

Comparison of methods for control allocation in multiple arm studies using response adaptive randomization

Clinical Trials
2020, Vol. 17(1) 52–60
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1740774519877836
journals.sagepub.com/home/ctj



Kert Viele¹, Kristine Broglio¹, Anna McGlothlin¹ and Benjamin R Saville^{1,2} 

Abstract

Background/Aims: Response adaptive randomization has many polarizing properties in two-arm settings comparing control to a single treatment. The generalization of these features to the multiple arm setting has been less explored, and existing comparisons in the literature reach disparate conclusions. We investigate several generalizations of two-arm response adaptive randomization methods relating to control allocation in multiple arm trials, exploring how critiques of response adaptive randomization generalize to the multiple arm setting.

Methods: We perform a simulation study to investigate multiple control allocation schemes within response adaptive randomization, comparing the designs on metrics such as power, arm selection, mean square error, and the treatment of patients within the trial.

Results: The results indicate that the generalization of two-arm response adaptive randomization concerns is variable and depends on the form of control allocation employed. The concerns are amplified when control allocation may be reduced over the course of the trial but are mitigated in the methods considered when control allocation is maintained or increased during the trial. In our chosen example, we find minimal advantage to increasing, as opposed to maintaining, control allocation; however, this result reflects an extremely limited exploration of methods for increasing control allocation.

Conclusion: Selection of control allocation in multiple arm response adaptive randomization has a large effect on the performance of the design. Some disparate comparisons of response adaptive randomization to alternative paradigms may be partially explained by these results. In future comparisons, control allocation for multiple arm response adaptive randomization should be chosen to keep in mind the appropriate match between control allocation in response adaptive randomization and the metric or metrics of interest.

Keywords

Multiple arm clinical trial, response adaptive randomization, control allocation

Background/Aims

Multiple arm clinical studies allow the investigation of multiple therapies within a single trial, sharing logistical and patient resources in evaluating therapies.¹ In addition to operational and statistical gains resulting from a shared control arm, further efficiencies can be achieved by reducing or eliminating allocation to poorly performing experimental arms during the trial. This has been implemented through strategies such as arm dropping,² multi-arm multi-stage designs,³ or response adaptive randomization.^{4,5}

The latter of these methodologies, response adaptive randomization (RAR), is highly controversial on both statistical and ethical grounds. For two-arm trials, with a single control and single treatment under comparison,

RAR performs poorly on a variety of metrics. RAR has been shown to have reduced power^{6,7} relative to fixed allocation, reflecting the theoretical result that 1:1 randomization is optimal for power and RAR deviates from that ratio. Thall et al.⁶ show that for a specific implementation of RAR, where adaptive allocation is begun very early in a trial, severely biased estimates

¹Berry Consultants LLC, Austin, TX, USA

²Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, TN, USA

Corresponding author:

Kert Viele, Berry Consultants LLC, 3325 Bee Caves Road, Suite 201, Austin, TX 78746, USA.

Email: kert@berryconsultants.net

and a meaningful probability of excessive allocation to an inferior arm may occur. While there are specific instances where RAR has optimality properties,⁸ typically these require a short patient horizon, meaning that a sizable proportion of the total population under consideration is treated within the clinical trial itself.⁹ Ethically, RAR has the appealing property of treating more patients in the trial on better performing arms and therefore increasing the expected number of desirable outcomes for patients within a trial,^{6,8} but RAR has also been criticized from an ethical standpoint.^{10,11} Hey and Kimmelman¹⁰ argue against the ethics of RAR on a number of fronts, including (1) the reduction of power in the two-arm setting represents an ethical harm by increasing the cost of research, (2) the complexity of the trial design can complicate the validity of informed consent if patients are unable to understand the trial conduct, (3) equipoise may be violated by allocating any number of patients to arms with meaningful but not conclusive evidence of being inferior, and (4) the disadvantage to patients enrolling early in a study compared to patients enrolling late when there is a higher probability of receiving a better performing arm. While some of these concerns apply generally to all adaptive designs, others such as the concern about equipoise are specific to RAR and have been raised in practice by Buyse et al.¹² in response to the use of RAR in the ISPY2 trial.¹³

RAR has been employed in settings with control and multiple treatment arms,^{4,14–17} including large-scale platform trials.¹⁸ Given this usage, it is imperative to understand the degree to which criticisms of RAR in the two-arm setting generalize to the multiple arm setting. Comparisons and explorations in this area have been highly varied, with results showing adequate performance of RAR^{2,19,20} in comparison to multi-arm multi-stage designs and arm dropping (Friedlin and Korn² and Trippa et al.⁴ have an additional aspect of their debate involving the operational complexity of arm dropping versus RAR) while others such as Wathen and Thall²¹ show RAR underperforming in the multiple arm setting in a similar manner to the two-arm setting.

