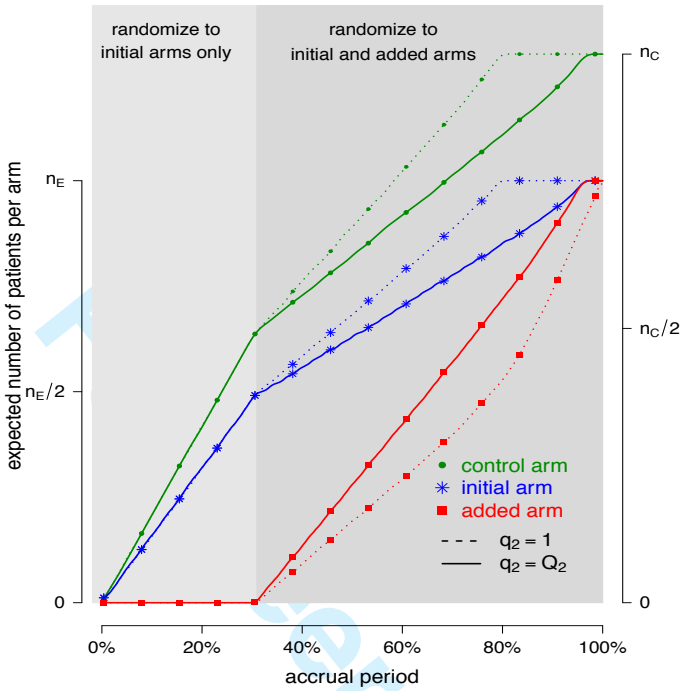

Fig. 1. Adding experimental arms to a multi-arm BR trial. We consider a trial with two initial $A_1 = 2$ and two added $A_2 = 2$ experimental arms. The graph shows the expected number of patients randomized to an arm during the accrual period for the control $a = 0$, one initial arm in $\mathcal{A}_1$ and one added arm in $\mathcal{A}_2$. The two additional arms were added after 50% of the initially planned sample size, at $M_2 = n_1/2$. Patients were initially randomized to the control or experimental arm with ratio $q_0 = 1.25$ to $q_1 = 1$. Dashed lines correspond to $q_2 = 1$. Solid lines correspond to the $q_2 = Q_2$, in this case all arms are expected to complete accrual at the same time.

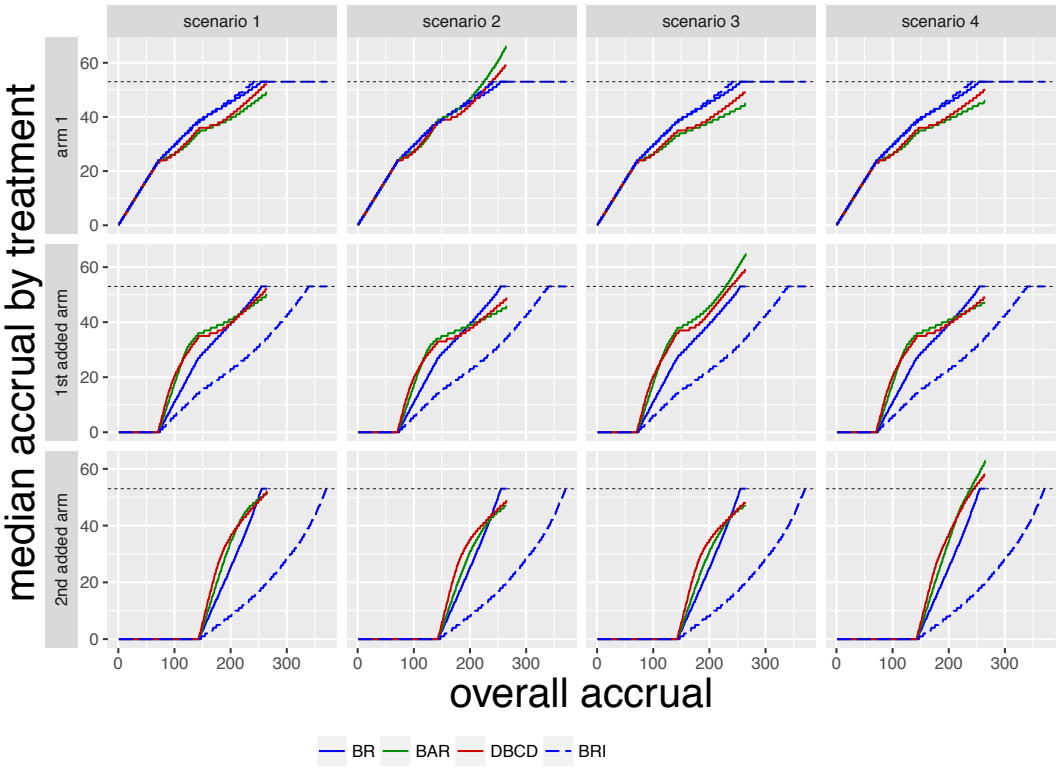

Fig. 2. Number of patients randomized to treatment arms during the accrual period, across 5000 simulations, under BRI, BR, BAR and DBCD for a study with 2 initial experimental arms and two arms that are added after the enrollment of $M_2 = 72$ and $M_3 = 144$ patients. For each arm a, the plotted graph (x, y) represents the median number of patients y assigned to arm a, after a total of x patients have been randomized. In scenario 1 all experimental arms are ineffective, whereas in scenarios 2 to 4 either arm 1, the first or the second added arm have a treatment effect, with a response probability of 0.5 compared to 0.3 for the control.

