Robust Response-Adaptive Randomization Design

Thevaa Chandereng\textsuperscript{1,2} and Rick Chappell\textsuperscript{1,2}

Abstract

In clinical trials, patients are randomized with equal probability among treatments to obtain an unbiased estimate of the treatment effect. However, response-adaptive randomization has been proposed due to ethical reasons, especially in rare diseases where randomization ratio is tilted to favor the better performing treatment. The substantial disagreement regarding time-trends in adaptive randomization is not fully addressed. The type I error is inflated in the traditional Bayesian adaptive randomization approach when time-trend is present. In our approach, patients are assigned in blocks and the randomization ratio is recomputed for blocks rather than traditional adaptive randomization where it is done per patient. We further investigate the design with a range of scenario for both frequentist and Bayesian design. We compare our method with equal randomization and with different number of blocks including traditional response-adaptive randomization design where randomization ratio is altered patient by patient basis. The analysis is done in stratification if there is two or more patient in each block. Having a large number of blocks or randomizing a few subjects into a block should be avoided due to the possibility of not acquiring any information from the block(s). On the other hand, response-adaptive randomization with a small number of blocks has a good balance between efficiency and treating more subjects to the better-performing treatment.

Keywords

adaptive design, play the winner rule, clinical trials, group-sequential design, Bayesian design

Introduction

Response-adaptive randomization

Randomization remains a pivotal methodology for advancement in medical knowledge. It removes any systematic bias and allows direct inference between the treatment group and outcome. Traditionally, fixed randomization scheme (1:1 or 2:1) is used to due to simplicity in design and execution of the trial. However, response-adaptive randomization (RAR) design utilizes accrual information to tilt the randomization ratio to the better performing treatment group. Patients enrolled in these trials are not only treated to obtain the effectiveness of treatment but also treated to the best way possible\textsuperscript{1}. Response-adaptive randomization design primarily targets to solve both the issues mentioned above at once with a good balance. The biasness in the trial can be greatly affected if one arm is very superior to the other arm(s) due to a drastic alteration in the randomization ratio.

Response-adaptive randomization has been highly advocated due to ethical reasons. The primary use of adaptive designs in clinical trials is that it improves the benefit/risk for the patients enrolled in a trial\textsuperscript{2}. For instance, in Zidovudine Treatment (AZT) trial conducted to test the reduction of maternal-infant transmission of HIV type 1, patients were randomized to 1:1\textsuperscript{3}. The equal randomization scheme used placed 239 women in the treatment group (AZT) and 238 patients in the placebo group. Out of the 238 women in the AZT group, 60 of the infants were transmitted with the HIV virus while only 20 infants contracted the HIV virus for the AZT group\textsuperscript{3}. The outcome of the trial confirmed that the new treatment works. Given the seriousness of the outcome of this study, it is reasonable to argue that 50-50 allocation was unethical. As the outcome of the trial becomes available, the randomization ratio should have tilted in favor of the AZT group. Response-adaptive randomization design could have reduced the number of infants that contracted HIV disease from their maternal.

On the other hand, opponents of RAR have argued that adaptive randomization challenges the whole notion of equipoise\textsuperscript{4}. Hey and Kimmelman has also argued that most new treatments offer small improvement standard treatment, thus they offer limited benefit and require a larger sample size\textsuperscript{2}. Hey and Kimmelman also suggested that equal randomization helps reduce the trial size and length, thus it benefits future patients rather than current patients enrolled in the trial\textsuperscript{2}. Korn and Friedlin measure the difference in non-responders under equal and adaptive randomization and found that adaptive randomization required a larger trial to achieve the same power and type I error\textsuperscript{5}. The major drawback to response-adaptive randomization design is that the trial needs to be short to be able to obtain the outcome of the trial for future randomization\textsuperscript{6}.

\textsuperscript{1}Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA
\textsuperscript{2}Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

Corresponding author:

Thevaa Chandereng, Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, 201 WARF, 610 Walnut Street, Madison, WI 53726.

Email: chandereng@wisc.edu
Time-trend issues

However, one of the main critics against response-adaptive randomization is the time-trend issue. This contributes to the main factor of why RAR is infrequently used. The type I error rate is usually not controlled at the nominal level under traditional Bayesian RAR design. Besides affecting type I error, studies have shown that there is a large bias in the estimation of treatment difference under traditional RAR design. Figure 1 shows that the increase in response in the treatment group tilts the randomization ratio in favor of the treatment. However, this is directly confounded by time.

