Contemporary Clinical Trials 62 (2017) 114–120

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Impact of adaptation algorithm, timing, and stopping boundaries on the performance of Bayesian response adaptive randomization in confirmative trials with a binary endpoint

Yunyun Jiang*, Wenle Zhao, Valerie Durkalski-Mauldin

Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC, USA

ARTICLE INFO

Keywords:
Adaptive design
Response adaptive randomization
Trial efficiency
Ethics
Allocation variability
Multi-arm trial

ABSTRACT

Despite the concerns of time trend in subject profiles, the use of Bayesian response adaptive randomization (BRAR) in large multicenter phase 3 confirmative trials has been reported in recent years, motivated by the potential benefits in subject ethics and/or trial efficiency. However three issues remain unclear to investigators: 1) among several BRAR algorithms, how to choose one for the specific trial setting; 2) when to start and how frequently to update the allocation ratio; and 3) how to choose the interim analyses stopping boundaries to preserve the type 1 error. In this paper, three commonly used BRAR algorithms are evaluated based on type 1 error, power, sample size, the proportion of subjects assigned to the better performing arm, and the total number of failures, under two specific trial settings and different allocation ratio update timing and frequencies. Simulation studies show that for two-arm superiority trials, none of the three BRAR algorithms has predominant benefits in both patient ethics and trial efficiency when compared to fixed equal allocation design. For a specific trial aiming to identify the best or the worst among three treatments, a properly selected BRAR algorithm and its implementation parameters are able to gain ethical and efficiency benefits simultaneously. Although the simulation results come from a specific trial setting, the methods described in this paper are generally applicable to other trials.

1. Introduction

Bayesian response adaptive randomization (BRAR) has been utilized in early phase trials motivated by potential benefits in trial efficiency and/or patient ethics [1,2]. However, its use in large confirmative phase 3 trials is controversial [3–9]. For two-arm trials, previous research indicated that BRAR may lead to power reduction due to treatment allocation imbalances [10], may assign more patients to the inferior treatment arm due to the allocation ratio variation [11], and may introduce bias in the treatment effect estimation [12]. For multi-arm trials, it is recognized that the performance of BRAR varies based on the adaptation algorithm and the trial setting, and under some scenarios it may provide benefits in patient ethics and trial efficiency simultaneously [13,14,15,16]. For example, Connor et al. reported a simulation study for a three-arm comparative trial, and found that BRAR could randomize a higher proportion of subjects to a better performing arm with a higher power and smaller sample size, as compared to the fixed randomization [14]. Nevertheless, in addition to the concern of a potential time trend in subject profiles during the long period of the trial [17], several important issues associated with the use of BRAR in large phase 3 trials remain unclear to investigators. First, there are several Bayesian response adaptation algorithms proposed in literature [13,14,18]. For a specific trial, how should the investigator choose one among them? Second, to implement a BRAR, when should the treatment allocation ratio update start and how frequently should the allocation ratio be updated? Some suggest to update the allocation ratio every 100 subjects [14], while some others recommended to update the allocation ratio for every one subject [11,18]. What is the impact of the timing and frequency of the allocation ratio update on the operating characteristics of the trial? Third, interim analyses are often conducted in phase III. How should the efficacy and futility stopping boundaries be specified so that the operating characteristics of different BRAR designs can be evaluated appropriately? It is also worth noticing that some multi-arm trials include a control arm [13,16], some others may not [13]. Some trials aim to identify the best treatment [16], while some others may target to identifying the best or the worst arm [14]. The evaluation of a BRAR design under a specific trial setting may not necessarily generalize to other trials settings. Wathen and Thall et al. [19] recently published on the use of a control in a multi-arm BRAR setting. They showed that a control arm is important to include, however, a true

* Corresponding author at: Department of Public Health Sciences, Medical University of South Carolina, 135 Cannon Street, Suite 305, Charleston, SC 29425, USA.
E-mail address: jiany@musc.edu (Y. Jiang).

http://dx.doi.org/10.1016/j.cct.2017.08.019
Received 10 March 2017; Received in revised form 19 July 2017; Accepted 29 August 2017
Available online 01 September 2017

control is not always feasible due to ethical reasons (i.e. life threatening conditions).

In this manuscript, we further explore the performance of three commonly used BRAR algorithms under the assumption of no time trend, and examine the impact of the allocation adaptation frequency and the burn-in period length via simulation studies. Performance of the BRAR is evaluated by the proportion of subjects assigned to the inferior arm, the total number of failures, type I error, power, and the final sample size. When applicable, both the expected value and the variance are included for each assessment. Section 2 reviews the rationale and commonly used algorithms of response adaptive randomization under the Bayesian framework. Section 3 presents the simulation design, followed by the result in Section 4, and discussions in Section 5.

2. Background

Response adaptive randomization was originally motivated by patient ethics [20], which are measured by the proportion of patients assigned to the better performing arm, known as the individual ethics, and the total number of failures, known as the population ethics.

2.1. Three commonly used BRAR allocations

Thompson et al. defined a BRAR scheme as a function of the posterior probability that one treatment is better than the other, i.e. $\pi_j = P(p_j > p_k, k = 1,2,\cdots,m; k \neq j)$, where p_j and p_k are the posterior success probabilities for jth and kth treatments respectively [21]. To reduce the allocation ratio variability, the square root transformation is employed. Generalized to m-arm trials, the randomization probability for arm j is:

$$r_j = \frac{\sqrt{\pi_j}}{\sum_{j=1}^{m}\sqrt{\pi_j}} \quad (j = 1,2,\ldots,m) \tag{1}$$

It is known as the *probability-weighted allocation* [13], and is referred to as BRAR (1/2) hereafter.

