Biostatistics (2018) **19**, 2, *pp.* 199–215
doi:10.1093/biostatistics/kxx030
Advance Access publication on July 17, 2017

Adding experimental arms to platform clinical trials: randomization procedures and interim analyses

STEFFEN VENTZ*

Department of Computer Science and Statistics, University of Rhode Island, 9 Greenhouse Road, Kingston, RI 02881, USA

ventzer@yahoo.de

MATTEO CELLAMARE

Harvard T. H. Chan School of Public Health, 655 Huntington Ave, Boston, MA 02115, USA, Department of Statistical Sciences, Sapienza University, Piazzale Aldo Moro 5, 00185 Roma, Italy, and Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Ave, Boston, MA, 02115, USA

GIOVANNI PARMIGIANI, LORENZO TRIPPA

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 450 Brookline Ave., Boston, MA 02115, USA and Department of Biostatistics, Harvard T. H. Chan School of Public Health, 655 Huntington Ave, Boston, MA, 02115, USA

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. A bottleneck for implementation of a multi-arm trial is the requirement that all experimental treatments have to be available at the enrollment of the first patient. New drugs are rarely at the same stage of development. These limitations motivate our study of statistical methods for adding new experimental arms after a clinical trial has started enrolling patients. We consider both balanced and outcome-adaptive randomization methods for experimental designs that allow investigators to add new arms, discuss their application in a tuberculosis trial, and evaluate the proposed designs using a set of realistic simulation scenarios. Our comparisons include two-arm studies, multi-arm studies, and the proposed class of designs in which new experimental arms are added to the trial at different time points.

Keywords: Bootstrap; Multi-arm clinical trials; Outcome-adaptive randomization; Platform trials.

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 study for each experimental treatment. Multi-arm studies test experimental treatments against a common control arm, whereas when experimental drugs

*To whom correspondence should be addressed.

are evaluated using two-arm studies the control arm is replicated in each study. This difference reduces the overall sample size for testing multiple experimental drugs in a single multi-arm study compared to using independents two-arm trials. The gain in efficiency is substantial and has been discussed by various authors (Freidlin *and others*, 2008; Wason *and others*, 2014).

The use of response-adaptive assignment algorithms can further strengthen the efficiency gain of multi-arm studies compared to two-arm studies (Berry *and others*, 2010; Trippa *and others*, 2012; Wason and Trippa, 2014; Ventz *and others*, 2017). As the trial progresses, adaptive algorithms typically increase randomization probabilities towards the most promising treatments. On average, this translates into larger 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.

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

Nonetheless, multi-arm studies constitute a small fraction of the ongoing early stage clinical studies. A major bottleneck in their implementation is the requirement that all therapies, often drugs from different pharmaceutical companies, must to be available for testing 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 (e.g. from a clinical trial for a different disease). Additionally, holdups in the supply chain are not uncommon. Investigators thus face a choice between delaying the start of the trial or testing only a subset of 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*, 2017). The study originally sought to evaluate eight experimental treatments. While designing the trial, it became clear that four drugs would have not been available for the initial 12 months of the study or longer. Because of the need to test an increasing number of experimental treatments (Berry *and others*, 2015) similar examples exist in several other disease 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.

Nonetheless, statistical studies of designs that allows the addition of arms to an ongoing trial are limited. A recent literature review of designs that involved the addition of experimental arms Cohen *and others* (2015) concluded that the statistical approaches remain mostly *ad hoc*: few guidelines are available for controlling and optimizing the operating characteristics of such studies, and the criteria for evaluating the designs remain 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 Yuan *and others* (2016).

We focus on randomization procedures and inference for trials during which new experimental arms are added. We discuss three randomization methods and study their operating characteristics. The first one is a balanced randomization (BR) algorithm. In this case the arm-specific accrual rates vary with the number of treatments available during the trial. We show that the approach yields substantial efficiency gains compared to separate two-arm studies. The other two methods use the outcome data to adaptively vary the randomization probabilities. 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). In all three cases the relevant difference between the designs that we consider and BR, BAR, or DBCD is the possibility of adding new experimental arms to an ongoing trial. We also introduce a Bootstrap procedure to test efficacy under the proposed platform designs. The algorithm extends previously introduced bootstrap schemes (Rosenberger and Hu, 1999; Trippa *and others*, 2012) to platform trial designs with group-sequential interim analysis (IA). The resampling method estimates sequentially stopping boundaries that correspond to pre-specified type-I error values at interim and final analyses.

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, these randomization procedures are combined with early stopping rules and a bootstrap algorithm for testing efficacy. Section 4 evaluates the proposed designs in a simulation study. In Section 5, we compare the performances of the three designs under scenarios tailored to the endTB trial. Section 6 concludes the article with a discussion.

2. Adding arms to an ongoing trial

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

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

2.1. *Balanced randomization*

A non-adaptive randomization algorithm for a multi-arm trial assigns patients to control and experimental arms with a ratio q_0/q_1. Here $q_\ell > 0, \ell = 0, 1$, are pre-specified non-negative weights, for instance $q_0/q_1 = 1/2$, that determine the ratio of patients assigned to the control arm compared to each experimental arm.

The overall sample size is $n_1 = n_C + A_1 \times n_E$, where the number of patients treated with the control arm n_C and each experimental arm n_E are selected based on targeted type I/II error probabilities. For the moment, we do not consider early stopping.