RAR represents a large class of designs, and varying factors such as the aggressiveness and timing of adaptation have been shown to have a significant impact on RAR performance.²² Comparisons between classes of designs should be premised on comparing exemplars of each class optimized to perform on the metrics of interest. Control allocation represents a particularly interesting feature in the generalization of two-arm RAR results to multiple arms, as many deficiencies of two-arm RAR, such as the loss of power, may be traced to the reduction of control allocation over the course of the trial. In contrast, the multiple arm setting allows for the possibility of maintaining or even increasing control allocation over the course of the trial while

employing RAR on the active arms, increasing the sample size on both the control and best arm relative to fixed allocation.

Our primary aim is to systematically compare different control allocations in a multiple arm RAR setting and explore the degree to which critiques of RAR in the two-arm setting generalize according to control allocation in the multiple arm setting. We also aim to explore the degree to which differing allocations to control may explain the differing results reported in comparisons between RAR and alternative paradigms (fixed designs, arm dropping, multiple arm multiple stage designs, etc.) as well as to suggest more optimal representatives of RAR for future comparisons.

We focus on statistical aspects of multiple arm RAR in this article, in particular power, which is reduced for two-arm RAR, and estimation of treatment effects as a risk of significant biases has been observed in some variants of two-arm RAR. We additionally consider arm selection. Arm selection is moot for two-arm trials but is a significant issue for multiple arm trials with the goal of identifying the best experimental arm. We do not focus on the ethical objections to RAR described above except to the degree that increased efficiency (higher power, a better estimation of treatment effects) in clinical trials itself represents an ethical gain by reducing the societal and patient burden of clinical trials. London²³ presents a potential counterargument to the equipoise concerns^{10,12} by focusing on a societal representation of equipoise, arguing potentially unequal randomization is ethical for any arm for which a reputable clinical sub-community would recommend that arm in practice.

We conduct this exploration with a simulation exploring multiple variants of RAR within a multiple arm trial with a dichotomous primary endpoint. In the “Methods” section, we describe the hypothetical trial, metrics for comparison, scenarios used in the simulation, and eight different control allocation schemes. We then describe results and provide conclusions and discussion in the following sections.

Methods

Clinical setting and modeling

We consider an experiment with control ($t = 0$) and three active treatment arms ($t = 1,2,3$) in a dichotomous setting. Let Y_i be the indicator of response for the i th patient with distribution

$$Y_i|t \sim \text{Bernoulli}(p_t)$$

where p_t is the underlying response rate for arm t . Define the log-odds of p_t

$$\log\left(\frac{p_t}{1-p_t}\right) = \theta_t$$

In what follows inference is based on independent normal priors on each arm

$$\theta_t \sim N(0, 1.82^2) \text{ for } t = 0, 1, 2, 3$$

These priors are roughly uniform when transformed back to p_t and thus minimally informative. The effective sample size of this prior distribution is approximately two patients' worth of information. The trial is considered successful at the final analysis if there is a high posterior probability that at least one arm has a higher rate than control

$$\max_t \Pr(p_t > p_0) > \delta$$

where δ is a threshold chosen to maintain familywise type I error for the study at one-sided 2.5%.

Metrics

We employ five metrics to characterize the performance of RAR. Three of these are standard statistical measures, including power, mean square error, and accuracy of arm selection. We also consider the number of responders within the trial. These metrics cover problematic areas for two-arm RAR designs, power, and estimation,^{6,7} as well as an area where RAR in the two-arm setting typically generates an improvement over fixed allocation, the number of responders in the trial.^{6,8,9} Finally, we introduce a combined power/arm selection metric described below that captures the expected response rate for patients treated after the trial is complete. This measure, quantifying the benefit to the broader patient population, is a complement to the expected number of responders within the trial. We present the mean and, where applicable, the cumulative distribution function of each of these metrics.

In detail, our five metrics are as follows:

1. Power: the probability of meeting the trial success condition.
2. Arm selection: the probability of choosing each arm as the best arm. If no experimental arm meets the criteria for declaring superiority to control, the control arm is viewed as the selected arm (as the trial may be unlikely to change clinical practice).
3. Mean square error: if an experimental arm is selected, we also produce estimates (posterior means) of the responder rate for that arm as well as the treatment effect for that arm compared to control. These estimates are relevant to payers, regulators, and anyone conducting future trials with the selected arm. Formally, we compute the mean square error of the selected arm rate $E[(\hat{p}_{t'} - p_{t'})^2]$ and the mean square error of the selected arm's treatment effect $E[\{(\hat{p}_{t'} - \hat{p}_0) - (p_{t'} - p_0)\}^2]$ where t' is the selected arm.

