Article

CLINICAL TRIALS

Clinical Trials
1–9

DOI: 10.1177/1740774517692302
journals.sagepub.com/home/ctj
SAGE

A simulation study of outcome adaptive randomization in multi-arm clinical trials

J Kyle Wathen[1] and Peter F Thall[2]

Abstract

Randomizing patients among treatments with equal probabilities in clinical trials is the established method to obtain unbiased comparisons. In recent years, motivated by ethical considerations, many authors have proposed outcome adaptive randomization, wherein the randomization probabilities are unbalanced, based on interim data, to favor treatment arms having more favorable outcomes. While there has been substantial controversy regarding the merits and flaws of adaptive versus equal randomization, there has not yet been a systematic simulation study in the multi-arm setting. A simulation study was conducted to evaluate four different Bayesian adaptive randomization methods and compare them to equal randomization in five-arm clinical trials. All adaptive randomization methods included an initial burn-in with equal randomization and some combination of other modifications to avoid extreme randomization probabilities. Trials either with or without a control arm were evaluated, using designs that may terminate arms early for futility and select one or more experimental treatments at the end. The designs were evaluated under a range of scenarios and sample sizes. For trials with a control arm and maximum same size 250 or 500, several commonly used adaptive randomization methods have very low probabilities of correctly selecting a truly superior treatment. Of those studied, the only adaptive randomization method with desirable properties has a burn-in with equal randomization and thereafter randomization probabilities restricted to the interval 0.10–0.90. Compared to equal randomization, this method has a favorable sample size imbalance but lower probability of correctly selecting a superior treatment. In multi-arm trials, compared to equal randomization, several commonly used adaptive randomization methods give much lower probabilities of selecting superior treatments. Aside from randomization method, conducting a multi-arm trial without a control arm may lead to very low probabilities of selecting any superior treatments if differences between the treatment success probabilities are small.

Keywords

Adaptive randomization, Bayesian design, play the winner, screening trial, simulation

Introduction

Outcome adaptive randomization (AR) has been proposed by many authors as an alternative to equal randomization (ER), for comparing treatments A and B. AR uses the interim outcome data to unbalance randomization probabilities in favor of the treatment arm, or arms, having currently higher empirical success rates. Proponents of AR consider it more ethical than ER for the patients enrolled in the trial because AR leads to sample sizes, N_A, and N_B, on average unbalanced in favor of the truly superior treatment. AR was proposed by Thompson[1] for binary outcomes. He suggested that, assuming success probabilities π_A and π_B following beta priors, the nth patient should receive treatment A with probability $r_{A,n} = \Pr(\pi_B < \pi_A \mid data_n)$ and B with probability $r_{B,n} = 1 - r_{A,n}$. Adaptive statistical criteria used to define AR probabilities similar to $r_{A,n}$ and $r_{B,n}$ sometimes are called "randomized play-the-winner" rules.[2,3] Many different AR methods have been proposed, and clinical trials have been conducted using various AR methods.[4–10]

[1] Model Based Drug Development, Statistical Decision Sciences, Janssen Research & Development, LLC, Titusville, NJ, USA
[2] Department of Biostatistics, MD Anderson Cancer Center, Houston, TX, USA

Corresponding author:
J Kyle Wathen, Statistical Modeling and Methodology, Quantitative Sciences, Janssen Research & Development, LLC, 1125 Trenton-Harbourton Road, Titusville, NJ 08560, USA.
Email: kwathen@its.jnj.com

Use of AR in clinical trials remains controversial. Critics argue that AR provides a small advantage in sample size imbalance in favor of the superior treatments, while introducing inferential problems that decrease benefit to future patients. Discussions of AR have been given by many authors.[11–18] Berry has argued that the greatest advantages of AR over ER may be obtained in multi-arm trials. Thall et al.[19] reported a simulation study, for two-arm trials, comparing several Bayesian AR methods to a group sequential design using ER.[20] Their simulations showed that, compared to ER, AR methods often have a much lower probability of selecting a truly superior treatment arm, much larger estimation bias, produce distributions of N_A and N_B with much greater variability and skewness, and have a nontrivial probability of unbalancing N_A and N_B in favor of the inferior treatment. Thus, only reporting mean sample sizes from simulations may be very misleading. The particular way an AR method is defined, and other aspects of a trial design, can greatly affect the overall design performance. Because there are numerous ways to design a randomized trial, and many different ways to define AR methods, statements about the comparative desirability of AR versus ER must be accompanied by detailed explanations of these design specifics.

In this article, we report a simulation study examining four AR methods and ER in multi-arm clinical trials. A multi-arm trial design may or may not (1) include a control arm, (2) restrict the randomization to a control arm if it is included, (3) involve various rules for between-arm comparisons or stopping an arm early, (4) enrich the remaining arms with larger sample sizes when some arms are terminated early, (5) select one best or possibly several experimental treatments, and (6) include two or more than two stages, or monitor continuously. Thus, to obtain reasonable comparisons of randomization methods, the underlying designs must have qualitatively identical structures, decision rules, and maximum sample size. To obtain results that are useful to practitioners, we evaluate several relatively simple clinical trial designs and AR methods, for five-arm trials that either do or do not include a control arm. We consider Bayesian designs for trials with binary outcomes that use either ER or one of four specific AR methods.