Here we describe a randomization scheme for adding new treatments, focusing on the case of $K = 2$ first. We 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 the initial experimental arms with probabilities proportional to $q_0 I\{N'_0(i) < n_C\}$ and $q_1 I\{N'_a(i) < n_E\}, a = 1, \ldots, A_1$. At the arrival of the M_2-th patient, arms $A_1 + 1, \ldots, A_1 + A_2$ are added, and the sample size is extended by $n_2 = n_{C,2} + n_E A_2$ patients, $n_{C,2} \geq 0$ for the control arm and n_E for each added arm. The remaining patients $i = M_2, \ldots, n_1 + n_2$ are

then randomized to the initial arms $a = 0, \ldots, A_1$ or to the added arms $a = A_1 + 1, \ldots, A_1 + A_2$, with probabilities

$$p[C_i = a|D_i] \propto \begin{cases} q_0 \times I\{N_0'(i) < n_C + n_{C,2}\} & \text{if } a = 0, \\ q_1 \times I\{N_a'(i) < n_E\} & \text{if } 0 < a \le A_1, \\ q_2 \times I\{N_a'(i) < n_E\} & \text{if } A_1 < a \le A_1 + A_2, \end{cases} \tag{2.1}$$

where $q_2 > 0$. At the completion of the study, n_E patients have been assigned to each experimental arm $a > 0$, and $n_C + n_{C,2}$ patients are assigned to the control arm. 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 added experimental arms. An additional $n_{C,2} > 0$ patients for the control arm may be necessary for longer studies with a slow accrual and potential drifts in the population. The parameter q_2 modulates the enrollment rate to the new arms after these arms have been added to the trial. The choice of q_2 should depend on $(q_0, q_1, M_2, A_1, A_2)$. 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 case $K \ge 2$ is similar. At the enrollment of the M_k-th patient ($2 \le k \le K$), A_k new arms are added; and the sample size is increased by $n_k = n_{C,k} + A_k n_E$ patients, $n_{C,k} \ge 0$ 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 and $M_1 = 1$. Patient $M_k \le i < M_{k+1}$ is assigned to an active arm a, with probability

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

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

$$Q_k = \frac{q_0 + \sum_{j=1}^{k-1} A_j Q_j}{(\sum_{j=1}^{k} n_j - M_k + 1)/n_E - A_k} \tag{2.3}$$

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

The step function $I\{N_a'(i) \le n_E\}$ 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 smoothly decreasing function.

Example 2.1 We consider a multi-arm trial with *four* experimental arms and a control arm with response probability of $\theta_0 = 0.3$ after 8 weeks of treatment. A multi-arm trial with ($q_0 = q_1 = 1$) and targeted type I/II error probabilities of 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 $n_C = n_E = 53$ patients. With an accrual rate of six patients per month, the trial duration is approximately 45 months. We can now introduce a departure from this setting. Two treatments $a = 3, 4$ become available approximately 12 and 24 months after the beginning of the trial ($M_2 = 72$, $M_3 = 144$, and $A_2 = A_3 = 1$). We describe three designs. (1) The first one uses all outcomes of the control arm available at completion of the study to evaluate arms $a = 1, \ldots, 4$. In this case, $n_{C,k} = 0$ and $q_k = Q_k$ for $k = 2, 3$ with definition of Q_k as in (2.3). (2) To avoid bias from possible population trends,

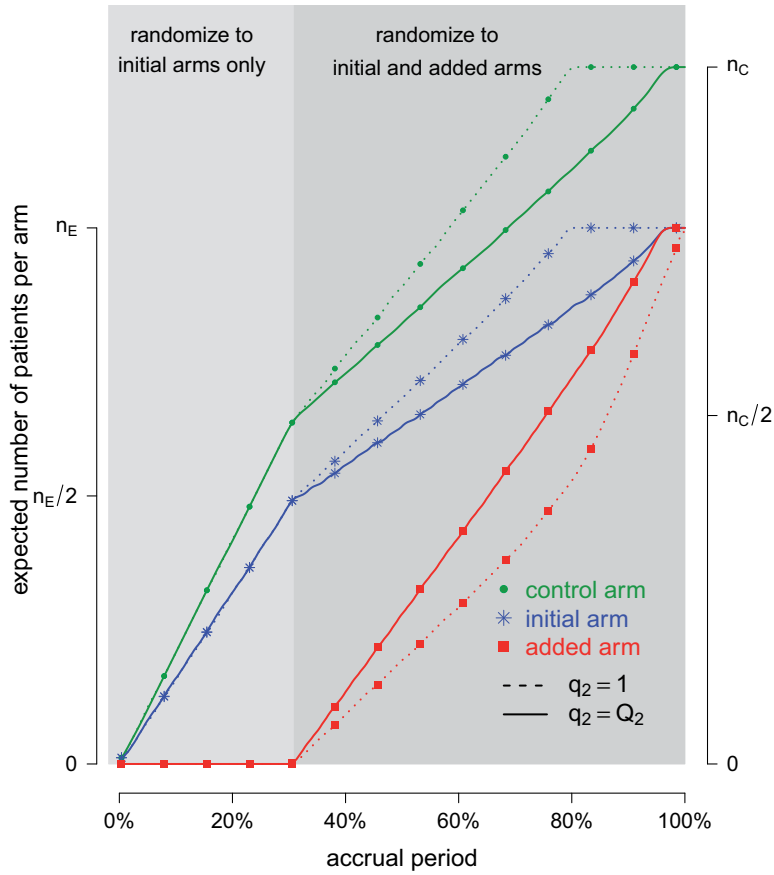

Fig. 1. Adding experimental arms to a multi-arm BR trial. We consider a trial with two initial arms $A_1 = 2$ and two added $A_2 = 2$ experimental arms. The graph shows the expected number of patients randomized to an arm during the accrual period for the control $a = 0$, one initial arm in $\mathcal{A}_1$ 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. Bold numbers are operating characteristics of effective experimental arms.

the second design estimates treatment effects of arm $a \in \mathcal{A}_k$ using only control outcomes of patients $i \geq M_k$ randomized to the control arm after the M_k-th enrollment. 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 $n_{C,2} = N_0'(M_2)$ and $n_{C,3} = N_0'(M_3) - n_{C,2}$ at the M_2-th and M_3-th arrival. (3) We also consider a third strategy with three independent trials; one for the initial experimental arms, and two additional two-arm trials for arm $a = 3$ and $a = 4$, each study has its own control arm. We assume again an average enrollment of 6 patients per month. Design 1 requires 265 patients, and the treatment effect estimates are available approximately 45 months after the first enrollment. Design 2, with $q_k = 1$, requires on average 307 patients. Treatment effects estimates are available approximately 37, 47, and 53 months after the first enrollment. The three independent trials in design 3 would instead require 371 patients and the effect estimates are available approximately 46, 60, and 64 months after the first patient is randomized.