4. Expected number of responders in the trial—a random number of responders are observed in each trial. We compute the expected value of this quantity, the average number of responders averaged over trials. This is a function of the number of patients on the arms with the highest true rates. Formally, we compute $E[\sum_i Y_i]$.
5. Ideal design percentage—a combination of arm selection and power. Let π_t be the probability arm t is selected ($t = 0, 1, 2$, or 3 as defined in the arm selection metric). The expected responder rate for the external patient population (outside the trial) is

$$\text{Expected Rate} = \sum_{t=0}^3 \{p_t \pi_t\}$$

In the worst design possible, we always pick the arm with the lowest response rate. In the ideal design (impossible in practice), we would always pick the arm with the highest response rate. The expected rate is somewhere between the lowest true responder rate and the highest true responder rate. The ideal design percentage measure is

$$\begin{aligned} \text{Ideal Design Percentage} &= 100 * \\ &(\text{Expected Rate} - \text{Min True Rate}) / \\ &(\text{Max True Rate} - \text{Min True Rate}) \end{aligned}$$

This metric quantifies where the design performance falls in the range from worst possible to best possible, combining arm selection and power and naturally measures the degree of any incorrect arm selections. For example, choosing the second-best arm when that arm is 1% worse than the best is different than choosing the second-best arm when that arm is 10% worse than the best. The ideal design percentage incorporates these differences in the expected value.

The relative value of these metrics within each specific trial must be considered in determining a design choice. Designs best for an external patient population may provide less benefit to patients within the trial, and vice versa. We consider power, arm selection, and ideal design percentage as measures meaningful to the broader patient population and the number of responders as the measure meaningful to the patients in the trial.

Scenarios

Table 1 shows the four clinical scenarios investigated. The null scenario is used to control type I error, with the remaining scenarios representing various degrees of effectiveness and varying numbers of effective arms. Note the order of the experimental arms is unimportant. The experimental arm labels can be interchanged with no change in design performance.

Table 1. Scenarios to be investigated.

Scenario	Control rate	Arm 1 rate	Arm 2 rate	Arm 3 rate
Null	0.35	0.35	0.35	0.35
Nugget	0.35	0.35	0.35	0.65
Two	0.35	0.35	0.65	0.65
Mixed	0.35	0.45	0.55	0.65

Designs

We consider eight designs, each with a total sample size of $N = 228$. This sample size was chosen to achieve approximately 80% power in the nugget scenario using equal randomization and a Bonferroni correction for multiplicities. Our first three designs employ fixed allocation in three ratios:

1. F25: an equal randomization design (1:1:1:1 allocation in blocks of 4), thus allocating 25% of the patients to control.
2. F40: a trial randomizing 2:1:1:1 in blocks of size 5, allocating 40% of the patients to control.
3. F50: a trial randomizing 3:1:1:1 in blocks of size 6, allocating 50% of patients to control.

Our remaining designs all employ RAR with some common elements. All have interims occurring at $N = 40, 80, 120, 160,$ and 200 patients. We employ a burn-in of 10 patients per arm for the first 40 patients. After each interim analysis, the randomization probabilities for the experimental arms are set to be proportional to the probability that each experimental arm has the maximal responder rate, $\Pr_t(\text{Max}) = \Pr(p_t = \max_i p_i)$.²¹ If an allocation probability for an active arm is below 10%, we truncate that allocation to zero and renormalize the remaining probabilities, resulting in temporary arm dropping.

We vary control allocation within RAR, investigating three designs with constant control allocation,¹⁴ a design which potentially reduces allocation to control throughout the trial,²¹ and one design which potentially increases control allocation throughout the trial, similar to Trippa et al.⁴

4. R25: 25% of patients on control using blocks of size 4 with 1 control per block, and the remainder, after burn-in, allocated in proportion to $\Pr_t(\text{Max})$ among the experimental arms.
5. R40: 40% of patients on control using blocks of size 5 with two controls per block.
6. R50: 50% of patients on control using blocks of size 6 with three controls per block.
7. RAdjCtrl: in this design, control allocation can increase or decrease (potentially to zero with thresholding) throughout the trial. All arms,

including control, are allocated in proportion to their probability of being the maximal arm. This is analogous to an approach evaluated by Wathen and Thall.²¹