Outcome AR methods

There are many ways to do AR.[2,6,7,21,22] The Bayesian AR methods considered here are similar to those previously studied for two-arm trials, generalized to accommodate multi-arm trials here.[1,19,23] Index treatments by $k = 1, \ldots, K$, and intermediate sample sizes by $n = 1, \ldots, N$, for maximum overall sample size N. Denoting response probabilities of the K treatments by $\pi_1, \ldots, \pi_K$, the AR probabilities are defined in terms of the K posterior probabilities

$$r_{k,n} = Pr(\pi_k = max\{\pi_1, \ldots, \pi_K\} \mid data_n), \quad k = 1, \ldots, K, \tag{1}$$

which sum to 1. Thus, $r_{1,n}, \ldots, r_{K,n}$ generalize the original definition 1 given for $K = 2$.

It is well known that using $\{r_{1,n}, \ldots, r_{K,n}\}$ as AR probabilities often leads to undesirable treatment assignments due to "stickiness," wherein an outcome adaptive treatment assignment rule assigns a suboptimal treatment to an undesirably large number of patients.[24] With the above AR probabilities, if a truly inferior treatment arm happens to have a higher early success rate, it is likely to receive a larger proportion of patients thereafter, and consequently, the trial design is not likely to identify a truly superior treatment. Various modifications of $r_{k,n}$ have been proposed to mitigate stickiness. We consider AR methods that use different combinations of three such modifications. The first is a "burn-in" wherein, initially, a fixed number of patients are randomized equally among the arms, with AR applied subsequently. The second replaces $r_{k,n}$ with

$$r^{(c)}_{k,n} = \frac{(r_{k,n})^c}{\sum_{j=1}^{K} (r_{j,n})^c} \tag{2}$$

for some $c>0$, with $c = 0.50$ used very commonly. This shrinks $r_{k,n}$ toward .50, so the AR method is more like ER, for which $c = 0$ and all $r^{(0)}_{k,n} \equiv 1/\mathrm{K}$. The third modification restricts $e \le r_{k,n} \le 1 - e$ for small $e>0$. If $r_{k,n} < e$, then the AR probability for arm k is set equal to e, and if $r_{k,n} > 1 - e$, the AR probability is set equal to $1 - e$, with the K resulting AR probabilities normalized so that they sum to 1. A method using $r^{(c)}_{k,n}$ restricted to $[e, 1 - e]$ will be denoted by $AR(c, e)$.

All designs include a burn-in with the first 50 patients randomized equally among the arms, with exactly 10 patients assigned to each arm. We first consider $AR(1, 0)$, which randomizes patients to arm k with probability $r_{k,n}$, a K-arm generalization of Thompson,[1] but imposing a burn-in. The second method, $AR(0.5, 0)$, randomizes patients to arm k with probability $r^{(0.5)}_{k,n}$ given by equation (2). $AR(0.5, 0)$ minimizes the expected number of non-responders.[11] The third method, $AR(n/2N, 0)$, generalizes Thall and Wathen's[23] two-arm trial method by applying equation (2) using $c = n/2N$, for current sample size $n = 1, \ldots, N$. The fourth method, $AR(1, 0.10)$, uses $r_{k,n}$ with the restriction $0.10 \le r_{k,n} \le 0.90$. We thus evaluate $AR(1, 0)$, $AR(0.5, 0)$, $AR(n/2N, 0)$, $AR(1,\ 0.10)$, and ER.

Trial designs

Each simulation case is determined by whether a control arm is included, the maximum sample size $N = 250$ or $N = 500$, decision rules, and randomization method. All cases are five-arm trials. When a control arm, C, is included, we index it by $k = C$ and the four experimental arms by $k = 1, 2, 3, 4$. When C is not included, we index the five experimental treatments by $k = 1, 2, 3, 4, 5$. For all designs, we assume the response probabilities, $\{\pi_k\}$, are independent with *beta*(0.20, 0.80.) priors. Each design requires one parameter, a_U, to define the treatment arm selection rule, determined via preliminary simulations under the null scenario where all fixed response probabilities equal 0.20.

When C is included, its response probability, π_C, is used as the comparator in the decision rules. These rules may stop randomization to an experimental arm E_k due to futility, or select an E_k as promising, based on the posterior of $\pi_k - \pi_C$. If no control arm is included, one possible approach is to use a fixed standard probability, p_C, for comparison. Unless p_C is completely arbitrary, this requires the assumption that there exists a standard treatment with response probability known to equal p_C, that is, $\Pr(\pi_C = p_C) = 1$. It also requires that the numerical value p_C, obtained in practice from previous trials or clinical experience, will remain a valid comparator during the trial. This implies there are no between-trial or trial-versus-historical effects. Because these are very unrealistic assumptions, we do not consider designs assuming a fixed standard. Thus, the designs without a control arm that we consider make decisions based on comparisons among the E_k's.