Thall et al. proposed a randomization algorithm by including the current sample size proportion in the power transformation:

$$r_j = \frac{(\sqrt{\pi_j})^{\frac{n}{N}}}{\sum_{j=1}^{m}(\sqrt{\pi_j})^{\frac{n}{N}}} \quad (j = 1,2,\ldots,m) \tag{2}$$

here n is the current sample size and N is the trial's maximum sample size [18]. This algorithm was named as the *natural lead-in allocation* by Bello et al. [22], and is referred to as BRAR(n/2N) hereafter. It tends to reduce the allocation variability in the early phase of the trial, and yield a better protection of the statistical power.

Conner et al. [13,14] developed an algorithm to incorporate the precision of the posterior estimate,

$$r_j = \frac{\sqrt{\pi_j Var(p_j)/n_j}}{\sum_{j=1}^{m}\sqrt{\pi_j Var(p_j)/n_j}} \quad (j = 1,2,\ldots,m) \tag{3}$$

here $Var(p_j)$ is defined as variance of the posterior probability of success for arm j. This scheme is named as the *information-weighted allocation*, and is referred as BRAR(1/2, σ^2) hereafter. It allows the allocation ratio to be directed toward the treatment arm with the higher observed success proportion, smaller sample size and lower precision (or greater variance) of the treatment estimation [13,14].

2.2. Implementation parameters

The performance of a BRAR can be affected by many implementation factors and trial design parameters; including the burn-in period length, the frequency of allocation ratio update, the number of arms in the trial, the maximum sample size, the number and timing of interim analyses, the efficacy and futility stopping boundaries, and the treatment efficacy profile. For instance, the pre-specified values of futility and efficacy stopping boundaries can affect the type I error and the power of the trial. The recently reported simulation studies for Bayesian adaptive designs use 99% [18] or 97.5% [14] as the efficacy stopping boundary, which may lead to different type I error rates under different trial design settings, making the power comparison of different trial design less convincing. In the simulation studies, calibrated type I error may be recommended so that the power comparison between different trial designs could have a common base.

This manuscript will closely examine the impact of the factors in the specific trial design settings.

3. Methods

3.1. Basic concepts for Bayesian adaptive design

Consider an m-arm trial with a binary endpoint. Let $x_{i,j}$ denote the outcome of the ith patient in arm j, with $i = 1,2,\ldots,n_j$ and $j = 1,2,\ldots,m$. Let $x_{i,j} = 1$ if the patient's response is a success, and $x_{i,j} = 0$ otherwise. The total number of successes is $s_j = \sum_{i=1}^{n_j} x_{i,j}$. The probability of success for arm j has a uniform non-informative prior distribution $p_j \sim Beta(1,1)$, assuming there is one success and one failure in the two subjects prior to the start of the trial. After observing s_j successes from n_j subjects, the posterior probability of success in arm j follows a Beta distribution $p_j | s_j, n_j \sim Beta(1+s_j, 1+n_j-s_j)$. If the objective of the study is to identify one most effective treatment among the m arms, a trial is being considered as 'success' if its probability of identifying the most (or least) effective treatment exceeds a cut-off value, the posterior probability of treatment j being the most effective arm exceeds γ, which is the pre-specified efficacy stopping boundary. We define an indicator function $\varphi(p_j)$, the decision rule for concluding trial efficacy is $\varphi(p_j) = \begin{cases} 1 & \Pr(p_j > p_k \mid j,k = 1,2,\ldots,m, j \neq k) > \gamma \\ 0 & \text{otherwise} \end{cases}$. Where γ is set at a high probability cut-off value such as 0.99. The power is defined as the expected probability of achieving a significant (positive) result, $\pi = E(\varphi(p_j)) = P(\varphi(p_j) = 1)$.

A trial can also be stopped for futility after an interim analysis, based on the predictive probabilities of having $\tilde{s}_j$ successes in the additional $\tilde{n}_j$ subjects for treatment $j = 1,2,\ldots,m$ prior to reaching the maximum sample size of the study. The number of additional successes follows a beta-binomial distribution $\tilde{s}_j \sim betabin(\tilde{n}_j, 1+s_j, 1+n_j-s_j)$. The value of $\tilde{n}_j$ can be determined by the remaining sample size and adaptive allocation based on observed patients' response outcomes. The trial is stopped if none of the treatment arms has a predictive probability greater than the pre-specified futility stopping boundary. If the trial is not stopped, the treatment allocation ratio will be updated based on the selected BRAR algorithm and the posterior probabilities for each treatment arm.

3.2. Simulation trial design

Computer simulation studies are designed to evaluate the performance of the three aforementioned BRAR algorithms (1–3) under two trial scenarios; a two-arm trial with a maximum sample size of 300 to identify the better performing arm and a three-arm trial with a maximum sample size of 720 to identify the most or the least effective treatment. The three-arm trial setting mimics the Established Status Epilepticus Treatment Trial (ESETT), a large multicenter phase 3 trial funded by the National Institution of Neurological Disorder and Stroke (NINDS) to determine the most or the least effective treatment among fosphenytoin, levetiracetam, and valproic acid in patients with benzodiazepine-refractory status epilepticus older than age 2 years [14]. The simulation procedure includes a burn-in period with a fixed equal allocation ratio. After that, interim analyses are conducted based on the pre-specified frequency, checking the efficacy and futility boundaries based on the posterior probabilities and the predictive probabilities for each arm. The efficacy boundaries for two-arm and three-arm trials are