Example of time trend issues in response-adaptive randomization (RAR) design includes:

- The disease itself can change, sometimes radically (e.g., AIDS in the early 1990s).
- Our definition of the disease can change due to new scientific discoveries or diagnostic methods (e.g., stage migration in nasopharyngeal carcinoma due to the introduction of CAT scans to Hong Kong 2005).
- Inclusion criteria can change, either formally (in which case we can stratify analysis on before vs. after the change) or informally due to recruiting zeal or other issues (in which case we cannot).
- Centers can change, such as when VAs enter the trial earlier or later than academic institutions.
- Patients within centers can change, especially but not only with chronic diseases, due to the phenomenon of a queue of desperate patients lining up at the door.
- In addition to these examples, an investigator who wants to game the system could cross his/her fingers that his favored treatment arm is ahead, then progressively enroll better prognosis patients over time.

In long duration trials, time-trends more likely to occur. Patient’s characteristics might be completely different throughout the trial or even at the beginning and end of the trial (which is also known as “patient drift”). Since most of the RAR design have adaptiveness on the fly, there is one important assumption made. The sequence of patients who arrive for entry into the trial represents samples drawn at random from two homogenous populations, with no drift in the probabilities of success. However, this assumption is usually violated. For example, there were more smokers enrolled in the latter part of the trial than the beginning of the trial in the Lung Cancer Elimination (BATTLE).

Given the serious flaw in RAR design, there is not much literature on this area to address the time-trend issue. Randomization test for adjusting type I error inflation was proposed by Simon and Simon using different RAR rules for double-arm trials. Jennison and Turnbull explored a group-sequential method for RAR with continuous outcome utilizing multi-stage terminology. Karrison et al. introduced a stratified group-sequential method with a simple example of altering the randomization ratio to address this issue. Coad used a very similar stratified analysis to Karrison et al. Rosenberger et al. introduced a covariate-adjusted RAR procedure where time mechanism can be added as a covariate in the model. Thall et al. investigated type I error under a linear time-trend induced in the traditional response-adaptive randomization design and showed that the type I error is significantly above the nominal level. Villar et al. explored the hypothesis testing procedure and adjusting for covariates for correcting type I error inflation and the effect on power in RAR design with time-trend effect added to the model for two-armed and multi-armed trials.

Time-trend can not only greatly affect the biasness in the difference in treatment effect but it can also wrongly reject a true null hypothesis. We propose a block (group-sequential) design where the randomization ratio is altered in a block level instead of a patient by patient basis using both frequentist and Bayesian approach. In each block, the randomization ratio is kept constant. The block design is similar to the stratified group design introduced by Karrison et al. Then, we further study the robustness in different block sizes using both frequentist and Bayesian approach. We also compare these results with traditional RAR design and with fixed (1:1) randomization.

Trial Design and Simulation

Block Design for RAR and Why?

In binary outcomes, events (success/failure) are observed within a short period from the beginning of the treatment. In a block design, patients are enrolled in a sequential manner. For instance, in two-arm design (treatment A and B) patients are enrolled in block with sample size of $n_{Ak}$ and $n_{Bk}$, for $k = 1, 2, 3, ..., K$, where $n_{Ak}$ and $n_{Bk}$ represent the sample sizes in treatment group A and B in block $k$. In this design, the randomization ratio is constant for patients within each block and the randomization ratio is altered at the block level. Unlike traditional RAR which alters the randomization ratio by patient basis, this method speeds up the process of RAR trials since randomization ratio is modified for a group of patients in the block. The initial randomization ratio is set to 1:1. Bashir et al. have implemented two-block design where patients are randomized 1:1 in a group of 10 and then based on the outcome, they randomized the next ten patient to the superior treatment. However, at the second block,
they randomized all the patient to the lower dosage because the probability of randomizing patient to the lower dosage was 0.9 compared to 0.1 to the higher dosage. This design should be considered a randomized control trial for the first block and an observational study for the second block. As asserted by Karrison et al., the block design will eliminate bias due to drift through stratification. However, the optimal block size remains unclear and it is highly dependent on the operating characteristics of a trial. Blocks with a large number of patients also help reduce the large probability of imbalances in the wrong direction, where more patients are assigned to the inferior treatment compared to traditional RAR. Blocks with fewer patients can reduce the power of the trial because if patients are not randomized to both treatments in a block, the block becomes uninformative.

**Simulation and Design**

We investigated the effect of a different number of blocks using simulations for both frequentist and Bayesian approach. The rules for altering the randomization ratio in both Bayesian and frequentist design are described in the subsection below. The target sample size of \( N = 200 \) subjects were considered with number of blocks (strata), \( K = 1, 2, 4, 5, 10, 20, 100, 200 \). When \( K = 1 \), this follows the traditional equal allocation design and when \( K = 200 \), this follows the traditional RAR design where stratification is ignored. Upon completion of enrollment and collection of data of the subjects in each block, the interim analyses are done to revise the allocation rule. The interim analysis also allows for early stopping and details of early stopping for both Bayesian and frequentist design is included below. 10,000 independent simulations were performed for each design yielding a Monte-Carlo standard error of 0.25%.

