
INFORMED BURN-IN DECISIONS IN RAR: HARMONIZING ADAPTIVITY AND INFERENTIAL PRECISION BASED ON STUDY SETTING

A PREPRINT

Lukas Pin^{*1}, Stef Baas¹, Gianmarco Caruso¹, David S. Robertson¹, and Sofía S. Villar¹

¹Efficient Study Design Group, MRC Biostatistics Unit, University of Cambridge, Cambridge, UK

^{*}Corresponding author: lukas.pin@mrc-bsu.cam.ac.uk

November 27, 2025

ABSTRACT

Response-Adaptive Randomization (RAR) is recognized for its potential to deliver improvements in patient benefit. However, the utility of RAR is contingent on regularization methods to mitigate early instability and preserve statistical integrity. A standard regularization approach is the “burn-in period, an initial phase of equal randomization before treatment allocation adapts based on accrued data. The length of this burn-in is a critical design parameter, yet its selection remains unsystematic and improvised, as no established guideline exists. A poorly chosen length poses significant risks: one that is too short leads to high estimation bias and type-I error rate inflation, while one that is too long impedes the intended patient and power benefits of using adaptation. The challenge of selecting the burn-in generalizes to a fundamental question: what is the statistically appropriate timing for the first adaptation? We introduce the first systematic framework for determining burn-in length. This framework synthesizes core factors total sample size, problem difficulty, and two novel metrics (reactivity and expected final allocation error) into a single, principled formula. Simulation studies, grounded in real-world designs, demonstrate that lengths derived from our formula successfully stabilize the trial. The formula identifies a “sweet spot that mitigates type-I error rate inflation and mean-squared error, preserving the advantages of higher power and patient benefit. This framework moves researchers from conjecture toward a systematic, reliable approach.

Keywords Expected Patient Outcomes, Fixed Randomisation, Neyman Allocation, Patient-benefit, Wald test

1 Introduction

Adaptive designs introduce flexibility to clinical trials by allowing modifications based on accumulating data, thereby yielding more efficient and patient-oriented studies [Pallmann et al., 2018]. Response-Adaptive Randomization (RAR) is one such design where allocation probabilities change according to observed outcomes, typically to favor better-performing arms and/or increase statistical power.

Despite the potential advantages of RAR, its implementation requires careful management to ensure statistical validity and operational stability. To this end, regularization methods such as tuning, clipping, and the burn-in period are almost universally employed. These techniques are essential for controlling statistical properties like the type-I error rate, mitigating bias, and preventing the premature convergence of allocation based on unreliable early data. While such methods are widely adopted, a critical gap exists: as highlighted by Du et al. [2017], there is limited guidance on how to systematically select or calibrate them for a given trial design. The lack of clear, established guidance on how to effectively regularize RAR designs represents a high practical barrier to their optimal use in cases where they are most clinically relevant.

This paper focuses on the most common yet least explored of these techniques: the burn-in period. This is an initial phase where allocation follows a fixed randomization scheme, typically equal allocation, before switching to a response-adaptive rule. Its widespread use is underscored by a recent systematic review, which found that 88% of RAR trials included a burn-in phase [Wilson et al., 2025]. However, the choice of its length is often not explicitly justified. A burn-in that is too short may fail to stabilize the trial against early data fluctuations, while one that is too long may undermine potential power and/or patient benefits of the adaptive design. This reflects a fundamental trade-off between exploration and exploitation that is central to the design’s success. While recent work by Tang et al. [2025] has begun to analyze the effect of the burn-in period on operating characteristics for the specific case using the Bayesian RAR (BRAR), a specific recommendation and generalized framework on how to choose the length of the burn-in period remains absent.

Here we address this gap by introducing the first systematic framework for determining the burn-in length in two-arm trials with binary outcomes. We analyze the fundamental components governing this decision: total sample size, intrinsic problem difficulty (standardized treatment effect), and two novel metrics—design reactivity and expected final allocation error. These metrics quantify the inherent exploration-exploitation trade-off. The framework culminates in a single, principled formula (Equation 8) that synthesizes these distinct factors, offering a practical tool for researchers.

The paper is structured as follows: Section 2 establishes the notation and RAR designs evaluated. Section 3 systematically develops our framework, introducing and analyzing the impact of the standardized treatment effect, sample size, reactivity, and allocation error. Section 4 presents our culminating formula and discusses its practical, adaptive implementation. Section 5 validates our approach in simulation studies based on two real-world trials. Finally, Section 6 discusses the implications, limitations, and future research directions.

2 Notation and considered designs

We first define the trial setting. Consider a trial with a total of n patients and two arms: a control ($k = 0$) and a treatment ($k = 1$). Let Y_{ki} be the potential outcome for patient i on arm k , and a_{ki} be the allocation indicator, where $\sum_{k=0}^1 a_{ki} = 1$ for all $i = 1, \dots, n$. We assume patients are enrolled sequentially and outcomes are observed without delay. Although this might not align with all practical scenarios where patients might enter trials in cohorts or where outcomes are delayed, we maintain these assumptions to preserve analytical tractability and clarity. Such assumptions are consistent with traditional RAR literature, which often avoids these complexities to enhance the understanding of methodological limits and potential gains over standard practices.

For the primary analysis, we consider binary endpoints where $Y_{ki} \sim \text{Bern}(p_k)$, with p_k being the unknown response probability for arm k . For a trial with $K > 1$ treatment arms, things might be even more complex since more treatment effects have to be taken into account. The multi-arm case is discussed in Section 6.

In the following Sections we compare some commonly known and novel RAR designs to an patient benefit benchmark design and equal randomization targeting equal allocation:

- **Equal Randomisation (ER):** This term refers to designs that enforce *equal allocation* ($n_0 = n_1$). It represents the non-adaptive baseline in our framework. As our simulations do not involve time trends, our ER results can be interpreted as representative of any design that ensures a final 1:1 balance, such as the *permuted block design*, *random allocation rule*, or *big stick design*. For more information and other designs see Berger et al. [2021].
- **Patient Benefit Benchmark Design (PBB):** Serving as a theoretical benchmark for patient benefit, this design knows the ground truth and allocates all subsequent patients to the superior arm following the burn-in period. We use the PBB even though it needs no learning phase to precisely analyze the burn-in’s isolated impact on key operating characteristics, particularly patient benefit.
- **Bayesian Response-Adaptive Randomization (BRAR (U)):** The untuned version of BRAR stems from Thompson [1933]. Also known as **Thompson Sampling**, this unregularized Bayesian design allocates the next patient to arm 1 with a probability equal to the current posterior probability that arm 1 is superior, i.e., $P(p_1 > p_0 \mid \text{data})$. We use a non-informative Beta(1,1) prior.
- **Bayesian Response-Adaptive Randomization (BRAR (T)):** The tuned version of BRAR was proposed by Thall et al. [2015]. This is a regularized Bayesian design. It tunes the allocation probabilities in a way that for earlier patients the probabilities closer towards 0.5 than the ones of BRAR (U).
- **Neyman Allocation for the Wald test (N_1):** Neyman allocation [Neyman, 1934] is derived to maximize statistical power for the Wald test (Z_1). The allocation proportion is targeted using the Efficient Randomized-Adaptive Design (ERADE) [Hu et al., 2009].

- **Neyman Allocation for the score test (N_0):** This is a novel power-optimizing allocation derived in Pin et al. [2025b], targeted using ERADE. This proportion is specifically designed to maximize statistical power when the final analysis is conducted using the more robust score test (Z_0) preventing type-I error inflation.
- **RSHIR Allocation for the Wald test (R_1):** This is the classical RSHIR allocation [Rosenberger et al., 2001], targeted with ERADE. This allocation proportion is derived to optimize for patient benefit by minimizing the expected number of failures in the trial, subject to a constraint on the variance of the Wald test (Z_1).
- **RSHIR Allocation for the score test (R_0):** This is a novel patient-benefit-optimizing allocation from Pin et al. [2025b], targeted with ERADE. It minimizes the expected number of failures subject to a constraint on the variance of the score test (Z_0).
- **The Play the Winner Rule (PTW) [Zelen, 1969]:** This is a deterministic allocation rule. After a success on an arm, the next patient is allocated to the same arm. After a failure, the next patient is allocated to the other arm.
- **The Randomized Play the Winner Rule (RPTW) [Wei and Durham, 1978]:** A “success-driven” urn-based design. An urn contains balls of two types (e.g., red and blue). Initially it contains one ball of each type. The next patient’s allocation is determined by drawing a ball from the urn. When a success is observed on an arm (e.g., red), a red ball is added to the urn. If a failure is observed, a blue ball is added.

The list of designs evaluated is representative, not exhaustive. The procedure we will define in the following sections is general and valid for any RA(R) algorithm, including those yet to be developed. We use ER and the PBB design as fundamental benchmarks to frame the analysis, as they represent the non-adaptive baseline and the theoretical ideal for patient benefit, respectively. The two Neyman allocations N_1 and N_0 are theoretically (and asymptotically for N_1) optimal in terms of power for the Wald and score test, respectively, but they do not always achieve this in finite samples [Pin et al., 2025a].

In the next section we investigate what influences the burn-in length b , where b is the number of patients allocated to each arm before we start adapting.

3 What influences the burn-in length and how?

3.1 Standardized treatment effect

Treatment effect We define the treatment effect as the simple difference between the experimental and control response probabilities: $p_1 - p_0$. A smaller treatment effect (i.e., a small difference) makes it more difficult to correctly identify the superior arm, even when one truly exists. This scenario creates greater initial uncertainty, suggesting that a longer burn-in period may be beneficial to ensure adequate exploration of all arms before adaptation begins. While our analysis focuses on this simple difference, other measures like the log relative risk or log odds ratio could be used, but they would alter the RAR design and the operating characteristics of the final inference method [Pin et al., 2024], see Section 6 for more details.

