WORTH ADAPTING? REVISITING THE USEFULNESS OF
OUTCOME-ADAPTIVE RANDOMIZATION

J. Jack Lee¹, Nan Chen¹ and Guosheng Yin¹,²

¹Department of Biostatistics, The University of Texas MD Anderson Cancer Center; ²Department of Statistics and Actuarial Science, The University of Hong Kong

Key words: Adaptive and fixed randomization, Bayesian clinical trial design, Efficacy and futility early stopping, Type I error and statistical power, Patient population, Sample size

Funding Support: Supported in part by NCI grants CA16672, CA97007, and the Hong Kong RGC grant.

Acknowledgement: The authors thank Ms. LeeAnn Chastain for editorial assistance and the reviewers and the Senior Editor, Dr. John Crowley, for constructive comments.

Running head: Usefulness of outcome-adaptive randomization revisited

Corresponding author: J. Jack Lee, Ph.D., Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Unit 1411, Houston, Texas 77030

There is no conflict of interest to declare for all authors.
ABSTRACT - (237 words, 250 words maximum)

Outcome-adaptive randomization (AR) allocates more patients to the better treatments as the information accumulates in the trial. Is it worth to apply outcome-AR in clinical trials? Different views permeate the medical and statistical communities. We provide additional insights to the question by conducting extensive simulation studies. Trials are designed to maintain the type I error rate, achieve a specified power, and provide better treatment to patients. Generally speaking, equal randomization (ER) requires a smaller sample size and yields a smaller number of non-responders than AR by controlling type I and type II errors. Conversely, AR produces a higher overall response rate than ER with or without expanding the trial to the same maximum sample size. When there exist substantial treatment differences, AR can yield a higher overall response rate as well as a lower average sample size and a smaller number of non-responders. Similar results are found for the survival endpoint. The differences between AR and ER quickly diminish with early stopping of a trial due to efficacy or futility. In summary, ER maintains balanced allocation throughout the trial and reaches the specified statistical power with a smaller number of patients in the trial. If the trial’s result is positive, ER may lead to early approval of the treatment. AR focuses on treating patients best in the trial. AR may be preferred when the difference in efficacy between treatments is large or when limited patients are available.
INTRODUCTION

The origin of randomization in experimental design can be dated back to its application in a psychophysics experiment published in 1885 (1-4). However, randomization was not widely recognized or accepted until Fisher applied it to agricultural research starting in the 1920’s (5,6). One of the first applications of randomization to clinical trials was the streptomycin trial published in 1948 (7). Since then, randomized trials have gradually evolved to become the gold standard for comparing the relative performance of treatments.

Randomization eliminates the bias in clinical trials that arises from the subjective assignment of treatments to individual patients. Properly implemented, randomization can reduce the confounding effect of both known and unknown prognostic factors, as well as the inherent heterogeneity of an experiment. Moreover, randomization provides a solid statistical foundation for valid inference in estimation and hypothesis testing (8-10).

To provide a fair ground for comparing the effect across different treatments, commonly-used randomization methods apply blocking, stratification, or covariate-adjusted methods such as minimization to achieve balance in the baseline characteristics (10-12). Equal randomization (ER) is the most widely-used procedure. Under the equipoise principle, which states that all treatments are likely to be equally effective, subjects are randomized equally across treatments. On the other hand, response- or outcome-adaptive randomization (AR) dynamically assigns patients to treatments with a probability based on the
currently observed outcomes. The general goal is to assign more patients to better treatments. This concept can be traced back to the work of Thompson (13) with an early implementation in the form of the randomized play-the-winner design, which assigns more patients to the current winner with a higher probability (14, 15). Many similar designs have been proposed in the literature (16, 17).