In each simulation, the success rate of treatment A (control group), \( p_A \), is set to 0.25. The alternatives (the success rate of treatment B, \( p_B \)) is set to 0.25 (null case), 0.35, 0.45. On top of simulating both Bayesian and frequentist design for \( p_A \) and \( p_B \) specified above, we also simulated RAR with drift effect. To examine the effects of drift, we increased both \( p_A \) and \( p_B \) linearly from their initial values to a final value of 0.25 larger. The drift was added onto each block rather than patient by patient basis. For instance, the success rate in strata \( k \) for treatment A is \( p_A(k) = 0.25 + 0.25((k - 1)/K) \), where \( K \) is the total number of strata. While both the \( p_A \) and \( p_B \), the treatment effect remains constant throughout the trial. For both frequentist and Bayesian design, the final analysis is done using stratification for a number of blocks, \( K \) where \( 1 < K < 200 \). For traditional RAR and traditional RAR with 1:1 allocation ratio, the standard analysis is used. Details of the analysis are attached in the Bayesian and frequentist approach section.

**Frequentist Approach**

The allocation rule for the frequentist approach is altered using optimal allocation ratio for 2-armed trial specified by Rosenberger et al. The allocation probability for treatment A in stratum \( j \) is defined as

\[
\pi_{j,A} = \frac{\sqrt{p_A}}{\sqrt{p_A} + \sqrt{p_B}},
\]

where \( \hat{p}_A \) is the estimated success rate of treatment A and \( \hat{p}_B \) is the estimated success rate of treatment B up to block \( j - 1 \). However, the allocation probability is only altered upon observed both event (success and no response) in both the arm. Simulation for early stopping is also included, the alpha spending approach was incorporated to stop for early success or failure. For traditional RAR and fixed allocation, one-sided chi-square test was used to analyze the outcome. However, for the block design, one-sided Cochran-Mantel-Haenszel test was utilized to deal with the stratification. Yates’s correction was not applied for both chi-square and Cochran-Mantel-Haenszel test due to the overly conservative nature of the correction. Treatment B is proclaimed superior to treatment A if the one-sided p-value < 0.05.

The computation of the mean proportion of treatment difference is computed as below for stratified design. In two treatment scenario (treatment A & B), suppose there are \( K \) strata, let \( p_{Ak} \) and \( p_{Bk} \) be the proportion of success in treatment A and stratum \( k \). Let \( n_{Ak} \) and \( n_{Bk} \) be the number of patients in treatment A and treatment B and stratum \( k \). Let

\[
\delta_k = p_{Bk} - p_{Ak}
\]

be the observed difference in proportion between treatment A and B in stratum \( k \). The observed proportion of treatment difference (\( \delta \)) is computed as below,

\[
\hat{\delta} = \frac{1}{K} \sum_{k=1}^{K} w_k \delta_k
\]

where \( w_k = \frac{(n_{Ak}+n_{Bk})^{-1}}{\sum_{k=1}^{K}(n_{Ak}+n_{Bk})^{-1}} \) is the weight of the stratum \( k \) and \( \sum_{k=1}^{K} w_k = 1 \). For non-stratified design, where the block size \( K = 1, 200 \), the difference in proportion between treatment B and A (\( \delta \)) is obtained as follows,

\[
\hat{\delta} = \hat{p}_B - \hat{p}_A
\]

where \( \hat{p}_B \) is the estimated proportion of success in treatment B and \( \hat{p}_A \) is the estimated proportion of success in treatment A.

**Bayesian Approach**

In the Bayesian approach, the Bayesian adaptive randomization (BAR(1/2)) method introduced by Thall and Wathen is employed. The probability of randomizing subjects to treatment A in stratum \( j \) is defined as

\[
\pi_{j,A} = \frac{(p_{A>B}(data))^{1/2}}{(p_{A>B}(data))^{1/2} + (p_{B>A}(data))^{1/2}},
\]

where \( p_{A>B}(data) \) is the posterior probability that treatment A has a higher success rate than treatment B and \( p_{A>B}(data) = 1 - p_{B>A}(data) \). The beta-Binomial conjugate prior is used for the estimation of the posterior probability. The posterior probability that the treatment A has a higher event rate than treatment B is

\[
p_{A>B}(data) \sim \text{Beta}(y_A + a_0, N_A - y_A + b_0) - \text{Beta}(y_B + a_0, N_B - y_B + b_0) > 0,
\]
where $y_A$ and $y_B$ denote the number of events in treatment A and B, $N_A$ and $N_B$ denote the number of subjects in treatment A and B, $a_0$ and $b_0$ denote the prior rate parameter of the beta distribution. Similar to Thall and Wathen’s approach, beta(0.5, 0.5) priors were assumed for both treatment A and B. Since there is no closed-form solution for the difference in beta distributed random variables, Monte Carlo simulations were performed to estimate the posterior of the treatment difference.

