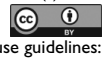
Special issue: Response-adaptive randomization in clinical trials

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

A burn-in(g) question: How long should an initial equal randomization stage be before Bayesian response-adaptive randomization?

Statistical Methods in Medical Research
1–21



DOI: 10.1177/09622802251411538
journals.sagepub.com/home/smm


Edwin YN Tang[1], Stef Baas[2], Daniel Kaddaj[3], Lukas Pin[2], David S Robertson[2] and Sofía S Villar[2]

Abstract

Response-adaptive randomization (RAR) can increase participant benefit in clinical trials, but also complicates statistical analysis. The burn-in period—a non-adaptive initial stage—is commonly used to mitigate this disadvantage, yet guidance on its optimal duration is scarce. To address this critical gap, this paper introduces an exact evaluation approach to investigate how the burn-in length impacts statistical operating characteristics of two-arm binary Bayesian RAR (BRAR) designs. We show that (1) commonly used calibration and asymptotic tests show substantial type I error rate inflation for BRAR designs without a burn-in period, and increasing the total burn-in length to more than half the trial size reduces but does not fully mitigate type I error rate inflation, necessitating exact tests; (2) exact tests conditioning on total successes show the highest average and minimum power up to large burn-in lengths; (3) the burn-in length substantially influences power and participant benefit, which are often not maximized at the maximum or minimum possible burn-in length; (4) the test statistic influences the type I error rate and power; (5) estimation bias decreases quicker in the burn-in length for larger treatment effects and increases for larger trial sizes under the same burn-in length. Our approach is illustrated by re-designing the ARREST trial.

Keywords

Conditional exact test, exact operating characteristics, binary outcomes, two-arm trial, unconditional exact test

1 Introduction

Clinical trials are important studies that evaluate the effects of new treatments on human health outcomes. Randomization is typically used in confirmatory clinical trials (and recommended where possible in Phase II settings) because it induces comparable treatment groups, mitigates selection bias and can provide a basis for statistical inference.[1] More often than not, modern-day clinical trials use a fixed randomization scheme (usually the permuted block design). Alternatively, statisticians can consider response-adaptive randomization (RAR), which allows allocation probabilities to change based on previous allocations and outcomes. A well-designed RAR procedure typically aims to balance the goal of drawing correct inferential conclusions with that of maximizing participant benefit.

Response-adaptive (RA) procedures have been used in exploratory or seamless phase II/III multi-arm trials (see, e.g., Berry and Viele[2]). Furthermore, a publicly available list summarizing RA clinical trials in the last 100 years currently reports 10 out of 30 RA trials were confirmatory.[3] Despite this, the fraction of confirmatory trials with an RA component remains relatively low. A reason for this could be the ongoing debates about the risks associated with RA designs (see, e.g., Robertson et al.[4] for an overview). While arguments in favour emphasize the prospect of improving participant benefit,[5]

[1]Department of Statistics, University of Warwick, Coventry, UK
[2]MRC Biostatistics Unit, University of Cambridge, Cambridge, UK
[3]Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Cambridge, UK

Corresponding author:
Stef Baas, University of Cambridge, MRC Biostatistics Unit, CB2 0SR, Cambridge, UK.
Email: stef.baas@mrc-bsu.cam.ac.uk

counterarguments highlight the non-negligible possibility of assigning more participants to the inferior arm.[6] The risk of substantial between-arm imbalances is especially worrisome when there is accrual bias (e.g., more severely ill patients being enrolled earlier on) or a temporal trend in prognostic baseline characteristics.[7] The use of an RA design also impacts the statistical analysis, since classical statistical methods may not maintain their desirable and well-understood properties such as type I error rate control.[8]

A common and ad hoc approach to alleviate the weaknesses mentioned above is the inclusion of a period of non-adaptive (fixed) allocation at the start of the RA design, which we will refer to from now on as a *burn-in period*. Computational results in Du et al.[9] show that in comparison to a fixed non-RA design with equal randomization, a suitable burn-in period length for a Bayesian RAR (BRAR) design allows more participants to be assigned to the superior arm on average due to some response-adaptiveness, but at the cost of a small decrease in statistical power.

Although the importance of a burn-in phase is generally recognized, few studies offer a rationale or provide practical guidance on its duration. The absence of a robust justification for burn-in period length is a notable deficiency in the current literature. Robertson et al.[4] only touch upon burn-in in their review, and Thorlund et al.'s[10] recommendation of 20–30 patients per arm may be overly broad, as the optimal length likely varies with the RA procedure, primary outcome type and sample size. Viele et al.[11,12] considered the effect of burn-in length for multi-arm RA designs, while Granholm et al.[13] considered general adaptive designs. Our approach differs from theirs in that we focus on type I error rate control across the null parameter set and numerically evaluate multiple metrics to inform burn-in recommendations for two-arm BRAR designs. As demonstrated in Supplemental Table 1, burn-in periods are common in BRAR trials, yet the existing reports rarely include a transparent explanation or rationale for the chosen duration.

To address the above gap, we concentrate on two-arm BRAR designs using a burn-in period as the sole tuning parameter and which use the posterior probability of control superiority (PPCS) to test for a treatment effect. The BRAR design is prevalent in implemented RA trials,[3] whereas this test statistic is prevalent in BRAR designs using a burn-in (Supplemental Table 1). To ensure a focused evaluation, we treat the burn-in period as the exclusive mechanism of regularization, thus isolating its influence on operating characteristics (OCs) from other design adjustments (such as clipping or tuning). Importantly, our exact analysis framework is generalizable to different test statistics and to other BRAR variations, such as batched allocation, clipping, and power transformations (see, for example, Du et al.[9]). In this context our contributions are as follows, we (1) exactly assess the effect of the burn-in length on the type I error rate, power, participant benefit, and the probability of an imbalance in the wrong direction (PIWD) in a BRAR clinical trial design. (2) We compute the point-wise, average, minimum, and maximum OCs over the parameter space, where the last three measures summarize the dependence of OCs on the burn-in length over the complete parameter space. Our approach avoids Monte Carlo error,[14] which allows us to, for example, compute the optimal burn-in length in terms of power, something that is much harder to estimate through simulation (due to non-smoothness). (3) We construct conditional and unconditional exact tests for BRAR designs with a burn-in and compare them to commonly used calibrated or asymptotic tests. Exact tests for RA designs, introduced in Wei et al.,[15] bound the type I error rate above by the target significance level for all possible parameters under the null hypothesis. This strong form of type I error rate control is often desired in confirmatory trials. Although such exact tests are not novel, they have mostly been limited to fully sequential RA procedures in the literature, whereas our paper considers BRAR designs with a burn-in period and a group-sequential BRAR design in the real-life application. (4) By default, exact inference is computationally more demanding than inference using asymptotic tests. Similar to Baas et al.,[8] we use the efficiency considerations outlined in Jacko[16] to efficiently compute design OCs and exact tests. As a result, we are able to evaluate RA designs with all possible burn-in lengths for trial sizes of up to 240 participants, while exact approaches for RA designs in the literature are often limited to less than 100 participants. (5) Finally, we use our novel findings to provide suggestions on how to choose the burn-in in clinical trials using a BRAR design.

This paper is structured as follows: Section 2 introduces the model and notation for a two-arm RA design with a burn-in period. Section 3 introduces statistical methods. Section 4.1 provides a numerical investigation to assess to what extent the burn-in period can be used to control type I error rates for commonly used tests, Section 4.2 investigates the added value of a burn-in period for BRAR designs using exact tests, Section 4.3 evaluates the impact of the burn-in length on participant benefit metrics, Section 4.4 evaluates the impact of the burn-in length on estimation bias, Section 4.5 gives recommendations for the burn-in length, and Section 4.6 provides a sensitivity analysis of our findings with respect to the choice of prior, where our default is the uniform prior. Section 5 applies our proposed methodology to a real-world BRAR clinical trial that used blocked allocation and early stopping (the ARREST trial). Section 6 summarizes the findings and presents directions for future research. The Supplemental Materials to this paper (available online) contain a summary table of BRAR trials with a burn-in period, results for the same analysis as in Section 4.1 and Section 4.2 but for the Wald test instead of the posterior probability test, critical values of the exact tests of Section 4.2, additional results for the ARREST trial analysis (Section 5), and tables for the sensitivity analysis in Section 4.6.

2 Two-arm RA design with a burn-in period

This section provides the model and notation for a two-arm RA clinical trial with binary outcomes and a burn-in period. In this paper, we will mainly follow the notation of Baas et al.[8] We consider the parametric model where $\boldsymbol{\theta} = (\theta_{\mathrm{C}}, \theta_{\mathrm{D}})$ are the unknown success probabilities of the control treatment (C) and the developmental treatment (D) respectively. In the remainder, the same ordering of the treatment indicators (i.e., first C then D) will be used to construct vectors. Let $\boldsymbol{Y}_{\mathrm{C}} = (Y_{\mathrm{C},i})_{i=1}^{\bar{i}}$ and $\boldsymbol{Y}_{\mathrm{D}} = (Y_{\mathrm{D},i})_{i=1}^{\bar{i}}$ be two sequences of independent Bernoulli random variables, where $\mathbb{P}_{\boldsymbol{\theta}}(Y_{a,i} = 1) = \theta_a$ for $a \in \{\mathrm{C}, \mathrm{D}\}$. The random variable $Y_{a,i}$ denotes the potential outcome under treatment a for trial participant i, while the natural number $\bar{i}$ denotes the fixed trial size.

In a two-arm RA clinical trial, participants $i \in \mathcal{I} = \{1, 2, \dots, \bar{i}\}$ arrive sequentially, and each participant is allocated to a treatment arm A_i, resulting in a response $Y_{A_i,i}$. Let $\boldsymbol{H}_i = (A_1, Y_{A_1,1}, \dots, A_i, Y_{A_i,i})$ be the trial history up to and including participant $i \in \mathcal{I}$. Denote the support set of all trial histories by $\mathcal{H} = \bigcup_{i=0}^{\bar{i}} \mathcal{H}_i$ where $\mathcal{H}_0 = \emptyset$ and

$$\mathcal{H}_i = \{(a_1, y_1, \dots, a_i, y_i) : \ y_w \in \{0,1\}, \ a_w \in \{\mathrm{C}, \mathrm{D}\} \ \ \forall w \in \{1, \dots, i\}\}.$$

An RA procedure is a function $\pi : \mathcal{H} \mapsto [0,1]$, where $\pi(\boldsymbol{H}_i) := \mathbb{P}^{\pi}(A_{i+1} = \mathrm{C} \mid \boldsymbol{H}_i)$. The joint probability measure on the outcomes and allocations induced by the RA procedure will henceforth be denoted by $\mathbb{P}^{\pi}_{\boldsymbol{\theta}}$. With the above notation we can now define the total successes and treatment group sizes up to participant i, defined respectively as:

$$S_{a,i} = \textstyle\sum_{i'=1}^{i} Y_{A_{i'},i'} \mathbb{I}(A_{i'} = a) \qquad N_{a,i} = \textstyle\sum_{i'=1}^{i} \mathbb{I}(A_{i'} = a), \quad a \in \{\mathrm{C}, \mathrm{D}\}.$$

Letting $\mathcal{I}_0 = \{0\} \cup \mathcal{I}$, where $i = 0$ represents the point at which no participant outcomes have been collected (i.e., the start of the trial), we define $\boldsymbol{X}_i = (\boldsymbol{S}_i, \boldsymbol{N}_i)$ as the tuple containing the total successes and treatment group sizes with support

$$\mathcal{X}_i = \{((s'_{\mathrm{C}}, s'_{\mathrm{D}}), (n'_{\mathrm{C}}, n'_{\mathrm{D}})) : \boldsymbol{s}', \boldsymbol{n}' \in \mathcal{I}_0^2, \boldsymbol{s}' \leq \boldsymbol{n}', \ n'_{\mathrm{C}} + n'_{\mathrm{D}} = i\}.$$

The tuple $\boldsymbol{X}_{\bar{i}}$, consisting of the total successes and treatment group sizes at the end of the trial, which are the sufficient (summary) statistics for the Bernoulli (exponential family) model, can be used to determine many estimators and test statistics (e.g., the Wald, score and likelihood ratio test statistic, as well as the maximum likelihood estimator for $\boldsymbol{\theta}$).

In this paper, we will consider the BRAR procedure with a burn-in length b per arm, denoted π_b^{B}. Under this procedure, the first $2b$ trial participants for a burn-in $b \in \{0, \dots, \bar{i}/2\}$ are allocated to treatment in an non-RA manner, such that the treatment group sizes deterministically equal b after allocating participant $2b$, that is, $N_{a,2b} = b$ for all $a \in \{\mathrm{C}, \mathrm{D}\}$. After the burn-in phase has been completed, that is, participant i has to be allocated for $i > 2b$, the procedure becomes RA and participants are allocated to the control treatment with probability equal to the posterior probability that the control treatment is superior. Assuming a common $\mathrm{Beta}(\alpha_0, \beta_0)$ prior for both arms, we have for $i > 2b$ and $\boldsymbol{x}_i \in \mathcal{X}_i$,

$$\pi_b^{\mathrm{B}}(\boldsymbol{x}_i) = \int_{\theta_{\mathrm{C}} \geq \theta_{\mathrm{D}}} \prod_{a \in \{\mathrm{C},\mathrm{D}\}} d\mathrm{Beta}(\theta_a \,;\, s_a(\boldsymbol{x}_i) + \alpha_0, \, n_a(\boldsymbol{x}_i) - s_a(\boldsymbol{x}_i) + \beta_0) \, d\boldsymbol{\theta}, \tag{1}$$

where $d\mathrm{Beta}$ denotes the density of the Beta distribution and the functions s_a, n_a are defined such that $n_a(\boldsymbol{X}_i) = N_{a,i}$ and $s_a(\boldsymbol{X}_i) = S_{a,i}$ for $a \in \{\mathrm{C}, \mathrm{D}\}$.