2.2. *Bayesian adaptive randomization*

BAR uses accumulating data during the trial to vary the randomization probabilities (Thall and Wathen, 2007; Lee *and others*, 2010). Initially, BAR randomizes patients with equal probabilities to each arm. As the trial progresses and information on efficacy becomes available, randomization favors the most promising treatments. This can translate into a higher power compared to balanced designs (Wason and Trippa, 2014).

We complete the outcome model with a prior $\theta_a \sim p[\theta_a|\nu]$ for the response probabilities of arm $a \geq 0$. We use a conjugate beta distributions with parameters $\nu = (\nu_1, \nu_2)$. To predict the response probabilities of new arms in the group $\mathcal{A}_k, k > 1$, even when no outcome data are available for treatments in $\mathcal{A}_k$, we leverage on 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

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h(i)} & \text{if } a \in \mathcal{A}_1, \\ c(i) \exp\left\{ - b \times \left[N_0'(i) - \max_{a \in \mathcal{A}_1} N_a'(i)\right]\right\} & \text{if } a = 0, \end{cases} \tag{2.4}$$

where $b > 0$, $c(i) = \sum_{a \in \mathcal{A}_1} P[\theta_a > \theta_0|D_i]^{h(i)}/A_1$ and the function $h(\cdot)$ is increasing in the number of enrolled patients (Thall and Wathen, 2007). Initially $h(\cdot)$ equals zero, and randomization is balanced. 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 (2.4) 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 extend 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 defined as

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h_1(i)} \times q_1(i) & \text{if } a \in \mathcal{A}_1, \\ p[\theta_a > \theta_0|D_i]^{h_2(i)} \times q_2(i) & \text{if } a \in \mathcal{A}_2 \text{ and } i \geq M_2, \\ c(i) \exp\left\{ - b \times \max_{\ell=1,2} s_\ell(i)\right\} & \text{if } a = 0, \end{cases} \tag{2.5}$$

where $s_\ell(i) = I(i \geq M_\ell)\left\{\left[N_0'(i) - N_0'(M_\ell)\right] - \max_{a \in \mathcal{A}_\ell} N_a'(i)\right\}$ and $c(i) = \sum_{k=1,2; a \in \mathcal{A}_k} I\{M_k \leq i\} \times q_k(i) \times p[\theta_a > \theta_0|D_i]^{h_k(i)}$. We introduce group-specific scaling and power functions $q_k(i)$ and $h_k(i)$. The power function $h_k(i)$ controls 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 treatments, which compete for 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

$$q_k(i) = r_0 + r_1 \exp\left\{ - \exp\left(N'^{(k)}(i) - m_k\right)\right\}, \tag{2.6}$$

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

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 (2.6) by the indicator $I\{N'_a(i) < n'_E\}$, where $n'_E > 0$ represents a desired maximum number of patients in each experimental arm.

We use a function $h_1(\cdot)$ that is increasing in the number of patients randomized to arms in $\mathcal{A}_1$ with a maximum $\beta > 0$ after n_1 enrollments. Similarly, for the added arms in $\mathcal{A}_2$, $h_2(\cdot)$ is increasing in the number of patients randomized to $\mathcal{A}_2$, with a maximum H at n_2. In particular $h_k(i)$ is equal to $\beta \times [N'^{(k)}(i)/n_k]^\gamma$ if $N'^{(k)}(i) \le n_k$ and β otherwise, where $\gamma \ge 0$ and, as explained above, n_k denotes the extension of the overall sample size after M_k enrollments.

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

$$p[C_i = a|D_i] \propto \begin{cases} p[\theta_a > \theta_0|D_i]^{h_\ell(i)} \times q_\ell(i) & \text{if } a \in \mathcal{A}_\ell \text{ and } M_\ell \le i, \\ c(i) \exp\left\{ - b \times \max_{\ell: M_\ell < i} s_\ell(i) \right\} & \text{if } a = 0, \end{cases} \tag{2.7}$$

where $s_\ell(i)$ and $c(i)$ are defined as in (2.5), and $q_\ell(i)$ is the Gompertz function defined in (2.6). 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 We consider the same trial as in Example 2.1, but use a BAR design instead. To simplify comparison to BR, we set the overall sample size to $n = 265$ as for BR. We can easily verify that if $(\beta, b, n'_E) = (0, 0, 53)$ and $q_k(i) = 1$, 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, either arm $a = 1$, or arm $a = 3$ added at $M_2 = 72$, or arm $a = 4$ added at $M_3 = 144$ have positive treatment effects, $(\theta_a, \theta_0) = (0.5, 0.3)$. In each scenario, the remaining 4 of the 5 arms, including the control, have identical response rates equal to 0.3. We tuned the parameters of the design to maximize power under the assumption that there is a single effective arm and $n_2 = n_3 = 53$, $(\beta, \gamma, b) = (3, 1.5, 0.5)$. 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 for each arm when $\theta_k = 0.3$ for all arms.