8. RMatch—define $V_t = \Pr_t(\text{Max})$. Allocation to control is determined by computing

$$V_0 = \min \left\{ \sum_{t=1}^3 V_t \frac{(n_t + 1)}{(n_0 + 1)}, \max(V_1, V_2, V_3) \right\}$$

and then renormalizing V_0, V_1, V_2, V_3 to sum to 1. Primarily this rule simply allocates the control in proportion to the best arm, setting $V_0 = \max(V_1, V_2, V_3)$ and renormalizing. This quantity is directly focused on $\Pr_t(\text{Max})$ as opposed to $\Pr(p_t > p_0)$ as in the article by Trippa et al.,⁴ which is similar in spirit. In addition, the first term in the minimization allows for the possibility that an arm with lower enrollment may become the arm with the highest V_t , in which case that arm is given higher allocation than control. In the extreme, suppose we had $N = 100$ on control, $N = 20$ on arms 1 and 2, and $N = 100$ on arm 3, with $V_1 = 0.1, V_2 = 0.8, V_3 = 0.1$. Arm 2 now has high probability of being the best arm, but the control allocation already significantly exceeds arm 2. Here $V_0 = 0.287$ and $V_2 = 0.800$, allowing arm 2 to catch up to control from its current limited enrollment.

Results

Type I error control

To ensure comparability among designs, type I error was controlled by simulating 100,000 trials in the null scenario for each design and choosing the δ required to achieve a one-sided 2.5% family-wise type I error. This value was then used to simulate 100,000 trials in each of the remaining scenarios. FACTS version 6.1 was used for all simulations. For the fixed allocation designs F25, F40, and F50 the required thresholds were between 0.9912 and 0.9924. For the RAR designs the required thresholds were between 0.9872 (for R25) and 0.9892 (for RMatch). RAR generally requires less conservative thresholds for trial success compared to fixed allocation. Type I errors can occur when there is a random low on the control arm and/or a random high on the experimental arms. RAR increases allocation to better performing experimental arms, providing more opportunity for random highs on the experimental arms to regress to their true mean and thereby reduce the risk of a type I error.

Operating characteristics for each metric

Figures 1 and 2 show the results for each metric. In each panel, the x -axis represents the scenario (nugget, two, or mixed). The y -axis gives the particular metric

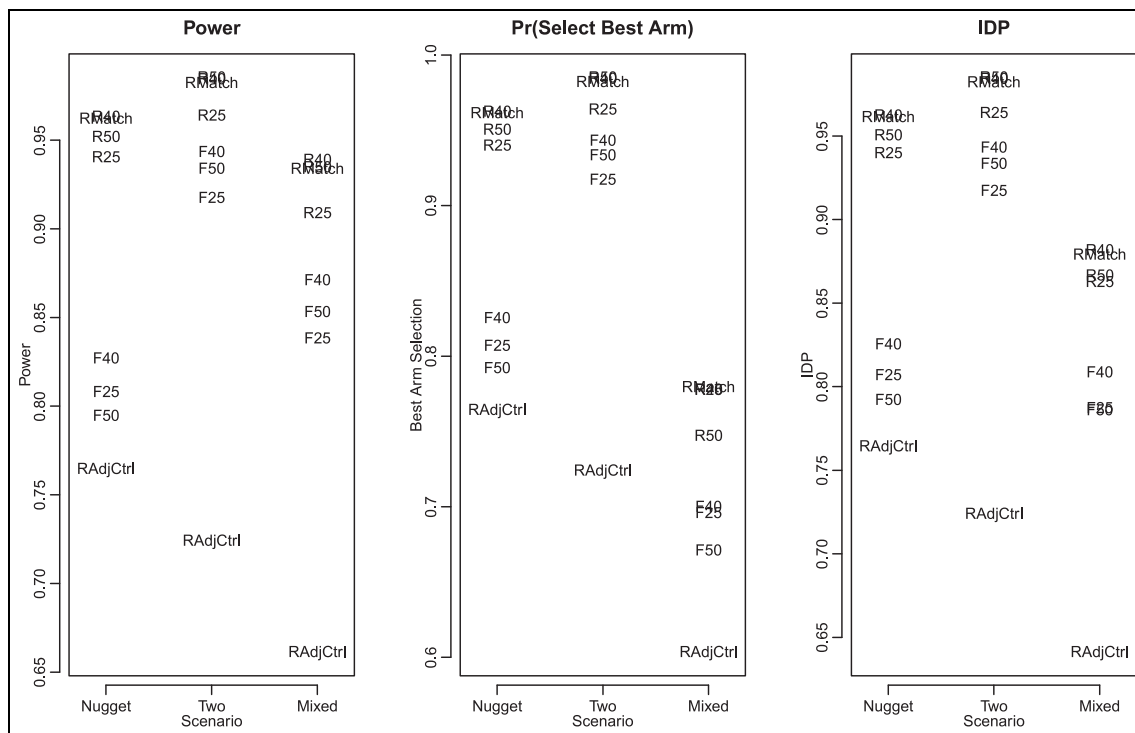


Figure 1. Results for external patient considerations. Left panel shows power, middle panel the probability of selecting the best arm, and the right panel ideal design percentage. The text string within each column indicates the design, either “F25,” “F40,” or “F50” for the fixed designs, “R25,” “R40,” and “R50” for the RAR designs with fixed control, and “RAdjCtrl” and “RMatch” for the RAR designs that alter control throughout the trial.