Outcome-AR is conceptually appealing. At the beginning of a study, not much is known about the difference in the treatment effect; hence, equal randomization (ER) is reasonable because of clinical equipoise. However, as the trial moves along and more information about the treatment difference accumulates, it makes sense to assign more patients to the better performing arms by aligning the randomization probability with treatment efficacy. When sufficient evidence is obtained, the trial can be stopped. With outcome-AR, patients enrolled in the trial can benefit from having a higher chance of being assigned to the better treatment, if any. In contrast, traditional clinical trials with ER have a main goal of providing information for a definitive comparison between treatments. Patients participating in trials contribute to the scientific knowledge to benefit the public in general. Such trials typically are designed to maximize the statistical power. When the variances of the treatment effect measures are equal between treatments and the total sample size is fixed, ER is the optimal design. Conversely, AR can be applied to increase the overall success for patients enrolled in the trial while controlling type I and type II errors as well. Excellent
discussions on the inherent competition between AR and ER on designing clinical trials can be found in the literature (10, 18, 19).

Two recent publications have reinvigorated the debate regarding the use of outcome-AR versus fixed randomization (FR) methods in clinical trials. (20, 21) Korn and Freidlin described AR as “inferior to 1:1 randomization in terms of acquiring information for the general clinical community and offers modest-to-no benefits to the patients on the trial. (20)” They recommended the use of ER or 2:1 FR when assigning more patients to the presumably better arm would increase the study’s accrual rate. While acknowledging the added complexity of AR, Berry contended that the benefits are limited but real in two-arm trials, that these benefits can be more evident in trials with more than two arms, and that AR can shorten the time of cancer drug development and better identify responding patient populations (21). Additional letters to the editor provided further support for ER over AR (22, 23). Three recently published books provide a comprehensive presentation of the use of randomization in clinical trials (10), a rigorous theoretical assessment of outcome-AR from the frequentist point of view, (17) and theoretical and practical overviews on a wide range of Bayesian adaptive methods applied to clinical trials (24). To gain a deeper understanding of the performance of the AR and FR trial designs, we compare their operating characteristics through extensive simulation studies.

METHODS
We base all inference in the simulated trials on the posterior probability under the Bayesian framework, and control the frequentist type I and type II error rates. We consider both binary and survival endpoints. When the endpoint is binary, patients either respond or do not respond to the treatment. In two-arm studies, we denote $p_1$ as the response rate of the control arm and $p_2$ as that of the experimental arm. We conclude that the experimental arm is better than the control arm if the posterior probability $\Pr(p_2 > p_1 \mid D) > \theta_r$, where $D$ denotes the observed data and $\theta_r$ is the cutoff value. The sample size and cutoff value are chosen to control the type I error rate at 10% under the null hypothesis of $p_1 = p_2 = 0.2$, and achieve 90% power under the alternative hypothesis of $p_1 = 0.2$ and $p_2 = 0.4$. We fix $p_1 = 0.2$ and vary $p_2$ from 0.05 up to 0.95. We take Beta(1,1) as the prior distribution for the response rates of both arms.

For AR, we take the probability for randomizing patients to Arm 2 (the experimental arm) as

$$\Pr(p_2 > p_1 \mid D)^c / \{\Pr(p_2 > p_1 \mid D)^c + \Pr(p_1 > p_2 \mid D)^c\},$$

Equation (1)

where $c$ is the tuning parameter controlling the degree of imbalance (25). A value of $c=0$ corresponds to ER; $c=\infty$ corresponds to the deterministic “play-the-winner” assignment (14). Thall and Wathen (25) recommended using a value between 0 and 1 for $c$, for example, $c = n / (2N)$, where $n$ is the current number of patients in the trial and $N$ is the total sample size. In this case, $c=0$ at the beginning of the trial and $c=0.5$ at the end of the trial, such that the variability in
the randomization rate is reduced in the early stage of the trial and the statistical power is preserved. To produce a larger contrast when comparing the performance of AR versus ER, we also investigate a case with \( c = (n/N)^{0.1} \).

Consequently, the value of \( c \) is 0 at the beginning of the trial, but quickly rises to 1: \( c = 0.87, 0.93, \) and 0.97 correspond to 25%, 50%, and 75% of the patients in the trial, respectively. The AR procedure is applied from the beginning of each trial. To prevent extreme patient allocation, we also restrict the randomization probabilities to values between 0.1 and 0.9.