As for BR, in all three scenarios, the trial completes accrual after approximately 45 months. In scenario 1, BAR randomizes on average 64 patients to arm 1 and to the control arm across 5000 simulations, while on average (43, 46, 47) patients are assigned to the ineffective arms 2, 3, 4 with standard deviations (SDs) of 4.7, 6.4, 7.4, 5.3, and 5.4 (see Table 1). The power increases to 85%—compared to 80% for BR—with an identical overall sample size. In scenarios 2 and 3, BAR randomizes on average 64 and 63 patients to arm $a = 3$ and $a = 4$, respectively. This translates into 86% and 85% power for the added arms 3 and 4, respectively, compared to 80% for BR.

2.3. *Doubly adaptive biased coin design*

The DBCD (Eisele, 1994) is a response adaptive randomization scheme that seeks to assign patients to treatments according to a vector of target proportions $\{\rho_a\}_a$ that depends on response rates. Examples include the Neyman allocation $\rho_a \propto \sqrt{\theta_a(1 - \theta_a)}$ and $\rho_a \propto \theta_a^{1/2}$ (Hu and Zhang, 2004). Since the response probabilities θ_a are unknown, the target allocation is estimated from the accumulated data by $\widehat{\rho}_a(i)$. For $K = 1$, patients are randomized to arm $a = 0, \ldots, A_1$ with probabilities

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

Table 1. *Expected sample size (E), standard deviation (SD) and power (Po) for experimental arm 1, the first added arm $a = 3$ and the second added arm $a = 4$ for a trial with two initial experimental arms, and two arms which are added after 12 and 24 month, $(M3, M_4) = (72, 144)$*

Scenario	Control		Arm 1			First added arm			Second 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**

Results are based on 5000 simulated trials under balanced randomization (BR), Bayesian adaptive randomization (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. Bold numbers are operating characteristics of effective experimental arms.

Here, $q_a(i) = \left(\widehat{\rho}_a(i) \times (i+1)/(N'_a(i)+1)\right)^{\beta}$ varies with the ratio of (i) the estimated target allocation proportion $\widehat{\rho}_a(i)$ and (ii) the current number of patients that are 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. As for BAR, we limit the maximum number of patients per arm by multiplying the correction $q_a(i)$ by the indicator $I\{N'_a(i) < n'_E\}$.

We now consider adding new experimental arms during the study. Until the M_2-th arriving patient, the target $\{\rho_a\}_{a=0}^{A_1}$ is a function of $\{\theta_a\}_{a=0}^{A_1}$, and it is estimated through the hierarchical Bayesian model in Section 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 (2.8). Then, at the enrollment of the M_k-th patient, $k \geq 2$, A_k arms are added, and the overall sample size is increased by n_k patients. Before observing any outcome for arm $a \in \mathcal{A}_k$, the target is re-defined to $\rho_a(\theta), 0 \leq a \leq A_1 + \cdots + A_k$, with $\theta = \{\theta_a; 0 \leq a \leq A_1 + \cdots + A_k\}$. The posterior distribution of the hierarchical model is used to compute $\widehat{\rho}_a(i) = E[\rho_a(\theta)|D_i]$ for all initial and added arms a. 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 $i \geq 1$ is randomized to the control arm $a = 0$ or to treatments $a \in \mathcal{A}_k$ for groups k added before the i-th arrival with probability

$$p(C_i = a|D_i) \propto \widehat{\rho}_a(i) \times q_a(i). \tag{2.9}$$

For treatments in $\mathcal{A}_k$, $1 \leq k \leq K$, the functions $q_a(i) = \left[\widehat{\rho}_a(i)(i+1)/(N'_a(i)+1))\right]^{h_k(i)}$ 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 (2.9) 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 $w(i) \propto 1 \Big/ \Big(1 + \sum_{k;a\in\mathcal{A}_k} I\{M_k \leq i, N'_a(i) < n'_E\}\Big)$, a decreasing function of the number of active arms. Also for the DBCD

design, the function $h_k(\cdot)$ increases during time with $h_k(i) = h_k + \beta \times (N'^{(k)}(i)/n_k)^\gamma$ if $0 \leq N'^{(k)}(i) < n_k$ and $h_k + \beta$ 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.

Example 2.3 We consider again the setting in Examples 2.1 and 2.2, and use a DBCD design for the trial. Following Hu and Zhang (2004) we use the target allocation $\rho_a(\theta) \propto \theta_a^{1/2}$ for $a > 0$. To preserve the power of the design, similarly to Example 2.2, we use $\rho_0(\theta) = \max_{a>0} \theta_a^{1/2}$ 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 and $n_2 = n_3 = 53$. If the response probabilities for all arms are 0.3, a DBCD with $(\beta, \gamma) = (3, 1)$ and $(h_1, h_2, h_3) = (0, 4, 5)$ randomizes on average 52 patients to each experimental arm, and 57 to the control (SD 3.3, 4.8, 4.8, 4.6, and 4.3). We consider the same scenarios as in Examples 2.1 and 2.2.

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 $a = 3$ and $a = 4$ (SD of 3.8 and 3.4). Compared to 80% and 85% for BR and BAR the power of the DBCD becomes 82%. Similarly in scenario 3, for arm 4 we have a power equal to 82% using the DBCD compared to 80% and 85% using BR and BAR.

3. Early stopping rules and hypothesis testing

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

3.1. *A bootstrap test for platform trials without early stopping for efficacy*

For a platform trial $\mathcal{T}$, if arm $a \in \mathcal{A}_k$ added after M_k enrollments is not stopped for futility, we compute a bootstrap P-value estimate at a pre-specified time τ_a, for example when $N_a(i)$ reaches n'_E, or at the completion of the trial $\mathcal{T}$. The bootstrap procedure is similar to the algorithms discussed in Rosenberger and Hu (1999) and in Trippa *and others* (2012). We use the statistic T_a, the standardized difference between the estimated response rate of arm $a > 0$ and the control, to test the null hypothesis $H_a : \theta_a \leq \theta_0$ at significance level α. Large values of T_a indicate evidence of a treatment effect. The algorithm estimates the distribution of T_a under the null hypothesis H_a and the platform design, which includes changes of the randomization probabilities when new experimental arms are added.