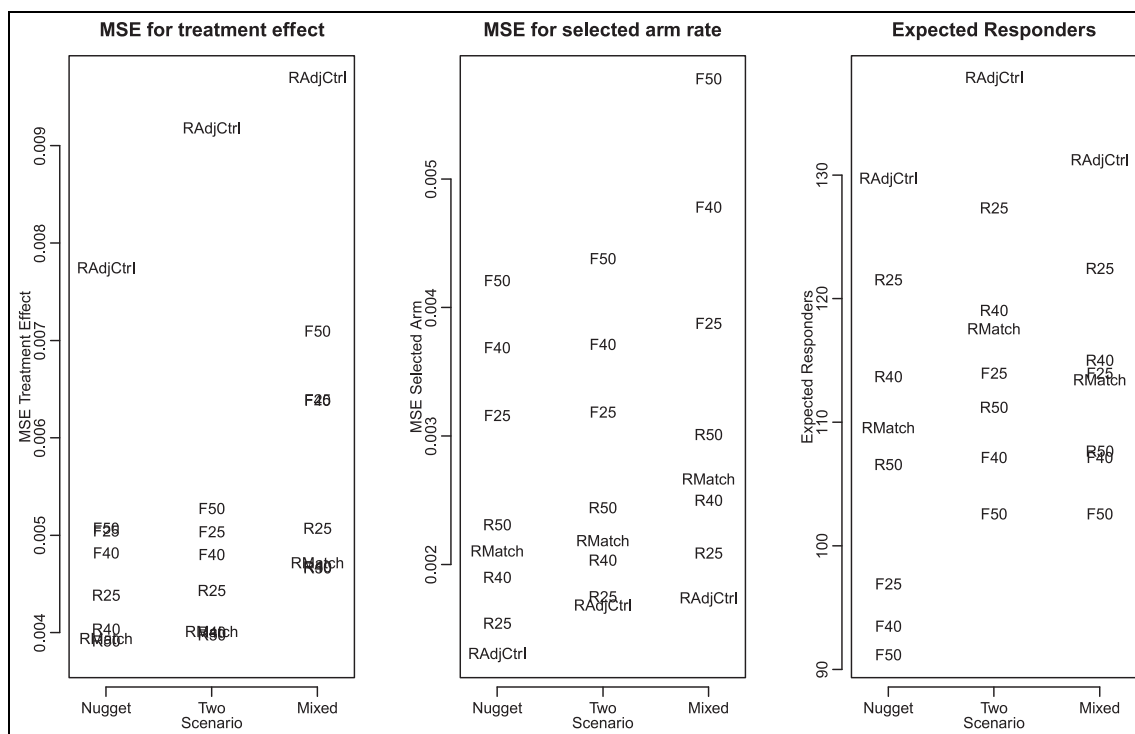


Figure 2. The left and middle panel show mean square error for estimated treatment effects (left) and the response rate on the selected arm (middle), while the right panel shows the expected number of responders. The text string indicates the design as shown in Figure 1.

of interest (power, arm selection, expected responders, etc.).

Figure 1 shows our primary metrics related to external patient considerations. The left panel shows power. The worst performing design is RAdjCtrl, which is consistent with results in the paper by Wathen and Thall²¹ that indicate power decreases when control allocation is reduced. All other RAR designs outperform all fixed designs, with R40 and RMatch producing the best results. Obtaining good performance with 40% allocation to control is not specific to RAR as it is close to the optimal $\sqrt{3}:1:1:1$ allocation among fixed designs.²⁴ The middle and right panels of Figure 1 show the results for arm selection and ideal design percentage, which indicate a similar qualitative picture. For external patients, R40 and RMatch give the best performance and are similar to each other, with R50 nearly identical or close behind. R40 might be preferred in practice to RMatch given its simpler description.