The RA procedure π_b^{B} is a *Markov RA procedure*,[17] which means that the RA procedure can be written as a function $\pi : \mathcal{X} \mapsto [0,1]$ where $\mathcal{X} = \cup_i \mathcal{X}_i$. Under a Markov RA procedure π, $(\boldsymbol{X}_i)_i$ is a Markov chain with initial state $\boldsymbol{X}_0 = \boldsymbol{x}_0 = ((0,0),(0,0))$, state space $\mathcal{X}$, and transition structure

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{X}_{i+1} = \boldsymbol{x}_{i+1} \mid \boldsymbol{X}_i = \boldsymbol{x}_i) = \begin{cases} \theta_{\mathrm{C}} \cdot \pi(\boldsymbol{x}_i), & \text{if } \boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \partial \boldsymbol{s}_{\mathrm{C}}, \\ (1 - \theta_{\mathrm{C}}) \cdot \pi(\boldsymbol{x}_i), & \text{if } \boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \partial \boldsymbol{f}_{\mathrm{C}}, \\ \theta_{\mathrm{D}} \cdot (1 - \pi(\boldsymbol{x}_i)), & \text{if } \boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \partial \boldsymbol{s}_{\mathrm{D}}, \\ (1 - \theta_{\mathrm{D}}) \cdot (1 - \pi(\boldsymbol{x}_i)), & \text{if } \boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \partial \boldsymbol{f}_{\mathrm{D}}, \\ 0, & \text{else}, \end{cases}$$

where $\partial \boldsymbol{s}_{\mathrm{C}} = ((1,0),(1,0))$ and $\partial \boldsymbol{f}_{\mathrm{C}} = ((0,0),(1,0))$ are the change in $\boldsymbol{X}_i$ after a success and failure for the control arm, and $\partial \boldsymbol{s}_{\mathrm{D}}$, $\partial \boldsymbol{f}_{\mathrm{D}}$ are defined similarly.

For Markov RA procedures, it was shown in Yi[17] that the likelihood can be written as

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{X}_i = \boldsymbol{x}_i) = g^{\pi}_i(\boldsymbol{x}_i) \prod_{a\in\{\mathrm{C,D}\}} \theta_a^{s_a(\boldsymbol{x}_i)}(1-\theta_a)^{n_a(\boldsymbol{x}_i)-s_a(\boldsymbol{x}_i)} \tag{2}$$

where g^{π}_i represents the part of the distribution of $\boldsymbol{X}_i$ found by summing the probabilities of all allocation paths that lead to $\boldsymbol{X}_i$ and is defined as $g^{\pi}_i(\boldsymbol{x}_i) = 0$ for $\boldsymbol{x}_i \in \mathbb{Z}^4 \backslash \mathcal{X}_i$ and otherwise recursively by

$$\begin{aligned} g^{\pi}_0(\boldsymbol{x}_0) &= 1, \\ g^{\pi}_i(\boldsymbol{x}_i) &= \sum_{\substack{a\in\{\mathrm{C,D}\} \\ \partial\boldsymbol{x}_a\in\{\partial s_a, \partial \boldsymbol{f}_a\}}} g^{\pi}_{i-1}(\boldsymbol{x}_i - \partial\boldsymbol{x}_a)\pi(\boldsymbol{x}_i - \partial\boldsymbol{x}_a)^{\mathbb{I}(a=\mathrm{C})}(1-\pi(\boldsymbol{x}_i - \partial\boldsymbol{x}_a))^{\mathbb{I}(a=\mathrm{D})}. \end{aligned}$$

Remark 1. Allocation method during burn-in period. There are multiple procedures to allocate participants during the burn-in period. For small burn-in lengths in particular, it can be very important to aim for a small probability of treatment imbalances during the burn-in period. Several allocation procedures, such as the truncated binomial design, big stick design, permuted block design, and the random allocation rule can be used, where each procedure has its advantages and difficulties.[18] In this remark, we want to emphasize that, while being a relevant topic in practice, it does not matter for our evaluation which allocation method is used during the burn-in period so long as the allocation method allocates b participants to each treatment arm. In that case, we have under the two-arm RA clinical trial model described above that:

$$g^{\pi^{\mathrm{B}}_b}_{2b}(\boldsymbol{x}_{2b}) = \binom{b}{s_{\mathrm{D}}(\boldsymbol{x}_i)}\binom{b}{s_{\mathrm{C}}(\boldsymbol{x}_i)} \cdot \mathbb{I}(n_{\mathrm{C}}(\boldsymbol{x}_{2b}) = b).$$

Hence, assuming the outcomes are i.i.d., the specific allocation procedure used during the burn-in period does not affect the distribution of $\boldsymbol{X}_{\bar{i}}$.

3 Statistical analysis

In this section, we first focus on tests for the null hypothesis

$$H_0 : \theta_{\mathrm{D}} = \theta_{\mathrm{C}} \quad \text{versus} \quad H_1 : \theta_{\mathrm{D}} \neq \theta_{\mathrm{C}},$$

after which we consider the exact calculation of trial OCs. In the main paper we will focus on tests for H_0 that use the PPCS, equal to the right-hand side of (1) with $i = \bar{i}$ (i.e, equal to $\pi^{\mathrm{B}}_0(\boldsymbol{x}_{\bar{i}})$ for all $\boldsymbol{x}_{\bar{i}} \in \mathcal{X}_{\bar{i}}$).

The expression of the likelihood (2) facilitates an exact analysis of the trial data, as well as the exact calculation of trial OCs. In the following two subsections, we first describe exact tests for testing H_0 (Section 3.1), after which we provide methods to efficiently calculate OCs based on (2) (Section 3.2).

3.1 Exact tests

3.1.1 Conditional exact test based on total successes. We first introduce the conditional test based on total successes and show that it is an exact test. The conditional test based on total successes generalizes Fisher's exact test[19] under a design with fixed treatment group sizes to RA designs. The conditional test constructs a critical value from the conditional distribution of the test statistic given the total sum of successes $\mathrm{S}(\boldsymbol{x}_{\bar{i}}) = \sum_{a\in\{\mathrm{C,D}\}} s_a(\boldsymbol{x}_{\bar{i}})$ in the trial, where the nuisance parameter $\theta = \theta_{\mathrm{C}} = \theta_{\mathrm{D}}$ under the null hypothesis is eliminated by conditioning on $\mathrm{S}(\boldsymbol{X}_{\bar{i}})$.

Definition 1 Conditional test based on total successes**.** Let $\mathcal{X}_S(s')$ be the pre-image of $s' \in \mathcal{I}$ under S. A conditional test based on S for test statistic function T, significance level $0 < \alpha < 1$, and RA procedure π rejects when $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \geq \bar{c}(\mathrm{S}(\boldsymbol{X}_{\bar{i}}))$ or $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \leq \underline{c}(\mathrm{S}(\boldsymbol{X}_{\bar{i}}))$ where, for $\bar{\alpha}$, $\underline{\alpha} > 0$ such that $\bar{\alpha} + \underline{\alpha} = \alpha$, we have for all $s' \in \mathcal{I}$

$$\bar{c}(s') = \min\left\{ c \in \bar{\mathrm{T}}(\mathcal{X}_{\mathrm{S}}(s')) : \binom{\bar{i}}{s'}^{-1} \sum_{\boldsymbol{x}_{\bar{i}}\in\mathcal{X}_{\mathrm{S}}(s'):\ \mathrm{T}(\boldsymbol{x}_{\bar{i}})\geq c} g^{\pi}_{\bar{i}}(\boldsymbol{x}_{\bar{i}}) \leq \bar{\alpha} \right\}, \tag{3}$$

and $\underline{c}$ is defined similarly using the left tail and $\underline{\alpha}$ (see, e.g., Baas et al.[8]). In the above, $\mathrm{T}(E)$ denotes the image of $E \subseteq \mathcal{X}_{\bar{i}}$ under T, while $\bar{\mathrm{T}}(E) = \mathrm{T}(E) \cup \{-\infty, \infty\}$.

The next result, proven in Baas et al.,[8] states that the conditional test based on S is exact under the model of the previous section, and will hence be denoted the CX-S test in the following. As the critical value of the conditional exact test based on total successes (CX-S) test is based on the range of the test statistic, this result holds without restrictions on the test statistic function (although the choice of statistic does influence OCs such as power).

Lemma 1. *For every parameter vector $\boldsymbol{\theta}$ satisfying the null hypothesis H_0 we have*

$$\mathbb{P}^{\pi}_{\boldsymbol{\theta}}\Big(T(\boldsymbol{X}_{\bar{i}}) \geq \bar{c}(S(\boldsymbol{X}_{\bar{i}})) \text{ or } T(\boldsymbol{X}_{\bar{i}}) \leq \underline{c}(S(\boldsymbol{X}_{\bar{i}}))\Big) \leq \alpha.$$

3.1.2 Unconditional exact test. In this subsection we discuss an unconditional test for RA designs, generalizing Barnard's test.[20] The unconditional test uses a critical value that bounds the highest rejection rate under the null hypothesis by the significance level.

Definition 2 Unconditional test**.** An unconditional test for null hypothesis H_0 given a test statistic function T, RA procedure π, and significance level $0 < \alpha < 1$ rejects when $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \geq \bar{c}$ or $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \leq \underline{c}$ where, for $0 < \bar{\alpha}$, $\underline{\alpha} < 1$ such that $\bar{\alpha} + \underline{\alpha} = \alpha$, we have (for $\mathbb{P}^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{X}_{\bar{i}} = \boldsymbol{x}_{\bar{i}})$ given in (2))

$$\bar{c} = \min\left\{ c \in \bar{\mathrm{T}}(\mathcal{X}_{\bar{i}}) : \max_{\substack{\boldsymbol{\theta}\in[0,1]^2,\\ \theta_{\mathrm{C}}=\theta_{\mathrm{D}}}} \sum_{\boldsymbol{x}_{\bar{i}}\in\mathcal{X}_{\bar{i}}:\, \mathrm{T}(\boldsymbol{x}_{\bar{i}})\geq c} \mathbb{P}^{\pi}_{\boldsymbol{\theta}}(\boldsymbol{X}_{\bar{i}} = \boldsymbol{x}_{\bar{i}}) \leq \bar{\alpha} \right\}, \tag{4}$$

and $\underline{c}$ is defined similarly using the left tail and $\underline{\alpha}$ see, for example, Baas et al.[8]

The next result, which follows immediately from Definition 2 as the maximum rejection rate over the null set bounds the rejection rate at any point in the null set, states that the unconditional test is exact under the model of Section 2. The unconditional test will be denoted by the *UX test* in the following.

Lemma 2. *Under H_0 it holds that $\mathbb{P}^{\pi}_{\boldsymbol{\theta}}\Big(T(\boldsymbol{X}_{\bar{i}}) \geq \bar{c}$ or $T(\boldsymbol{X}_{\bar{i}}) \leq \underline{c}\Big) \leq \alpha$, where $\bar{c}$, $\underline{c}$ are as given in Definition* 2.

Algorithm 2 in Baas et al.[8] can be used to calculate $\bar{c}, \underline{c}$ up to a desired precision.

The UX test is defined similarly to the commonly-used *calibrated* test, where the distribution of the test statistic under a parameter configuration $\boldsymbol{\theta}'$ under H_0 is used to determine a critical value.

Definition 3 Calibrated test**.** A calibrated test given a test statistic function T and parameter vector $\boldsymbol{\theta}' \in [0,1]^2$ such that $\theta'_{\mathrm{C}} = \theta'_{\mathrm{D}}$, RA procedure π, and significance level $0 < \alpha < 1$ rejects when $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \geq \bar{c}$ or $\mathrm{T}(\boldsymbol{X}_{\bar{i}}) \leq \underline{c}$ where, for $0 < \bar{\alpha}$, $\underline{\alpha} < 1$ such that $\bar{\alpha} + \underline{\alpha} = \alpha$, we have

$$\bar{c} = \min\left\{ c \in \bar{\mathrm{T}}(\mathcal{X}_{\bar{i}}) : \sum_{\boldsymbol{x}_{\bar{i}}\in\mathcal{X}_{\bar{i}}:\, \mathrm{T}(\boldsymbol{x}_{\bar{i}})\geq c} \mathbb{P}^{\pi}_{\boldsymbol{\theta}'}(\boldsymbol{X}_{\bar{i}} = \boldsymbol{x}_{\bar{i}}) \leq \bar{\alpha} \right\}, \tag{5}$$

and $\underline{c}$ is defined similarly using the left tail and $\underline{\alpha}$, while $\mathbb{P}^{\pi}_{\boldsymbol{\theta}'}(\boldsymbol{X}_{\bar{i}} = \boldsymbol{x}_{\bar{i}})$ is given in (2).

The calibrated test is often applied to ensure type I error rate control for the PPCS test under an assumed parameter vector $\boldsymbol{\theta}'$ (see, for example, Du et al.,[9] Yannopoulos et al.,[21] Viele et al.[11,12]). As there is no type I error rate guarantee for this test when the true parameter vector is different from the parameter $\boldsymbol{\theta}'$ the test is calibrated for, the calibrated test is not exact under H_0, though it can be seen as an exact test for the null hypothesis $H'_0 : \boldsymbol{\theta} = \boldsymbol{\theta}'$. Although not considered in this paper, one can also calibrate a test to a strict subset $\tilde{\Theta}_0 \subset [0,1]$ which leads to an (intermediate) exact test for all success rates in $\tilde{\Theta}_0$.