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 hypothesis H_a. If $\widehat{\theta}_a > \widehat{\theta}_0$, we use the following bootstrap procedure, which is also summarized in Algorithm 1:

(i) For all arms a' that enrolled patients before time τ_a, we compute maximum likelihood estimates (MLEs) $\widehat{\theta}_{a'}$. For arm a and the control, we restrict the MLE to $\theta_0 = \theta_a$.

(ii) The algorithm then simulates C trials $\{\mathcal{T}_{a,c}\}_{c=1}^C$, from the first enrollment until time $\tau_{a,c}$. Here $\tau_{a,c}$ is defined identical to τ_a and corresponds to simulation $\mathcal{T}_{a,c}$. In each simulation $\mathcal{T}_{a,c}$, A_j arms are added to the study after M_j enrollments, for all $M_j < \tau_{a,c}$, and randomization probabilities are updated as

described in Section 2. Patients in these simulations respond to treatments with probabilities $\widehat{\theta}$ and the simulations' accrual rate is identical to the accrual rate of the actual trial $\mathcal{T}$.

(iii) For each simulation $\mathcal{T}_{a,c}$ we compute the test statistics $T_{a,c}$ at time $\tau_{a,c}$, and set $S_{a,c}$ equal to zero if arm a was stopped early for futility and equal to one otherwise. The simulations generate test statistics $T_{a,c}$ under the null hypothesis H_a and the platform design. We can therefore estimate the P-value by $\widehat{p(T_a)} = \sum_{c=1}^{C} I\{T_{a,c} \geq T_a, S_{a,c} = 1\}/C$ and reject H_a at level α if $\widehat{p(T_a)} \leq \alpha$.

3.2. *A bootstrap test for platform trials with early stopping for efficacy*

We extend the procedure described in the previous section and include early stopping for efficacy. In this case, there is a connection between the α-spending method of Lan and DeMets (1983) and our algorithm. We consider J IA, conducted after a pre-specified set of observed outcomes. At each IA, the arms that are evaluated for efficacy may vary, for instance because arms have been added or removed from the trial. We partition the type I error probability $\alpha = \sum_{j=1}^{J} \alpha_a^{(j)}$ into pre-specified values $\alpha_a^{(j)} \in [0, 1]$ for each IA j. For each initial and added arm $a \in \mathcal{A}_k, k \geq 1$, the algorithm estimates the thresholds $t_a^{(j)}$ defined by the following target. Under the platform design and the unknown combination $(\theta_0, \ldots, \theta_{a-1}, \theta_0, \theta_{a+1}, \ldots)$, where we replace θ_a with θ_0, the probability of stopping arm a for efficacy at the IA j is $\alpha_a^{(j)} \approx p\big(T_a^{(j)} \geq t_a^{(j)}, S_a^{(j)} = 1\big)$. Here $S_a^{(j)} = 0$ if arm a is stopped before the j-th IA and equals 1 otherwise, while $T_a^{(j)}$ is the test statistics computed at IA j. In what follows, simulations under the null hypothesis 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 following algorithm for up to six IA and eight arms:

`First IA`: We compute the MLEs $\hat{\theta}$ and the statistics $T_a^{(1)}$ for all initial and added arms $a \in \mathcal{A}_k$ that enrolled patients before the first IA. Then, separately for each of these arms a:

(i) We generate C platform trials $\{\mathcal{T}_{a,c}\}_{c=1}^{C}$ under H_a, from the first enrollment until the first IA. In these simulations, all arms $a' \in \mathcal{A}_{k'}$ which have been added to the actual trial $\mathcal{T}$ before the first IA are successively added to the simulated trial $\mathcal{T}_{a,c}$, and patients respond to treatment a' with probability $\widehat{\theta}_{a'}$. By adding these arms to the trial $\mathcal{T}_{a,c}$ before the first IA we account for and mimic in simulation $\mathcal{T}_{a,c}$ the variations of the randomization probabilities after the new arms have been added.

(ii) We then compute the test statistics and the indicator variable $\left(T_{a,c}^{(1)}, S_{a,c}^{(1)}\right)$ for each simulated trial $\mathcal{T}_{a,c}$, with definitions identical to those of $\left(T_a^{(1)}, S_a^{(1)}\right)$ for the actual trial $\mathcal{T}$. The threshold $t_a^{(1)}$ is then estimated by $\widehat{t}_a^{(1)} = \min_t \left\{t : \sum_c I\{T_{a,c}^{(1)} \geq t, S_{a,c}^{(1)} = 1\}/C \leq \alpha_a^{(1)}\right\}$ and arm a is stopped for efficacy at the first IA if $T_a^{(1)} \geq \hat{t}_a^{(1)}$ and $S_a^{(1)} = 1$.

`Second IA`: We recompute the MLEs $\hat{\theta}$ using the data available at the second IA.

(i) Separately for each arm added before the second IA we re-estimate $\hat{t}_a^{(1)}$ using a new set of simulations $\{\mathcal{T}_{a,c}\}_{c=1}^{C}$ under H_a from the first enrollment until the first IA.

(ii) In these new simulations $\mathcal{T}_{a,c}$, if for any arm a' that enrolled patients before the first IA the statistics $T_{a',c}^{(1)} > \hat{t}_{a'}^{(1)}$ and $S_{a',c}^{(1)} = 1$, then arm a' is stopped for efficacy in the simulated trial $\mathcal{T}_{a,c}$. This part of the algorithm creates, for each arm a that enrolled patients before the first IA, C simulations under H_a that cover the time window from the first enrollment until the first IA.