Figure 2 shows the results for the statistical and internal patient metrics. The left and middle panels show the mean square error for the treatment effect and the selected arm response rate. Unlike the other measures, lower values of mean square error are desirable. With the exception of RAdjCtrl, all RAR choices outperform fixed allocation for estimation of treatment effect and response rate for the selected arm. The design RAdjCtrl has highly varied performance, producing the worst mean square error for the treatment effect and the best mean square error for the response rate on the selected arm. RAdjCtrl minimizes control allocation when an effective arm is available. This produces a small control sample size for estimating the control rate

and thus, a component of the treatment effect is estimated poorly. In contrast, RAdjCtrl produces the highest allocation to the selected arm, resulting in good mean square error for the response rate. More limited reversals can be seen in the relative ordering of the RAR designs, with R25 having the second-worst mean square error among RAR designs for the treatment effect, but the second-best mean square error for the response rate on the selected arm.

The right panel of Figure 2 shows the expected number of responders. Here again, RAdjCtrl, with its aggressiveness toward minimizing control allocation in the presence of an apparently effective arm, produces the largest number of expected responders, followed by R25. The other RAR designs outperform their fixed counterparts, typically by 10 or more responders.

Full distribution of each metric

The average behavior of a design may not fully reflect a design’s performance, in that undesirable events may still occur with meaningful probability. For example, Thall et al.⁶ show that a specific form of RAR may obtain a meaningful probability of allocating a larger number of subjects to an inferior arm, even when the overall average allocation is in favor of the superior arm. Figure 3 shows the empirical cumulative distribution functions for the number of responders within each of the 100,000 simulated trials in each scenario (left panel) and the distribution of the squared error of estimation for the treatment effect (middle panel, with a “zoomed in” version in the right panel). These results are only shown for the “mixed” scenario, but they are

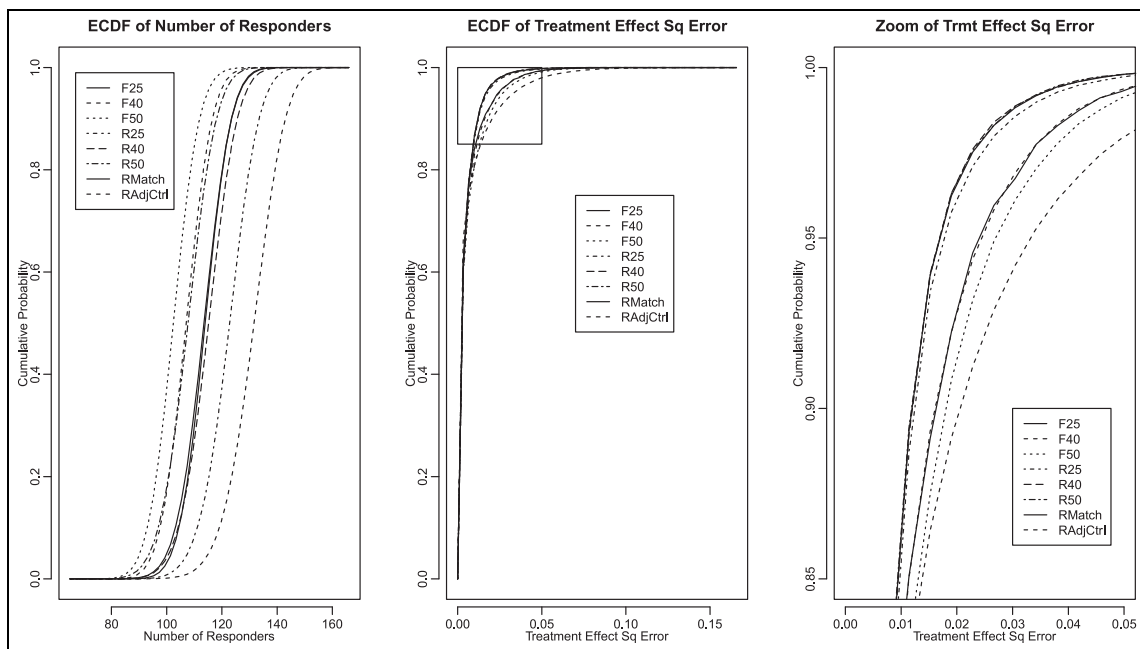


Figure 3. Empirical cumulative distribution functions for the number of responders and the treatment effect mean square error.

Table 2. Performance of each design on arm selection in the mixed scenario.

	Pr(pick arm 1 or better) (%)	Pr(pick arm 2 or better) (%)	Pr(pick arm 3) = Pr(pick best arm) (%)
RMatch	93.5	92.4	78.0
R25	90.9	90.2	77.9
R40	93.9	92.8	77.8
R50	93.5	91.8	74.8
F40	87.1	85.5	69.9
F25	83.9	82.8	69.6
F50	85.4	83.4	67.1
RAdjCtrl	66.1	65.8	60.3

qualitatively similar to the other scenarios. The relative ordering of the designs in expected value matches the relative ordering of the designs throughout the range of the distribution.