3.2 OCs and their exact calculation

Apart from exact tests, the likelihood (2) also allows for the calculation of exact OCs, such as the power or type I error rate. An OC can be written as $\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})]$ for a function $f : \mathcal{X}_{\bar{i}} \times [0,1]^2 \mapsto \mathbb{R}$. The OCs considered in the paper will be:

- **Rejection rate**:
  This OC is calculated as $\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})]$ for $f(\boldsymbol{x}_{\bar{i}}, \boldsymbol{\theta}) = \mathbb{I}(\mathrm{T}(\boldsymbol{x}_{\bar{i}}) \leq \underline{c}(\boldsymbol{x}_{\bar{i}})$ or $\mathrm{T}(\boldsymbol{x}_{\bar{i}}) \geq \bar{c}(\boldsymbol{x}_{\bar{i}}))$. For $\delta \in (0, 1)$ let $\Theta_\delta = \{\boldsymbol{\theta} : \theta_{\mathrm{D}} - \theta_{\mathrm{C}} = \delta\}$, then this OC is the type I error rate when $\boldsymbol{\theta} \in \Theta_0$ and power when $\boldsymbol{\theta} \in \Theta_\delta$ for $\delta \neq 0$ under the test based on test statistic T and lower and upper critical value functions $\underline{c}, \bar{c}$. For a significance level α it is desired to have the type I error rate bounded by α, while higher power is better. We denote one minus the type I error rate by the true negative rate (TNR).
- **Expected proportion of allocations on the superior arm (EPASA)**:
  This OC equals
  $$\sum_a \mathbb{E}^{\pi}_{\boldsymbol{\theta}}[n_a(\boldsymbol{X}_{\bar{i}})/\bar{i}]\mathbb{I}(\theta_a = \max_{a'} \theta_{a'}) - \mathbb{I}(\theta_{\mathrm{C}} = \theta_{\mathrm{D}})/2.$$
  The OC represents the proportion of participants on the superior arm. Higher values of EPASA are better.
- **PIWD(φ)**:
  For $\boldsymbol{\theta} \in \Theta_\delta$ and $\delta \neq 0$ this OC, also considered in Thall et al.,[6] equals
  $$\mathbb{P}^{\pi}_{\boldsymbol{\theta}}(n_{\mathrm{C}}(\boldsymbol{X}_{\bar{i}})/\bar{i} > n_{\mathrm{D}}(\boldsymbol{X}_{\bar{i}})/\bar{i} + \varphi)\mathbb{I}(\theta_{\mathrm{D}} > \theta_{\mathrm{C}}) + \mathbb{P}^{\pi}_{\boldsymbol{\theta}}(n_{\mathrm{D}}(\boldsymbol{X}_{\bar{i}})/\bar{i} > n_{\mathrm{C}}(\boldsymbol{X}_{\bar{i}})/\bar{i} + \varphi)\mathbb{I}(\theta_{\mathrm{C}} > \theta_{\mathrm{D}})$$
  and represents the probability of allocating a proportion $\varphi \in [0, 1]$ more participants to the inferior arm than to the superior arm. Lower values of this OC are better. We denote one minus PIWD(φ) by the probability of no imbalance in the wrong direction (PNIWD(φ)).
- **Bias of the treatment effect estimator**:
  This OC equals
  $$\mathbb{E}^{\pi}_{\boldsymbol{\theta}}\left[\hat{\theta}_{\mathrm{D}}(\boldsymbol{X}_{\bar{i}}) - \hat{\theta}_{\mathrm{C}}(\boldsymbol{X}_{\bar{i}})\right] - (\theta_{\mathrm{D}} - \theta_{\mathrm{C}}),$$
  where $\hat{\theta}_a(\boldsymbol{X}_{\bar{i}}) = (s_a(\boldsymbol{X}_{\bar{i}}) + \iota(\boldsymbol{X}_{\bar{i}}))/(N_a + 2\iota(\boldsymbol{X}_{\bar{i}})))$ and $\iota(\boldsymbol{X}_{\bar{i}}) = \mathbb{I}(\min_a n_a(\boldsymbol{X}_{\bar{i}}) = 0)$. This OC represents the expected error in the treatment effect estimate $\hat{\theta}_{\mathrm{D}}(\boldsymbol{X}_{\bar{i}}) - \hat{\theta}_{\mathrm{C}}(\boldsymbol{X}_{\bar{i}})$.

We have from (2):

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})] = \sum_{\boldsymbol{x}_{\bar{i}} \in \mathcal{X}_{\bar{i}}} f(\boldsymbol{x}_{\bar{i}}, \boldsymbol{\theta}) g^{\pi}_{\bar{i}}(\boldsymbol{x}_{\bar{i}}) \prod_{a \in \{\mathrm{C},\mathrm{D}\}} \theta_a^{s_a(\boldsymbol{x}_{\bar{i}})}(1 - \theta_a)^{n_a(\boldsymbol{x}_{\bar{i}}) - s_a(\boldsymbol{x}_{\bar{i}})}, \tag{6}$$

which can be written as $(\boldsymbol{g}^{\pi}_{\bar{i}})^{\top}(\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{p}_{\boldsymbol{\theta}})$ with $\circ$ the Hadamard product and $\boldsymbol{p}_{\boldsymbol{\theta}}$ containing the product-term on the right in the above expression. Hence, we can store $\boldsymbol{g}^{\pi}_{\bar{i}}$ and then only have to take the inner product with $(\boldsymbol{f}_{\boldsymbol{\theta}} \circ \boldsymbol{p}_{\boldsymbol{\theta}})$ for different vectors $\boldsymbol{\theta}$ when we want to calculate (6) for different values of $\boldsymbol{\theta}$.

While in practical applications one might have a more specific idea of the realistic parameter range, we aim to offer a robust picture to highlight the potential variation between parameters even with the same treatment effect difference δ. Hence in the following, we discuss two measures that describe the behaviour of the OCs over the complete parameter space, namely the average over OCs and minimum/maximum over OCs.

3.2.1 Average over OCs. Let $|\Theta|$ be the area or length of Θ and let f be an OC function not depending on $\boldsymbol{\theta}$. Based on (6) the average of the OC $\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}})]$ over Θ is represented by $\mathbb{E}^{\pi}_{\Theta}[f(\boldsymbol{X}_{\bar{i}})]$ and defined as

$$\begin{aligned}\frac{1}{|\Theta|}\int_\Theta \mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}})]d\boldsymbol{\theta} &= \sum_{\boldsymbol{x}_{\bar{i}} \in \mathcal{X}_{\bar{i}}} f(\boldsymbol{x}_{\bar{i}}) g^{\pi}_{\bar{i}}(\boldsymbol{x}_{\bar{i}}) \underbrace{\frac{1}{|\Theta|}\int_\Theta \prod_{a \in \{\mathrm{C},\mathrm{D}\}} \theta_a^{s_a(\boldsymbol{x}_{\bar{i}})}(1 - \theta_a)^{n_a(\boldsymbol{x}_{\bar{i}}) - s_a(\boldsymbol{x}_{\bar{i}})} d\boldsymbol{\theta}}_{\boldsymbol{p}_\Theta(\boldsymbol{x}_{\bar{i}})} \\ &= \sum_{\boldsymbol{x}_{\bar{i}} \in \mathcal{X}_{\bar{i}}} f(\boldsymbol{x}_{\bar{i}}) g^{\pi}_{\bar{i}}(\boldsymbol{x}_{\bar{i}}) \boldsymbol{p}_\Theta(\boldsymbol{x}_{\bar{i}}).\end{aligned} \tag{7}$$

For instance, we have

$$\boldsymbol{p}_{\Theta_\delta}(\boldsymbol{x}_{\bar{i}}) = \int_0^{1-\delta} \frac{\prod_{a \in \{\mathrm{C},\mathrm{D}\}}(\theta_{\mathrm{C}} + \delta \cdot \mathbb{I}(a = \mathrm{D}))^{s_a(\boldsymbol{x}_{\bar{i}})}(1 - \theta_{\mathrm{C}} - \delta \cdot \mathbb{I}(a = \mathrm{D}))^{n_a(\boldsymbol{x}_{\bar{i}}) - s_a(\boldsymbol{x}_{\bar{i}})}}{1 - \delta} d\theta_{\mathrm{C}}, \tag{8}$$

$$\boldsymbol{p}_{\Theta_0}(\boldsymbol{x}_{\bar{i}}) = \mathrm{B}(s(\boldsymbol{x}_{\bar{i}}) + 1, \bar{i} - s(\boldsymbol{x}_{\bar{i}}) + 1), \tag{9}$$

where B is the Beta function. The average OC generalizes the OC at a single parameter vector (as we can take $\Theta = \{\boldsymbol{\theta}\}$) and represents the average value of the OC over a specific part of the parameter space. It hence better represents the behaviour of the OC under RA procedure on average over the part of the parameter space that is of interest. Although not interpreted as a Bayesian measure, the average OC equals the Bayesian average value $\int_{[0,1]^2} \mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})]p(\boldsymbol{\theta})d\boldsymbol{\theta}$ of the OC for a prior density p on $[0,1]^2$ where p equals the uniform prior on Θ_δ (i.e., $p(\boldsymbol{\theta}) = \mathbb{I}(\theta \in \Theta_\delta)/|\Theta_\delta|$). Using the average power or type I error rate to objectively evaluate the performance of a statistical test in the case where there is no prior information, or as a long-term evaluation measure, has been argued for in, for example, Rice,[22] Andrés and Mato,[23] and Best et al.[24] where the first and last paper propose the average OC in a more general Bayesian context.

3.2.2 Grid-approximated minimum and maximum over OCs. In order to find the grid-approximated minimum and maximum OCs, we discretize the set Θ_δ to a finite set $\hat{\Theta}_\delta$. The minimum OC $\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})]$ over $\hat{\Theta}_\delta$ (approximating the minimum OC over Θ_δ) equals $\min_{\theta \in \hat{\Theta}_\delta} \mathbb{E}^{\pi}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_{\bar{i}}, \boldsymbol{\theta})]$. The maximum OC is calculated in a similar vein. The minimum power to indicate the worst-case behaviour of a test has also been considered in Haber[25] although not in addition to the average.

4 The effect of the burn-in length in a Bayesian RA design

In this section, we investigate the effect of the burn-in length on the type I error rate for the calibrated test based on the PPCS (Section 4.1), consider what the added value of using a burn-in period is when using an exact test in a BRAR design (Section 4.2), consider how EPASA and PIWD(0.1) change with the burn-in length (Section 4.3), evaluate the effect of the burn-in length on treatment effect estimation bias (Section 4.4), give a recommendation for choosing the burn-in length (Section 4.5), and provide a sensitivity analysis with respect to the prior used for BRAR, where our default is the uniform prior (Section 4.6). As indicated in Section 1, to purely investigate the effect of the burn-in period length, we limit the focus on two-arm BRAR designs using a burn-in period as the sole tuning parameter, while the trial application in Section 5 considers a trial with blocked allocation, early stopping, and clipped allocation probabilities.

We consider two specific trial sizes $\bar{i} \in \{60, 240\}$, which could represent an early-stage exploratory trial and a small confirmatory trial, respectively. These two numbers were chosen due to their large number of divisors, making them more suitable and probable to be used in designs using blocked allocation than, for example, trial sizes 50 and 250. Following Thall et al.,[6] we take $\varphi = 0.1$ to define PIWD. In the remainder, we set $\hat{\Theta}_\delta = \{\boldsymbol{\theta} \in [0,1]^2 : \theta_C = \theta_D - \delta,\ \theta_D \in \{\delta, \delta + 0.01, \ldots, 1.00\}\}$ and $\underline{\alpha} = \bar{\alpha} = 0.025$. All calculations in this paper are exact and not simulation-based.

The allocation probabilities π^B_b were calculated using Gauss—Kronrod quadrature using the *QuadGK Julia* package see the QuadGK package documentation[26] with absolute tolerance 10^{-3}. Computation of the allocation probabilities for all states $\boldsymbol{x} \in \mathcal{X}$ for $\bar{i} = 240$ took 2950 seconds on a standard laptop (1.7 Ghz, 10 cores, 32 GB RAM). This vector can also be used for $\bar{i} = 60$ and different burn-in parameters b, hence this vector only needs to be calculated once. Computation of $g_{\bar{i}}^{\pi^B_b}$, where we loop over the same set of states but perform a simpler calculation than numerical integration, took 173 seconds for $\bar{i} = 240$ and $b = 0$ (which is the value of b with the longest computation time). The Gauss–Kronrod quadrature with a (default) relative tolerance $\sqrt{\epsilon}$ (where ϵ is the machine epsilon in Julia, for example, for one device this was around $2.2 \cdot 10^{-16}$) was used to compute the values $p_{\boldsymbol{\Theta}_\delta}(\boldsymbol{x}_{\bar{i}})$ for every state $\boldsymbol{x}_{\bar{i}} \in \mathcal{X}_{\bar{i}}$ and $\delta \in \{0.1, 0.2, 0.4\}$, while (9) was used to compute average OCs under H_0. For $\bar{i} = 240$ the former calculations took 46 , 44 , and 40 seconds, respectively, while the latter calculation took 2 seconds (as no numerical integration is needed). The PPCS statistic for each final state was calculated with an absolute tolerance 10^{-6}, which took 183 seconds for $\bar{i} = 240$. Based on Jacko[16] the amount of values $\pi^B_b, g_i^{\pi^B_b}$ to calculate (equal to $|\mathcal{X}|$) is of order $\mathcal{O}(\bar{i}^4)$, while the amount of end-states (hence the amount of average probabilities and PPS values to calculate) grows with order $\mathcal{O}(\bar{i}^3)$.

4.1 Issues arising from the application of commonly used tests for BRAR designs

We will consider the calibrated test based on the PPCS defined in Section 3.1, where we calibrate this test to the parameter configuration $\theta_C = \theta_D = 0.5$ which induces the highest outcome variance under H_0.

Figure 1 shows the type I error rate profile for the calibrated test for RA procedure π^B_b. When no burn-in is used, the calibrated critical value does not control the type I error rate well, reaching a type I error rate around 14.53% for common success rates above 0.9 and $\bar{i} = 60$, almost three times the nominal significance level. Designs with a larger burn-in, for example, those where $b \geq \bar{i}/4$, show a more balanced type I error rate profile reaching a lower maximum value. For $\bar{i} = 60$ the type I error rate is under control when $b = \bar{i}/2$ for all evaluated parameter values under the null, while this is not the case for $\bar{i} = 240$.