(iii) Simulations then continue beyond the first IA. After the thresholds $\hat{t}_a^{(1)}$'s have been recomputed, we extend, for each arm a, the new simulations $\{\mathcal{T}_{a,c}\}_c$ in time to cover the window between the first and the second IAs. Importantly, these simulations include early stopping at the first IA. Analogous

to the first IA, if new arms (e.g. $a' \in \mathcal{A}_2$) are added between the first and second IAs, then, starting from the M_2-th enrollment, all simulations will include the added arms to mimic the actual platform trial.

(iv) We estimate $\hat{t}_a^{(2)} = \min_t \left\{ t : \sum_c I\{T_{a,c}^{(2)} \geq t, S_{a,c}^{(2)} = 1\}/C \leq \alpha_a^{(2)} \right\}$ and then stop arm a at the `second IA` for efficacy if $T_a^{(2)} \geq \hat{t}_a^{(2)}$ and $S_a^{(2)} = 1$.

`jth IA` The same procedure is iterated similarly to $j = 2$ for all other IAs $j = 3, \ldots, J$. In some simulations of the multi-arm study under H_a, where $a \in \mathcal{A}_k$, arm a might not appear because the experimental arms have been all dropped and the trial stopped before M_k enrollments. To account for this, all 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.

We refer to the Supplementary material available at *Biostatistics* online for an example (Example S1.1) of the described algorithm. We also include in the Supplementary material available at *Biostatistics* online, a discussion on hypothesis testing and departures from the error rate α when the patient population varies during the trial.

4. Simulation study

We continue Examples 2.1, 2.2, and 2.3 using four scenarios. In scenario 1 no experimental arm has a treatment effect. Whereas in scenarios 1–3 either the initial arm $a = 1$, the first added arm $a = 3$, or the second added arm $a = 4$ is effective with response rate of 0.5. The remaining experimental arms have response rates equal to the control rate of 0.3. The initial sample size is $n_1 = 159$ and $n_2 = n_3 = 53$, and the type I error is controlled at 10%.

Both, BAR and DBCD, can assign at most $n'_E = 69 \approx 1.3 \times n_E$ patients to each experimental arm, and all three designs use outcomes from patients randomized to the control before and during the accrual period of the added arms to define randomization probabilities and for hypothesis testing. In Section S2 of supplementary available at *Biostatistics* online, we outline possible modifications of the designs when trends on the patient population during the trial represent a concern.

We used the same parameters to define the randomization probabilities as in Examples 2.1, 2.2, and 2.3. For BR the scaling parameters equal $q_k = Q_k$, with Q_k defined in (2.3), and $n_{C,k} = 0$. The parameters of the Gompertz function in BAR equal $(r_0, r_1) = (1, 3)$, $(m_1, m_2, m_3) = (20, 30, 45)$ and $(\beta, \gamma, b) = (3, 1.5, 0.5)$. For DBCD we used $(\beta, \gamma) = (3, 1)$, and $(h_1, h_2, h_3) = (0, 4, 5)$, and the randomization probabilities for active arms have been restricted to values larger than $w(i) = 1 \Big/ \left[3\left(1 + \sum_{k;a\in\mathcal{A}_k} I_a(i)\right)\right]$. Here $I_a(i) = 1$ if $M_k \leq i, N'_a(i) < n'_E$ and arm a has not been stopped before the enrollment of the i-th patient, and the indicator $I_a(i)$ is equal to zero otherwise.

We first summarize the operating characteristics of the three designs without early stopping to illustrate the performance of the randomization schemes. The three designs are compared to three independent trials; one trial for the initial two experimental arms, and two independent two-arm studies for the added arms 3 and 4, each with their own control arms. We indicate them as balanced randomized and independent trials (BRI). The overall rate of accrual of the three independent trials in BRI is set to six patients per month, and is assumed to be identical for the competing studies.

Figure 2 shows the median number of patients randomized to arms 1, 3, and 4 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 $x = 1, 2, \ldots, 265$ (371 for BRI) patients have been enrolled to the trial (x-axis). Under BRI, 106 additional patients are necessary for the two additional control arms. This prolongs the trials and slows down the accrual to experimental arms.