Multi-arm trials with a control arm

For each experimental arm, E_k, $k = 1, 2, 3, 4$, after the initial burn-in, the following decision rules are applied continuously during the trial.

FUTILITY. For each $k = 1, 2, 3, 4$, arm E_k is terminated early due to futility if

$$Pr(\pi_k > \pi_C \ + \ 0.20 \mid data_n) < 0.01$$

If all four experimental arms are terminated, the trial is stopped.

ENRICHMENT. If an E_k is terminated early for futility, the remaining patients, up to N, are randomized among the remaining open arms.

SELECTION. If E_k is not terminated early, then at the end of the trial E_k is selected if

$$Pr(\pi_k > \pi_C + 0.20 \mid data_n) > a_U \tag{3}$$

The design thus allows more than one E_k to be selected. It is typical practice to require a new treatment to provide a minimal clinically significant improvement, here specified to be $\delta = 0.20$. The futility rule decreases the number of patients randomized to an E_k that is very unlikely to achieve the targeted improvement over C, and thus enriches the sample sizes of arms having larger success probabilities. For each design, the numerical value of a_U is determined to ensure overall false-positive probability 0.05 for the trial, with a false positive defined as selecting any E_k in the null case where all true $p_k = 0.20$. The numerical value of a_U depends on the randomization method, the value of N and the initial burn-in. Supplementary Table S1 gives the numerical value of the cut-off a_U used by each design's selection rule in each case. An alternative to deriving a_U in this way is to set it equal to a fixed value, such as $a_U = 0.95$. We chose to determine a_U for each design to obtain the same overall false-positive probability 0.05 in order ensure fair comparisons among the randomization methods in terms of per-arm selection probabilities, stopping probabilities, and sample size distributions.

Multi-arm trials without a control arm

For trials without a control arm, the decision rules are as follows:

FUTILITY. For each $k = 1, 2, 3, 4, 5$, accrual to E_k is terminated due to futility if

$$Pr(\pi_k > max\{\pi_r : \ r \neq k\} \mid data_n) < 0.01$$

ENRICHMENT. If an E_k is closed early for futility, the remaining patients, up to maximum sample size N, are randomized among the remaining open arms.

SELECTION. If E_k is not terminated early, at the end of the trial E_k is selected if

$$Pr(\pi_k > max\{\pi_r : \ r \neq k\} \mid data_n) > a_U \tag{4}$$

At the end of the trial, the designs with a control arm may select more than one E_k, whereas the designs without a control arm may select at most one E_k. While one might question why at most one E_k may be selected in trials without a control arm, it is extremely unlikely that two different π_k's both will satisfy the criterion (4) for any reasonably large a_U. Moreover, in the cases of no control arm, there is no required improvement, such as the value $\delta = 0.20$ that is used in the selection rule. If the selection criterion (4) was replaced by

$$Pr(\pi_k > max\{\pi_r : \ r \neq k\} + \delta \mid data_n) > a_U$$

for $\delta = 0.15$ or 0.20, our simulations show that, for N= 250 or 500 in a five-arm trial, this design would be extremely unlikely to correctly select any E_k in many scenarios where there actually are substantive differences among the p_ks.

Simulation study design

Under the Bayesian formulation, the probabilities, $\pi_C, \pi_1, \ldots, \pi_4$ in the case with a control arm, or

Table 1. Simulation results for designs with a control arm in the null scenario with all $p_k = 0.20$, for $N = 250$. $\bar{n}$ = mean per-arm sample size. Each $\eta_m = \Pr(N_C > N_k + m)$, the probability that the number of patients randomized to arm C is at least m larger than the number randomized to arm E_k. Values in the row E_1-E_4 are per-arm.

Method	Arm	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1,0)	C	–	–	33 (10, 63)	–
	E_1-E_4	0.01	0.67	42 (10, 135)	0.4, 0.27, 0.15
			Total	202 (70, 250)	
AR($\frac{1}{2}$,0)	C	–	–	40 (13, 69)	–
	E_1-E_4	0.01	0.74	40 (10, 109)	0.44, 0.32, 0.19
			Total	200 (70, 250)	
AR($\frac{n}{2N}$,0)	C	–	–	38 (13, 65)	–
	E_1-E_4	0.01	0.73	40 (10, 116)	0.43, 0.3, 0.18
			Total	199 (70, 250)	
AR(1, 0.1)	C	–	–	38 (20, 63)	–
	E_1-E_4	0.01	0.74	41 (10, 124)	0.44, 0.3, 0.16
			Total	200 (70, 250)	
ER	C	–	–	58 (17, 98)	–
	E_1-E_4	0.01	0.81	36 (10, 82)	0.6, 0.45, 0.33
			Total	200 (70, 250)	