Similar to the frequentist design, simulation for the possibility of early stopping are included. If treatment B is selected to be better than treatment A, then if $P_{B>A}(\text{data}) > .99$, the trial is stopped early for success and if $P_{B>A}(\text{data}) < .01$, the trial is stopped early for failure. For non-stratified design, where $K = 1, 200$, if the final posterior probability $P_{B>A}(\text{data}) > .95$, then treatment B is declared superior to treatment A. The mean estimate of the proportion of treatment difference is estimated using Monte Carlo simulation. The mean value of $P_{B>A}(\text{data})$ is used to estimate the proportion of treatment difference.

For stratified design by block, Bayesian logistic regression was implemented to estimate the posterior probability of treatment difference. The logistic regression model is defined as below

$$Logit(y_{ij}) \sim \beta_0 + \beta_{\text{trt}}x_{ij} + \sum_{j=1}^{K} \beta_j x_{ij},$$

where $y_{ij}$ is the outcome of treatment $i$ and stratum $j$, $x_{ij}$ is the indicator variable of treatment $i$ and stratum $j$, $\beta_0$ is the intercept term, $\beta_{\text{trt}}$ is the treatment effect, $\beta_j$ is the stratum effect of stratum $j$. The uninformative prior applied in the model are as defined

$$\beta_0, \beta_{\text{trt}}, \beta_j \sim N(0, \sigma_2^2), \quad \sigma_2^2 \sim Inv - \chi^2(1, 2.5).$$

The posterior value of $\beta_{\text{trt}}$ is used to estimate the proportion difference in treatment. The R package arm with quasi family and $\mu(1 - \mu)$ link was used to fit Bayesian logistic regression and obtain the posterior samples of $\beta_{\text{trt}}$. The proportion difference in treatment is estimated using the mean posterior value of $\beta_{\text{trt}}$. If difference of proportion between B and A is estimated using $\beta_{\text{trt}}$, then treatment B is declared superior if $E(\beta_{\text{trt}} > 0) > 0.95$.

Results

For all cases, the simulation was replicated for 10,000 times. The simulation is reproduced for both Bayesian and frequentist design with and without time trends. The number of blocks ($K$) chosen are as following, $K = 1, 2, 4, 5, 10, 20, 100, 200$. For all the simulation, the power (treatment B is superior to treatment A), bias, probability of sample imbalance of more than 20 patients assigned to treatment A over treatment B, the mean, 2.5% and 97.5% percentile of difference between $N_B - N_A$ (sample size imbalance favoring treatment B over treatment A) are reported. The power reflects the proportion of 10,000 trial that declares treatment B superior to treatment A. The type I error (false-positive) is simulated by setting both $p_A, p_B = 0.25$. The bias computes the estimated difference in proportion of treatment estimate, bias = $\delta - \delta_{20}$, the probability of assigning 20 or more subjects to treatment A over treatment B ($P(N_A - N_B > 20)$) is included in the simulation. Due to nontrivial possibility of assigning more patients to the inferior arm by chance, it is vital to analyze $\pi_20$. To highlight the main advantage of RAR compared to equal randomization, the difference in sample size between the superior arm and inferior arm is presented. Since the difference between subjects assigned to treatment B and treatment A ($N_B - N_A$) is skewed and dispersed as illustrated in Figure 2, the mean, 2.5% and 97.5% of $N_B - N_A$ are reported.