Arm-specific response variances The more variable the response of an arm, the more observations it requires to reliably estimate its expected outcome. Ideally, when arm-specific response variances are high, we might want to guarantee that the arms are “sufficiently” explored through a longer burn-in period.

Thus, rather than considering the simple treatment effect, we suggest considering a standardized treatment effect, i.e.,

$$\delta = \frac{|p_1 - p_0|}{\sqrt{p_0(1 - p_0) + p_1(1 - p_1)}}, \quad (1)$$

which measures the difference in response probabilities scaled by the total variability in the two arms. The metric is unbounded and approaches infinity as the treatment effect $|p_1 - p_0|$ approaches 1 while both arm-specific variances approach 0 (e.g., as $p_0 \rightarrow 0$ and $p_1 \rightarrow 1$, or vice-versa). The quantity is only undefined in degenerate cases where both responses are deterministic (i.e. $p_0, p_1 \in \{0, 1\}$). In these situations, we define

$$\delta = \begin{cases} 0, & \text{if } p_1 = p_0, \\ \infty, & \text{if } p_1 \neq p_0. \end{cases}$$

The standardized effect δ therefore lies in the range $[0, \infty]$. The lower bound $\delta = 0$ occurs when there is no treatment effect, i.e., $p_1 = p_0$, provided that p_0 and p_1 are not both 0 or both 1. The metric is technically undefined if $p_0 = p_1 = 0$

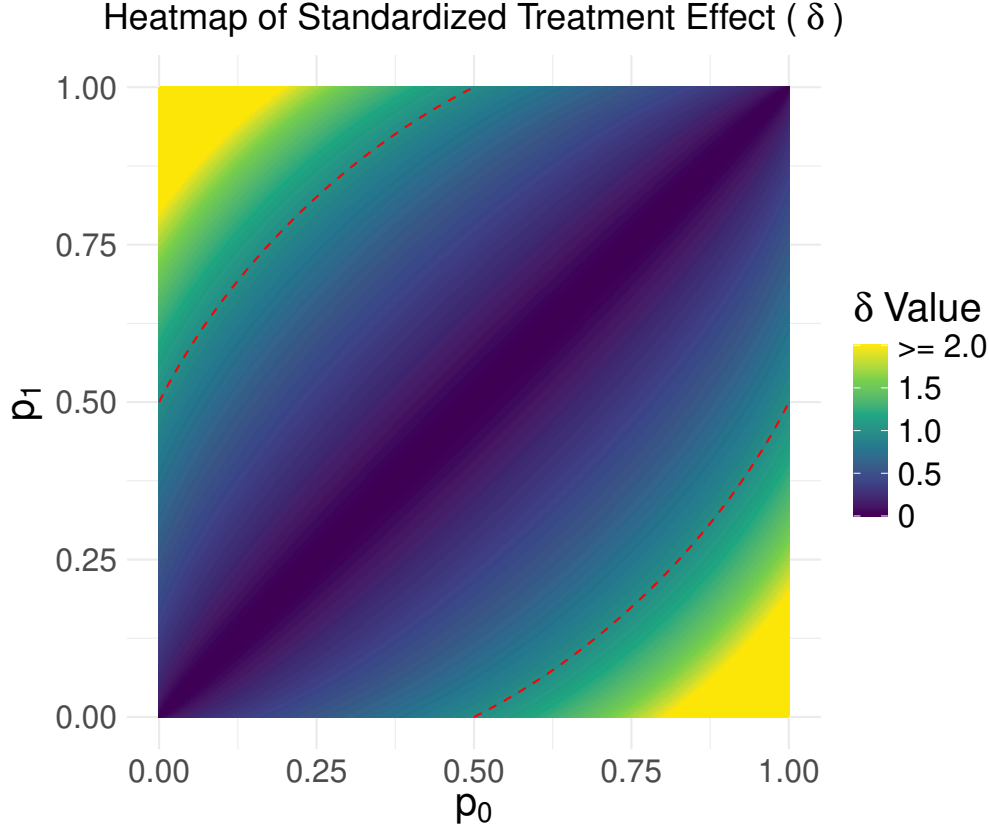


Figure 1: Heatmap of the standardized treatment effect (δ) across the full parameter space. The dashed red contour marks $\delta = 1$. The color scale is capped at 2 for visualization, as δ approaches infinity in the top-left ($p_0 \rightarrow 0, p_1 \rightarrow 1$) and bottom-right ($p_0 \rightarrow 1, p_1 \rightarrow 0$) corners.

or $p_0 = p_1 = 1$, as this results in an indeterminate $0/0$ form. A larger δ signifies a clearer, more easily detectable difference between the arms relative to their variability.

The standardized treatment effect accounts for the fact that response probabilities closer to 0.5 are associated to a higher variability in the arm response than response probabilities which are close to the boundaries 0 and 1. For example, let us consider two scenarios, each one characterized by different pairs of response probabilities, i.e. $p_0 = 0.7, p_1 = 0.9$ (first scenario) and $p_0 = 0.4, p_1 = 0.6$ (second scenario). Although the treatment effect is 0.2 in both scenarios, the second scenario is associated with higher variability than the first one and, thus, we would require a higher exploration (i.e. a higher burn-in) under this scenario. Indeed, the higher total variability of the responses of the second scenario is reflected in a smaller value of the standardized treatment effect ($\delta \approx 0.29$) than the first scenario ($\delta \approx 0.37$).

While δ is theoretically unbounded (ranging from 0 to ∞), Figure 1 reveals that $\delta < 1$ for the majority of the parameter space. The red contour ($\delta = 1$) highlights the boundary of this region. A key observation is that δ can only exceed 1 if the absolute treatment difference, $|p_1 - p_0|$, is greater than 0.5. This minimum difference is only sufficient at the boundaries (e.g., $p_0 = 0, p_1 = 0.5$). As the probabilities move away from 0 or 1 and towards 0.5, the treatment difference required to maintain $\delta = 1$ increases, as shown by the red curve's shape. This effect is most pronounced when one probability equals 0.5. In that case ($p_0 = 0.5$ or $p_1 = 0.5$), δ reaches its maximum possible value of 1, regardless of how extreme the other probability becomes.

In practice, δ can be calculated using the response probabilities specified for the trial's power analysis. Sample size determination typically requires an assumed treatment effect and a baseline success probability, p_0 , often derived from previous studies. If p_0 is specified as a range rather than a single value, one could use the midpoint of this range. Alternatively, one might select the p_0 value within the range that results in the highest standardized treatment effect, although this would represent a more optimistic scenario for detection.

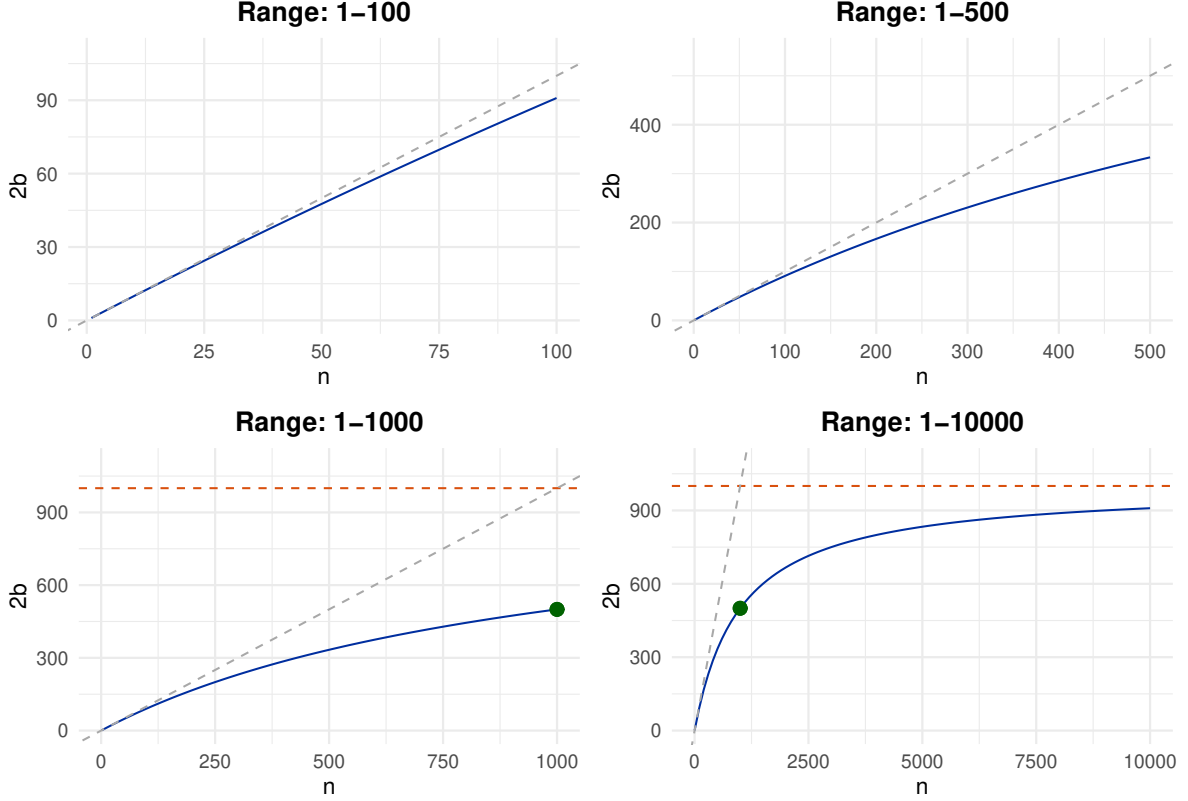


Figure 2: Total available burn-in proportion (largest $2b$ possible) based on sample size (n) and $\frac{n \cdot n_{1/2}}{n + n_{1/2}}$ with $n_{1/2} = 1000$. The gray dashed line reflects the largest burn-in budget upper bound ($2b = n$).