$\pi_1, \ldots, \pi_5$ in the case without a control arm, are random. We distinguish between these random quantities and corresponding assumed fixed probabilities, denoted using p_k in place of π_k, that are used to define scenarios and simulate data. In all simulation scenarios, we assumed fixed null response rate 0.20. We consider three scenarios. In the null scenario, all $p_k = 0.20$. Given fixed targeted improvement $\delta = 0.20$, the least favorable configuration has one experimental $p_k = 0.20 + \delta$ and all other $p_k = 0.20$.Thus, $p_C = p_1 = \cdots = p_3 = 0.20$ and $p_4 = 0.20 + \delta = 0.40$ if there is a control arm, and $p_1 = \cdots = p_4 = 0.20$ with $p_5 = 0.20 + \delta = 0.40$ if there is no control arm. The least favorable configuration is determined, in the case with a control arm, by assuming that (1) no experimental p_k is between p_C and $p_C + \delta$ and (2) at least one experimental arm has $p_k \geq p_C + \delta$. The least favorable configuration is the vector of $p_1, \ldots, p_K$ values that minimizes the probability, under (1) and (2), that at least one E_k for which $p_k \geq p_C + \delta$ is selected. The name "least favorable configuration" is somewhat misleading, since the requirements (1) and (2) are quite strong, and they ensure that it is relatively easy to identify the one E_k providing a δ improvement over p_C. This motivates the third, more realistic "staircase" scenario, for which the p_k's are 0.20, 0.25, 0.30, 0.35, and 0.40.

Simulation results for trials with a control arm

In Tables 1 and 2, $\bar{n}$(95% CI) denotes the mean and (2.5th and 97.5th) percentiles of each per-arm sample size distribution.

In practice, an AR-based design with a large sample size imbalance favoring a superior arm is unlikely to be used if it has substantially lower probability of correctly selecting E_4 than ER. Table 2 shows that, under the least favorable configuration with $p_4 = 0.40$ and $N = 250$, AR(1,0), AR($\frac{1}{2}$,0) and AR($\frac{n}{2N}$,0) all have very low probability of correctly selecting E_4, between 0.44 and 0.48, compared to AR(1, 0.1) and ER, which have probability of correct selection values 0.67 and 0.66. One reason for this large loss in probability of correct selection for AR(1,0), AR($\frac{1}{2}$,0) and AR($\frac{n}{2N}$,0) is that each gets stuck randomizing patients to E_4 very early in the trial, resulting in a smaller $\bar{n}$ for C. The AR methods have $\bar{n}$ ranging from 23 to 35, with the widest 95% CI (11, 70) for AR($\frac{1}{2}$,0), compared to $\bar{n}$ = 72 and 95% CI (37, 110) for ER. AR(1, 0.1) provides a favorable sample size imbalance, with $\bar{n}$ = 127 for E_4 compared to $\bar{n}$ = 70 with ER. To ensure false-positive probability 0.05, the cut-off a_U in the selection rule (3) must be larger for AR(1,0), AR($\frac{1}{2}$,0) and AR($\frac{n}{2N}$,0) compared to AR(1, 0.1) or ER, resulting in much smaller Pr(Select E_4) for the first three AR methods .

Thall and Wathen[23] and Thall et al.[19] showed that, in the two-arm case, there is a significant risk that AR(1,0) and AR($\frac{1}{2}$,0) will get stuck randomizing more patients to the inferior treatment arm. To determine whether this holds in the multi-arm case, we calculate $\eta_m = Pr(N_C > N_k + m)$ for m = 10, 20, or 30 for each method. When some E_k is superior, η_m is the probability that a method will randomize at least m more patients to the inferior control arm than to E_k. An AR procedure having η_m much larger than that obtained with ER is undesirable. Under the least favorable configuration with $p_4 = 0.40$, using AR(1, 0.1), on average, 127 patients are treated with E_4 compared to 70 using ER, so an additional 57 patients are treated with E_4 as a result of using AR(1, 0.1), which has $\eta_{10} = 0.05$ compared to 0.23 for ER. The reason that ER has larger

Table 2. Simulation results for designs with a control arm in least favorable configuration scenario, $p_C = p_1 = p_2 = p_3 = 0.20, p_4 = 0.40$, for $N = 250$. $\bar{n}$ = mean per-arm sample size. Each η_m = Pr($N_C > N_k$ + m), the probability that the number of patients randomized to arm *C* is at least *m* larger than the number randomized to arm E_k. Values in the row $E_1 - E_3$ are per-arm.