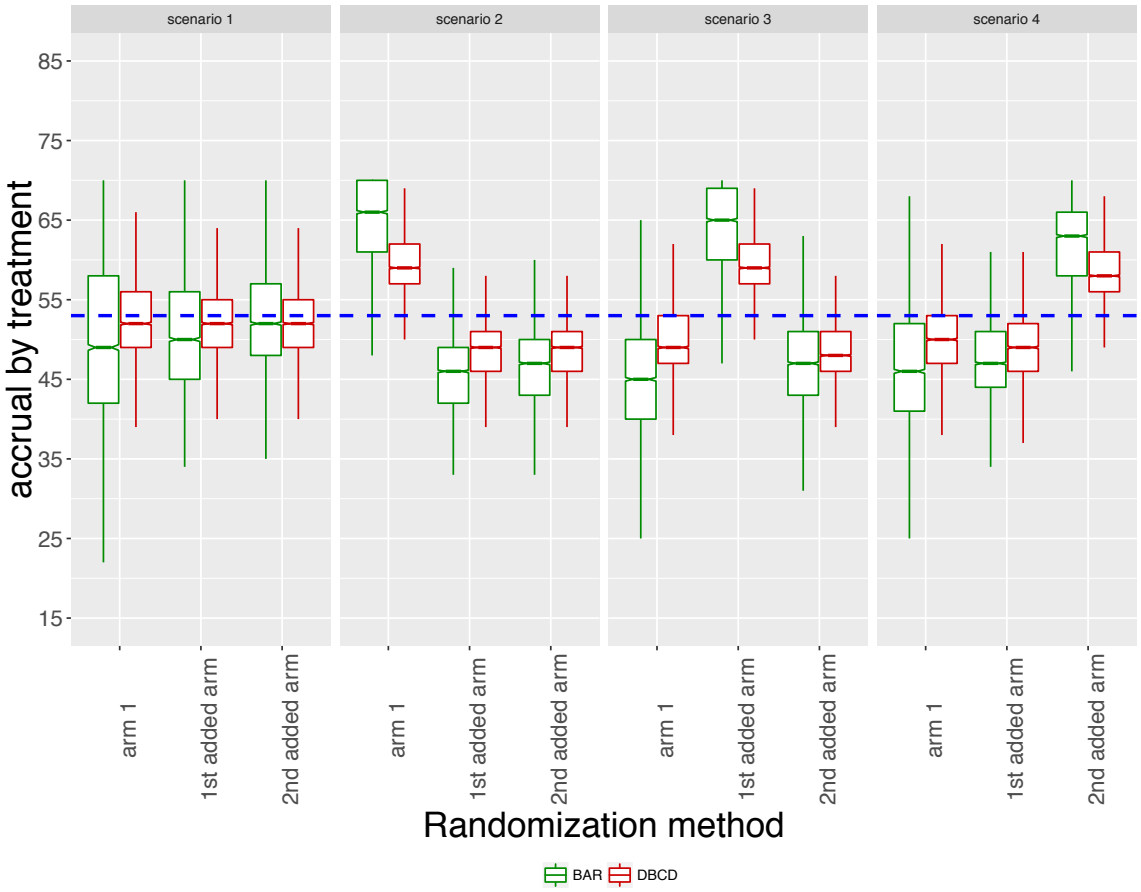

Fig. 3. Boxplots of the number of patients randomized to each treatment arm under BR, BAR and DBCD across 5000 simulations for a trial with 2 initial arms and two arms that are added after the enrollment of $M_2 = 72$ and $M_3 = 144$ patients. The dashed line shows the number of patients randomized to each arm under BR.

scenario	control		arm 1			1st added arm			2nd added arm		
	E	SD	E	SD	Po	E	SD	Po	E	SD	Po
BR 1	53	0.0	53	0.0	0.11	53	0.0	0.10	53	0.0	0.10
2	53	0.0	53	0.0	0.80	53	0.0	0.11	53	0.0	0.11
3	53	0.0	53	0.0	0.11	53	0.0	0.80	53	0.0	0.10
4	53	0.0	53	0.0	0.11	53	0.0	0.11	53	0.0	0.80
BAR 1	62	3.7	50	10.1	0.10	51	7.6	0.10	52	6.9	0.10
2	64	4.7	64	6.4	0.85	46	5.3	0.11	47	5.4	0.10
3	64	4.3	45	7.8	0.10	64	5.8	0.86	47	5.7	0.11
4	62	3.4	46	8.2	0.10	48	5.9	0.11	62	5.5	0.85
DBCD 1	57	3.3	52	4.8	0.10	52	4.6	0.10	52	4.3	0.10
2	60	3.9	59	4.0	0.82	48	4.4	0.10	49	4.2	0.10
3	58	3.6	49	4.6	0.10	59	3.8	0.82	48	4.1	0.10
4	58	3.4	50	4.5	0.09	49	4.3	0.10	58	3.4	0.83