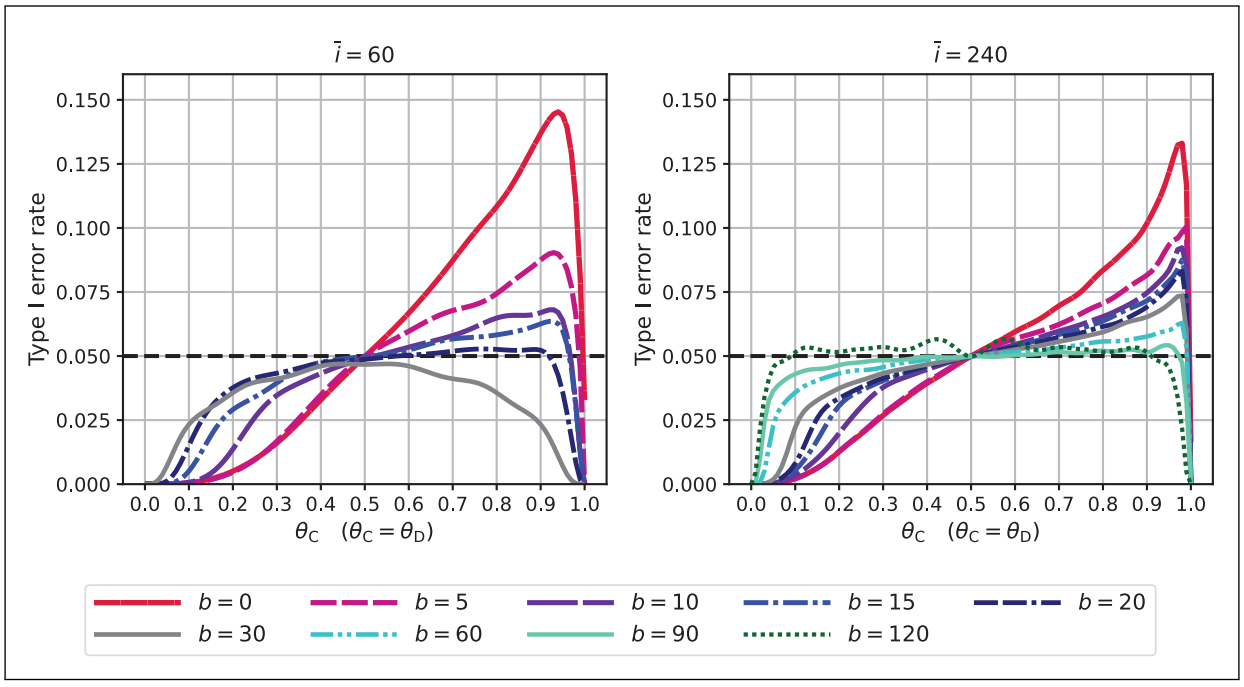


Figure 1. Type I error rate profiles for the Bayesian response-adaptive randomization design with calibrated test (calibrated for $\theta_C = \theta_D = 0.5$) based on the posterior probability of control superiority for trial sizes $\bar{i} = 60$ and $\bar{i} = 240$, across $\theta_C = \theta_D \in \{0.00, 0.01, \ldots, 1.00\}$ for different burn-in lengths b. The significance level α was set to 0.05 (indicated by the horizontal dashed line).

Table 1. Maximum type I error rate (in %) across the whole parameter space (using a grid approximation) of the calibrated test (calibrated for $\theta_C = \theta_D = 0.5$) based on the posterior probability of control superiority (significance level $\alpha = 0.05$) under the fully sequential BRAR design as the burn-in length varies.

BP	0	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$\bar{i} = 20$	**12.72**	**13.71**	**10.68**	9.54	8.02	5.69	6.20	5.13	6.22	5.43	5.00
$\bar{i} = 40$	**14.59**	**12.48**	8.55	7.07	6.23	6.39	6.02	5.84	5.40	4.91	6.09
$\bar{i} = 60$	**14.53**	**11.04**	8.59	7.20	6.76	6.36	5.47	5.62	5.15	4.99	4.69
$\bar{i} = 80$	**15.07**	9.83	8.37	7.42	6.60	6.21	5.72	5.31	5.47	5.23	5.04
$\bar{i} = 100$	**14.56**	9.44	8.10	7.23	6.95	6.29	5.78	5.49	5.62	5.23	5.07
$\bar{i} = 240$	**13.31**	9.01	7.82	7.23	6.64	6.30	5.64	5.38	5.17	5.09	5.66

BP is the proportion of the trial within the burn-in stage, given by BP$= 2b/\bar{i}$. BP$= 1.0$ corresponds to equal allocation. Type I error rates above 6% are indicated in bold, with the severity of the inflation emphasized by a gradient (red being worst). BP: burn-in proportion; BRAR: Bayesian response-adaptive randomization.

The variability of the type I error rates over different common success rates can be explained by the discreteness of the binary PPCS test coupled with the relatively small trial sizes considered. The asymmetry present for low burn-in lengths b can be explained intuitively. The PPCS will be close to 1/2 when the common success rate is small since the RA procedure is likely to switch between treatments when a failure is recorded (inducing balance). On the other hand, for high common success rates, we are more likely to (erroneously) favour one treatment because we keep on recording successes for that treatment. In cases where almost all successes are on one arm, the PPCS roughly equals the common success rate or one minus this value (one of the success rates has a uniform distribution, while the other distribution has low variance) hence the rejection rate grows in the common success rate.

To keep the type I error rate roughly under 6%, Figure 1 suggests the burn-in length should be more than a quarter of the trial size (i.e., $b > \bar{i}/4$). For $\bar{i} \geq 60$ we show that this indeed seems to be a valid rule of thumb in the case of a two-arm BRAR and when testing H_0 using the calibrated PPCS test. Table 1 shows the grid-approximated maximum type I error rate of the calibrated PPCS test versus b for $\bar{i} \in \{20, 40, 60, 80, 100, 240\}$. Table 1 shows that when the burn-in proportion (BP) is higher than or equal to 0.6, the type I error rate is controlled at 6% for $\bar{i} \geq 60$.

In conclusion, the use of calibrated tests leads to a substantial risk of type I error inflation under parameter misspecification, and this problem cannot be fully eliminated by increasing the burn-in length. Due to regulatory demands for strict type I error rate control, we study burn-in length's effect on OCs with an exact test.

Section 2 of the Supplemental Materials shows the same evaluation when instead of considering the PPCS as the test statistic, we use the Wald statistic with the Agresti-Caffo adjustment as defined in Equation (14) Baas et al.[8] We use a standard asymptotic two-sided Wald test (significance level 5%) instead of a calibrated test, common in theory but rare in BRAR trials with burn-in. (see, e.g., Supplemental Table 1). The finite-sample and asymptotic properties of the Wald test under RA procedures are discussed in Baldi Antognini et al.[27]

The main differences between the results for PPCS and the Wald statistic are that the type I error rate profile for the Wald test is more symmetrical, while the maximum type I error rate inflation for $b = 0$ is slightly lower than for PPCS, around 10%-12% (Supplemental Figure 1), due to the more symmetrical type I error rate profile. The commonalities with the PPCS test are that the rule $b > \bar{i}/4$ works for controlling the type I error rate at 6% (in this case, even for all values of $\bar{i}$, see Supplemental Table 2).

4.2 The added value of a burn-in when exact tests are applied in BRAR designs

This section presents the results for the exact tests given in Section 3.1 using the PPCS statistic under RA procedure π_b^{B} for different burn-in lengths b. Note that for each burn-in length b and type of exact test, a different critical value is derived. To give some intuition, we provide some of these critical values in Section 3 of the Supplemental Materials, where we also include critical values for the Wald test. We provide a power comparison of the CX-S test and UX test, as both of them control the type I error rate.

The graphs in the first row of Figure 2 show the average, and grid-approximated minimum and maximum type I error rate across different burn-in lengths. As expected with exact tests, both the UX and CX-S tests maintain strict type I error control at 5% across all burn-in lengths, ensuring a maximum type I error rate of 5%. As b increases, the average type I error rate for the UX test roughly increases, while it roughly decreases for the CX-S test. Hence, the UX (CX-S) test is likely overly conservative when the burn-in length is small (large). The minimum and maximum type I error rate for the CX-S test shows a highly non-smooth behaviour, which follows from the fact that the critical values for these tests are different for each burn-in length, and by the discreteness of binary tests in general. Increasing b exacerbates this behaviour, leading to a narrower range of treatment group sizes. Figure 2 shows that the type I error rate for the UX test varies between 0% and 5% for all burn-in lengths, yet the CX-S test shows much less variability over the parameter space and is less conservative than the UX test for smaller burn-in lengths, where the minimum type I error rate is around or higher than 3% for b up to $\bar{i}/4$. The maximum type I error rate for the CX-S test decreases for larger values of b, where the maximum type I error rate for CX-S at $b = \bar{i}/2$ roughly equals the average type I error rate of the UX test for $b = \bar{i}/4$ for both $\bar{i} \in \{60, 240\}$, hence the CX-S test is likely overly conservative for larger burn-in lengths. We note that for $b = \bar{i}/2$, the CX-S test equals Fisher's exact test,[19] while the UX test equals Barnard's test.[20] Fisher's exact test is known to be more conservative than Barnard's test in the case of equal treatment group sizes and relatively small sample sizes.[28]

The three bottom rows in Figure 2 show how the minimum, maximum and average power to under treatment effects $\delta \in \{0.1, 0.2, 0.4\}$ vary across different burn-in lengths. The average power for $b = \bar{i}/2$ is higher than for $b = 0$ for both exact tests, however, the average power does not increase monotonically in the burn-in length. For the CX-S test, the average power can furthermore attain a maximum at $b < \bar{i}/2$, for example, when $\bar{i} = 240$ the maximum occurs around $b = 100$.

Figure 2 shows that for $\bar{i} = 60$ the average power for the CX-S test is higher than that of the UX test when $b < \bar{i}/4$, while it is lower for $b \geq \bar{i}/4$. We note that, generally, the power differences between the two tests are smaller in the latter situation ($b \geq \bar{i}/4$) than in the former situation ($b < \bar{i}/4$). For $\bar{i} = 240$ and $\delta \in \{0.1, 0.2\}$ the CX-S test has higher average and minimum power up to $b = 80$. This agrees with Baas et al.,[8] showing CX-S test's higher power over the UX test for many RA procedures with a high degree of response-adaptiveness. A potential explanation of this phenomenon lies in the high dependency (for more aggressive RA procedures) of the distribution of treatment group sizes on total successes. On the contrary, for larger burn-in lengths this is not the case and the power of the CX-S test suffers from the higher discreteness of the conditional distribution of the test statistic. Figure 2 shows a large spread in the power for the UX test for, e.g, $\delta = 0.2$ and $\bar{i} = 60$ and $\delta = 0.1$ and $\bar{i} = 240$, as indicated by the maximum and minimum power values. This could be due to the highly asymmetrical type I error rate profile for the PPCS test, hence the UX PPCS test is overly conservative for certain parameter values. Lastly, Figure 2 shows that the minimum power for the CX-S test is higher than the minimum power for the calibrated test for at least a few burn-in lengths (indicated by dashed lines). This outperformance mainly happens for low success rates, where the calibration test was also shown to be conservative. Hence, the CX-S test has, for some parameter configurations, higher power than the commonly used calibration test with the added benefit of being exact.

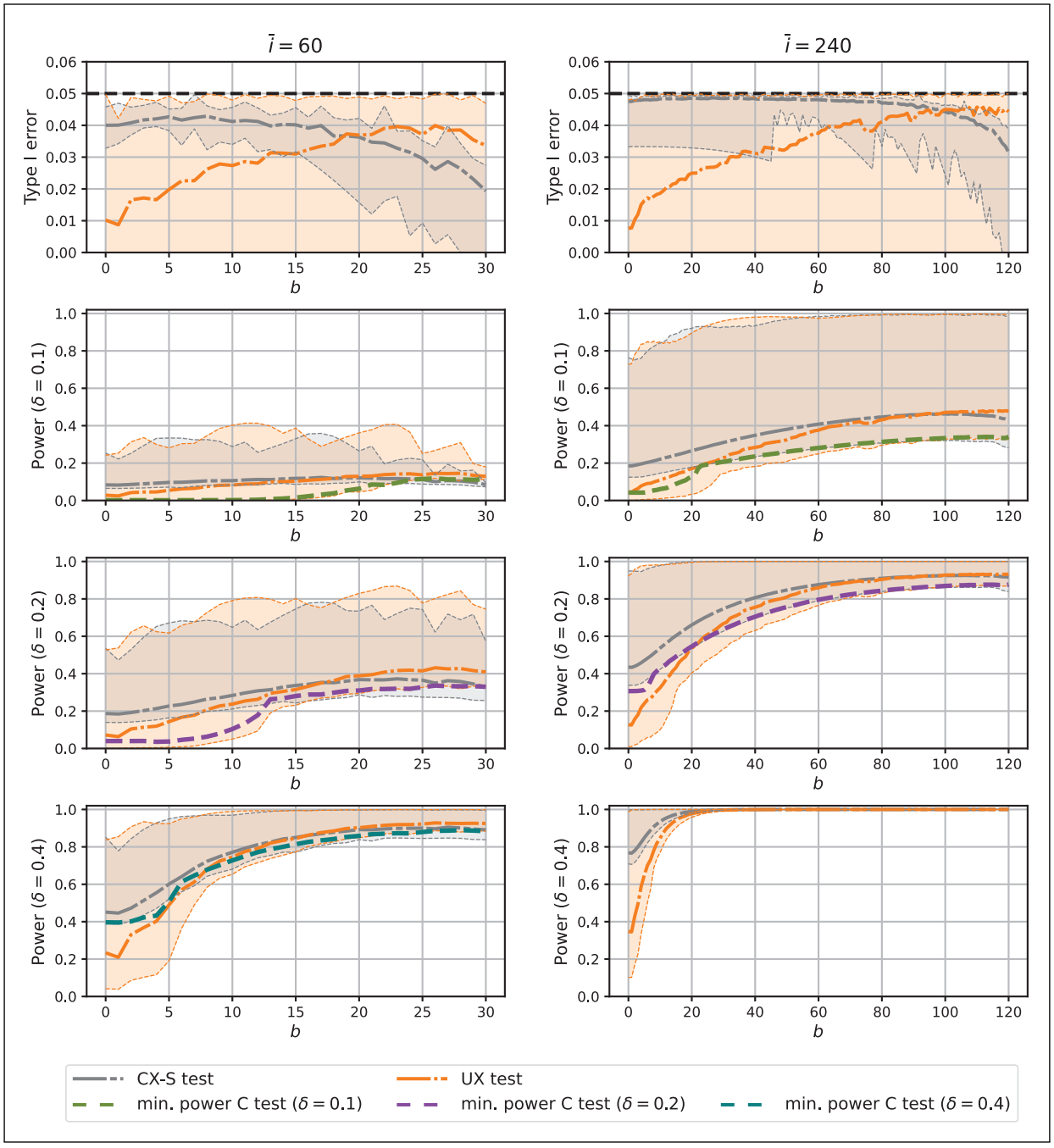


Figure 2. Type I error rate and power under treatment effects $\delta \in \{0.1, 0.2, 0.4\}$ for the Bayesian response-adaptive randomization design using the UX test and CX-S test based on the posterior probability of control superiority with trial size $\bar{I} = 60$ and $\bar{I} = 240$ across different burn-in lengths. For comparison, we have plotted the (grid-approximated) minimum power for the calibrated (C) test for $\bar{I} = 60$, $\delta \in \{0.1, 0.2, 0.4\}$ and $\bar{I} = 240$, $\delta \in \{0.1, 0.2\}$. The averages for each treatment effect δ and fixed burn-in length are represented by thick lines, while the minimum and maximum are presented by thin dotted lines and the ribbons. The significance level α was set to 0.05. UX: unconditional exact; CX-S: conditional exact test based on total successes.