Simulation with No Time-Trend

Table 1 and Table 2 displays the results of simulations for a number of blocks, $K = 1, 2, 4, 5, 10, 20, 100, 200$ using the frequentist and Bayesian approach with no drift applied and not stopping early for success or failure. The frequentist approach (Table 1) manages to control the type I error unlike the Bayesian design in Table 2. Even though stratified RAR design has higher type I error, the type I error is still controlled under the nominal level (Table 1). However, type I error is high for 4, 5, 10 and 20 number of blocks under the Bayesian design (Table 2). Block design with 2, 4, 5 and 10 number of blocks provides a higher power when $p_B = 0.35, 0.45$ (Table 1). In the Bayesian design, the fixed randomization ratio provides the best power. Although the type I error is close to the nominal level under $K = 2, 4$ and 5, with a slight increase in sample size it could be lowered to the nominal level. A small increase in sample size might still be favorable if more subjects are treated with the best possible cure. Block design with a small number of subjects in a block should not be considered due to low power in
the design as shown in Table 1 and 2 with 100 blocks. This poor performance is due to the reality that some of the blocks can be noninformative if subjects in the block are randomly assigned to the small treatment group. Ethically, this design (2 subjects per block) places more subjects at risk without contributing to the advancement in science.

| pB  | Block | Power | Bias | π20  | N datasets |
|-----|-------|-------|------|------|------------|
| 0.25 | 1     | 0.03  | 0.00 | 0.06 | 0.00 (-28, 28) |
| 2    | 0.05  | 0.00  | 0.03 | 6.97 (-22.00, 36) |
| 4    | 0.05  | 0.00  | 0.02 | 11.80 (-18, 44) |
| 5    | 0.05  | 0.00  | 0.01 | 13.01 (-18, 46) |
| 10   | 0.05  | 0.00  | 0.01 | 14.63 (-16, 48) |
| 20   | 0.05  | 0.00  | 0.01 | 15.02 (-16, 48) |
| 100  | 0.05  | 0.00  | 0.01 | 15.24 (-16, 48) |
| 200  | 0.03  | 0.00  | 0.01 | 15.11 (-16, 48) |
| 0.35 | 1     | 0.33  | 0.00 | 0.07 | 0.15 (-28, 28) |
| 2    | 0.45  | 0.00  | 0.02 | 9.94 (-20, 42) |
| 4    | 0.45  | 0.00  | 0.01 | 15.83 (-16, 52) |
| 5    | 0.46  | 0.00  | 0.01 | 16.83 (-16, 52) |
| 10   | 0.44  | 0.00  | 0.01 | 19.28 (-14, 58) |
| 20   | 0.42  | 0.00  | 0.01 | 19.97 (-14, 58) |
| 100  | 0.20  | 0.00  | 0.01 | 20.26 (-12, 58) |
| 200  | 0.35  | 0.00  | 0.01 | 20.61 (-14, 58) |
| 0.45 | 1     | 0.85  | 0.00 | 0.07 | 0.04 (-28, 28) |
| 2    | 0.91  | 0.00  | 0.01 | 14.99 (-16, 48) |
| 4    | 0.90  | 0.00  | 0.00 | 23.26 (-10, 60) |
| 5    | 0.91  | 0.00  | 0.00 | 25.26 (-8, 62) |
| 10   | 0.90  | 0.00  | 0.00 | 28.62 (-6, 66) |
| 20   | 0.88  | 0.00  | 0.00 | 29.03 (-6, 68) |
| 100  | 0.47  | 0.00  | 0.00 | 29.75 (-4, 68) |
| 200  | 0.84  | 0.00  | 0.00 | 29.80 (-6, 70) |

Table 1. RAR using frequentist approach with no early stopping criteria and no drift applied. The pA is set to 0.25 for all cases. Bias = δ - δ. π20 denote the probability of assigning more than 20 patients in the inferior treatment group. P(N datasets > 20). N datasets and N datasets denote the number of patient assigned to treatment A and B. The mean (2.5%, 97.5%) of N datasets - N datasets is reported in the last column. 10,000 simulation was done for each case.

The estimation of the actual treatment effect is unbiased under the frequentist design regardless of the number of blocks used. However, the estimation of treatment effect is biased for most stratified RAR design and traditional RAR as shown in Table 2. Unlike the Bayesian approach (Table 2), the variability in sample size assigned to treatment A and B is smaller in the frequentist approach (Table 1). The mean difference in the subject’s assignment of treatment is also smaller in the frequentist design compared to the Bayesian design. Thus, the frequentist design is a little conservative of assigning patients to the better-performing treatment, unlike the Bayesian design. At the largest difference in proportion pB = 0.45, pA = 0.25, there is still room for imbalance in the wrong direction in the frequentist design compared to the Bayesian design (Table 1 and 2). The imbalance in sample size favoring the inferior treatment is small under all scenario as shown by π20 in Table 1. The difference in sample size (N datasets - N datasets) is relatively small and close to 0 for the frequentist design. On the other hand, the imbalance is large for the Bayesian design as illustrated by π20 in Table 2. The large difference in sample size (more than half the total sample size) is seen in the Bayesian design with the most extreme difference in the outcome of treatment A and B.