We compare the performance of different designs by examining their operating characteristics including the number of non-responders, the averaged sample size, and the overall response rate. To achieve a fair comparison of the overall response rate, we expand the sample size such that the total sample size is the same across all designs. In this case, a larger overall response rate corresponds to a smaller number of non-responders. For the expansion cohort, if the null hypothesis is rejected, all the remaining patients are assigned to the better arm. Otherwise, they would be assigned to the control arm.

In a three-arm randomized trial, we compare two experimental treatments (Arms 2 and 3) and one control treatment (Arm 1). Under this setup, we would reject the null hypothesis if at least one experimental treatment is superior to the control as indicated by \( \Pr(p_k > p_1 | D) > \theta_r \), where \( k = 2 \) or 3. The sample size and cutoff value \( \theta_r \) are chosen to control the 10% type I error rate under the null
hypothesis of $p_1 = p_2 = p_3 = 0.2$, and achieve 90% power under the alternative hypothesis of $p_1 = 0.2$ and $p_2 = p_3 = 0.4$. Similar to the two-arm trials, we take Beta$(1,1)$ to be the prior distribution for the response rates of all arms.

In the Bayesian AR, we compare the response rate of each treatment with the average response rate of the three arms, i.e., we compute $\Pr(p_k > \bar{p} | D)$ where $\bar{p} = (p_1 + p_2 + p_3)/3$. In order to obtain the posterior distribution of $\bar{p}$, we sample $p_k$ for each arm $k$ and compute the average of the three arms using 2000 posterior samples. The probability of assigning a patient to arm $k$ is

$$\frac{\{\Pr(p_k > \bar{p} | D)\}^c}{\sum_{i=1}^{3} \{\Pr(p_i > \bar{p} | D)\}^c},$$

Equation (2)

which reduces to Equation (1) for a two-arm trial.

Furthermore, we study the performance of each trial by incorporating early stopping for futility and efficacy. Evidence of efficacy is defined as for any treatment arm, if $\Pr(p_k > p_1 | D) > \theta_H$, the trial is stopped and the null hypothesis is rejected. Evidence of futility is defined as $\Pr(p_k > p_1 | D) < \theta_L$ for all $k$, at which point the trial is stopped and the null hypothesis is accepted. For each configuration, we carried out 500,000 simulations.

We also consider survival endpoints, for which we assume that the failure time follows an exponential distribution, $\exp(-t/\mu)$, with mean $\mu$. We take a
conjugate prior of an inverse-gamma distribution with parameters (0.01, 0.01), and thus the posterior distribution of $\mu$ also follows an inverse-gamma distribution.

We apply an AR scheme that is similar to that used in the case of a binary endpoint. The probability of randomizing a patient to Arm 2 is

$$\Pr(\mu_2 > \mu_1 \mid D)^c / \{\Pr(\mu_2 > \mu_1 \mid D)^c + \Pr(\mu_1 > \mu_2 \mid D)^c\}.$$ 

We specify a threshold for the ratio of the mean survival times between two arms, $\tau$ ($\tau$ is set at 1.2), and a threshold value, $\theta_T$. After the trial is completed (trial duration = 5 years), if

$$\Pr(\mu_1 / \mu_2 > \tau \mid D) > \theta_T \text{ or } \Pr(\mu_2 / \mu_1 > \tau \mid D) > \theta_T,$$

we reject the null hypothesis.

The accrual rate is 60 patients per year and the cutoff value of $\theta_T$ is calibrated to control the 10% type I error rate when $\mu_1 = \mu_2 = 1$ and achieve 80% power when $\mu_1 = 1$ and $\mu_2 = 1.5$. We carry out 10,000 simulations for the survival endpoints. All simulation results are given in the tables with the design parameters listed in the table footnotes.