Supplemental Figure 2 shows the evaluation above for the Wald test instead of the PPCS test. As the type I error rate profile of the asymptotic Wald test is more regular than that of the calibrated PPCS test for small burn-in lengths, the UX Wald test is less conservative for small burn-in lengths than the UX PPCS test. Comparing the CX-S Wald and CX-S PPCS tests, the type I error results are very similar. The CX-S Wald test often shows higher maximum power than the UX Wald test, while less often showing higher minimum power than the asymptotic Wald test (only for $\delta = 0.10, \bar{I} = 60$) than under

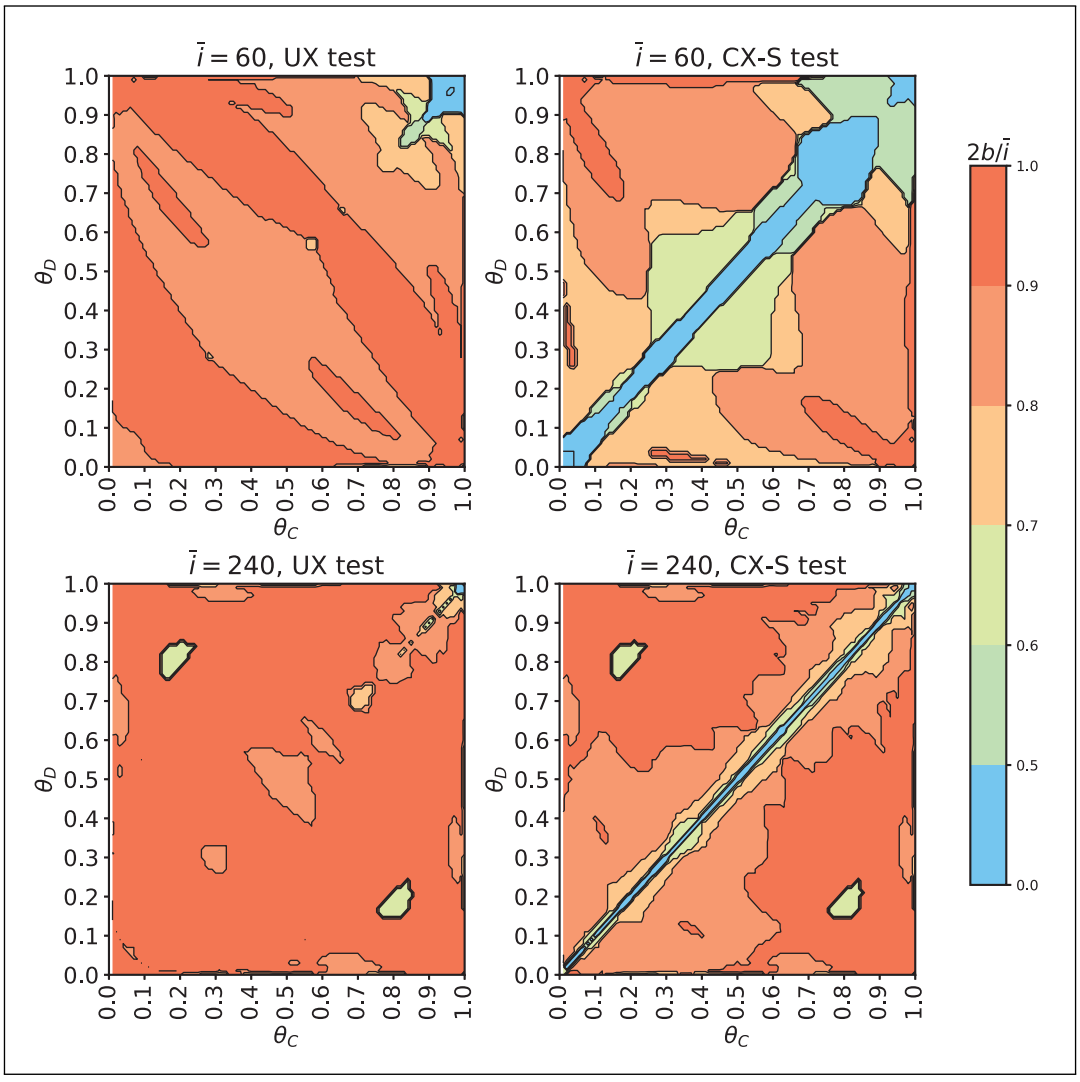


Figure 3. Optimal burn-in proportions ($2b/\bar{i}$) in terms of power of the Bayesian response-adaptive randomization design using the UX and CX-S test based on the posterior probability of control superiority for every parameter configuration in the grid $\theta_C, \theta_D \in \{0.00, 0.01, \ldots, 1.00\}$. The considered trial sizes are $\bar{i} = 60$ and $\bar{i} = 240$ participants. UX: unconditional exact; CX-S: conditional exact test based on total successes.

the calibration PPCS test comparison. The CX-S test again shows higher average power than the UX test when $b < \bar{i}/4$ and vice versa for $b \geq \bar{i}/4$, and maximum average power values are again found for $b < \bar{i}/2$.

4.2.1 Optimal burn-in proportion across the parameter space. Figure 3 shows the *optimal burn-in proportion in terms of power* (P-OBP) for the UX and CX-S test for each parameter configuration $\theta_C, \theta_D \in \{0.00, 0.01, \ldots, 1.00\}$.

As the calibrated test does not control type I error rates for every parameter configuration, this test was not considered in this evaluation. Note that the P-OBP can only be exactly computed using our calculation method, finding these maxima using simulation is prohibitive through Monte Carlo errors combined with the fluctuating and non-smooth behaviour of the power in the burn-in length.

The patterns in Figure 3 are hard to explain. Due to the discreteness of the binary tests, with changing critical values as the burn-in length changes, the power curve as a function of the burn-in length has a highly non-smooth behaviour and can suddenly jump to a high value. What first stands out in Figure 3 is that the P-OBP for power is often less than 1.0 for $\bar{i} = 60$. For the UX test it is often higher than 0.8, whereas for the CX-S test the P-OBP is often lower (especially around the diagonal). For $\bar{i} = 240$ the P-OBPs are higher; the values for the UX test are closer to 1.0 than for $\bar{i} = 60$ and the values for the CX-S test are again lower (especially around the diagonal).

Supplemental Figure 3 contains the P-OBP plots when the Wald statistic is used instead of PPCS, where overall the findings are the same as for PPCS. The subfigures for the CX-S test are very similar to the ones in Figure 3 where, upon inspecting the results, numerical differences were seen although they were very small. This could be explained by the fact that the CX-S test fixes total successes, hence the behaviour of this test is less sensitive to the choice of the test statistic. Larger differences are indeed seen for the UX test, where for $\bar{i} = 60$ lower P-OBPs are seen for the Wald test than for the PPCS test. In conclusion, for the considered trial sizes the P-OBP for power depends heavily on the choice of test and is often not equal to one.

4.3 Impact of burn-in length on participant benefit metrics

One of the the main arguments in the literature for using RA procedures is the potential to allocate more participants to the better treatment. This section investigates the behaviour of EPASA and PIWD(0.1) as we vary the burn-in length. Note these metrics are not inference-related, hence test-independent.

4.3.1 EPASA and PIWD(0.1) for different burn-in lengths. The top row in Figure 4 shows the average and grid-approximated minimum and maximum EPASA versus the burn-in length for treatment effects $\delta \in \{0.1, 0.2, 0.4\}$. As expected, for every treatment effect δ, lower burn-in lengths give a higher average, minimum and maximum EPASA, and increasing the sample size and increasing δ raises EPASA. A surprising observation here is that EPASA does not necessarily decrease as the burn-in length increases, for example, the maximum average EPASA occurs for $b = 1$ when $\bar{i} = 60, \delta = 0.1$. The average, minimum and maximum EPASA increases initially, then decreases slowly, and then it decreases linearly as b increases. One explanation is that a longer burn-in will more likely identify the better arm, leading to almost all participants being allocated to the best arm in the RA phase, tracing out a straight line on the EPASA graph.

The bottom row in Figure 4 shows that the average and grid-approximated minimum and maximum PIWD(0.1) decrease as we increase the burn-in length. The lines are in agreement with Robertson et al.,[4] who state that more aggressive RA procedures are more likely to have higher probabilities of imbalances in the wrong direction, which in our case would correspond to a lower burn-in length. As expected, for any fixed δ, lower burn-in lengths give a higher PIWD(0.1), while increasing sample size and increasing δ reduces this OC. Initially for small burn-in lengths, PIWD(0.1) decreases in a roughly linear fashion, then towards the end, the PIWD(0.1) drops steeply towards 0. This is because for $b > \bar{i}(1 - \varphi)/2$ it is impossible to achieve imbalance in the wrong direction.

For both metrics, the variation over the parameter space for $\delta = 0.1$ is larger than the variation for $\delta \in \{0.2, 0.4\}$. It is difficult for the RA procedure to detect a small treatment effect. That said, for $\delta \geq 0.2$ and $\bar{i} = 240$ the PIWD(0.1) is less than 1% regardless of the burn-in length, so we almost certainly improve in-trial participant benefit when the treatment effect is not too small given the trial size.

While it may seem alarming that the average PIWD(0.1) for $\bar{i} = 60, \delta = 0.1$ can get as high as 20% for small burn-in lengths, we highlight that the interpretation of the PIWD metric is not straightforward. For instance, for $\bar{i} = 60, \theta_C = 0.4, \theta_D = 0.5$ we find PIWD(0.1) = 0.22, meaning that the probability of having at least 33 participants allocated to the control treatment, leading to at least 0.3 more expected treatment failures in total than under fixed equal allocation, is 22%. The severity of this imbalance still depends on the probability of having an even worse misallocation, such as 65% of participants being allocated to the worst arm. If this is zero, then the situation would not be severe after all, since for fixed equal allocation and $\delta = 0.1$ the probability of having 10% less expected successes for 50% of participants is 100%. Hence, although we are only considering one imbalance measure, we recommend looking at different imbalance measures when choosing a burn-in length (e.g., PIWD(φ) for different values φ).

4.3.2 PIWD(0.1) across the parameter space. Figure 5 presents a heatmap of values of PIWD(0.1) across the parameter space, where we restrict to a grid of values $\theta_C, \theta_D \in \{0.01, 0.02, \ldots, 1.00\}$ with $\theta_C \neq \theta_D$ which gives a clearer picture on where the imbalance occurs.

As also indicated in Robertson et al.[4] (who restricted to the case $\theta_C = 0.25$, $\theta_D \geq \theta_C$ and $\bar{i} = 200$), Figure 5 shows that the PIWD(0.1) is highest near the diagonal $\theta_C = \theta_D$. The graphs on the bottom row correspond to higher burn-in lengths and show a smaller PIWD(0.1) compared to the top row. Hence, the maximum of PIWD(0.1) shown in Figure 4 happens around the point $(1/2-\delta/2, 1/2+\delta/2)$. One of the limitations of the PIWD(0.1) metric is that it is large when the difference in treatments is smallest, and so the cost of allocating to the wrong treatment is lowest, as noted in Robertson et al.[4] Inspection of the numerical results shows that the heatmaps are not symmetrical around the line $\theta_C + \theta_D = 1$, which is not noticeable from visual inspection. An explanation could be that the arm with the highest variance switches when comparing a parameter vector with its reflection along the line $\theta_C + \theta_D = 1$, hence for one of these scenarios it is more difficult to identify the best arm. For example, comparing the vector $(0.4, 0.5)$ with $(0.5, 0.6)$ the superior arm has the highest variance for the first parameter vector, while for the second it has the smallest variance.