Table ?? and Table ?? displays the results of simulations for a number of blocks, K = 1, 2, 4, 5, 10, 20, 100, 200 using the frequentist and Bayesian approach and with early stopping criteria for success or failure implemented. Parallel to the earlier results, the drift is higher in Table ?? and Table ??, Table ?? emphasizes that K = 10, 20, 100 and 200 have a inflated type I error of 0.12, 0.16 and 0.18 similar to Table 4. As seen in type I error and power in Table ?? and ??, a large number of stratum with Bayesian design should not be considered for clinical studies.

#### Simulation with Time-Trend

Table 3 and Table 4 displays the results of simulation for number of blocks, K = 1, 2, 4, 5, 10, 20, 100, 200 using the frequentist and Bayesian approach with 0.25 drift applied and not stopping early for success or failure. With time drift, the false positive rate is still under the nominal level for the frequentist design with all block sizes as seen in Table 3. However, the type I error is inflated for traditional RAR, 10 and 20 blocks in the Bayesian design. Having a 2, 4, 5, 10 and 20 blocks still remain the most powerful design with the most extreme difference in outcome of pA and pB. In the Bayesian design, block size with 50 subjects is still comparable to the traditional fixed randomization ratio design, since the type I error is controlled under 0.05 and it has the highest power of 0.91 under the maximum difference between pA and pB.
Large number of blocks still remains as a poor design as illustrated earlier in Table 2. The estimation of treatment difference remains the same to the simulation with no time-trend except the bias is a little higher under traditional RAR design as presented in Table 4. The difference in sample size, the imbalance in the wrong direction and the variability between the arms (N_B – N_A) remains comparable to the simulation without time drift applied.

Table ?? and Table ?? displays the results of simulation for number of blocks, K = 1, 2, 4, 5, 10, 20, 100, 200 using the frequentist and Bayesian approach and with early stopping criteria for success or failure implemented. The output in Table ?? is similar to Table 3 except the treatment different is slightly biased. However, it is shown that clinical trials with early stopping are slightly biased compared to trials without early stopping criteria imposed. Table ?? emphasizes that K = 10, 20, 100 and 200 have a inflated type I error of 0.14, 0.17 and 0.11 similar to Table 4. As seen in type I error and power in Table 4 and ??, large number of stratum design should not be considered for clinical studies.

### Table 3. RAR using frequentist approach with no early stopping criteria and with 0.25 drift applied. The p_A is set to 0.25 for all cases. $Bias = \delta - \pi_{20}$ denote the probability of assigning more than 20 patients in the inferior treatment group. N_A and N_B denote the number of patient assigned to treatment A and B. The mean (2.5%, 97.5%) of N_B – N_A is reported in the last column. 10,000 simulation was done for each case.

| p_B | Block | Power | Bias | $\pi_{20}$ | N_B – N_A |
|-----|-------|-------|------|-----------|-----------|
| 0.25 | 1 | 0.03 | 0.00 | 0.07 | 0.00 (-28, 28) |
| 2 | 0.05 | 0.00 | 0.04 | 5.28 (-24, 34.00) |
| 4 | 0.05 | 0.01 | 0.03 | 9.56 (-20, 40) |
| 5 | 0.05 | 0.00 | 0.02 | 10.81 (-20, 42) |
| 10 | 0.05 | 0.00 | 0.01 | 12.90 (-16, 44) |
| 20 | 0.05 | 0.00 | 0.01 | 13.21 (-16, 46) |
| 100 | 0.05 | 0.00 | 0.01 | 13.28 (-16, 46) |
| 200 | 0.03 | 0.00 | 0.01 | 13.23 (-16, 46) |
| 0.35 | 1 | 0.30 | 0.00 | 0.07 | 0.06 (-28, 28) |
| 2 | 0.40 | 0.00 | 0.03 | 7.09 (-22, 38) |
| 4 | 0.41 | 0.00 | 0.02 | 12.16 (-18, 44) |
| 5 | 0.41 | 0.00 | 0.02 | 13.55 (-18, 48) |
| 10 | 0.40 | 0.00 | 0.01 | 16.36 (-14, 52) |
| 20 | 0.38 | 0.00 | 0.01 | 16.92 (-14, 50) |
| 100 | 0.18 | 0.00 | 0.01 | 17.30 (-14, 52) |
| 200 | 0.29 | 0.00 | 0.01 | 17.61 (-14, 52) |
| 0.45 | 1 | 0.82 | 0.00 | 0.07 | -0.07 (-28, 28) |
| 2 | 0.89 | 0.00 | 0.02 | 11.02 (-20, 40) |
| 4 | 0.88 | 0.01 | 0.03 | 18.41 (-14, 52) |
| 5 | 0.88 | 0.01 | 0.03 | 20.40 (-12, 56) |
| 10 | 0.88 | 0.00 | 0.03 | 23.70 (-10, 58) |
| 20 | 0.85 | 0.00 | 0.00 | 24.32 (-8, 60) |
| 100 | 0.46 | 0.00 | 0.00 | 25.02 (-8, 60) |
| 200 | 0.80 | 0.00 | 0.00 | 25.31 (-8, 62) |