Method	Arm	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1, 0)	C	–	–	23 (10, 58)	–
	E_1-E_3	0.02	0.40	23 (10, 69)	0.28, 0.15, 0.08
	E_4	0.44	0.07	152 (11, 201)	0.04, 0.03,0.02
			Total	244 (140, 250)	
AR($\frac{1}{2}$, 0)	C	–	–	34 (11, 70)	–
	E_1-E_3	0.02	0.56	29 (10, 70)	0.42, 0.29, 0.18
	E_4	0.46	0.07	123 (10, 177)	0.05, 0.04,0.02
			Total	243 (130, 250)	
AR($\frac{n}{2N}$, 0)	C	–	–	31 (12, 63)	–
	E_1-E_3	0.02	0.52	27 (10, 68)	0.39, 0.25, 0.13
	E_4	0.48	0.07	132 (11, 179)	0.04, 0.03,0.02
			Total	243 (120, 250)	
AR(1, 0.1)	C	–	–	35 (23, 60)	–
	E_1-E_3	0.03	0.58	27 (10, 66)	0.44, 0.26, 0.11
	E_4	0.67	0.07	127 (10, 177)	0.05, 0.04,0.02
			Total	243 (130, 250)	
ER	C	–	–	72 (37, 110)	–
	E_1-E_3	0.02	0.78	34 (10, 71)	0.73, 0.64, 0.56
	E_4	0.66	0.08	70 (10, 109)	0.23, 0.07, 0.03
			Total	243 (130, 250)	

Table 3. Simulation results for designs with a control arm in staircase scenario, $(p_C, p_1, p_2, p_3, p_4) = (0.20, 0.25, 0.30, 0.35, 0.40)$ for $N = 250$. $\bar{n}$ = mean per-arm sample size. Each η_m = Pr($N_C > N_k$ + m), the probability that the number of patients randomized to arm *C* is at least *m* larger than the number randomized to arm E_k.

Method	Arm	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1, 0)	C	–	–	20 (10, 51)	–
	E_1	0.05	0.24	27 (10, 74)	0.17, 0.08, 0.04
	E_2	0.11	0.16	40 (10, 113)	0.12, 0.06, 0.03
	E_3	0.22	0.10	62 (10, 158)	0.08, 0.04, 0.02
	E_4	0.40	0.06	101 (10, 185)	0.05, 0.03, 0.02
AR($\frac{1}{2}$, 0)	C	–	–	28 (11, 60)	–
	E_1	0.06	0.33	32 (10, 73)	0.25, 0.16, 0.08
	E_2	0.14	0.22	44 (10, 93)	0.17, 0.11, 0.06
	E_3	0.26	0.13	60 (10, 121)	0.10, 0.07, 0.04
	E_4	0.45	0.06	84 (11, 149)	0.05, 0.04, 0.03
AR($\frac{n}{2N}$, 0).	C	–	–	26 (11, 55)	–
	E_1	0.06	0.31	31 (10, 72)	0.24, 0.13, 0.06
	E_2	0.13	0.21	42 (10, 97)	0.16, 0.09, 0.04
	E_3	0.25	0.12	61 (10, 130)	0.09, 0.06, 0.03
	E_4	0.45	0.07	90 (10, 156)	0.05, 0.04, 0.02
AR(1, 0.1)	C	–	–	33 (23, 54)	–
	E_1	0.10	0.36	31 (10, 70)	0.29, 0.16, 0.05
	E_2	0.21	0.22	41 (10, 99)	0.19, 0.10, 0.04
	E_3	0.38	0.13	58 (10, 131)	0.11, 0.07, 0.03
	E_4	0.61	0.06	86 (11, 151)	0.06, 0.04, 0.02
ER	C	–	–	58 (40, 93)	–
	E_1	0.08	0.49	39 (10, 65)	0.50, 0.38, 0.31
	E_2	0.22	0.29	46 (10, 72)	0.36, 0.24, 0.19
	E_3	0.44	0.15	51 (10, 79)	0.26, 0.14, 0.10
	E_4	0.66	0.07	55 (11, 86)	0.21, 0.07, 0.05

η_{10} than AR(1, 0.1) is that, if E_4 is dropped and the trial continues, ER assigns more patients to Cs than AR(1, 0.1). Thus, results of the two-arm case cannot be extended to the multi-arm setting. In this case, AR(1, 0.1) achieves a very favorable patient imbalance in favor of E_4 compared to ER while maintaining

Table 4. Simulation results for designs with a control arm comparing N = 250 and N = 500 in least favorable configuration scenario, $p_1 = p_2 = p_3 = p_C = 0.20$, $p_4 = 0.40$. Values of $\bar{n}$ = mean per-arm sample size and 95% CI are for E_4.

		N = 250	N = 500
AR(1,0)	Pr(Select E_4)	0.44	0.53
	$\bar{n}$(95% CI)	152(11, 201)	369 (11, 444)
AR($\frac{1}{2}$,0)	Pr(Select E_4)	0.46	0.67
	$\bar{n}$(95% CI)	123(11, 177)	319 (11, 413)
AR($\frac{n}{2N}$,0)	Pr(Select E_4)	0.48	0.77
	$\bar{n}$(95% CI)	132(11, 179)	321 (11, 406)
AR(1, 0.1)	Pr(Select E_4)	0.67	0.87
	$\bar{n}$(95% CI)	127(11, 177)	313 (11, 403)
ER	Pr(Select E_4)	0.66	0.85
	$\bar{n}$(95% CI)	70(10, 109)	175 (13, 238)

Table 5. Simulation results for designs without a control arm in the null scenario $p_1 = \cdots = p_5 = 0.20$, for N = 250. Each $\eta_m = \Pr(N_{E_1} > N_{E_k} + m)$, the probability that the number of patients randomized to arm C is at least m larger than the number randomized to arm E_k. All values are per-arm.