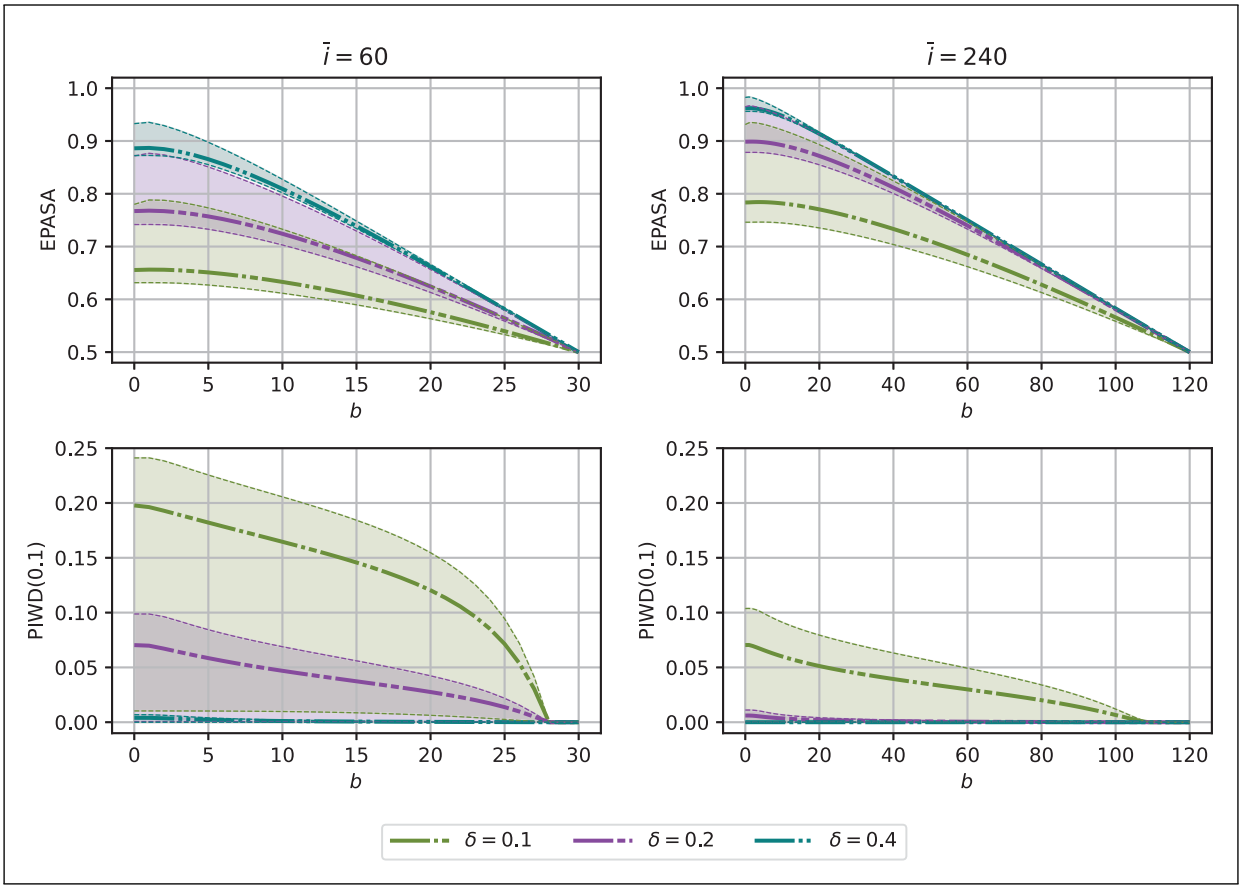


Figure 4. EPASA and PIWD(0.1) under different burn-in lengths b and treatment effects $\delta \in \{0.1, 0.2, 0.4\}$.The averages for each treatment effect δ and fixed burn-in length are represented by thick lines, while the minimum and maximum are presented by thin dotted lines and the ribbons. EPASA: expected proportion of allocations on the superior arm; PIWD: probability of an imbalance in the wrong direction.

4.4 Bias of treatment effect estimate

This section investigates the bias of the treatment effect estimator (Section 3.2) versus the burn-in length. Figure 6 displays the average and grid-approximated minimum and maximum bias versus the burn-in length for treatment effects $\delta \in \{0.1, 0.2, 0.4\}$. As one might expect from Bowden and Trippa,[29] the treatment effect estimator largely has a positive bias, while negative values also occur for $\bar{i} = 60$ when the developmental treatment has a larger outcome variance than the control treatment (i.e., when $\theta_D \leq 0.5$). Surprisingly, the maximum and average bias, while starting at higher values for $b = 0$, decreases faster in b for larger treatment effects δ. In addition, a lower bias is seen for $b = 0$ than for $b = 1$ which might be due to the case distinction made in the estimator when $\min_a n_a(\boldsymbol{X}_{\bar{i}}) = 0$, which decreases the variance of $\hat{\theta}(\boldsymbol{X}_{\bar{i}})$. Figure 6 shows that, to reduce bias, setting a larger burn-in is better (in case we assume $b > 0$). For a fixed burn-in length b, if we increase $\bar{i}$, the bias tends to be larger on average and in the maximum case, indicating that it might be a good idea to make the burn-in length a function of the trial size $\bar{i}$ to make it satisfy the same upper bound. We note that, for a trial setting at hand, bias reduction techniques such as the ones proposed in Bowden and Trippa[29] can be used, which require one to make a (case specific) choice of debiasing technique and bias-variance trade-off.

4.5 Recommendation for burn-in

The general takeaway from our analysis is that while the addition of a burn-in period yields a more balanced type I error rate profile for common non-exact tests (calibration and asymptotic), it cannot guarantee type I error rate control across the parameter space. Non-exact tests applied with insufficient burn-in lengths can exhibit substantial type I error rate inflation (approaching three times the nominal significance level). Therefore, practitioners must exercise caution when using calibrated or asymptotic tests, unless their statistical properties have been thoroughly investigated through simulation

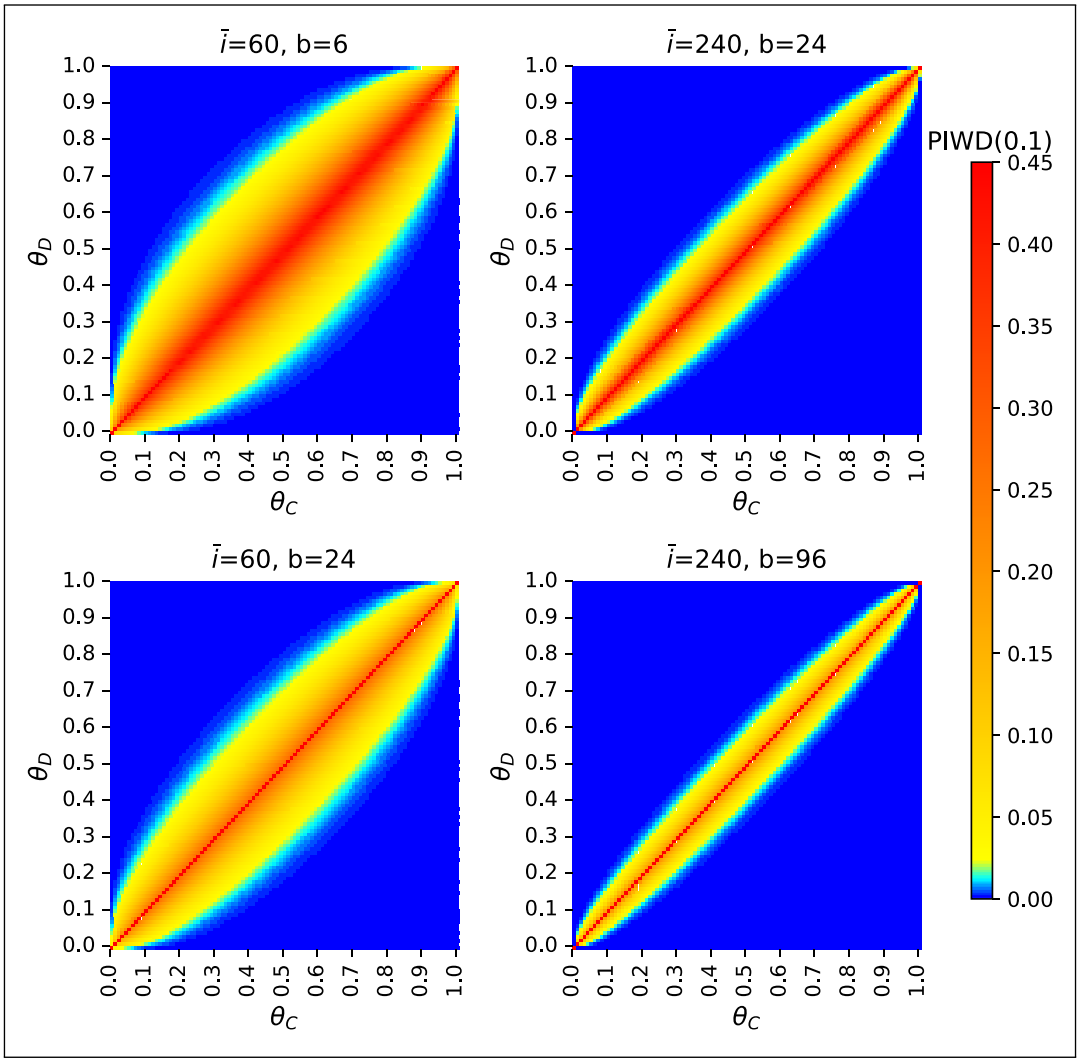


Figure 5. Probability of an imbalance in the wrong direction (PIWD) across the parameter space for $\theta_C, \theta_D \in \{0.00, 0.01, \ldots, 1.00\}$. burn-in lengths $b \in \{0.1\bar{i}, 0.4\bar{i}\}$ are considered, as well as trial sizes $\bar{i} \in \{60, 240\}$. For continuity purposes, PIWD(0.1) is set to the maximum value 0.45 on the diagonal $\theta_C = \theta_D$.

or exact methods. For exploratory settings where strict type I error rate control is not paramount, a practical rule of thumb is to set $b = \bar{i}/4$ for the BRAR design, which leads to a balanced type I error rate and robust power. If estimation quality is the primary concern, we recommend choosing the smallest burn-in length greater than $b = \bar{i}/4$ that yields a desired bound on the maximum bias.

In settings where strict type I error control is required (e.g., in confirmatory settings), we recommend considering exact tests. In such settings, type I error rate control is guaranteed, and the burn-in length can be set such that a sufficient power is reached. As we saw that for burn-in lengths at least up to $\bar{i}/4$ the CX-S test has higher average power than the UX test, the CX-S test could be preferred in designs that at least target a moderate amount of response-adaptivity. The above guidelines do not consider the PIWD(0.1) due to the difficulties with the interpretation of this OC (as indicated in Section 4.3). We note that if type I error rate and power are not deemed important, it might still be better to use a burn-in, as the maximum average EPASA for our considered trial sizes occurs at small but positive burn-in lengths.

The guidelines above mainly focus on the average OCs. Since different trials have different priorities, practitioners may also set threshold values for the OCs which suit their needs, and then investigate which burn-in length satisfies the conditions. As we saw that the power and EPASA can have a non-monotonic behaviour in the burn-in length, it is recommended to inspect more burn-in lengths than just $b = 0$ and $b = \bar{i}/2$, and inspect at least the values of b close to these two endpoints. This recommendation is further supported by the optimal burn-in proportions presented in Figure 3

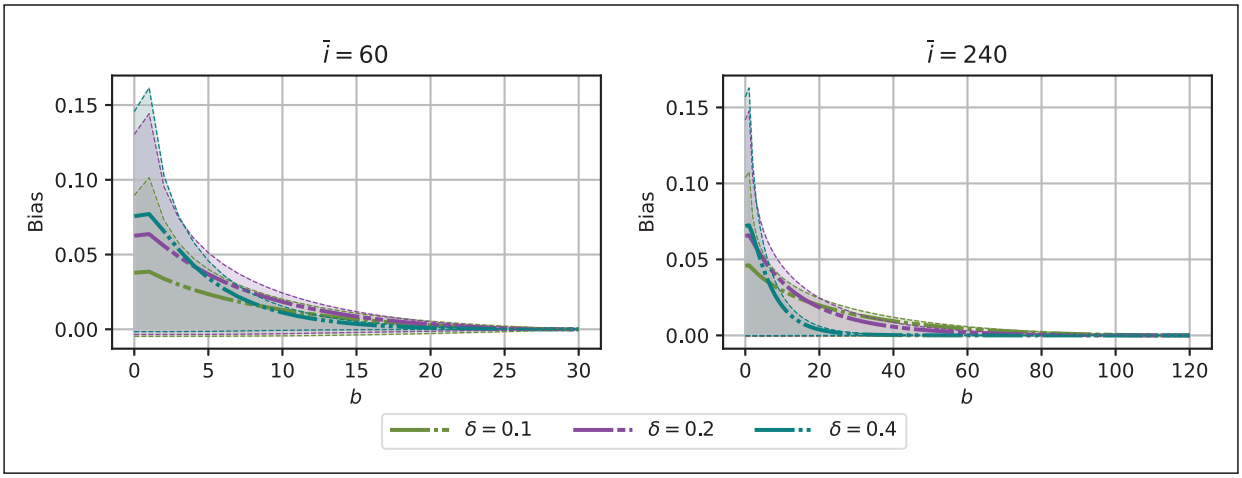


Figure 6. Bias under treatment effects $\delta \in \{0.1, 0.2, 0.4\}$ for the Bayesian response-adaptive randomization design with trial size $\bar{i} = 60$ and $\bar{i} = 240$ across different burn-in lengths b. The averages for each treatment effect δ and fixed burn-in length are represented by thick lines, while the minimum and maximum across the parameter space are presented by thin dotted lines and the ribbons.

and Supplemental Figure 3, where many burn-in lengths are found that maximize power which are strictly between 0 and $\bar{i}/2$.

Let us now give an example in a more specific setting. Suppose we want to run a trial with $\bar{i} = 240$ participants, hoping to achieve a minimum power of 80% under a treatment effect $\delta = 0.2$ at a 5% significance level, while keeping PIWD(0.1) $< 5\%$ in $\Theta_{0.2}$. If strict type I error control is required, then we can use exact tests. Figure 2 shows that minimum power can be attained using the CX-S test with $b = 63$ or UX test with $b = 68$. If we are lenient with type I error control, we can use the calibrated test with maximum type I error rate $< 6\%$. A numerical evaluation then shows that $b = 63$ would be sufficient. For $\delta = 0.2$, PIWD(0.1) is very close to zero for $b \geq 60$, which does not put further restrictions on b. Thus we can recommend $b = 63$ for the calibrated test, $b = 63$ for the CX-S, and $b = 68$ for the UX test as it maximizes the EPASA while fulfilling all the threshold values for the OCs. We would prefer to use the CX-S test in this scenario, which yields the same participant benefit, while it has a stronger type I error rate control than the calibrated test. From this example, we see that even when type I error rate control is not of highest importance, an exact test might give more desirable OCs. As indicated in Figure 6, this burn-in value yields a maximum bias (over $\delta \in \{0.1, 0.2, 0.4\}$) of around 0.51%, which might be acceptable in practical settings.