Impact on burn-in A smaller expected standardized treatment effect suggests that a larger burn-in period is required. This is because a small δ indicates that the difference between the arms’ response probabilities is difficult to detect. This difficulty can be caused by either a small true treatment effect (a small numerator) or high response variability (a large denominator). The δ metric effectively captures both of these challenges in a single variable.

3.2 Impact of sample size on burn-in

The burn-in serves two primary functions: (1) to stabilize estimators at the beginning of the trial, and (2) to improve inferential properties, such as type-I error rate and power. We assume that for a given treatment effect, these objectives are typically met in a standard, non-adaptive trial (using only ER) once the total sample size n is sufficiently large. We posit that this same level of stability and performance can be achieved within the burn-in phase of our adaptive trial, provided the total burn-in size ($2b$) is large enough. Once this sufficient burn-in size is reached, additional burn-in patients provide diminishing or no benefit. Consequently, the proportion of patients allocated to the burn-in period ($BP = 2b/n$) should decrease as the total trial size n increases. To ensure BP decreases appropriately, we define a non-linear term

$$\frac{n \cdot n_{1/2}}{n + n_{1/2}}, \quad (2)$$

where $n_{1/2}$ is a sample size parameter that defines the non-linear decrease. This term functions as an available sample size for the burn-in, which grows more slowly than n . By setting the sample size parameter $n_{1/2}$ equal to 1000, we achieve a largest total burn-in ($2b$) of 500 for a trial with $n = 1000$, see Figure 2.

3.3 Reactiveness of a RAR design

Here, we present an operational definition for reactivity of a RAR design. With reactivity we describe a characteristic of adaptive designs that has been referred to as ‘aggressiveness’ without formally defining it, see e.g. Villar et al. [2018], Viele [2025], Baas et al. [2025a,b].

The allocation proportion to the experimental treatment after allocation of participant i in the trial can be calculated based on the current patient allocation, $n(i) = (n_0(i), n_1(i))$, as $n_1(i)/i$. The following assumption is made:

Assumption 1 *The RA(R) procedure is such that $\lim_{i \rightarrow \infty} n_1(i)/i = \rho$ exists for almost every $\{p_0, p_1\} \in [0, 1]^2$.*

This assumption holds for most RA(R) procedures. Note that “almost every $\{p_0, p_1\} \in [0, 1]^2$ ” excludes the case $p_0 = p_1$ as this line has no (Lebesgue) measure in $[0, 1]^2$. This detail enforces that Assumption 1 holds for procedures such as unregularized Bayesian RAR (i.e., Thompson sampling) where $n_1(i)/i$ may converge in distribution to a uniform random variable when $p_0 = p_1$ [See, e.g., Proposition 1 in Zhang et al., 2020, showing this result for the normal-normal model]. For RA(R) procedures such as the randomized play-the-winner rule (RPTW), BRAR or targeting optimal proportions, ρ is known. If the theoretical limit is unknown we can estimate the limiting proportions through simulations.

Given the limiting proportion ρ for the RA(R) procedure, we can estimate the expected speed of convergence of $n_1(i)/i$ to ρ . We assume for $c > 0$ that

$$n_1(i)/i \approx \rho + (1/2 - \rho)i^{-c}. \quad (3)$$

The approximation (3) is based on the assumption that the smallest burn-in of two patients per arm is employed, $\lim_{i \rightarrow \infty} n_1(i)/i = \rho$, and furthermore $n_1(i)/i = \rho + O(n^{-c})$ for $c > 0$ (see, e.g., Yi and Li [2018], for a similar assumption). From (3) we obtain

$$c \approx \begin{cases} -\log(|n_1(i)/i - \rho|/|1/2 - \rho|)/\log(i), & \text{if } \rho \neq 1/2, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

While c can be determined exactly from one realization (e.g., $n_1(n)/n$) of an allocation proportion when Equation (3) is a perfect fit (i.e., for deterministic procedures with perfect knowledge), it is likely the case for practical RAR procedures that (3) is only an approximation, hence we will estimate the geometric slope c from the complete trial realization with total trial size n , i.e., using

$$\hat{c}(\rho) = \frac{1}{n} \sum_{i=2}^n -\log(|n_1(i)/i - \rho|/|1/2 - \rho|)/\log(i) \cdot \mathbb{I}(\rho \neq 1/2).$$

In the above, the exclusion of $i=1$ follows from the fact that $\log(1) = 0$. The integral of the geometric slope, independent of the parameters p_0 and p_1 , equals $\int_{[0,1]^2} c \cdot d\mathbf{p}$ (where \mathbf{p} is the vector (p_0, p_1)) and can be estimated using Monte Carlo sampling:

$$\tilde{r}(\rho) = \frac{1}{n_{\text{sim}}} \sum_{m=1}^{n_{\text{sim}}} \hat{c}_m(\rho_m) \quad (5)$$

where ρ_m is the target proportion following from parameter vector $\mathbf{p}_m \sim U([0, 1]^2)$ for all $m \in \{1, \dots, n_{\text{sim}}\}$.

In practice, it was found that the proportion $n_1(i)/i$ can get stuck around zero or one for very reactive RAR procedures, even if this is not the limit proportion. This is because, in such settings where most participants are allocated to one treatment, little is learned about the actual differences in the treatment groups. Due to this, it may be that at the end of the trial the difference $|n_1(n)/n - \rho|$ is larger than $|n_1(n)/n - 1/2|$ and hence we get $\hat{c}_m < 0$. To this end, we estimate the reactivity parameter, denoted r , as

$$r = \max(0, \tilde{r}(\rho), \tilde{r}(1), \tilde{r}(0)), \quad (6)$$

i.e., the reactivity is restricted to be nonnegative, and defined as the maximum rate at which the allocation proportion goes to zero, one, or the theoretical limit. For RAR procedures, the reactivity parameter r is expected to take on values between 0 and 0.5, this is because sampling with a fixed biased coin design (i.e., where the target proportion is known) leads to an asymptotic convergence rate of 1/2 (indicated by the central limit theorem). For RAR procedures, i.e. procedures that are learning the target proportion sequentially during the trial, this convergence will likely be slower. For instance, Hu and Rosenberger [2006] show that the central limit theorem holds with slower rates than 1/2 for RPTW and rates equal to 1/2 for RAR based on sequential estimation such as ERADE. We note that that these are asymptotic rates, and when the reactivity is estimated based on a finite horizon as in (4), the estimated reactivity parameter will likely be lower. This is because the allocated proportion under a RAR procedure may not have converged to the asymptotic limit proportion.

Table 1 reports comparisons of the reactivity parameters across several RAR procedures (listed in Section 2) for $n = 200, 500, 1000$, and 2000. The slope defined in Equation (5) is computed independently of the parameters p_0, p_1 . We observe that the resulting estimates change over the different choices of n .

Table 1: Estimated reactiveness parameters (r and ϵ_ρ) and resulting burn recommendation (b) and burn-in proportion (BP), independent of parameters p_0, p_1 , for trial sizes $n \in \{200, 500, 1000, 2000\}$ based on $n_{\text{sim}} = 1000$ simulations.