Method	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1, 0)	0.01	0.19	50 (10, 137)	0.42, 0.35, 0.28
			Total	250 (250, 250)
AR($\frac{1}{2}$, 0)	0.01	0.26	50 (10, 110)	0.41, 0.33, 0.26
			Total	249 (250, 250)
AR($\frac{n}{2N}$, 0)	0.01	0.25	50 (10, 118)	0.41, 0.34, 0.27
			Total	249 (250, 250)
AR(1, 0.1)	0.01	0.24	50 (10, 128)	0.41, 0.34, 0.26
			Total	250 (250, 250)
ER	0.01	0.32	50 (10, 97)	0.32, 0.23, 0.2
			Total	248 (250, 250)

Pr(Select E_4) and reducing the likelihood of randomizing patients to inferior treatments.

In the staircase scenario, it is much more difficult to discriminate among the E_k's. Table 3 summarizes the simulations in this case for trials including C with N = 250. Compared to ER, AR(1, 0.1) has sightly smaller probabilities of selecting E_3 or E_4, which have $p_3 = 0.35$ and $p_4 = 0.40$. This is due to the fact that E_1, E_2, and E_3 remain in the trial longer because these treatments provide some improvement over C, limiting the number of patients treated with E_4, and reducing the probability that any AR method will select E_4. Still, AR(1, 0.1) assigns more patients to the better treatment arms, on average. Additionally, η_{10}, η_{20}, and η_{30} each are smaller for AR(1, 0.1) compared to ER. Compared to AR(1, 0.1) or ER, the probabilities of selecting the best arms E_4 or E_5 are much smaller for AR(1,0), AR($\frac{1}{2}$,0) and AR($\frac{n}{2N}$,0).

Tables 2 and 3 show that, for designs with a control arm and N = 250 patients, in the least favorable configuration or staircase scenarios, the highest probabilities of selecting the best arm are 0.66 or 0.67, obtained by AR(1, 0.1) or ER. A trial probably would not be conducted if there were only a 66% chance of selecting an E_k achieving the targeted improvement. In practice, one would either increase N, increase the false-positive rate, or both. Supplementary Tables S1–S3 summarize the simulations in the three scenarios for N = 500 with a control arm. Table S3 shows that N = 500 gives much larger probabilities of selecting superior E_k's in the staircase scenario, with E_4 selected with probabilities 0.84 by AR(1, 0.1) and 0.86 by ER, while the other three AR methods have substantially inferior performance. Tables S2 and S3 show that, under the least favorable configuration, for N = 500, the probability of stopping superior arm E_4 is 0.08–0.09 for AR(1,0.01). If desired, these Pr(Stop) values may be made smaller by reducing the futility stopping rule cut-off to a value smaller than .01, such as .005, but the price would be smaller per-arm sample sizes for E_4 and consequently lower Pr(Select) values.

Table 4 compares Pr(Select E_4) for N = 250 and N = 500 under the least favorable configuration when $p_4 = 0.40$. When N = 500, AR(1, 0.1) and ER have Pr(Select E_4) values 0.87 and 0.85. Compared to ER, although AR(1, 0.1) has a much more disperse subsample size distribution for E_4, on average AR(1, 0.1) randomizes many more patients to E_4. The Pr(Select E_4) values 0.77, 0.67, and 0.53 for AR($\frac{n}{2N}$,0), AR($\frac{1}{2}$,0), and AR(1,0) are much smaller. AR($\frac{1}{2}$,0)would require

Table 6. Simulation results for designs with no control arm in the least favorable configuration scenario $p_1 = p_2 = p_3 = p_4 = 0.20$ and $p_5 = 0.40$, for $N = 250$. $\bar{n}$ = mean per-arm sample size. Each $\eta_m = \Pr(N_{E_1} > N_{E_k} + m)$, the probability that the number of patients randomized to arm E_1 is at least m larger than the number randomized to arm E_k. Values in the row $E_1 - E_4$ are per-arm.