4.6 Sensitivity analysis to the choice of prior

This paper considers BRAR using the Beta(1,1) prior. This section provides a sensitivity analysis of our findings with respect to this prior choice. Under the Beta(1,1) prior our main findings are (1) standard (non-exact) tests show substantial type I error rate inflation which can be reduced but not fully mitigated through a larger burn-in length; (2) exact tests can demonstrate superior power over standard tests, with the conditional (unconditional) exact test showing higher power for smaller (larger) burn-in lengths; (3) the choice of test statistic affects power and type I error rate; (4) statistical bias decreases more quickly in the burn-in length for large treatment effects and is higher for larger trial sizes; (5) power and EPASA are not always maximized at the largest and smallest burn-in lengths, respectively.

Supplemental Table 5 shows average and maximum type I error rates for the calibrated and exact PPCS tests, as well as the asymptotic Wald test for different priors for both treatment arms (both used for BRAR and testing) and trial sizes $\bar{i} \in \{60, 240\}$. The average and maximum type I error rate for the calibrated PPCS test under the Beta(0.01, 0.01) prior substantially increase in comparison to the uniform prior (which likely stems from its heavy prior mass near 0 and 1, together with the reasoning given in the third paragraph of Section 4.1), showing that the type I error rate under the BRAR design with the calibrated PPCS test is sensitive to the prior. As the maximum type I error rate remains substantially larger than 6%, finding (1) no longer holds in this setting. The other priors show a lower degree of additional type I error inflation: Beta(0.5, 0.5) and Beta(1.4,0.6) result in a maximum type I error rate above 7% for $\bar{i}/4$ under $\bar{i} = 60$. This increase in type I error rate inflation may be due to to a higher probability mass given to large common success rates under the Beta(0.01, 0.01), Beta(0.5, 0.5) and Beta(1.4,0.6) priors. Under the asymptotic Wald test, using the prior only to determine the allocation probabilities, the differences are smaller and the rule $b \geq \bar{i}/4$ shows maximum type I error

rates below 7%, agreeing with finding (1). For all priors except Beta(0.01, 0.01) we find agreement with finding (2): the CX-S test is less (more) conservative than the UX test on average for smaller (larger) burn-ins. The conservativeness of the UX test for Beta(0.01, 0.01) comes from correcting the substantial type I error rate inflation under this prior. The type I error rates vary greatly when comparing the calibrated PPCS and asymptotic Wald tests, agreeing with finding (3). The maximum type I error rates are most stable for the uniform prior, and the average type I error rates under the CX-S PPCS test are robust to the chosen prior.

Supplemental Table 6 provides the minimum power of the calibrated PPCS test, exact PPCS tests, and asymptotic Wald test, as well as average EPASA and bias under a BRAR design for different prior configurations, trial sizes, treatment effects, and burn-in lengths. Comparing the minimum power values for $\bar{i} = 60$, the minimum power of the CX-S PPCS test is often higher than that of the other tests, except for Beta(0.01, 0.01) (where substantial type I error inflation was found), agreeing with finding (2). Power improvements for $\bar{i} = 240$ under the CX-S test are only seen across PPCS tests, with the asymptotic Wald test showing higher power, agreeing with finding (3). Looking at the bias, noticing that $\bar{i} \cdot 4/10 = 24$ for $\bar{i} = 60$ and $\bar{i}/10 = 24$ when $\bar{i} = 240$, we find agreement with finding (4) across prior configurations. In agreement with finding (5), the average EPASA attains a maximum for $\bar{i}/10$ for the Beta(0.01, 0.01) prior for $\delta = 0.2$. In general, differences between OCs decrease for larger burn-in lengths. The prior Beta(1.4, 0.6) often improves power and decreases bias in comparison to the uniform prior (at the cost of EPASA), which might be because the BRAR procedure is less sensitive to early successes in this setting.

5 Real-world application: ARREST trial

In this section, we consider the effect of the burn-in length on the OCs of the Advanced R^2Eperfusion STrategies for Refractory Cardiac Arrest (ARREST) trial described in Yannopoulos et al.,[21] where we also consider the use of an UX test. The CX-S optional stopping threshold (OST) is omitted here, as the combination of this test with early stopping is not straightforward. We analyze the effect of different burn-in lengths on the unconditional OCs following from the trial design, that is, we do not condition on the realised size of the trial.

In the ARREST trial, extracorporeal membrane oxygenation (ECMO) facilitated resuscitation (developmental) was compared to standard advanced cardiac life support (control) in adults with an out-of-hospital cardiac arrest and refractory ventricular fibrillation. The inferential goal of the trial was to test $\theta_D = \theta_C$ versus $\theta_D \geq \theta_C$, where a success ($Y_{a,i} = 1$) represented survival to hospital discharge.

Participants were allocated treatment in groups of 30 under a permuted block design, with a control group allocation probability equal to the posterior probability (based on independent uniform priors) that the control treatment is superior, restricted between 0.25 and 0.75 that is, the clip method as defined in Du et al.[9] If at any of the interim analyses either the PPCS or one minus the PPCS became higher than an OST of 0.986, the recommendation was made to stop the trial early for futility or superiority, respectively. This OST was calibrated based on a success rate of 0.12 under the null hypothesis, and controlled type I error rate at 0.05 based on a simulation study of 10,000 samples. The considered alternative hypothesis $\theta_C = 0.12$ and $\theta_D = 0.37$ (based on a treatment effect of 0.25) yielded a power around 90%. The ARREST trial stopped for efficacy of the ECMO treatment after allocating the first group of 30 participants, with posterior probability of superiority of 0.9861.

In Baas et al.[8] this trial was re-analyzed, where the calibrated (C) OST was compared to an UX OST. The UX OST was computed based on an extension of the Markov chain $(\boldsymbol{X}_t)_t$ defined in Section 2, where the extension of $(\boldsymbol{X}_t)_t$ also models the optional stopping component of the trial. It was shown that while the UX OST bounds type I error rate by 0.05 over the whole of the parameter space, it also leads to a decrease in power and EPASA. Due to the optional stopping component EPASA, assuming $\theta_D \geq \theta_C$, is defined as:

$$\mathbb{E}^{\pi}_{\boldsymbol{\theta}}[(N_{D,i_U} + (\bar{i} - i_U)\mathbb{I}(\text{Optional stopping in favour of D}))/\bar{i}]$$

where U is the interim analysis at which the trial stops, the EPASA hence considers the fraction of participants allocated to the developmental treatment before and after optional stopping. It is hypothesised that a larger burn-in length will mitigate the large differences seen in EPASA and power across OSTs.

The ARREST trial had a maximum trial size of 150 participants and considered blocked allocation with blocks of 30 participants, hence we can only have $b = 15, 30, 45, 60, 75$ (noting that in the original design, the allocation probability could only be changed after 15 participants were allocated per arm). During the burn-in period, the target allocation proportion is equal to 0.5 (as before) and early stopping is only possible after allocating participant $i \geq 2b$. The Markov chains for computing the UX OST and the OCs described in Baas et al.[8] were adapted to account for a longer burn-in length by decreasing the number of interim analyses and setting the first interim analysis at $i = 2b$. The original C OST

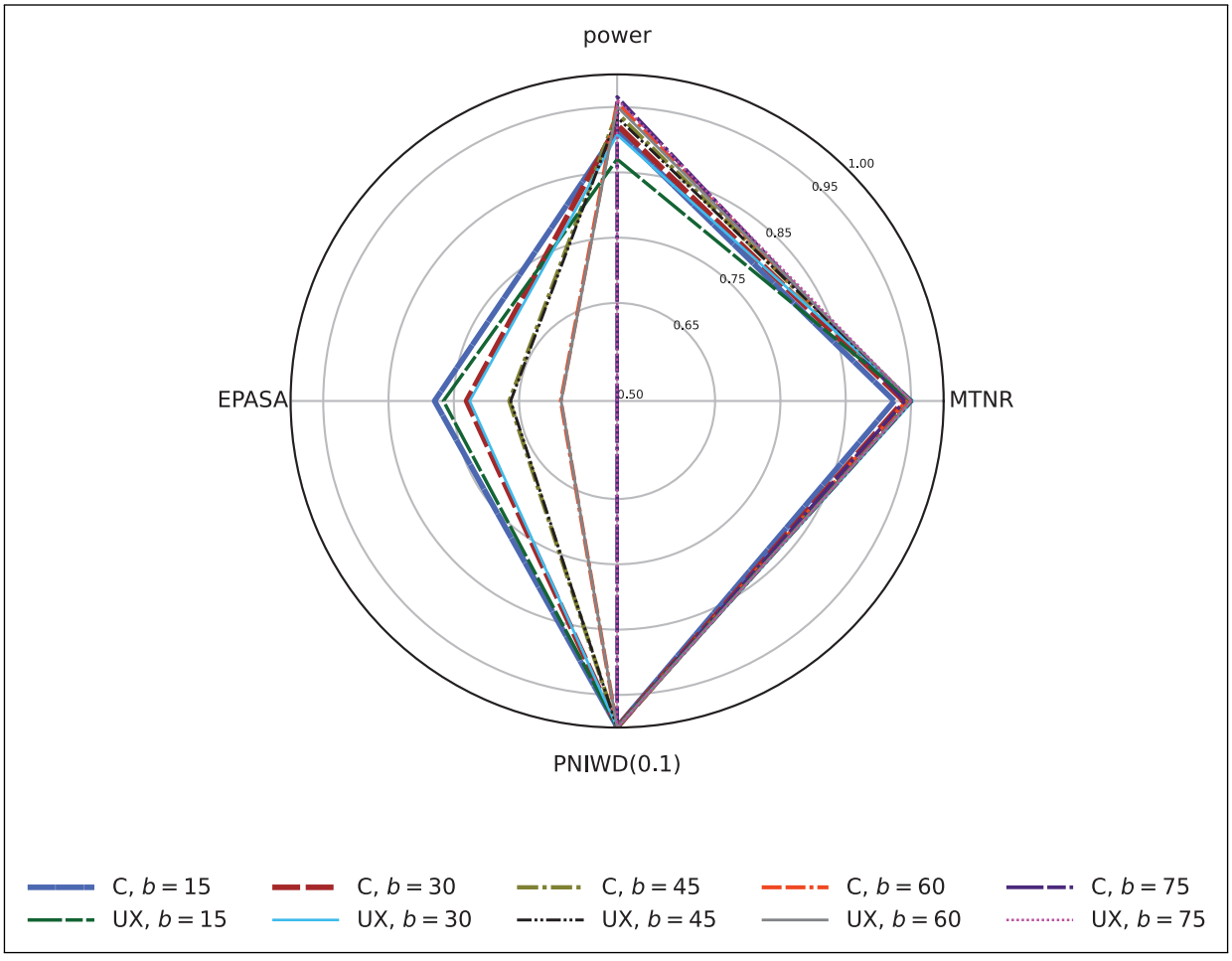


Figure 7. Star plot comparing the EPASA, power, MTNR over interval $\theta_C = \theta_D \in \{0.02, 0.03, \ldots, 0.22\}$, and PNIWD(0.1) for the ARREST trial when using a calibrated (C) and UX optional stopping threshold combined with a burn-in b of 15, 30, 45, 60, and 75 participants per arm. EPASA: expected proportion of allocations on the superior arm; MTNR: minimum true negative rate; PNIWD: probability of no imbalance in the wrong direction; UX: unconditional exact.

0.986 was used for $b = 15$, while for $b > 15$ the C OST was determined using the procedure outlined in Section 5 Baas et al.[8] where the only value under the null hypothesis considered was $\theta_C = \theta_D = 0.12$. The same procedure (taking all possible success rates into account) was used to determine the UX OST.

Figure 7 shows a star plot of the EPASA, power, minimum TNR (MTNR, minimum value of one minus the type I error rate) and PNIWD(0.1). For all OCs, a higher value is better. The EPASA, power, and PNIWD(0.1) were calculated under the alternative $\theta_C = 0.12$, $\theta_D = 0.37$, while the MTNR was calculated over the range $\theta_C = \theta_D \in \{0.020, 0.03, \ldots, 0.22\}$.

Figure 7 shows that the EPASA decreases in b, while the power increases. The EPASA for the C OST designs is higher than for the UX OST designs due to the higher power under the C OSTs, which means that the trial stops earlier for superiority under the alternative. For $b < 75$, the type I error rate is not controlled under the C OST for $\theta_C = \theta_D \in \{0.02, 0.03, \ldots, 0.22\}$, corresponding to an MTNR less than 0.95, whereas for the other designs, the type I error rate is under control. Figure 7 shows that the PNIWD(0.1) is very similar across different designs, it ranges from 0.999 to 1.00, in agreement with Figure 5 which shows that PIWD(0.1) is largest when treatment effects are small. As hypothesized, the differences in the OCs between the C OST and UX OST, mainly EPASA and power, decrease in b.

The star plot can be used to propose designs based on trade-offs in the OCs and to make a choice of clinical design, including burn-in length, based on multiple trial objectives. If type I error rate control at level 0.05 is not of the highest importance, Figure 7 shows that a good option could be to use the C OST with $b = 15$. If the type I error rate should be controlled, one design which stands out is the UX OST design with $b = 30$, showing a lower EPASA but similar power to the previously mentioned design. If EPASA is of higher importance than power, then the UX OST design with $b = 15$

might be best. If EPASA is not important at all, the C OST design with $b = 75$ would be the best option, yielding a power higher than 95%.

Supplemental Figure 4 shows the same evaluation for a range of alternative hypotheses $\theta_C = \theta_D - 0.25 \in \{0.02, 0.03, \ldots, 0.22\}$. The behaviour of the minimum PNIWD(0.1) (MPNIWD(0.1)) in Supplemental Figure 4 is similar to the PNIWD(0.1) in Figure 7. The power seems to be the most sensitive to the scenario considered, as the minimum power (MPow) is substantially lower than in Figure 7. The MPow for the UX test with $b = 15$ is less than the (usually chosen threshold of) 0.8, which might be a reason not to opt for this design despite it leading to exact type I error control and high (minimum) EPASA.