n	Design	r (x100)	ϵ_ρ (x100)	$r + \epsilon_\rho$ (x100)	b	BP (x100)
200	ER	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	-	-
	PBB	65.77 +/- 0.00	0.00 +/- 0.00	65.77 +/- 0.00	65 +/- 0.82	64.94 +/- 0.82
	BRAR (U)	34.25 +/- 1.17	0.83 +/- 0.31	35.09 +/- 1.19	51 +/- 0.96	50.13 +/- 0.96
	BRAR (T)	9.72 +/- 0.41	0.36 +/- 0.12	10.08 +/- 0.41	30 +/- 1.29	29.55 +/- 1.29
	N_0	18.91 +/- 1.09	1.53 +/- 0.18	20.43 +/- 1.10	37 +/- 1.39	36.63 +/- 1.39
	N_1	13.62 +/- 1.03	32.41 +/- 1.06	46.03 +/- 1.21	54 +/- 1.26	53.64 +/- 1.26
	R_0	19.48 +/- 1.12	1.18 +/- 0.13	20.65 +/- 1.12	37 +/- 1.43	36.37 +/- 1.43
	R_1	25.12 +/- 1.15	16.97 +/- 1.12	42.09 +/- 1.48	54 +/- 1.11	53.91 +/- 1.11
	PTW	36.15 +/- 1.52	1.29 +/- 0.15	37.44 +/- 1.49	52 +/- 1.08	51.88 +/- 1.08
500	RPW	25.41 +/- 1.29	1.79 +/- 0.21	27.20 +/- 1.25	43 +/- 1.25	42.30 +/- 1.25
	ER	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	-	-
	PBB	72.02 +/- 0.00	0.00 +/- 0.00	72.02 +/- 0.00	137 +/- 1.46	54.46 +/- 0.58
	BRAR (U)	38.61 +/- 1.24	1.07 +/- 0.36	39.68 +/- 1.24	109 +/- 1.68	43.30 +/- 0.67
	BRAR (T)	11.93 +/- 0.47	0.12 +/- 0.06	12.05 +/- 0.46	63 +/- 2.37	24.81 +/- 0.95
	N_0	23.05 +/- 1.07	0.90 +/- 0.12	23.95 +/- 1.07	80 +/- 2.78	31.73 +/- 1.11
	N_1	21.01 +/- 1.19	33.94 +/- 1.06	54.95 +/- 1.44	119 +/- 2.59	47.58 +/- 1.04
	R_0	24.69 +/- 1.14	0.85 +/- 0.11	25.54 +/- 1.14	82 +/- 2.75	32.46 +/- 1.10
	R_1	33.84 +/- 1.19	17.38 +/- 1.15	51.22 +/- 1.66	118 +/- 2.15	46.92 +/- 0.86
1000	PTW	38.76 +/- 1.35	0.85 +/- 0.10	39.61 +/- 1.33	109 +/- 1.83	43.42 +/- 0.73
	RPW	26.74 +/- 1.23	1.25 +/- 0.14	27.99 +/- 1.20	91 +/- 2.29	36.13 +/- 0.92
	ER	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	-	-
	PBB	75.50 +/- 0.00	0.00 +/- 0.00	75.50 +/- 0.00	210 +/- 2.02	41.93 +/- 0.40
	BRAR (U)	43.92 +/- 1.22	0.60 +/- 0.26	44.52 +/- 1.21	170 +/- 2.10	33.88 +/- 0.42
	BRAR (T)	13.23 +/- 0.49	0.10 +/- 0.06	13.32 +/- 0.49	102 +/- 3.31	20.38 +/- 0.66
	N_0	26.03 +/- 1.03	0.57 +/- 0.07	26.59 +/- 1.04	128 +/- 4.04	25.48 +/- 0.81
	N_1	28.11 +/- 1.23	34.54 +/- 1.05	62.65 +/- 1.50	192 +/- 3.66	38.21 +/- 0.73
	R_0	29.64 +/- 1.10	0.62 +/- 0.07	30.27 +/- 1.09	135 +/- 3.77	26.98 +/- 0.75
2000	R_1	37.76 +/- 1.20	18.41 +/- 1.18	56.17 +/- 1.78	188 +/- 3.31	37.42 +/- 0.66
	PTW	40.25 +/- 1.26	0.58 +/- 0.08	40.83 +/- 1.25	165 +/- 2.57	32.80 +/- 0.51
	RPW	30.39 +/- 1.21	0.85 +/- 0.12	31.24 +/- 1.19	138 +/- 3.20	27.59 +/- 0.64
	ER	0.00 +/- 0.00	0.00 +/- 0.00	0.00 +/- 0.00	-	-
	PBB	78.24 +/- 0.00	0.00 +/- 0.00	78.24 +/- 0.00	288 +/- 2.28	28.72 +/- 0.23
	BRAR (U)	45.72 +/- 1.18	0.51 +/- 0.24	46.23 +/- 1.16	235 +/- 2.66	23.45 +/- 0.27
	BRAR (T)	15.46 +/- 0.55	0.09 +/- 0.05	15.55 +/- 0.54	144 +/- 4.65	14.31 +/- 0.46
	N_0	28.98 +/- 0.94	0.40 +/- 0.04	29.37 +/- 0.93	177 +/- 5.35	17.62 +/- 0.53
	N_1	34.87 +/- 1.28	34.35 +/- 1.07	69.22 +/- 1.67	274 +/- 4.62	27.36 +/- 0.46
	R_0	31.65 +/- 1.07	0.39 +/- 0.05	32.04 +/- 1.07	186 +/- 5.05	18.51 +/- 0.51
	R_1	42.11 +/- 1.17	17.84 +/- 1.17	59.95 +/- 1.86	258 +/- 4.43	25.73 +/- 0.44
	PTW	41.76 +/- 1.17	0.39 +/- 0.04	42.16 +/- 1.16	221 +/- 3.20	22.10 +/- 0.32
	RPW	31.05 +/- 1.13	0.69 +/- 0.10	31.74 +/- 1.11	190 +/- 4.15	18.91 +/- 0.41

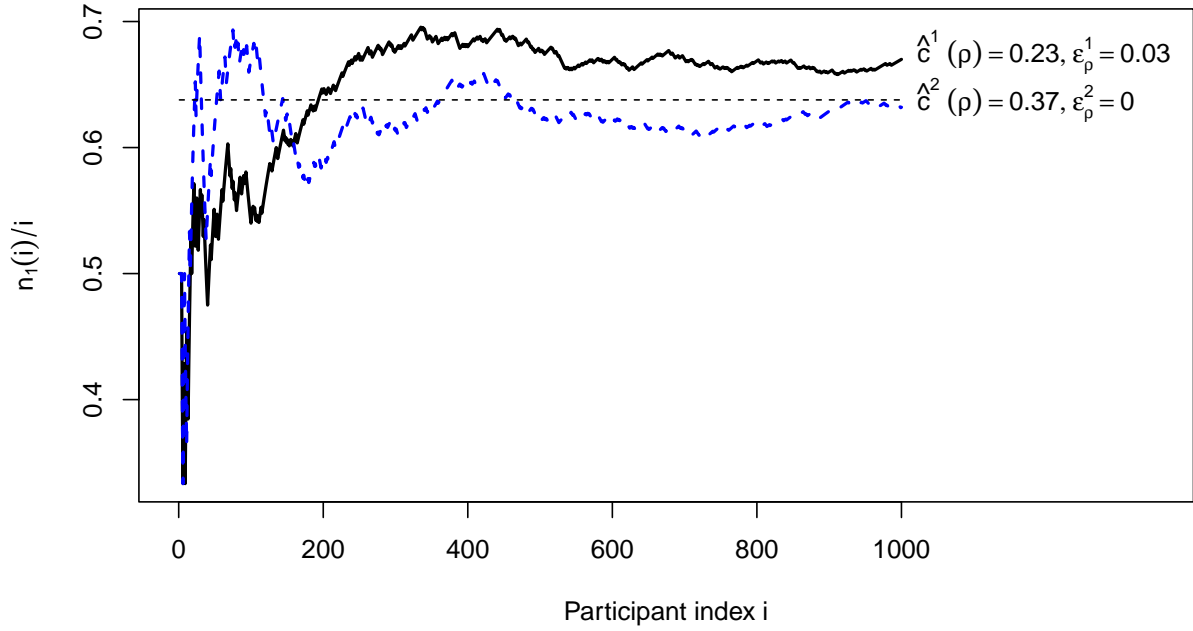


Figure 3: Two sampled paths (1000 participants) for the R_0 design under $p_0 = 0.6$, $p_1 = 0.8$ leading to $\rho = 0.64$. The solid path displays a slower trend than the dashed path and has not converged yet, ending up at a value higher than ρ ; hence, the estimated geometric slope is 0.23, while the error ϵ_p^1 is around 0.03. The dashed path has converged to a value around ρ and hence has a higher slope $\hat{c}_2(\rho) = 0.37$ than the solid path and error $\epsilon_p^2 = 0.00$ as the final proportion is below ρ .

3.3.1 Reactiveness in practice

In practical settings, the treatment difference $(p_1 - p_0)$, the baseline response probability p_0 , and the targeted sample size n are typically given in advance, as they are determined by power considerations during trial planning. If only a range for p_0 is known we can choose the central value in this range or the value that yields the largest δ . Thus, instead of evaluating reactivity across randomly drawn parameter pairs $(p_0, p_1) \sim U([0, 1]^2)$, we adapt the measure to reflect the specific planned trial characteristics.

For specific parameter configurations, we define reactivity as

$$\frac{1}{n_{\text{sim}}} \sum_{m=1}^{n_{\text{sim}}} \hat{c}_m(\rho), \quad (7)$$

where n denotes the planned trial size, n_{sim} the number of simulated trials and ρ the optimal proportion for the specific parameter combination. In this formulation, simulations are carried out using fixed parameter values p_0 and p_1 , rather than sampling them from a uniform distribution. This yields a reactivity measure that reflects how the design behaves under the *expected* conditions of the planned trial, rather than across a broad parameter space.

Impact on burn-in A design with high reactivity requires a larger burn-in. This is because the algorithm moves away from a balanced allocation quickly, and a burn-in is necessary to prevent this rapid convergence from being based on misleading early estimates. This need for a burn-in is further amplified if the design's limiting allocation (ρ) is far from 0.5. In such cases, the burn-in is not only protecting against premature convergence but also ensuring sufficient exploration before the design commits to a heavily imbalanced allocation. Therefore, the greatest need for a burn-in occurs when a design is both converging quickly and targeting an extreme allocation (ρ far from 0.5).

3.4 Expected final allocation error

The reactivity measure r , defined in Section 3.3, measures the speed of convergence and the extremeness of the limiting allocation, see Equation (6). This new measure, ϵ_ρ , addresses a different question: How accurately does the algorithm's final allocation match what it is supposed to do?

While r evaluates the reactivity of the allocation, it does not penalize an algorithm for converging to the wrong target (e.g., converging to 0.1 when the target ρ is 0.8). We introduce ϵ_ρ to specifically evaluate the accuracy of the final allocation proportion, $(n_0, n_1)/n$, relative to its theoretical target ρ . The goal is to isolate this accuracy error from the separate concept of reactivity, thereby avoiding a double penalty for designs that are intentionally reactive, such as BRAR.

The error for a single trial, $\epsilon_{\rho,n}$, is defined by the following logic:

- If $\rho \geq 0.5$, the target interval is $[0.5, \rho]$.
 - If $n_1/n \in [0.5, \rho]$, $\epsilon_{\rho,n} = 0$.
 - If $n_1/n > \rho$, $\epsilon_{\rho,n} = n_1/n - \rho$.
 - If $n_1/n < 0.5$, $\epsilon_{\rho,n} = 0.5 - n_1/n$.
- If $\rho < 0.5$, the target interval is $[\rho, 0.5]$.
 - If $n_0/n \in [\rho, 0.5]$, $\epsilon_{\rho,n} = 0$.
 - If $n_0/n < \rho$, $\epsilon_{\rho,n} = \rho - n_0/n$.
 - If $n_0/n > 0.5$, $\epsilon_{\rho,n} = n_0/n - 0.5$.