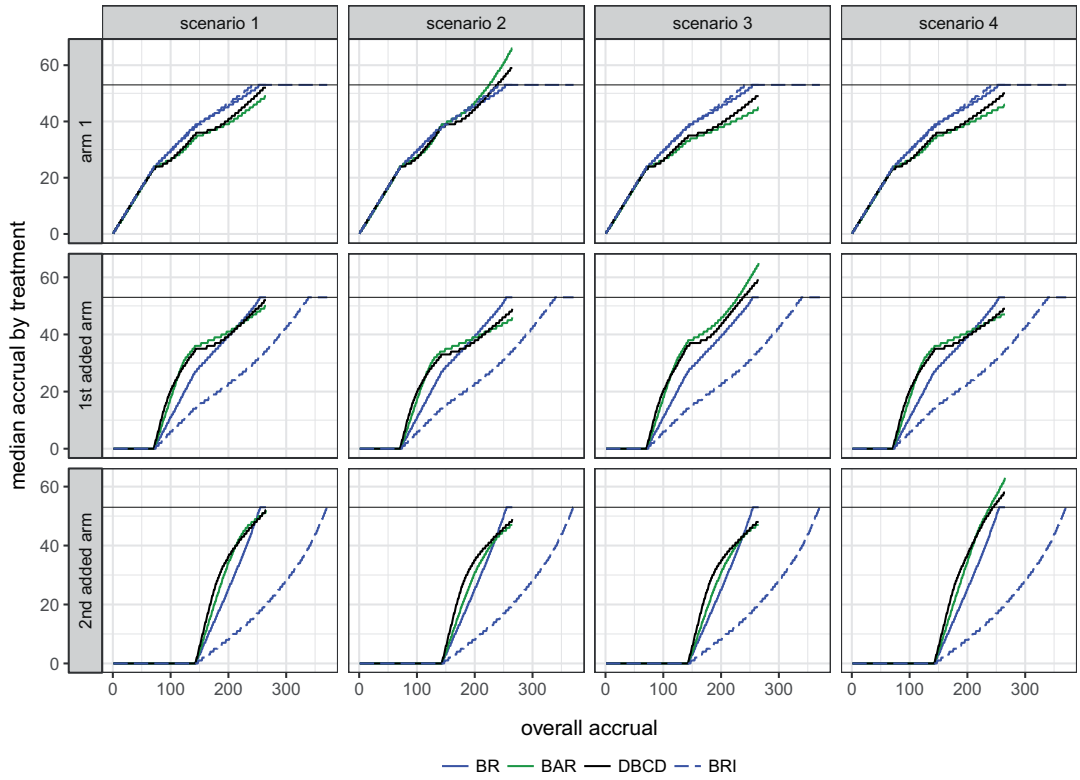

Fig. 2. Number of patients randomized to treatment arms during the accrual period of the study, for a trial with two initial experimental arms and two arms that are added after the enrollment of $M_2 = 72$ and $M_3 = 144$ patients. BRI corresponds to a design that uses three balanced and independent trials—one trial for the initial arms and one two-arm trial for each added arm, BR, BAR, and DBCD denote balanced randomization, Bayesian adaptive randomization and the doubly adaptive biased coin design. 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 2–4 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.

Figure S1 of supplementary available at *Biostatistics* online shows the variability of treatment assignments at the end of the trial. In scenario 1, DBCD has a median accrual of 52 patients for all 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), and for the two added arms the median equals 50 and 52 with IQs of (48, 57) and (45, 56). 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 1). In scenario 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 1). 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 assign 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) is selected such that the probability of stopping an effective initial arm early for futility is approximately 1%, $(f, g) =$

Table 2. *Expected sample size (E), standard deviation (SD) and power (Po) for experimental arm 1, the first added arm $a = 3$, and the second added arm $a = 4$, for a trial with two initial experimental arms, and two arms which are added after 12 and 24 months, with futility and efficacy early stopping*

Scenario	Control		Arm 1			First added arm			Second 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**

Two IA for efficacy are planned after 100, 200 patients have been enrolled. Results are based on 5000 simulated trials under balanced randomization (BR), Bayesian adaptive randomization (BAR) and doubly adaptive biased coin design (DBCD). The initial planned sample size is 159, which is then extended by 53 patients for each added arm. Bold numbers are operating characteristics of effective experimental arms.

$(0.25, 1.5)$ 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 arms $a = 3, 4$.

Table 2 shows the average sample size, SD 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 IA 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 a 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. 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 for multi-drug resistant Tuberculosis (MD-TB) (Cellamare *and others*, 2017). The trial tests five experimental treatments under a response-adaptive BAR design that is similar to the one described in Section 2.2. We initially designed the trial

Table 3. *Expected sample size (E), standard deviation (SD) and power (Po) 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*

Scenario	Control arm		Initial arms Arm 1			First added group Arm 5			Second 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**

In *Scenario 1*, all experimental arm has response rates identical to the control of 0.55, whereas in Scenarios 2 and 3 experimental arm 1 and 5 (Scenarios 2) or experimental arm 1 and 7 (Scenarios 3) are superior to the control with probability of response equal to 0.7 and 0.75. Lastly, in Scenario 4 experimental arms 1, 5 and 7 are effective with probability of response equal to 0.7, 0.75 and 0.7 compared to 0.55 for the control. Bold numbers are operating characteristics of effective experimental arms.

with eight experimental arms, but we were later informed that four of these treatments would have not been available at the activation of the trial. Thus the investigators wanted to know if the treatments could be added later during the study. Previous trials showed response probabilities of approximately 0.55 after 6 months of treatment with the control therapy. A response probability of 0.7 for experimental arms was considered a relevant improvement. The study was designed with an expected accrual rate of 10 patients per month.

We present a simulation with four 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 $M_3 = 300$ enrollments, and sample size is increased each time by $n_2 = n_3 = 200$ patients. The type I error is controlled at the $\alpha = 5\%$ level. We consider the four scenarios. Experimental arms without treatment effects have response rates identical to the control of 0.55. In scenario 1, all arms have identical response rates equal to 0.55. 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, arms 1, 2, and 7 are effective, with response probabilities 0.7, 0.75, and 0.7.

For BR we use the scaling parameters $q_k = Q_k$, with $n_{c,2} = n_{c,3} = 0$, and $(f, g) = (0.3, 1.5)$ for futility stopping. For BAR we use $(r_0, r_1) = (1, 3)$, $(m_1, m_2, m_3) = (200, 125, 135)$ and $(\beta, \gamma, b) = (3, 1.5, 0.5)$. The DBCD utilizes the parameters $(h_1, h_2, h_3) = (0.5, 2, 2)$ and $(\beta, \gamma) = (3, 1)$. For BAR and the DBCD the futility stopping rules are implemented with $(f, g) = (0.2, 1.5)$ and we limit the accrual to each experimental arm to $n'_E = 140$ patients.