6 Discussion and recommendations

This paper considers the effect of an initial burn-in phase on the OCs of the BRAR, that is, Thompson sampling-based, design, where testing is performed either using a calibrated, conditional, or UX test. The analyses were based on the minimum, average, and maximum values of the OCs over the parameter space.

Our numerical evaluation revealed several key insights. Firstly, calibrated or asymptotic tests exhibited significant type I error rate inflation compared to fixed designs with equal allocation, particularly in BRAR designs without a burn-in period. While increasing burn-in length offered partial mitigation, it did not eliminate the inflation entirely. Secondly, exact tests demonstrated superior power over calibration or asymptotic tests in certain parameter settings, notably with a trial size of 60 participants and varying burn-in lengths. Furthermore, the conditional exact test displayed a consistent performance profile across the parameter space and was less conservative than calibrated and asymptotic tests, even with extended burn-in periods. Third, the choice of test statistic significantly impacted statistical OCs; the Wald test, for instance, yielded a more balanced type I error rate profile and greater power than tests based on the PPCS. Fourth, our paper considered the effect of a burn-in on estimation bias. We saw that the expected bias decreases more quickly in the burn-in length for large treatment effects and becomes higher for larger trial sizes (under the same burn-in length); hence, for controlling the bias, we advice to make the burn-in a function of the trial size, same as for the type I error rate. Our prior sensitivity analysis indicated that these finding hold across different choices of beta priors.

The evaluation in this paper uses the approach in Baas et al.,[8] where this method was applied to RA clinical trials without a burn-in period (and mainly fully sequential trials) with up to around 1000 participants, and where it was noted that the main computational limit came from computation of the coefficients $g_i^{\pi_b^B}$. Section 4 states that the number of values $g_i^{\pi_b^B}$ to calculate is of order $\mathcal{O}(\bar{i}^4)$ for b fixed, however the number of burn-in lengths b to consider equals $\bar{i}/2 + 1$, making for an overall order $\mathcal{O}(\bar{i}^5)$ of coefficients to compute when determining an optimal burn-in length. As $1000^{4/5}$ is roughly 251, we expect that 240 participants is close to the current computational limit, however some speedups are made by parallelizing the computation of coefficients $g_i^{\pi_b^B}$ over burn-ins and using the formula for $g_{2b}^{\pi_b^B}$ given in Remark 1.

Some general guidelines for choosing the burn-in length follow from our analysis. If type I error rate control is not of the highest priority, then a good rule of thumb might be to choose the burn-in length per arm roughly equal to the trial size divided by four for the BRAR design. This burn-in length led to a more balanced type I error rate profile, while power was similar for higher burn-in lengths. In settings where type I error rate control is strictly required (e.g., in confirmatory settings), we recommend considering exact tests. In this case, the burn-in length could be chosen such that a sufficient power value is reached, possibly at a lower value than the trial size over four. The conditional exact test would be preferred, as our numerical evaluation showed that this test has higher power than the UX test in terms of average (and often minimum) power for small to moderate burn-in lengths.

The optimal burn-in length in terms of power or participant benefit is often different from the minimum or maximum possible value, hence our recommendation is to inspect more burn-in lengths than just the minimum (i.e., zero) and the maximum (i.e., trial size over two) and inspect at least the values of b close to these two endpoints, in contrast to the guidelines given in Du et al.[9] The above guidelines focus on the behaviour of the average OCs, whereas in specific settings expert opinion might better guide the burn-in length, for example, through a desired minimum power or maximum allowed type I error rate.

The penultimate section of this paper considers an illustrative application to a real-world Bayesian adaptive clinical trial with optional stopping. The trial had a complex nature, where allocation was performed in blocks based on a permuted block design and the trial stopped when, at interim, the posterior probability of superiority crossed a calibrated OST. Based on the Markov chain modelling framework in Baas et al.,[8] we evaluated the performance of the calibrated and unconditional OSTs under different burn-in lengths. As it was assumed that after early stopping, the remaining participants were allocated the treatment that was deemed superior, the participant benefit depends on the OST of choice. The

most sensitive OCs were participant benefit and power, and it was concluded that low burn-in lengths yielded a balanced trade-off.

We considered the imbalance measure, denoted PIWD(0.1), defined by Thall et al.,[6] which quantifies the probability of substantial allocation toward the inferior arm. Consistent with existing literature, our results show that, while PIWD(0.1) increases as the absolute treatment effect decreases, the negative impact on participant benefit simultaneously diminishes. This reveals a key limitation of the measure—its dichotomous definition (imbalance vs. no imbalance) fails to account for the magnitude of the clinical effect on expected outcomes. Although fixed equal allocation minimizes PIWD(0.1), we provide a method (Section 4.5) to control PIWD(0.1) while concurrently maximizing overall participant benefit. Future research may explore alternative imbalance metrics or different parameter values.

Some interesting areas are left for future research, such as an adaptive burn-in length instead of a fixed deterministic burn-in length, and the consideration of other clinical trial designs, for example, multi-outcome or multi-arm designs or different outcome types. It is not immediately clear whether all our conclusions generalize to the latter two settings (multiple arms and other outcome types) because these alternative settings come with their own intricacies (e.g., in the multi-arm setting, one can perform an omnibus test, or test all arms separately, while for normal outcomes the variance is not a function of the expectation). Zhang et al.[30] showed type I error rate inflation under the Z-test for normal outcomes under the BRAR procedure, which we hypothesize can be mitigated but not removed completely by using a burn-in period, type I error rate control is for example guaranteed by exact tests, but which exact test performs best, and how to compute exact tests for other outcome types than binary is not established.

Data collected during a burn-in period is not representative of data collected later on in the trial if there is a strong time trend. In such settings, the combination with other regularization procedures such as the clipping of allocation probabilities or a power transformation[9] may be more appropriate, which could be explored in future research. We hypothesize that, in a setting without time trends, such *tuning* procedures, as well as blocked allocation, will have a similar effect on overall trial OCs such as type I error rate, power, and participant benefit to using a burn-in period, hence for each configuration of these design aspects, another burn-in length might be optimal or a burn-in might not be needed at all.

Although we focused on BRAR designs using the posterior probability for testing, our exact evaluation method is readily applicable to analyze the effect of burn-in length across other RAR procedures for example, those targeting optimal proportions, such as in Pin et al.[31] where large type I error rate inflation was also observed. Furthermore, a comparative analysis of different calculation methods for the posterior probability of superiority would provide valuable and complementary insights to this research.

While this paper addresses the underexplored issue of optimal burn-in length through numerical evaluation, underpinning these findings with theoretical results remains a key area for future research. Current theoretical literature on BRAR, largely derived from multi-armed bandit models see, for example, Agrawal and Goyal,[32] often considers large-sample behavior where the effect of a finite burn-in is negligible. Furthermore, this body of work typically optimizes a single OC, contrasting with our multi-objective approach. Further exploring the null hypothesis behavior under BRAR extending the work of Zhang et al.[30] and multi-objective optimization extending the work of Qin and Russo[33] would provide valuable theoretical support for our findings.

ORCID iDs

Edwin YN Tang https://orcid.org/0009-0007-6562-2192
Stef Baas https://orcid.org/0000-0002-5890-1165
Lukas Pin https://orcid.org/0009-0003-9512-681X
David S Robertson https://orcid.org/0000-0001-6207-0416
Sofía S Villar https://orcid.org/0000-0001-7755-2637

Ethical considerations

Ethical approval was not required.

Consent to participate

Not applicable.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the UK Medical Research Council [grant numbers MC_UU_00002/15 and MC_UU_00040/03, MC_UU_00002/14] (SB, DSR, SSV), MRC Biostatistics Unit Core Studentship, the Cusanuswerk e.V. (LP), University of Warwick Chancellor's International Scholarship (EYNT), and the DPMMS PhD Studentship (DK).

Declaration of conflicting interests

The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: Sofía S. Villar declares being part of the advisory board of PhaseV.

Data availability statement

Code for obtaining the results presented in Sections 4 and 5 is available at https://github.com/SB11S/exact_tests_burn_in_BRAR.

Supplemental material

Supplemental material for this article is available online.

References

1. Rosenberger WF and Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. 2nd ed, Hoboken, NJ: John Wiley & Sons, Inc, 2016.
2. Berry SM and Viele K. Comment: Response adaptive randomization in practice. *Stat Sci* 2023; **38**: 229–232.
3. Pin L, Neubauer M, Robertson D, et al. Clinical trials using response adaptive randomization. https://github.com/lukaspinpin/RA-ClinicalTrials. (2025a, accessed on 19 January 2025).
4. Robertson DS, Lee KM, López-Kolkovska BC, et al. Response-adaptive randomization in clinical trials: From myths to practical considerations. *Stat Sci* 2023; **38**: 185–208.
5. Rosenberger WF, Sverdlov O and Hu F. Adaptive randomization for clinical trials. *J Biopharm Stat* 2012; **22**: 719–736.
6. Thall P, Fox P and Wathen J. Statistical controversies in clinical research: Scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015; **26**: 1621–1628.
7. Proschan MA and Evans S. Resist the temptation of response-adaptive randomization. *Clin Infect Dis* 2020; **71**: 3002–3004.
8. Baas S, Jacko P and Villar SS. Exact statistical analysis for response-adaptive clinical trials: A general and computationally tractable approach. *Comput Stat Data Anal* 2025; **211**: 108207.
9. Du Y, Cook JD and Lee JJ. Comparing three regularization methods to avoid extreme allocation probability in response-adaptive randomization. *J Biopharm Stat* 2018; **28**: 309–319.
10. Thorlund K, Haggstrom J, Park JJH, et al. Key design considerations for adaptive clinical trials: A primer for clinicians. *BMJ* 2018; **360**: 1–5.
11. Viele K, Broglio K, McGlothlin A, et al. Comparison of methods for control allocation in multiple arm studies using response adaptive randomization. *Clinical Trials* 2020a; **17**: 52–60.
12. Viele K, Saville BR, McGlothlin A, et al. Comparison of response adaptive randomization features in multiarm clinical trials with control. *Pharm Stat* 2020b; **19**: 602–612.
13. Granholm A, Kaas-Hansen BS, Lange T, et al. An overview of methodological considerations regarding adaptive stopping, arm dropping, and randomization in clinical trials. *J Clin Epidemiol* 2023; **153**: 45–54.
14. Koehler E, Brown E and A. Haneuse SJ-P. On the assessment of Monte Carlo error in simulation-based statistical analyses. *Am Stat* 2009; **63**: 155–162.
15. Wei LJ, Smythe RT, Lin DY, et al. Statistical inference with data-dependent treatment allocation rules. *J Am Stat Assoc* 1990; **85**: 156–162.
16. Jacko P. BinaryBandit: An efficient Julia package for optimization and evaluation of the finite-horizon bandit problem with binary responses. In: *Management science working paper, lancaster university management school*, 2019 pp.1–13. https://eprints.lancs.ac.uk/id/eprint/136340/1/Jacko2019_binarybandit_wp.pdf (accessed 13 July 2024).
17. Yi Y. Exact statistical power for response adaptive designs. *Comput Stat Data Anal* 2013; **58**: 201–209.
18. Berger V, Bour LJ, Carter K, et al. A roadmap to using randomization in clinical trials. *BMC Med Res Methodol* 2021; **21**: 1–24.
19. Fisher RA. *Statistical methods for research workers*. 5th ed, Edinburgh, England: Oliver & Boyd, 1934.
20. Barnard GA. A new test for 2x2 tables. *Nature* 1945; **156**: 177.
21. Yannopoulos D, Bartos J, Raveendran G, et al. Advanced reperfusion strategies for patients with out-of-hospital cardiac arrest and refractory ventricular fibrillation (ARREST): A phase 2, single centre, open-label, randomised controlled trial. *The Lancet* 2020; **396**: 1807–1816.
22. Rice WR. A new probability model for determining exact p-values for 2 x 2 contingency tables when comparing binomial proportions. *Biometrics* 1988; **44**: 1–22.
23. Andrés A and Mato A. Choosing the optimal unconditioned test for comparing two independent proportions. *Comput Stat Data Anal* 1994; **17**: 555–574.
24. Best N, Ajimi M, Neuenschwander B, et al. Beyond the classical type I error: Bayesian metrics for Bayesian designs using informative priors. *Stat Biopharm Res* 2024; **17**: 183–196.

25. Haber M. A comparison of some conditional and unconditional exact tests for 2x2 contingency tables. *Commun Stat - Simul Comput* 1987; **16**: 999–1013.
26. QuadGK package documentation. URL https://juliamath.github.io/QuadGK.jl/latest/ (accessed 20 March 2025).
27. Baldi Antognini A, Novelli M and Zagoraiou M. A simple solution to the inadequacy of asymptotic likelihood-based inference for response-adaptive clinical trials: Likelihood-based inference for RA trials. *Stat Papers* 2022; **63**: 157–180.
28. Mehrotra DV, Chan ISF and Berger RL. A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. *Biometrics* 2003; **59**: 441–450.
29. Bowden J and Trippa L. Unbiased estimation for response adaptive clinical trials. *Stat Methods Med Res* 2017; **26**: 2376–2388.
30. Zhang K, Janson L and Murphy S. Inference for batched bandits. In: Larochelle H, Ranzato M, Hadsell R, Balcan M, and Lin H, (ed), *Advances in neural information processing systems*, 2020, Vol. 33, pp.9818–9829. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf.
31. Pin L, Villar SS and Rosenberger WF. Revisiting optimal proportions for binary responses: Insights from incorporating the absent perspective of type-I error rate control, 2025b. *Biometrics* 2025b; **81**: ujaf114.
32. Agrawal S and Goyal N. Analysis of Thompson sampling for the multi-armed bandit problem. In: *Proceedings of the 25th annual conference on learning theory*, 2012, pp.39.1–39.26. https://proceedings.mlr.press/v23/agrawal12.html.
33. Qin C and Russo D. Optimizing adaptive experiments: A unified approach to regret minimization and best-arm identification. *arXiv preprint arXiv:2402.10592*, 2024. URL https://arxiv.org/abs/2402.10592.