Method	Arm	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1, 0)	E_1-E_4	0	0.58	24 (10, 72)	0.27, 0.15, 0.08
	E_5	0.78	0.02	141 (11, 199)	0.03, 0.02,0.02
			Total	236 (90, 250)	
AR($\frac{1}{2}$, 0)	E_1-E_4	0	0.74	29 (10, 74)	0.32, 0.21, 0.13
	E_5	0.81	0.02	102 (14, 164)	0.03, 0.02,0.02
			Total	217 (80, 250)	
AR($\frac{n}{2N}$, 0)	E_1-E_4	0	0.71	27 (10, 69)	0.31, 0.19, 0.1
	E_5	0.82	0.02	107 (13, 169)	0.02, 0.02,0.01
			Total	216 (70, 250)	
AR(1, 0.1)	E_1-E_4	0	0.68	26 (10, 71)	0.30, 0.17, 0.08
	E_5	0.80	0.02	123 (17, 184)	0.03, 0.02,0.02
			Total	228 (80, 250)	
ER	E_1-E_4	0	0.79	36 (10, 89)	0.34, 0.26, 0.19
	E_5	0.75	0.02	61 (13, 103)	0.07, 0.02,0.01
			Total	205 (70, 250)	

$N = 500$ patients to obtain the same Pr(Select E_4) as AR(1, 0.1) and ER with only $N = 250$. A trial utilizing AR(1, 0) would require more than double the sample size to obtain the same Pr(Select E_4) as AR(1, 0.1) or ER. A general conclusion is that AR(1, 0.1) provides more patients with superior treatment while maintaining acceptable Pr(Select E_4), for $N = 500$ in a five-arm trial with a control.

Simulation results for trials without a control arm

Each design without a control arm was calibrated to have a 1% chance of selecting each treatment in the null scenario (Table 5). In the least favorable configuration scenario with $p_5 = 0.40$ and $N = 250$, Table 6 shows that all methods provide Pr(Select E_4) for E_4 ranging from 0.75 to 0.82, and all of the η_m values are relatively small for E_4. If the only cases considered were the null and the least favorable configuration, then it might seem that running a multi-arm trial including a control arm is foolish. However, the opposite is true. Table 7 shows that, in the staircase scenario, for $N = 250$, the probabilities of selecting the best treatments are extremely low, ranging from 0.19 to 0.26, compared to approximately 0.65 when a control arm is included. The main reason for this large drop is that, without a control arm, comparison among the E_k's is extremely difficult if the differences between the p_ks are small. Supplementary Table S6 shows that, in the staircase scenario, even if the overall maximum sample size is increased to $N = 500$, the selection probabilities for E_5 range from 0.33 to 0.39 for any randomization method, with selection probabilities at most 0.04 for any of $E_1, \ldots, E_4$.

Discussion

A general conclusion is that, for multi-arm trials, AR(1, 0), AR($\frac{1}{2}$, 0), and AR($\frac{n}{2N}$, 0) should not be used. If one wishes to use some AR method in a multi-arm trial, if an initial burn-in is imposed, the superior performance of AR(1, 0.1) indicates that it is important to restrict the domain of possible AR probabilities by bounding them away from 0 and 1. Given the apparent popularity of AR(1,0) and AR(0.50, 0), this is a very important result. While we have not examined other hybrid methods, such as AR(.50, .10) or AR(n/2N, .10), the simulations suggest that these may perform well compared to AR(1, .10) or ER. The numerical limit e cannot be arbitrary, since, for example, AR(0.50, 0.20) would be close to ER in a five-arm trial. ER does the best job of selecting treatments having p_k's that are superior but close to each other.

In practice, it is not unlikely that two or more p_k's may be close to each other, so the staircase scenario may be closer to reality than the least favorable configuration. When the p_k's are close to each other, it is very difficult to select any E_k if no C is included as a comparator. The simulations in the staircase scenario indicate that conducting a multi-arm trial without a control arm may be a waste of resources, for any randomization method, and it is best to include a control arm in a multi-arm selection trial.

Many elaborations and alternative cases are possible, including time-to-event or multivariate outcomes, accounting for covariates, and evaluating AR methods for multi-arm trials in the presence of drift. This latter issue is closely related to so-called platform designs, which allow experimental arms to enter a trial after it has started.[25] These are important areas for future simulation study.

Table 7. Simulation results for designs with no control arm in the staircase scenario, $(p_1, p_2, p_3, p_4, p_5) = (0.20, 0.25, 0.30, 0.35, 0.40)$, for $N = 250$. $\bar{n}$ = mean per-arm sample size. $\eta_m = \Pr(N_{E_1} > N_{E_j} + m)$, the probability that the number of patients randomized to E_1 is at least m larger than the number randomized to E_j.