**RESULTS**

Table 1 shows the operating characteristics in a two-arm trial with binary endpoints for the 1:1 and 1:2 FR designs and for the AR 1 with $c = n / (2N)$ and AR 2 with $c = (n / N)^{0.1}$, respectively. The sample size required to achieve 90% power and the 10% type I error rate is the smallest for ER ($N=134$) and the largest for AR 2 ($N=184$). In terms of the number of non-responders for the given
sample size, ER has the least when $p_2 = 0.05$ or 0.2 but AR 1 does the best for $p_2 \geq 0.4$. AR 2 yields the highest overall response rate for all $p_2$ with or without expanding to the same sample size compared with the other three designs. The increase in the overall response rate for AR 2 is more evident when the difference between $p_1$ and $p_2$ increases. The relative gains in the overall response rates for AR 2 over ER are 18%, 12%, 20%, and 24% for $p_2 = 0.05, 0.4, 0.6,$ and 0.8, respectively.

The 1:2 FR yields poor results when the experimental arm is worse than the control arm. In all settings including that the experimental treatment is better than the control, AR 1 performs better than the 1:2 FR. One desirable feature of AR is that the randomization ratio is determined by the observed data instead of being prefixed. When $p_2 = 0.05, 0.2, 0.4,$ and 0.6, the rates of randomizing patients to Arm 2 are 32.5%, 50%, 67.5%, 76.2% for AR 1 and 19.3%, 50%, 80.6%, 85.9% for AR 2, respectively (Supplementary Table S1). The results illustrate a “learning” feature of adaptive randomization: the larger the response rate for the experimental arm compared to that for the control, the greater the number of patients assigned to the experimental arm.

Figure 1 shows the allocation rate on Arm 2 (denoted as $r_2$), which changes over time for both ER and AR 2 when $p_1 = 0.2$ and $p_2 = 0.4$. In particular, we show individual trial results for 10 randomly selected trials (gray for AR and light blue
for ER), as well as the averages (maroon for AR and dark blue for ER) over the entire 500,000 simulations. Without early stopping, panel (A) shows that $r_2$ remains at 50% for $n=0$ to 134, then jumps to 100% in 9 trials (as the null hypothesis is rejected) and drops to 0% for one trial (as the null hypothesis is not rejected) under ER. Under AR, the zigzag pattern of the allocation rate over time illustrates the adaptive, learning nature of the design. We see an initial delay while waiting for the response outcomes for the first 8 patients, then $r_2$ rises from 50% to 90% as the trial progresses. Due to the AR’s stochastic nature, $r_2$ could dip below 50%, particularly in the early stage of the trial. However, the trend corrects itself when data accumulate. The average rate of allocation to Arm 2 reaches about 80% for $n=50$ and 85% for $n=100$.

Table 2 shows the results when early stopping rules for futility and efficacy are imposed for comparing the treatment effect. The maximum sample size under ER, AR 1, and AR 2 are 190, 208, and 274, respectively. The early efficacy and futility stopping rates are similar between ER and AR designs (Supplementary Table S2). Due to early stopping, the actual sample sizes are typically smaller than those originally planned. When $p_1 = 0.2$ and $p_2 = 0.05, 0.2, 0.4,$ and 0.6, the average sample sizes for the AR 2 design ($N=134.7, 237.5, 110.0,$ and 39.2 in these four respective settings) are considerably larger than those for the ER design ($N=85.5, 162.8, 84.0,$ and 35.5). In contrast, the average sample size for the AR 1 design is similar to that of the ER design. The corresponding rates of randomization to Arm 2 are (43%, 50%, 57.1%, 52.8%) and (22.9%, 50%, 74.5%, 70.5%) for AR 1 and AR 2, respectively (Supplementary Table S2).
For the number of non-responders, ER produces the smallest for $p_2 = 0.05, 0.2, \text{ and } 0.4$ but AR 2 does the best for $p_2 \geq 0.6$. The overall response rate for the AR 2 design is higher than that of the AR 1 design, which is higher than the ER design in all settings. The differences among the designs, however, are smaller in settings with early stopping than those without early stopping. When there is a large difference in the response rates between treatments, the trial may be stopped very early, even before the advantages of using AR had been seen. In this case, the role of AR is substantially mitigated by early stopping. When the sample size is expanded to 274, the relative gains in the overall response rate for AR over ER are reduced to less than 5% in all settings. Note that in an extreme setting of $p_1 = 0.2$ and $p_2 = 0.95$, AR 2 has a smaller average sample size and a smaller number of non-responders compared to ER. Figure 1B is similar to Figure 1A but includes early stopping rules. With early stopping, the average $r_2$ for the AR design is consistently higher than that of the ER design across the entire accrual period.