A key feature of this metric is its behavior under the null hypothesis (i.e., when $\rho = 0.5$). In this scenario, the target interval $[0.5, \rho]$ collapses to the single point $[0.5, 0.5]$. The definition correctly simplifies to penalize any deviation from a balanced 0.5 allocation, and it does so symmetrically. The error becomes $\epsilon_{0.5,n} = |n_1/n - 0.5|$.

We then report the *Expected Final Allocation Error* ϵ_ρ by averaging $\epsilon_{\rho,n}$ across n_{sim} simulations. To illustrate, let us assume $\rho \geq 0.5$, for example $\rho = 0.6$. The target interval is $[0.5, 0.6]$.

- If the final allocation lies in the interval, e.g. $n_1(n)/n = 0.55$, then $\epsilon_{\rho,n} = 0$.
- If the final allocation $n_1(n)/n = 0.7$ (an “overshoot”), the error is $\epsilon_{\rho,n} = 0.7 - 0.6 = 0.1$.
- If the final allocation $n_1(n)/n = 0.3$ (an “undershoot”), the error is $\epsilon_{\rho,n} = 0.5 - 0.3 = 0.2$.

This example demonstrates that deviations in the wrong direction are penalized more heavily. Although both 0.3 and 0.7 are 0.2 away from 0.5, 0.7 lies closer to the boundaries of the target interval, so its penalty is smaller. Tang et al. [2025] concluded that the Probability of an Imbalance in the Wrong Direction (PIWD), as suggested by Thall et al. [2015], is not adequate to inform the burn-in length because it fails to capture the magnitude of the impact on the final allocation. Our Expected Final Allocation Error (ϵ_ρ) incorporates this idea, penalizing allocations towards the wrong arm based on the actual number of patients, not just the probability of an error.

Impact on burn-in A high Expected Final Allocation Error ϵ_ρ suggests that the design is frequently “confused”, exhibiting high bias or instability, and failing to converge within the prudent target interval. This confusion can even lead to the algorithm allocating a substantial proportion of patients to the wrong arm. This indicates a clear need for a larger burn-in. This “protects” the trial from being misled by early, random data, thereby reducing the risk of high bias and lowering the final expected error ϵ_ρ .

4 Formula for burn-in length

To determine an appropriate length for the burn-in period for a specific RAR design in a particular trial with n available patients, we synthesize the metrics introduced in Section 3. We propose a rule that balances the problem difficulty (δ and n) against the design's specific risks (r and ϵ_ρ) and scales with the trial size (n). We propose using the following rule to determine the number of burn-in patients b per arm:

$$b = \max \left\{ 2, \left\lfloor 0.5 \cdot \frac{n \cdot n_{1/2}}{n + n_{1/2}} \cdot (r + \epsilon_\rho)^\delta \right\rfloor \right\} \quad (8)$$

where the formula is floored at 2, representing the smallest burn-in. The $\lfloor \cdot \rfloor$ operation is used to round down the calculated value to the nearest integer. The components are:

- $0.5 \cdot \frac{n \cdot n_{1/2}}{n + n_{1/2}} \in [1, \min\{n, n_{1/2}\})$ is the available burn-in budget, where
 - we multiply with 0.5 because we split the budget among 2 arms,
 - $n \in [1, \infty)$ is the total estimated trial size,
 - and $n_{1/2} \in [1, \infty)$ is the saturation parameter from Section 3.2, which makes the influence of the trial size non-linear.
- $r \in [0, 1/2]$ is the reactivity parameter (design speed) from Section 3.3.
- $\epsilon_\rho \in [0, 1/2]$ is the expected final allocation error (design bias/error) from Section 3.4.
- $\delta \in [0, \infty)$ is the standardized treatment effect (problem difficulty) from Section 3.1.

Interaction of components The Formula (8) captures the interplay between these factors. The total burn-in $2b$ is a product of three conceptual parts:

1. **Available Sample Size:** The term $\frac{n \cdot n_{1/2}}{n + n_{1/2}}$ serves as the “burn-in budget”. As established in Section 3.2, this term scales non-linearly with n , ensuring that the burn-in proportion $(2b/n)$ decreases for very large trials.
2. **Design Risk:** The term $(r + \epsilon_\rho) \in [0, 1]$ is the “risk multiplier”. It combines the risk of a design that converges to fast to an extreme limit (r) with the risk of a design that is biased (ϵ_ρ). A perfectly stable and “non-adaptive” design (like ER) would have $r = 0$ and $\epsilon_\rho = 0$, causing the formula to default to the smallest burn-in of 2.
3. **Problem Difficulty:** The standardized treatment effect δ acts as an *exponent* on the Design Risk term. This creates the most important interaction:
 - When the problem is difficult (i.e., δ is small, close to 0), the exponent is small. Since $(r + \epsilon_\rho)$ is a fraction in $[0, 1]$, a small exponent inflates this risk term (e.g., $0.5^{0.3} \approx 0.81$). This significantly increases the burn-in, reflecting the high need for caution.
 - When the problem is easy (i.e., δ is large), the exponent is large. This shrinks the risk term (e.g., $0.5^{2.0} = 0.25$). This reduces the burn-in, as the strong signal (high δ) means even a risky design (high r) is less dangerous and will likely converge correctly.

In summary, the rule first establishes a burn-in budget based on trial size, then multiplies it by a risk factor derived from the design’s reactivity and instability. This result is then critically amplified or dampened based on the difficulty of the specific clinical problem, as captured by δ .

For any given RAR design and trial, one can estimate the reactivity parameter r , the expected final allocation error ϵ_ρ , and the standardized treatment effect δ to determine the appropriate burn-in. This framework is general, as these metrics can be estimated for any RA(R) procedure, including those not investigated in this publication or novel designs yet to be developed.

5 Simulation studies

We present two examples with differing problem difficulty (δ) and sample size (n) to illustrate that adaptive designs are not off-the-shelf methods they require careful adjustment. Our burn-in formula adapts to each scenario, yielding distinct results for the early-phase and late-phase examples.

5.1 Metrics affected by the burn-in

The burn-in length affects operating characteristics such as

- **Type-I error rate:** The probability of incorrectly rejecting the null hypothesis $H_0 : p_0 = p_1$ when it is true.
- **Statistical power:** The probability of correctly rejecting H_0 when a specific alternative hypothesis $H_1 : p_0 \neq p_1$ is true.
- **Proportion of patients allocated to the best arm:** A measure of in-trial patient benefit, calculated as

$$\max \{E[n_0(n)/n \mid p_0 \geq p_1], E[n_1(n)/n \mid p_1 > p_0]\}.$$

- **Mean Squared Error (MSE):** A measure of estimation accuracy for the treatment effect $\Delta = p_1 - p_0$. It is the expected squared difference between the estimator ($\hat{\Delta} = \hat{p}_1 - \hat{p}_0$) and its true value: $E[(\hat{\Delta} - \Delta)^2 \mid p_0, p_1]$.

RAR designs, particularly those targeting optimal proportions or BRAR, can suffer from substantial type-I error rate inflation when no burn-in period is used [Pin et al., 2025b, Tang et al., 2025]. Tang et al. [2025] specifically investigates the role of the burn-in length, showing that increasing the burn-in reduces this type-I error inflation, but does not fully mitigate it. Furthermore, they demonstrate that the burn-in length has a substantial influence on both statistical power and participant benefit (such as the % of patients allocated to the best arm). Crucially, these operating characteristics are often not maximized at the minimum or maximum possible burn-in length, suggesting a complex trade-off where an intermediate burn-in duration may be optimal.

We will investigate our proposed formula against three fixed burn-in baselines: the minimal burn-in, $b = 2$; a “mid-point” fixed burn-in, $b = n/3$; and the maximal burn-in, $b = n/2$, which corresponds to a full ER design. The $b = 2$ and $b = n/2$ choices represent the most extreme (and common) arbitrary selections. We include the $b \approx 0.33n$ comparator as a fixed alternative, inspired by literature on optimal timing for other adaptive trials (e.g., early stopping and sample size re-estimation often place the interim analysis between one-third and one-half of the total sample size) [Togo and Iwasaki, 2013, Ohn, 2011]. Our simulation aims to demonstrate that our formula achieves a better balance of operating characteristics than any of these fixed choices.

5.2 Statistical tests for final analysis

For the final analysis of the trial, we test the null hypothesis $H_0 : p_0 = p_1$ against the two-sided alternative $H_1 : p_0 \neq p_1$ at a significance level $\alpha = 0.05$. Let n_k be the total number of patients allocated to arm $k \in \{0, 1\}$, and let S_k be the number of successes observed on arm k . The success probability on each arm is estimated as $\hat{p}_k = S_k/n_k$.

5.2.1 Wald test

The (unpooled) Wald test is a standard test for this hypothesis. The test statistic is given by:

$$Z_1 = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}}} \quad (9)$$

H_0 is rejected if $|Z_1| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

5.2.2 Score test

The (pooled) score test is often recommended for its more robust performance, particularly in RAR designs, as it maintains the nominal type-I error rate more effectively [Eberhardt and Fligner, 1977]. It is derived under the null hypothesis H_0 , assuming a common success probability p . This common probability is estimated by the pooled estimator:

$$\bar{p} = \frac{S_0 + S_1}{n_0 + n_1} = \frac{S_0 + S_1}{n} \quad (10)$$

The score test statistic is then:

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\bar{p}(1-\bar{p}) \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}} \quad (11)$$

H_0 is rejected if $|Z_0| > z_{1-\alpha/2}$.

5.3 ARREST trial

Our first case study is based on the parameters from the ARREST (Advanced R²Eperfusion STRategies for Refractory Cardiac Arrest) trial [Yannopoulos et al., 2020]. This trial investigated a novel ECMO-facilitated resuscitation (experimental arm) against standard advanced cardiac life support (control arm) for adults experiencing refractory out-of-hospital cardiac arrest. The primary endpoint was survival to hospital discharge.