Table 2 shows the cumulative probability of picking successively better arms in the mixed scenario, ordered by their probability of picking the best arm. While RMatch, R25, and R40 have similar probabilities of picking the best arm, RMatch and R40 appear to perform slightly better at selecting the second-best arm if they miss the best. RAdjCtrl is particularly poor at arm selection. It has the lowest chance of picking the best arm, and if it misses the best arm has a far lower chance of picking the second-best arm compared to the alternative designs. This explains the large separation of RAdjCtrl from other designs on ideal design percentage, which averages over this arm selection.

Performance in the null scenario

Performance in the null scenario is a special case, as several metrics become irrelevant for comparison. We have purposely calibrated each design to have a common 2.5% familywise type I error, and any arm is “the best” rendering arm selection and ideal design percentage equal across all designs. Given a common 35% null rate in all arms, all designs result in a common Binomial($n = 228, 0.35$) distribution on the number of responders with expectation 79.8. Figure 4 shows the mean squared error for both the treatment effect and response rate estimate for the selected arm. The ordering is identical to the results for the alternative scenarios.

Conclusion and discussion

Our results indicate that in the multiple arm setting, if RAR is implemented where control allocation may be significantly reduced during the trial (RAdjCtrl), the design performance most closely approximates the features and deficiencies of RAR in the two-arm setting. Compared to equal allocation, RAdjCtrl increased the

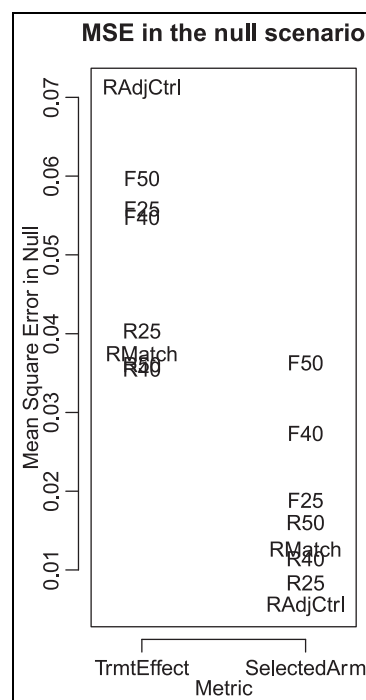


Figure 4. Mean square error on treatment effect and response rate within the null scenario (all other metrics are equal for all designs in the null scenario).

expected number of responders but at the cost of reduced power and poorer estimation of the treatment effect. In addition, this approach has significantly reduced accuracy in arm selection. These results are consistent with Wathen and Thall²¹ who report poor performance on external patient metrics such as power, and also consistent with Berry and Eick⁸ and Lee⁹ who emphasize good performance on internal patient population metrics. The presence of this tradeoff between internal and external patients also is reflective of the importance of the size of the broader patient population for this variant of RAR, as discussed in the commentary by Lee.⁹

In contrast, variants of multiple arm RAR which maintain or increase allocation to control mitigate or reverse many critiques of RAR within the two-arm setting, at least with respect to the simulation scenarios and design variants considered here. These variants of RAR (R25, R40, R50, and RMatch) produce higher power, superior arm selection, and improved mean square error of estimation compared to their fixed allocation counterparts. While not as beneficial on some of the internal patient population metrics as the RAdjCtrl strategy, maintaining or increasing allocation to control avoids many statistical critiques of RAR while maintaining an advantage over fixed allocation over all metrics considered. In design comparisons where external patient metrics such as power are a primary consideration, one of these variants should be preferred as an exemplar of RAR as opposed to RAdjCtrl. This may

explain the improved performance in comparisons such as in the articles by Lin and Bunn¹⁹ and Wason and Trippa.²⁰ The differences between RAdjCtrl and alternative RAR measures are striking, for example, nearly 30% differences in power, indicating that the RAR variant in any comparison needs to be carefully matched to the metric or metrics of primary interest.

The advantages of any design choice must incorporate operational aspects of conducting the trial. Any adaptive design requires interim analyses, and thus, the potential gains in efficiency should be weighed against the cost of conducting the interim analyses. This cost–benefit analysis may differ between sponsors depending on the availability of infrastructure designed for adaptive trials. In addition, within RAR, the performance of R40 closely approximated RMatch. R40 is a simpler design to describe and monitor and may be a preferred choice to increasing control. Alternative methods of increasing control allocation may produce improved results.