Table 1. Expected sample size (E), standard deviation (SD) and power (Po) for experimental arm 1, the 1st added arm $a = 3$ and the 2nd added arm $a = 4$ for a trial with two initial experimental arms, and two arms which are added after 12 and 24 month, $(M3, M_4) = (72, 144)$. Results are based on 5000 simulated trials under balanced (BR), Bayesian adaptive (BAR) and a doubly adaptive biased coin design (DBCD) without early stopping rules. The initial planned overall sample size is 159, which is then extended by 53 patients for each added arm.

scenario	control		arm 1			1st added arm			2nd added arm		
	E	SD	E	SD	Po	E	SD	Po	E	SD	Po
BR 1	51	3.4	47	9.3	0.10	49	7.0	0.11	51	5.2	0.09
2	51	3.4	52	3.0	0.79	48	8.0	0.10	51	5.7	0.10
3	51	3.2	47	9.7	0.11	52	2.6	0.79	51	5.0	0.10
4	51	3.2	46	9.8	0.10	49	7.4	0.11	52	3.9	0.78
BAR 1	62	5.1	48	12.8	0.10	50	9.4	0.10	52	8.7	0.09
2	65	5.2	54	13.2	0.84	49	9.8	0.10	51	8.8	0.10
3	65	4.9	45	11.2	0.10	62	6.3	0.84	48	7.8	0.09
4	63	4.5	46	11.6	0.11	49	8.3	0.11	59	6.9	0.84
DBCD 1	57	4.4	51	5.9	0.09	51	5.7	0.10	52	5.6	0.09
2	61	4.1	60	4.5	0.81	48	5.1	0.08	49	4.9	0.09
3	59	4.1	48	5.2	0.10	61	4.5	0.81	48	4.8	0.10
4	59	4.0	48	5.4	0.09	49	4.8	0.10	60	4.2	0.81

Table 2. Expected sample size (E), standard deviation (SD) and power (Po) for experimental arm 1, the 1st added arm $a = 3$, and the 2nd added arm $a = 4$, for a trial with two initial experimental arms, and two arms which are added after 12 and 24 months, $(M3, M_4) = (72, 144)$ with futility and efficacy stopping. Two interim analyses for efficacy are planned after 100, 200 patients have been enrolled. Results are based on 5000 simulated trials under balanced (BR), Bayesian adaptive (BAR) and a doubly adaptive biased coin designs (DBCD) . The initial planned sample size is 159, which is then extended by 53 patients for each added arm.

scenario	control	arm 1	arm 5	arm 7
1	0.55	0.55	0.55	0.55
2	0.55	**0.70**	**0.75**	0.55
3	0.55	**0.70**	0.55	**0.75**
4	0.55	**0.70**	**0.75**	**0.70**

Table 3. Simulation scenarios for the endTB trial. The trial starts with 4 experimental arms plus the control, with a planned sample size of $n_1 = 500$ patients. Two arms are then added after $M_2 = 200$ enrolled patients and 2 more arms after $M_3 = 300$ enrolled patients. The overall sample size is extended to 700 patients and subsequentely to 900 patients.

scenario	control		initial arms arm 1			1st added group arm 5			2nd added group arm 7		
	E	SD	E	SD	Po	E	SD	Po	E	SD	Po
BR 1	98	5.1	79	25.3	0.05	80	25.5	0.05	82	24.4	0.05
2	99	3.4	99	8.3	0.70	100	2.9	0.90	82	24.6	0.05
3	99	3.5	99	7.4	0.70	81	25.2	0.05	100	3.1	0.92
4	99	3.4	99	8.7	0.70	100	2.7	0.90	99	5.2	0.70
BAR 1	134	10.9	93	26.3	0.05	95	23.9	0.05	97	21.4	0.05
2	137	8.8	127	15.8	0.80	132	11.8	0.97	88	16.1	0.05
3	137	9.1	127	15.7	0.80	86	18.4	0.05	130	12.5	0.97
4	134	9.3	121	17.1	0.79	128	13.3	0.96	118	16.0	0.79
DBCD 1	106	5.8	99	6.4	0.05	98	6.3	0.05	98	6.2	0.05
2	110	4.7	106	4.9	0.73	108	4.4	0.93	95	4.9	0.05
3	109	4.8	106	5.0	0.73	96	5.2	0.05	108	4.4	0.93
4	109	4.7	105	4.9	0.72	107	4.5	0.93	103	4.5	0.73

Table 4. Expected accrual (E), standard deviation of accrual (SD) and statistical power for initial arm 1, arm 5 (added at $M_2 = 200$), and arm 7 (added at $M_3 = 300$) based on 5000 simulations under balanced randomization (BR), Bayesian adaptive randomization (BAR) or the doubly adaptive biased coin design (DBCD), with an initial planned sample size of $n_0 = 500$ patients and an extension of the overall sample size by 200 patients at time M_2 and M_3.