We adopt the parameters from the trial’s alternative hypothesis for our simulations: a control success rate of $p_0 = 0.12$ and an experimental success rate of $p_1 = 0.37$. The total sample size is set to $n = 86$ to yield approximately 80% power under ER for the Wald test. This scenario represents a case with a large absolute treatment difference ($\Delta = 0.25$) and a moderate standardized treatment effect ($\delta \approx 0.3095$).

Table 2: Estimated reactivity parameters and burn-in recommendations for the ARREST trial where $p_0 = 0.12$, $p_1 = 0.37$ and $n = 86$ and CALISTO trial where $p_0 = 0.941$, $p_1 = 0.991$ and $n = 360$ (1000 simulations). CIR stands for confidence interval radius (95%, based on a normal approximation).

ARREST					
Design	r (x100)	ϵ_ρ (x100)	$r + \epsilon_\rho$ (x100)	b	BP (x100)
ER	0.00	0.00	0.00	-	-
PBB	57.60	0.00	57.60	32	72.66
BRAR (U)	22.19	0.16	22.36	20	46.40
BRAR (T)	6.62	0.06	6.68	12	27.12
N_0	27.94	2.05	29.99	23	52.63
N_1	4.83	32.86	37.69	26	59.02
R_0	19.65	1.76	21.40	20	45.61
R_1	13.66	29.06	42.73	27	62.27
PTW	31.95	0.96	32.91	24	54.64
RPW	18.82	1.44	20.26	18	41.70
max CIR	1.01	0.89	0.96	0.49	1.13

CALISTO					
Design	r (x100)	ϵ_ρ (x100)	$r + \epsilon_\rho$ (x100)	b	BP (x100)
ER	0.00	0.00	0.00	-	-
PBB	70.03	0.00	70.03	124	68.55
BRAR (U)	25.64	0.51	26.15	98	54.09
BRAR (T)	4.13	0.02	4.15	69	37.87
N_0	22.00	4.11	26.12	98	54.38
N_1	24.83	27.04	51.87	116	63.95
R_0	7.26	0.21	7.47	61	33.44
R_1	7.42	0.21	7.64	67	36.84
PTW	35.10	2.77	37.86	106	58.80
RPW	7.56	5.66	13.22	79	43.76
max CIR	1.05	0.53	1.15	2.03	1.13

5.3.1 Burn-in considerations

The sample size budget for the burn-in is ≈ 85.27 , standardized treatment effect is $\delta \approx 0.3095$ and the values for the reactivity r and expected final allocation error ϵ_ρ are displayed in Table 2 for the different designs alongside the estimated recommended burn-in b .

The PBB design acts as a benchmark with a burn-in of $b = 32$, corresponding to a BP of 72.66%. For the adaptive designs evaluated, the recommended burn-in b varies significantly, ranging from a minimum of 12 for BRAR (T) to a maximum of 27 for R_1 . PTW is the most reactive ($r = 31.95$), apart from the PBB design, but notably maintains a low allocation error ($\epsilon_\rho = 0.96$). In contrast, R_1 is only moderately reactive ($r = 13.66$) but, together with N_1 , it incurs the highest expected final allocation error ($\epsilon_\rho = 29.06$ and $\epsilon_\rho = 32.86$, respectively). This high error results in R_1 having the largest combined $r + \epsilon_\rho$ score (42.73) among all tested designs and therefore the largest burn-in.

5.3.2 Simulations results

The simulation results for the ARREST trial, with ER serving as the baseline (type-I error rate = 5.93% for Z_1 and Z_0 , Power $\approx 80\%$, MSE = 0.0078), reveal the critical importance of the burn-in period.

The most striking observation is the failure of highly reactive designs when implemented with the smallest burn-in of $b = 2$. Designs such as PBB, N_1 , R_1 , and BRAR (U) exhibit extreme type-I error rate inflation for Z_1 (e.g., 78.24%, 89.80%, 89.65%, and 20.86%, respectively) - as pointed out by Pin et al. [2025b] for N_1 and R_1 - and high MSE. This demonstrates that immediate adaptation is nonviable for these methods.

Implementing a larger burn-in is essential. The comparison between our flexible b (e.g., $b = 26$ for N_1 , $b = 32$ for PBB) and the fixed $b \approx n/3$ baseline ($b = 29$) is nuanced. The $b = 29$ rule is a strong performer for controlling type-I error rate (Z_1) and power (Z_0) in many designs. However, our flexible b formula often finds a better balance for the Score test, securing better type-I error rate (Z_0) control in several key designs (PBB, BRAR(U), BRAR(T), RPW).

For the BRAR designs, our flexible b ($b = 20$ and $b = 12$) offers the best type-I error rate (Z_0) control (2.70% and 4.21%, respectively). The PTW design emerges as a special case where $b = 2$ is surprisingly effective, securing the

Table 3: ARREST trial where $p_0 = 0.12$, $p_1 = 0.37$ and $n = 86$ (10000 simulations). **Color Scheme:** Compares each value to the baseline **ER** row or a fixed target. **Type-I Error** (Z_1, Z_0) compares to an ideal 5 (Dark Green: 4–5, Light Green: 0–4 or 5–6, Yellow: 6–10, Red: >10). **Power** (Z_1, Z_0) compares to ER values (80.88, 79.94) (Dark Green: >ER, Light Green: within 2 points, Yellow: 2–5 points below, Red: >5 points below). n_1/n compares to 0.500 (Yellow: 0.475–0.525, Dark Green: >0.525, Red: <0.475). **MSE** compares to ER (0.0078) ± 0.001 (Dark Green: 0.0070–0.0090, Red: >0.0090, Yellow: <0.0070).

Bolding: For a given design, indicates the **best** value among the three burn-in options ($b=2$, flexible b , $b=N/3$). “Best” is defined as: type-I error rate closest to 5.0 within [0, 5.2], Power highest, n_1/n highest, and MSE lowest.

Design	Burn-In	Type-I error		Power		n_1/n	MSE
		Z_1	Z_0	Z_1	Z_0		
ER	-	5.93	5.93	80.88	79.94	0.500	0.0078
PBB	2	78.24	5.83	78.47	0.42	0.977	0.0591
	32	6.66	4.17	79.56	74.34	0.628	0.0078
	29	4.82	7.06	71.86	79.71	0.663	0.0078
BRAR (U)	2	20.86	0.64	74.82	40.71	0.835	0.0113
	20	13.83	2.70	77.97	66.21	0.735	0.0079
	29	4.38	9.90	73.80	80.14	0.649	0.0078
BRAR (T)	2	9.95	4.10	80.78	74.03	0.691	0.0081
	12	9.19	4.21	80.31	73.79	0.685	0.0080
	29	4.40	7.67	76.38	80.65	0.615	0.0079
N_0	2	-	5.94	-	79.52	0.393	0.0090
	23	-	5.52	-	79.15	0.399	0.0091
	29	-	3.59	-	78.27	0.415	0.0087
N_1	2	89.80	-	94.79	-	0.714	0.0530
	26	12.16	-	81.86	-	0.591	0.0077
	29	4.96	-	79.04	-	0.581	0.0077
R_0	2	-	5.50	-	77.74	0.442	0.0082
	20	-	5.75	-	79.07	0.442	0.0080
	29	-	3.93	-	79.25	0.446	0.0078
R_1	2	89.65	-	94.84	-	0.753	0.0347
	27	11.37	-	82.07	-	0.609	0.0075
	29	5.36	-	78.42	-	0.598	0.0077
PTW	2	5.12	4.51	81.01	78.84	0.578	0.0075
	24	5.61	4.41	81.38	79.12	0.536	0.0076
	29	4.60	5.87	79.23	81.65	0.526	0.0075
RPW	2	6.05	4.42	80.26	76.87	0.578	0.0077
	18	5.84	4.52	81.18	78.45	0.559	0.0077
	29	4.80	5.66	79.08	81.42	0.536	0.0077

best type-I error rate for both Z_1 (5.12%) and Z_0 (4.51%), as well as the best MSE (0.0075). For RPW, our flexible b ($b = 18$) provides a good balance, with optimal type-I error rate (Z_0) (4.52%) and power (Z_1) (81.18%).

In general, increasing the burn-in period from $b = 2$ to either the flexible b or the $b \approx n/3$ fixed value consistently maintains or improves MSE and leads to better type-I error rate control.

5.4 CALISTO trial

As a second case study to illustrate the application of the proposed methods, we use the CALISTO trial [Decousus et al., 2010]. This study investigated the efficacy of Arixtra (treatment) compared to a placebo in patients with acute symptomatic lower limb thrombophlebitis. The primary efficacy endpoint was a composite measure; a successful outcome was defined as the absence of death, symptomatic pulmonary embolism, symptomatic deep-vein thrombosis, or symptomatic recurrence/extension of the thrombosis by day 47. The study reported high success rates of 99.1% for the Arixtra group and 94.1% for the placebo group. This corresponds to a small absolute treatment difference ($\Delta = 0.05$) and, due to the high variance near the boundary, a very small standardized treatment effect ($\delta \approx 0.1515$).

While the original study enrolled 1502 patients (of a planned 3002), we adopt its parameters to redesign a new hypothetical trial. For our simulations, we set the total sample size to $n = 360$, as this is the size required to achieve approximately 80% power with ER.

5.4.1 Burn-in considerations

Given a sample size budget of approximately 347.5 and a standardized treatment effect of $\delta \approx 0.1515$, the trade-offs between design reactivity r , expected final allocation error ϵ_ρ , and the recommended burn-in b are displayed in Table 2. The PBB design serves as a benchmark with $b = 124$ and a BP of 68.55%. The recommended burn-in b for the adaptive designs spans from 61 for R_0 to 116 for N_1 . The PTW design is the most reactive ($r = 35.10$) while maintaining a relatively low allocation error ($\epsilon_\rho = 2.77$). Conversely, the N_1 design, despite having moderate reactivity ($r = 24.83$), incurs by far the highest expected final allocation error ($\epsilon_\rho = 27.04$), resulting in the largest combined $r + \epsilon_\rho$ score (51.87) among all designs.