The generalizability of our results to alternative endpoints, sample sizes, and effect sizes is unclear. Binomial distributions with moderate sample sizes are approximately normally distributed, with the allocation probabilities and final decisions governed by properties of the approximate joint multivariate normal distributions. This suggests that similar relationships between standardized effect sizes and information fractions may be present in RAR as in group sequential designs.⁵

We have only explored the effect of control allocation in this article, without exploration of other features such as interim frequency, aggressiveness of RAR, or the thresholding to 0 employed for allocation probabilities less than 10%. For example, RAR may be conducted where $\text{Pr}_i(\text{Max})$ is raised to a power before normalizing the allocation probabilities, with that power adjusting over the course of the trial. Further work is required to assess the interaction of these features to determine an optimal RAR that can be compared to alternative design paradigms.


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) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Benjamin Saville  <https://orcid.org/0000-0001-6058-7171>

References

1. Woodcock J and LaVange LM. Master protocols to study multiple therapies, multiple diseases, or both. *N Engl J Med* 2017; 377(1): 62–70.
2. Freidlin B and Korn EL. Adaptive randomization versus interim monitoring. *J Clin Oncol* 2013; 31(7): 969–970.
3. Magirr D, Jaki T and Whitehead J. A generalized Dunnett test for multi-arm multi-stage clinical studies with treatment selection. *Biometrika* 2012; 99(2): 494–501.
4. Trippa L, Lee E, Wen P, et al. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol* 2013; 30(26): 3258–3263.
5. Hu F and Rosenberger W. *The theory of response-adaptive randomization in clinical trials*. Hoboken, NJ: John Wiley, 2006.
6. Thall P, Fox P and Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Ann Oncol* 2015; 26(8): 1621–1628.
7. Korn E and Freidlin B. Outcome-adaptive randomization: is it useful. *J Clin Oncol* 2011; 29(6): 771–776.
8. Berry D and Eick SG. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Stat Med* 1995; 14(3): 231–246.
9. Lee J. Commentary on Hey and Kimmelman. *Clin Trials* 2015; 12(2): 110–112.
10. Hey S and Kimmelman J. Are outcome-adaptive allocation trials ethical. *Clin Trials* 2015; 12(2): 102–106.
11. Van der Graaf R, Roes KC and van Delden JJ. Adaptive trials in clinical research: scientific and ethical issues to consider. *JAMA* 2012; 307(22): 2379–2380.
12. Buyse M, Saad E and Burykowski T. Letter to the editor on adaptive randomization of neratinib in early breast cancer. *N Engl J Med* 2016; 375: 1591–1594.
13. Harrington D and Parmigiani G. I—SPY 2—a glimpse of the future of phase 2 drug development. *N Engl J Med* 2016; 375(1): 7–9.
14. Meinzer C, Martin R and Suarez JJ. Bayesian dose selection design for a binary outcome using restricted response adaptive randomization. *Trials* 2017; 18(1): 420.
15. Berry S, Petzold E, Dull P, et al. A response adaptive randomization platform trial for efficient evaluation of Ebola virus treatments: a model for pandemic response. *Clin Trials* 2016; 13(1): 22–30.
16. Connor J, Elm J and Broglio K. Bayesian adaptive trials offer advantages in comparative effectiveness trials: an example in status epilepticus. *J Clin Epidemiol* 2013; 66(8): S130–S137.
17. Lewis R, Viele K, Broglio K, et al. An adaptive, phase II, dose-finding clinical trial design to evaluate L-carnitine in the treatment of septic shock based on efficacy and predictive probability of subsequent phase III success. *Crit Care Med* 2013; 41(7): 1674–1678.
18. Alexander B, Ba S, Berger M, et al. Adaptive global innovative learning environment for glioblastoma: GBM AGILE. *Clin Cancer Res* 2018; 24(4): 737–743.
19. Lin J and Bunn V. Comparison of multi-arm multi-stage design and adaptive randomization in platform clinical trials. *Contemp Clin Trials* 2017; 54: 48–59.

20. Wason J and Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat Med* 2014; 33(13): 2206–2221.
21. Wathen J and Thall P. A simulation study of outcome adaptive randomization in multi-arm clinical trials. *Clin Trials* 2017; 14(5): 432–440.
22. Jiang Y, Zhao W and Durkalski-Mauldin V. Impact of adaptation algorithm, timing, and stopping boundaries on the performance of Bayesian response adaptive randomization in confirmative trials with a binary endpoint. *Contemp Clin Trials* 2017; 62: 114–120.
23. London A. Learning health systems, clinical equipoise and the ethics of response adaptive randomisation. *J Med Ethics* 2018; 44(6): 409–415.
24. Dunnett C. A multiple comparison procedure for comparing several treatments with a control. *J Amer Stat Assoc* 1955; 50: 1096–1121.