### Table 4. RAR using Bayesian approach with no early stopping criteria and with drift applied. The $p_A$ is set to 0.25 for all cases. $Bias = \delta - \pi_{20}$ denote the probability of assigning more than 20 patients in the inferior treatment group. N_A and N_B denote the number of patient assigned to treatment A and B. The mean (2.5%, 97.5%) of N_B – N_A is reported in the last column. 10,000 simulation was done for each case.

| p_B | Block | Power | Bias | $\pi_{20}$ | N_B – N_A |
|-----|-------|-------|------|-----------|-----------|
| 0.25 | 1 | 0.03 | -0.01 | 0.03 | 0.44 (-21, 28) |
| 2 | 0.07 | 0.00 | 0.29 | -0.30 (-80, 77) |
| 4 | 0.05 | -0.01 | 0.41 | -5.84 (-96, 86) |
| 5 | 0.06 | 0.00 | 0.43 | -1.66 (-117, 113) |
| 10 | 0.12 | 0.01 | 0.34 | 3.32 (-148, 148) |
| 20 | 0.20 | 0.02 | 0.33 | 12.40 (-137, 140) |
| 100 | 0.01 | 0.00 | 0.36 | 4.18 (-130, 141) |
| 200 | 0.17 | 0.02 | 0.32 | 12.74 (-106, 138) |
| 0.35 | 1 | 0.51 | 0.00 | 0.10 | -1.74 (-32, 26) |
| 2 | 0.42 | 0.01 | 0.06 | 38.02 (-35, 97) |
| 4 | 0.35 | 0.01 | 0.07 | 52.78 (-87, 140) |
| 5 | 0.46 | 0.03 | 0.06 | 68.38 (-48, 148) |
| 10 | 0.45 | 0.03 | 0.03 | 72.88 (-34, 149) |
| 20 | 0.51 | 0.02 | 0.11 | 73.24 (-50, 162) |
| 100 | 0.11 | -0.07 | 0.11 | 74.40 (-80, 162) |
| 200 | 0.59 | 0.05 | 0.07 | 75.92 (-60, 162) |
| 0.45 | 1 | 0.91 | 0.00 | 0.08 | -0.10 (-35, 28) |
| 2 | 0.93 | 0.02 | 0.00 | 68.28 (20, 111) |
| 4 | 0.91 | 0.04 | 0.01 | 105.34 (26, 148) |
| 5 | 0.79 | 0.03 | 0.01 | 105.96 (5, 160) |
| 10 | 0.82 | 0.04 | 0.00 | 119.52 (27, 173) |
| 20 | 0.75 | 0.03 | 0.01 | 119.74 (13, 177) |
| 100 | 0.22 | -0.13 | 0.00 | 123.12 (34, 175) |
| 200 | 0.89 | 0.06 | 0.00 | 125.48 (54, 179) |

**Conclusion**

RAR design has a lot of appealing properties mainly assigning more patients to better-performing treatment. However, trialist needs to be careful with issues of time-trend and methods to alter the randomization ratio. Time-trends can significantly impact the type I error rate which can affect the validity of clinical studies. As a statistician, we cannot emphasize more on the importance of controlling the false-positive rate in a clinical setting. Thall et al. have shown that methods that are able to control type-I error have failed to detect the true treatment difference.

Besides controlling for false positive rate, trialist needs to make sure the method of altering the randomization ratio is not too extreme which can greatly affect the bias and question the notion of randomization in clinical studies. Thall et al. have emphasized on the difference in simulation between BAR(1) and BAR(1/2), with BAR(1) having a large imbalance in the wrong direction and larger false-positive rate. Zelen’s play-the-winner rule was implemented to an extracorporeal membrane oxygenation (ECMO) trial where the first patient was assigned to both control and treatment group. Due to the failure in the control and success in the treatment group, all subsequent patient’s were randomized to the treatment group. However, it was later discovered that the first patient assigned to the control group was much sicker than all the patients randomized to the treatment groups.

On the other hand, scientists need to be aware of the risk that RAR poses which includes having assigned more patients to the inferior treatment. As highlighted by Thall et al., “The practical and ethical point is that AR may behave pathologically in that it carries a nontrivial risk of creating a large sample size imbalance in favor of the inferior treatment.” Large imbalance in the wrong direction can also...
be controlled by methods that do not alter the randomization ratio rapidly.