Compared to the ARREST trial, several differences are notable. Due to the larger sample size, the total range of the recommended burn-in b is larger (61–116 vs. 12–27). The largest BP , represented by the PBB, has decreased to 68.55% (from 72.66%), which can be attributed to non-linear sample size influences. Furthermore, the design requiring the largest burn-in has changed from R_1 in ARREST to N_1 in this trial. This shift is partially due to the different treatment effect and sample size, but also reflects a change in the relative orders of reactivity r and error ϵ_ρ across the designs.

5.4.2 Simulations results

The CALISTO trial simulations, with an ER baseline showing near-ideal type-I error rate control ($Z_1 = 4.91\%$, $Z_0 = 4.90\%$) and 79.5% power, demonstrate that the $b = 2$ burn-in is completely nonviable in this setting. It leads to catastrophic Z_1 Type-I error inflation (e.g., 96.23% for N_1) and massively inflated MSE.

This establishes that a substantial burn-in is mandatory. The core trade-off for any large burn-in is sacrificing patient benefit (n_1/n) to gain type-I error rate control and MSE stability. The $b \approx n/3$ ($b = 120$) rule and our flexible b rule both successfully fix the type-I error rate and MSE issues that $b = 2$ failed. The crucial difference lies in the efficiency of this trade-off.

The fixed $b = 120$ rule is a strong performer for the Z_1 (Wald) test, as it is essential for “rescuing” the Z_1 power from its $b = 2$ collapse. However, our flexible b formula is overwhelmingly superior for the Z_0 (Score) test, securing the best Z_0 power in 8 of the 9 adaptive designs. This $b = 120$ rule’s “win” on Z_1 power comes at a high cost, as it consistently reduces patient benefit (n_1/n) and Z_0 power compared to the flexible b .

For the BRAR designs, our flexible b ($b = 98$ and $b = 69$) offers the best Z_0 power (80.63% and 81.89%) and higher patient benefit. The fixed $b = 120$ rule, while boosting Z_1 power, cuts Z_0 power significantly (e.g., to 73.15% for BRAR(T)) and reduces n_1/n . For RPW, our flexible b ($b = 79$) provides the best overall balance: it has the best Z_0 power (80.12%), excellent type-I error rate control for both tests ($Z_1 = 5.07\%$, $Z_0 = 5.24\%$), the best MSE (0.0004), and higher patient benefit. The $b = 120$ rule’s “win” on Z_1 power is trivial (79.97% vs 79.39%) and not worth the cost to other metrics.

In conclusion, increasing the burn-in period from $b = 2$ is mandatory. The fixed $b \approx n/3$ rule is a blunt instrument that prioritizes Z_1 (Wald) power at the expense of Z_0 (Score) power and patient benefit. Our flexible b provides a far superior and more balanced trade-off, especially for the trialist who values patient benefit and intends to use the more robust Z_0 test for their final analysis.

6 Discussion

This paper introduces the first systematic framework for determining the burn-in length in RAR trials. Moving beyond arbitrary justifications, we explicitly incorporate key factors governing the burn-in decision: the non-linear impact of sample size (Section 3.2), two novel metrics to characterize a design’s behavior: its reactivity (r) (Section 3.3) and its expected final allocation error (ϵ_ρ) (Section 3.4) and the problem’s difficulty via the standardized treatment effect (δ). The result of this framework is the principled, general-purpose formula (Equation (8)) that combines these components in a meaningful, interactive way.

Simulation studies (Section 5) provide a clear and expected warning: implementing highly reactive designs (N_1 , R_1 , PBB) with the smallest burn-in ($b = 2$) results in severe performance issues, including type-I error rate inflation and inflated MSE. This demonstrates that insufficient initial exploration makes these algorithms vulnerable to misleading early data, leading to statistically unsound results, particularly in challenging scenarios, e.g small true effects (CALISTO) or high variability.

The primary finding is that our recommended burn-in b acts as an effective stabilizer, it consistently and effectively mitigated the severe type-I error inflation and brought the MSE back in line with the non-adaptive ER baseline. We

Table 4: CALISTO Trial with $p_0 = 0.941$, $p_1 = 0.991$ and $n = 360$ (10000 simulations)

Color Scheme: Compares each value to the baseline **ER** row or a fixed target. **Type-I Error** (Z_1, Z_0) compares to an ideal 5 (Dark Green: 4–5, Light Green: 0–4 or 5–6, Yellow: 6–10, Red: >10). **Power** (Z_1, Z_0) compares to the ER values (79.53, 79.53) (Dark Green: >ER, Light Green: within 2 points, Yellow: 2–5 points below, Red: >5 points below). n_1/n compares to 0.500 (Yellow: 0.475–0.525, Dark Green: >0.525, Red: <0.475). **MSE** compares to ER (0.0004) \pm 0.0002 (Dark Green: 0.0002–0.0006, Red: >0.0006, Yellow: <0.0002).

Bolding: For a given design, indicates the **best** value among the three burn-in options ($b=2$, flexible b , $b=N/3$). “Best” is defined as: type-I error rate closest to 5.0 within [0, 5.2], Power highest, n_1/n highest, and MSE lowest.

Design	Burn-In	Type-I error		Power		n_1/n	MSE
		Z_1	Z_0	Z_1	Z_0		
ER	-	4.91	4.90	79.53	79.53	0.500	0.0004
PBB	2	88.58	11.74	35.58	12.15	0.994	0.0289
	124	5.45	4.71	71.10	81.57	0.656	0.0005
	120	4.79	5.24	79.57	67.91	0.667	0.0005
BRAR (U)	2	2.62	16.94	5.13	76.12	0.881	0.0176
	98	2.93	6.55	56.24	80.63	0.708	0.0005
	120	5.94	3.07	82.80	71.10	0.655	0.0005
BRAR (T)	2	3.38	6.39	63.66	81.21	0.701	0.0007
	69	3.31	6.02	67.31	81.89	0.681	0.0006
	120	5.29	3.70	82.12	73.15	0.620	0.0005
N_0	2	-	5.25	-	79.97	0.723	0.0016
	98	-	5.17	-	79.86	0.663	0.0005
	120	-	3.51	-	70.65	0.624	0.0004
N_1	2	96.23	-	94.17	-	0.158	0.0009
	116	7.92	-	83.94	-	0.363	0.0003
	120	4.80	-	75.50	-	0.371	0.0003
R_0	2	-	5.75	-	81.37	0.672	0.0007
	61	-	6.32	-	81.39	0.661	0.0005
	120	-	3.48	-	71.09	0.599	0.0004
R_1	2	5.65	-	78.79	-	0.506	0.0024
	67	5.28	-	79.05	-	0.505	0.0004
	120	4.94	-	79.45	-	0.505	0.0004
PTW	2	2.72	6.12	9.56	74.53	0.847	0.0049
	106	4.23	5.52	69.23	80.31	0.636	0.0005
	120	5.40	4.41	80.30	72.88	0.607	0.0004
RPW	2	4.85	5.43	67.77	75.95	0.582	0.0006
	79	5.07	5.24	79.39	80.12	0.519	0.0004
	120	5.28	5.38	79.97	79.45	0.510	0.0004

observed that while the score test Z_0 generally provides better type-I error control than the Wald test Z_1 , a larger burn-in can lead to improved type-I error rate control for both tests across certain designs.

This stabilization must be compared against reasonable heuristics, such as a fixed $b \approx n/3$ rule. Our simulations revealed this fixed approach acts as a “blunt instrument”, enforcing a long ER period. While this successfully stabilizes MSE and type-I error rate, and rescued the Wald test’s power in challenging settings like CALISTO. This “win” came at a high and often unnecessary cost: reduced patient benefit and, critically, degraded power for the more robust score test (Z_0). This is where our flexible formula demonstrates its advantage. It performs an efficient balance, not just stabilization. For example, in the RPW design in CALISTO, our formula found a “sweet spot” missed by fixed rules. It achieved near-optimal type-I error rate, MSE, and Z_0 power, whereas the $b \approx n/3$ rule’s trivial gains in Z_1 power were achieved at the expense of all other key metrics. Our framework thus provides a more nuanced tool that performs robustly across both test statistics while better preserving the patient-benefit and efficiency benefits of adaptation.

The introduction of the novel metrics, reactivity (r) and expected final allocation error (ϵ_p), opens new avenues for future research. This framework now allows the first quantitative comparison of the inherent speed and stability of any two-arm RAR algorithm for binary outcomes, applicable even to future or specialized designs. More broadly, the core insights derived here are valuable for other common adaptive designs, like those involving sample size re-estimation or early stopping. Since these decisions are similarly vulnerable to early data variance, understanding the “reactiveness” of their stopping or re-estimation rules is crucial. Applying a principled stability analysis, analogous to

the one developed here, could help ensure these adaptations are triggered based on reliable information, but not too late.

A key practical recommendation for future research is the concept of *burn-in re-estimation*. While our simulations used a fixed burn-in b , in practice, clinicians could update the estimates of the standardized treatment effect δ using Maximum Likelihood Estimates (MLEs) from the accumulated data (Equation (1)) at the end of the initial burn-in phase. If the estimated effect is smaller or the variability is higher than expected (indicating a more challenging scenario), extending the burn-in period would be warranted before proceeding to the adaptive phase. No adjustment would be needed if the estimates align with prior assumptions.