Method	Arm	Pr(Select)	Pr(Stop)	$\bar{n}$(95% CI)	$\eta_{10}, \eta_{20}, \eta_{30}$
AR(1, 0)	E_1	0	0.61	19 (10, 52)	–
	E_2	0	0.44	27 (10, 81)	0.17, 0.08,0.03
	E_3	0	0.30	39 (10, 115)	0.12, 0.06,0.03
	E_4	0.04	0.16	63 (10, 160)	0.08, 0.04,0.02
	E_5	0.21	0.07	101 (10, 184)	0.04, 0.02,0.01
AR($\frac{1}{2}$, 0)	E_1	0	0.79	22 (10, 58)	–
	E_2	0	0.60	31 (10, 75)	0.2, 0.11,0.06
	E_3	0.01	0.39	44 (10, 100)	0.13, 0.08,0.04
	E_4	0.03	0.21	62 (10, 130)	0.08, 0.05,0.03
	E_5	0.21	0.07	87 (10, 152)	0.04, 0.03,0.02
AR($\frac{n}{2N}$, 0)	E_1	0	0.76	21 (10, 54)	–
	E_2	0	0.58	29 (10, 75)	0.19, 0.1,0.05
	E_3	0	0.38	42 (10, 103)	0.13, 0.07,0.03
	E_4	0.04	0.20	62 (10, 136)	0.07, 0.04,0.02
	E_5	0.26	0.07	91 (10, 158)	0.04, 0.02,0.01
AR(1, 0.1)	E_1.	0	0.75	21 (10, 52)	–
	E_2	0	0.55	29 (10, 77)	0.19, 0.1,0.04
	E_3	0	0.34	41 (10, 113)	0.12, 0.06,0.02
	E_4	0.04	0.18	62 (10, 149)	0.08, 0.04,0.02
	E_5	0.23	0.07	94 (10, 172)	0.04, 0.03,0.01
ER	E_1	0	0.89	25 (10, 69)	–
	E_2	0	0.71	36 (10, 82)	0.21, 0.14,0.09
	E_3	0	0.45	50 (10, 101)	0.14, 0.08,0.06
	E_4	0.03	0.22	62 (10, 108)	0.08, 0.05,0.03
	E_5	0.19	0.08	68 (10, 109)	0.05, 0.03,0.02

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Peter Thall's research was supported by NIH NCI Grant 2RO1 CA083932.

References

1. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. *Biometrika* 1933; 25: 285–294.
2. Wei LJ and Durham S. The randomized play-the-winner rule in medical trials. *J Am Stat Assoc* 1978; 73: 840–843.
3. Zelen M. A new design for randomized clinical trials. *N Engl J Med* 1979; 300: 1242–1246.
4. Cheung YK, Inoue LYT, Wathen JK, et al. Continuous Bayesian adaptive randomization based on event times with covariates. *Stat Med* 2006; 25: 55–70.
5. Thall PF and Wathen JK. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Stat Med* 2005; 24: 1947–1964.
6. Hu F and Rosenberger WF. The theory of response-adaptive randomization in clinical trials (Wiley series in probability and statistics). Hoboken, NJ: John Wiley & Sons, 2006.
7. Sverdlov O. Modern adaptive randomized clinical trials: statistical and practical aspects. Boca Raton, FL: CRC Press, 2015.
8. Giles FJ, Kantarjian HM, Cortes JE, et al. Adaptive randomized study of idarubicin and cytarabine versus troxacitabine and cytarabine versus troxacitabine and idarubicin in untreated patients 50 years or older with adverse karyotype acute myeloid leukemia. *J Clin Oncol* 2003; 21: 1722–1727.
9. Maki RG, Wathen JK, Hensley ML, et al. An adaptively randomized phase III study of gemcitabine and docetaxel versus gemcitabine alone in patients with metastatic soft-tissue sarcomas. *J Clin Oncol* 2007; 25: 2755–1763.
10. Kim ES, Herbsy RS, Wistuba II, et al. The battle trial: personalizing therapy for lung cancer.*Cancer Discov* 2011; 1: 44–53.
11. Chappell R and Karrison T. Letter to the editor. *Stat Med* 2007; 26: 3046–3056.
12. Korn EL and Freidlin B. Outcome-adaptive randomization: is it useful? *J Clin Oncol* 2011; 29: 771–776.
13. Yuan Y and Yin G. On the usefulness of outcome-adaptive randomization. *J Clin Oncol* 2011; 29: 390–392.
14. Lee JJ, Chen N and Yin G. Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clin Cancer Res* 2012; 18: 4498–4507.
15. Rosenberger WF, Sverdlov O and Hu F. Adaptive randomization for clinical trials. *JBS* 2012; 22: 719–36.
16. Buyse M. Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 119–121.

17. Lee JJ. Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12: 110–112.
18. Hey SP and Kimmellman J. Are outcome-adaptive allocation trials ethical? *Clin Trials* 2015; 12: 102–106.
19. Thall PF, Fox PS and Wathen JK. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann of Oncol* 2015; 26: 1621–1628.
20. Berry DA. Adaptive clinical trials: the promise and the caution. *J Clin Oncol* 2011; 29: 606–609.
21. Rosenberger WR and Lachin JM. The use of response-adaptive designs in clinical trials. *Control Clin Trials* 1993; 14: 471–484.
22. Karrison TG, Huo D and Chappell R. A group sequential, response-adaptive design for randomized clinical trials. *Control Clin Trials* 2003; 24: 506–522.
23. Thall PF and Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *EJC* 2007; 43: 860–867.
24. Sutton RS and Barto AG. Reinforcement learning: an introduction. Cambridge, MA: MIT Press, 1998.
25. Berry SM, Connor JT and Lewis RJ. The platform trial: an efficient strategy for evaluating multiple treatments. *JAMA* 2015; 313: 1619–1620.