Table 3 shows the mean number of patients randomized to the control arm, and to arms 1, 5, and 7 across 5000 simulations, together with the SD and the power. Under scenario 1, BR randomizes on

average 98 and 79 patients to the control arm and the initial arms, respectively, and 80 and 82 patients to arms in the second and third group (SD 5.1, 25.4, 25.4, and 24.4) 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 a treatment effect for arms $a = 1$ and $a = 5$ with response rates 0.7 and 0.75, respectively. BAR and DBCD have 10% and 3% higher power for arm $a = 1$ (80% and 73%), and 7% and 3% higher power for arm $a = 5$ (97% and 93%). The gain in power of BAR and DBCD compared to BR is associated with an increase of the average number of randomized patients to arm $a = 1$ and $a = 5$. 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%, 80%, and 74% for BR, BAR and DBCD. For the added arm $a = 7$, BR has 92% power compared to 97% and 93% for BAR and DBCD, with mean accruals of 100, 130, and 108 under BR, BAR, and DBCD, respectively. Lastly, in scenario 4, arms 1, 5, and 7 are effective with response rates 0.7, 0.75, and 0.7. Here BR randomizes an average (99, 100, 99) patients to these arms (SD 8.7, 2.7, and 5.2) with 70%, 90%, and 70% power. In comparison BAR and DBCD randomize on average (121, 128, 118) 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 areas focuses increasingly on targeted patient populations defined by biological pathways. Drugs that target 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 hormone receptor positive metastatic breast cancer 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 new experimental therapies. Sharing an active control arm among multiple experimental treatments reduces the proportion of patients allocated to the control.

We explored 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 substantially improve efficiency. BAR and DBCD require simulations for parameter tuning, but can potentially increase the power of the 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.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

SV and MC have no funding to declare. The work of G.P. was supported by the National Cancer Institute grant 4P30CA006516-51. LT has been funded by the Claudia Adams Barr Program in Innovative Cancer Research and the Burroughs Wellcome Fund in Regulatory Science.

REFERENCES

BARKER, A. D., SIGMAN, C. C., KELLOFF, G. J., HYLTON, N. M., BERRY, D. A. AND ESSERMAN, L. J. (2009). I-spy 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics* **86**, 97–100.

BERRY, S. M., CARLIN, B. P., LEE, J. J. AND MULLER, P. (2010). *Bayesian Adaptive Methods for Clinical Trials*. Boca Raton, FL: CRC Press.

BERRY, S. M., CONNOR, J. T. AND LEWIS, R. J. (2015). The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA* **313**, 1619–1620.

BURNETT, A. K., RUSSELL, N. H., HILLS, R. K., HUNTER, A. E., KJELDSEN, L., YIN, J., GIBSON, B. E., WHEATLEY, K. AND MILLIGAN, D. (2013). Optimization of chemotherapy for younger patients with acute myeloid leukemia: results of the medical research council AML15 trial. *Journal of Clinical Oncology*, **31**, 3360–3368.

CELLAMARE, M., VENTZ, S., BAUDIN, E., MITNICK, C. D. AND TRIPPA, L. (2017). A Bayesian response-adaptive trial in tuberculosis: the endTB trial. *Clinical Trials* **14**, 17–28.

COHEN, D. R., TODD, S., GREGORY, W. M. AND BROWN, J. M. (2015). Adding a treatment arm to an ongoing clinical trial: a review of methodology and practice. *Trials* **16**, 179.

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

ELM, J. J., PALESCH, Y. Y., KOCH, G. G., HINSON, V., RAVINA, B. AND ZHAO, W. (2012). Flexible analytical methods for adding a treatment arm mid-study to an ongoing clinical trial. *Journal of Biopharmaceutical Statistics* **22**, 758–772.

US FDA. (2013). Guidance for industry: codevelopment of two or more new investigational drugs for use in combination. FDA. https://www.fda.gov/downloads/drugs/guidances/ucm236669.pdf.

FREIDLIN, B., KORN, E. L., GRAY, R. AND MARTIN, A. (2008). Multi-arm clinical trials of new agents: some design considerations. *Clinical Cancer Research* **14**, 4368–4371.

HILLS, R. K. AND BURNETT, A. K. (2011). Applicability of a "Pick a Winner" trial design to acute myeloid leukemia. *Blood* **118**, 2389–2394.

HOBBS, B. P., CHEN, N. AND LEE, J. J. (2016). Controlled multi-arm platform design using predictive probability. *Statistical Methods in Medical Research*, pii: 0962280215620696. [Epub ahead of print].

HU, F. AND ZHANG, L.-X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Annals of Statistics*, **32**, 268–301.

LAN, K. K. G. AND DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659–663.

LEE, J. J., GU, X. AND LIU, S. (2010). Bayesian adaptive randomization designs for targeted agent development. *Clinical Trials* **7**, 584–596.

LIEBERMAN, J. A., STROUP, T. S., MCEVOY, J. P., SWARTZ, M. S., ROSENHECK, R. A., PERKINS, D. O., KEEFE, R. S., DAVIS, S. M., DAVIS, C. E., LEBOWITZ, B. D. *and others*. (2005). Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. *NEJM* **353**, 1209–1223.

ROSENBERGER, W. F. AND HU, F. (1999). Bootstrap methods for adaptive designs. *Statistics in Medicine* **18**, 1757–1767.

THALL, P. F. AND WATHEN, J. K. (2007). Practical Bayesian adaptive randomisation in clinical trials. *European Journal of Cancer* **43**, 859–866.

TRIPPA, L., LEE, E. Q., WEN, P. Y., BATCHELOR, T. T., CLOUGHESY, T., PARMIGIANI, G. AND ALEXANDER, B. M. (2012). Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *JCO* **30**, 3258–3263.

VENTZ, S., BARRY, W. T., PARMIGIANI, G. AND TRIPPA, L. (2017). Bayesian response-adaptive designs for basket trials. *Biometrics*, doi: 10.1111/biom.12668. [Epub ahead of print].

WASON, J. AND TRIPPA, L. (2014). A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Statistics in Medicine* **33**, 2206–2221.

WASON, J. M., STECHER, L. AND MANDER, A. P. (2014). Correcting for multiple-testing in multi-arm trials: is it necessary and is it done? *Trials* **15**, 364.

YUAN, Y., GUO, B., MUNSELL, M., LU, K. AND JAZAERI, A. (2016). MIDAS: a practical Bayesian design for platform trials with molecularly targeted agents. *Statistics in Medicine* **35**, 3892–3906.

[*Received July 26, 2016; revised April 20, 2017; accepted for publication May 7, 2017*]