This research contains several limitations that open avenues for future work. First, our formula contains a “meta-parameter”, $n_{1/2}$, which we set to 1000 to define a reasonable saturation curve (see Figure 2). Future work could explore the formula’s sensitivity to this choice and allow practitioners to tune its cautiousness based on risk tolerance. Moreover, practitioners may wish to tune the formula’s cautiousness based on their risk tolerance. This could be done by scaling the final recommendation (e.g., $b' = C \times b$ for $C = 1.2$) or by scaling the exponent (e.g., $b' \propto (r + \epsilon_\rho)^{C \cdot \delta}$) to explicitly modify the formula’s sensitivity to the problem’s difficulty. Second, our framework is currently specified for two-arm trials. A non-trivial but necessary extension would be to adapt it for multi-arm ($K > 2$) trials. This would likely involve generalizing the 0.5 scaling factor to $1/K$ and, more complexly, redefining δ based on a global null hypothesis and adapting r and ϵ_ρ to handle a K -dimensional allocation vector. Third, for response-adaptive clinical trials with a blocked structure (where the adaptive allocation probability is fixed within blocks [Proper et al., 2021]), we recommend using Equation (8). The parameters r and ϵ_ρ should be calculated based on the full RAR procedure, including the specific block randomization. Practically, we suggest setting the first block size equal to the total recommended burn-in length or ensuring the first few blocks maintain a 1:1 randomization ratio until the total participant count covers the recommended burn-in length. Future research could investigate this decision rule’s performance in blocked designs. Fourth, one could explore how other measures of interest, such as the relative risk or odds ratio, affect the burn-in Formula (8). Pin et al. [2024] derived how optimal designs change for different measures of interest. Analogously, we would need to redefine the standardized treatment effect δ based on the measure of interest and transformed arm response variances. Fifth, our formula is a principled heuristic designed for practical application. A more complex alternative would be to frame the choice of b as a formal optimization problem. This would involve defining a utility function that weights all key operating characteristics: Power, Patient Benefit (n_{best}/n), type-I Error, and MSE. One could then use extensive simulations to find the b that maximizes this utility. This is a computationally extensive task. Our method is lower-computation in comparison, as it only requires estimating r and ϵ_ρ to arrive at a direct recommendation, which can serve as a starting point for any finer-grained search. As an alternative to a single utility function, a two-stage optimization may better reflect regulatory priorities: (i) identify the range of b that satisfies hard constraints (e.g., type-I error ≤ 0.05 and $\text{MSE} \leq \text{MSE}_{\text{ER}}$), and (ii) find the b within that valid range that maximizes a weighted sum of power and patient benefit. Finally, our work does not consider the impact of delayed outcomes or cohort-based enrollment, which are common practical challenges. Further research could explore how these factors interact with the burn-in phase.

In conclusion, this paper challenges the long-standing practice of selecting burn-in lengths by “guesswork”. We provide a practical, data-driven, and generalizable tool to move this decision from ad-hoc art to a principled, scientific choice, ultimately fostering the design of safer, more reliable, and more efficient adaptive trials.

Acknowledgements

The authors acknowledge the use of large language models (LLMs) to assist with generating figures and refining grammar and wording in this paper. The LLMs were not used for data analysis, interpretation, or original scientific writing. All content has been carefully reviewed and verified by the authors, who take full responsibility for the integrity and accuracy of the work.

The authors acknowledge funding and support from the UK Medical Research Council (grants MC UU 00002/19 (GC), MC UU 00002/15 and MC UU 00040/03 (SSV, DSR, SB)), as well as an MRC Biostatistics Unit Core Studentship (LP) and the Cusanuswerk e.V. (LP). SSV is part of PhaseV’s advisory board.

References

S. Baas, P. Jacko, and S. S. Villar. Exact statistical analysis for response-adaptive clinical trials: A general and computationally tractable approach. *Computational Statistics & Data Analysis*, 211:108207, 2025a. ISSN 0167-9473. doi:<https://doi.org/10.1016/j.csda.2025.108207>.

- S. Baas, L. Pin, S. S. Villar, and W. F. Rosenberger. A computational method for type i error rate control in powermaximizing responseadaptive randomization. *arXiv preprint arXiv:2509.12448*, 2025b. doi:10.48550/arXiv.2509.12448.
- V. W. Berger, L. J. Bour, K. Carter, J. J. Chipman, C. C. Everett, N. Heussen, et al. A roadmap to using randomization in clinical trials. *BMC Medical Research Methodology*, 21:168, 2021. ISSN 1471-2288. doi:10.1186/s12874-021-01303-z.
- H. Decousus, P. Prandoni, P. Mismetti, R. M. Bauersachs, Z. Boda, B. Brenner, et al. Fondaparinux for the treatment of superficial-vein thrombosis in the legs. *New England Journal of Medicine*, 363:1222–1232, 2010.
- Y. Du, J. D. Cook, and J. J. Lee. Comparing three regularization methods to avoid extreme allocation probability in response-adaptive randomization. *Journal of Biopharmaceutical Statistics*, 28(2):309–319, 2017. doi:10.1080/10543406.2017.1293077.
- K. R. Eberhardt and M. A. Fligner. A comparison of two tests for equality of two proportions. *The American Statistician*, 31(4):151–155, 1977. doi:10.1080/00031305.1977.10479225.
- F. Hu and W. F. Rosenberger. *The Theory of ResponseAdaptive Randomization in Clinical Trials*. Wiley, 2006. ISBN 9780471653967. doi:10.1002/047005588X.
- F. Hu, L. X. Zhang, and X. He. Efficient randomized-adaptive designs. *Annals of Statistics*, 37:2543–2560, 2009. ISSN 00905364. doi:10.1214/08-AOS655.
- J. Neyman. On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558, 1934. ISSN 09528385. doi:10.2307/2342192.
- C. F. Ohn. *Group Sequential and Adaptive Methods Topics with Applications to Clinical Trials*. PhD thesis, University of Bath, 2011. URL https://purehost.bath.ac.uk/ws/portalfiles/portal/187958538/UnivBath_PhD_2011_C_Ohn.pdf. Accessed: 20251114.
- P. Pallmann, A. W. Bedding, B. Choodari-Oskooei, M. Dimairo, L. Flight, L. V. Hampson, et al. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Medicine*, 16(1):29, 2018. doi:10.1186/s12916-018-1017-7.
- L. Pin, S. S. Villar, and W. F. Rosenberger. Response-adaptive randomization designs based on optimal allocation proportions. In D. G. Chen, editor, *Biostatistics in Biopharmaceutical Research and Development*. Springer, 2024. doi:10.1007/978-3-031-65948-5_12. URL https://doi.org/10.1007/978-3-031-65948-5_12.
- L. Pin, S. Baas, D. S. Robertson, and S. S. Villar. Is 1:1 always most powerful? why careful determination of allocation ratios matters in trial design. *arXiv preprint arXiv:2507.13036*, 2025a. doi:10.48550/arXiv.2507.13036.
- L. Pin, S. S. Villar, and W. F. Rosenberger. Revisiting Optimal Allocations for Binary Responses: Insights from Considering Type-I Error Rate Control. *Biometrics*, 81(3):ujaf114, 08 2025b. ISSN 0006-341X. doi:10.1093/biomet/ujaf114.
- J. Proper, J. Connett, and T. Murray. Alternative models and randomization techniques for Bayesian response-adaptive randomization with binary outcomes. *Clinical Trials*, 18(4):417–426, 2021. doi:10.1177/17407745211010139.
- W. F. Rosenberger, N. Stallard, A. Ivanova, C. N. Harper, and M. L. Ricks. Optimal adaptive designs for binary response trials. *Biometrics*, 57:909–913, 2001. ISSN 0006341X. doi:10.1111/j.0006-341X.2001.00909.x.
- E. Y. N. Tang, S. Baas, D. Kaddaj, L. Pin, D. S. Robertson, and S. S. Villar. A burn-in(g) question: How long should an initial equal randomization stage be before bayesian response-adaptive randomization?, 2025. URL <https://arxiv.org/abs/2503.19795>.
- P. Thall, P. Fox, and J. Wathen. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology*, 26:1621–1628, 2015. ISSN 09237534. doi:10.1093/annonc/mdv238.
- W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933. ISSN 0006-3444. doi:10.1093/biomet/25.3-4.285.
- K. Togo and M. Iwasaki. Optimal timing for interim analyses in clinical trials. *Journal of Biopharmaceutical Statistics*, 23(5):1067–1080, 2013. doi:10.1080/10543406.2013.813522.
- K. Viele. Response-adaptive randomization in clinical trials: Current opinion considering recent publications. Blog post on Berry Consultants website, Apr. 2025. URL <https://docs.berryconsultants.com/blog/posts/2025-04-22.html>. Accessed: 2025-11-14.
- S. S. Villar, J. Bowden, and J. Wason. Responseadaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical Statistics*, 17(2):182–197, 2018. doi:10.1002/pst.1845.

- L. J. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73:840–843, 1978. ISSN 0162-1459. doi:10.1080/01621459.1978.10480109.
- I. Wilson, S. A. Julious, C. Y. Yap, S. Todd, and M. Dimairo. Response adaptive randomisation in clinical trials: Current practice, gaps and future directions. *Statistical Methods in Medical Research*, 34(9):1851–1874, 2025. ISSN 0962-2802. doi:10.1177/09622802251348183. Epub 2025 Jun 18.
- D. Yannopoulos, J. Bartos, G. Raveendran, E. Walser, J. Connett, T. A. Murray, 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*, 396(10265):1807–1816, 2020. doi:10.1016/S0140-6736(20)32338-2.
- Y. Yi and X. Li. Response adaptive designs with asymptotic optimality. *Canadian Journal of Statistics*, 46(3):458–469, 2018. doi:10.1002/cjs.11460.
- M. Zelen. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64: 131–146, 1969.
- K. Zhang, L. Janson, and S. Murphy. Inference for batched bandits. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9818–9829. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6fd86e0ad726b778e37cf270fa0247d7-Paper.pdf.