It is shown that a small number of blocks (K = 2, 4 and 5) has a good tradeoff between efficiency and ethically treating patients to the best known superior treatment. A large number of blocks should be clearly avoided for both ethical reason and poor design. Traditional RAR does not only delay the trial but also affects the clinical conclusion achieved. We have not considered the multiple treatment design with more than 2-arms. The design would be much more complex and it should be examined further.

An R package (blockRAR), for the frequentist and Bayesian models, is implemented in R and released as open source software under the MIT license. The blockRAR package is available at Comprehensive R Archive Network (CRAN) and at https://thevaachandereng.github.io/blockRAR/. We used blockRAR version 1.0.0 for all analyses.

Supplemental material
Supplementary material are available.

Declaration of conflicting interests
None declared.

Funding
None declared.

Acknowledgements
We are grateful to Tom Cook for his helpful comments and suggestions.

References
1. Wei L and Durham S. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* 1978; 73(364): 840–843.
2. Hey SP and Kimmelman J. Are outcome-adaptive allocation trials ethical? *Clinical trials* 2015; 12(2): 102–106.
3. Connor EM, Sperling RS, Gelber R et al. Reduction of maternal-infant transmission of human immunodeficiency virus type 1 with zidovudine treatment. *New England Journal of Medicine* 1994; 331(18): 1173–1180.
4. Begg CB. Ethical concerns about adaptive randomization, 2015.
5. Korn EL and Freidlin B. Outcome-adaptive randomization: is it useful? *Journal of Clinical Oncology* 2011; 29(6): 771.
6. Karrison TG, Huo D and Chappell R. A group sequential, response-adaptive design for randomized clinical trials. *Controlled Clinical Trials* 2003; 24(5): 506–522.
7. Thall P, Fox P and Wathen J. Statistical controversies in clinical research: scientific and ethical problems with adaptive randomization in comparative clinical trials. *Annals of Oncology* 2015; 26(8): 1621–1628.
8. Chappell R and Karrison T. Continuous bayesian adaptive randomization based on event times with covariates by cheung et al., statistics in medicine 2006; 25: 55–70. *Statistics in medicine* 2007; 26(15): 3050–3052.
9. Liu S and Lee JJ. An overview of the design and conduct of the battle trials. *Chinese clinical oncology* 2015; 4(3).
10. Simon R and Simon NR. Using randomization tests to preserve type i error with response adaptive and covariate adaptive randomization. *Statistics & probability letters* 2011; 81(7): 767–772.
11. Jennison C and Turnbull BW. Group sequential methods with applications to clinical trials. Chapman and Hall/CRC, 1999.
12. Coad D. A comparative study of some data-dependent allocation rules for bernoulli data. *Journal of statistical computation and simulation* 1992; 40(3-4): 219–231.
13. Rosenberger WF, Vidyashankar A and Agarwal DK. Covariate-adjusted response-adaptive designs for binary response. *Journal of biopharmaceutical statistics* 2001; 11(4): 227–236.
14. Villar SS, Bowden J and Wason J. Response-adaptive designs for binary responses: How to offer patient benefit while being robust to time trends? *Pharmaceutical statistics* 2018; 17(2): 182–197.
15. Bashir Q, Munsell MF, Giralt S et al. Randomized phase ii trial comparing two dose levels of thymoglobulin in patients undergoing unrelated donor hematopoietic cell transplant. *Leukemia & lymphoma* 2012; 53(5): 915–919.
16. Rosenberger WF, Stallard N, Ivanova A et al. Optimal adaptive designs for binary response trials. *Biometrics* 2001; 57(3): 909–913.
17. Demets DL and Lan KG. Interim analysis: the alpha spending function approach. *Statistics in medicine* 1994; 13(13-14): 1341–1352.
18. Haviland MG. Yates’s correction for continuity and the analysis of 2×2 contingency tables. *Statistics in medicine* 1990; 9(4): 363–367.
19. Thall PF and Wathen JK. Practical bayesian adaptive randomisation in clinical trials. *European Journal of Cancer* 2007; 43(5): 859–866.
20. Gelman A and Su YS. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models, 2018. URL https://CRAN.R-project.org/package=arm. R package version 1.10-1.
21. Zelen M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association* 1969; 64(325): 131–146.
22. Mike V, Krauss AN and Ross GS. Neonatal extracorporeal membrane oxygenation (ecmo): clinical trials and the ethics of evidence. *Journal of medical ethics* 1993; 19(4): 212–218.