We also compare the results of ER, AR 1, and AR 2 in a three-arm clinical trial that incorporates early stopping rules for both futility and efficacy (Table 3). The ER design requires the enrollment of up to 231 patients, the AR 1 and AR 2 designs require a maximum sample size of up to 255 and 321 patients, respectively. As before, we set $p_1 = 0.2$ in all configurations. When the experimental treatments are worse than the control (the first row), the futility early
stopping probabilities are 0.91, 0.95, and 0.99 for ER, AR 1, and AR 2, respectively. When at least one experimental arm is better than the control, the efficacy early stopping rule kicks in. The efficacy stopping rates are 0.76, 0.82, and 0.91 for ER, AR 1, and AR 2 when $p_1 = p_2 = 0.2$, and $p_3 = 0.4$. The early stopping probabilities are comparable for the three designs in all other scenarios (Supplementary Table S3).

A desirable design should have the smallest average sample size and the smallest number of non-responders. These two features generally go together. Among the three designs, ER is the best in 6 of the 11 scenarios in Table 3. The exceptions are: (1) When both experimental arms are worse ($p_2 = p_3 = 0.05$), AR 1 is the best, and (2) When a large difference is seen across treatments (in four scenarios: $p_2 = 0.4, p_3 = 0.8$; $p_2 = 0.1, p_3 = 0.6$; $p_2 = 0.2, p_3 = 0.6$; and $p_2 = 0.2, p_3 = 0.8$), AR 2 is the best. Similar to the two-sample scenarios, in all the alternative cases, the overall response rate under AR 2 is always higher than those under AR 1 and ER, with or without expansion to the maximum sample size. When the sample size is expanded to 321, the relative gains in the overall response rate for AR 2 over ER are reduced to less than 5% in all settings with early stopping.

In Table 4 we present the simulation results with the survival endpoint. The true median survival time of patients assigned to the control arm is fixed at 0.69 year. For the experimental arm, it varies from 0.35 to 2.08 years corresponding to a
hazard ratio of 0.5 to 3. To achieve the same 10% type I error rate and 80% power, ER requires 170 patients; whereas the AR 1 and AR 2 schemes require 180 and 218 patients, respectively. We compare the performance of the three designs using the average sample size and average median survival time of the patients in the trial. The average sample size is the smallest for ER, followed by AR 1, and AR 2 is the largest. Conversely, when comparing the median survival time, AR 2 outperforms AR 1 followed by ER, with or without expanding to N=218. For hazard ratios of 0.5 and 3, the gains for AR 2 in the median survival time are 15% and 20%, respectively, over ER without expanding the sample size. When the sample size is expanded to 218 for ER, the advantage of AR 2 remains but the relative gain in the median survival time is reduced 7% or less. Similar as before, for AR, more patients are randomized to the better arm in all cases. For hazard ratios of 0.5, 1.5, and 3, the percentage of patients being randomized to the better treatment arm are 75%, 70%, and 72%, respectively. Both the efficacy and futility stopping rates are higher in AR 2 compared to those in ER (Supplementary Table S4).

DISCUSSION

Despite its critical role in designing experiments and clinical trials, the principle of randomization was not well received initially. It was not until decades after its early use that the need for and the value of randomization became widely accepted in clinical trials. The modern debate is not focused on whether to use randomization, but rather on how to do it and which type of randomization
scheme is the most appropriate. One must examine the performance of ER and AR in their totality and determine their relative strengths and weaknesses.

Our extensive simulation studies show that, for a binary endpoint, ER typically results in a smaller sample size and a smaller number of non-responders than AR to control the type I and type II errors. ER has a smaller average sample size than AR when there is no difference in the response rates. AR consistently achieves a higher overall response rate by allocating more patients to more effective treatments during the course of the trial. By expanding the sample size to the same number across all designs, AR yields a higher overall response rate (a lower number of non-responders as well) than ER. In particular, when the experimental treatment is unexpectedly worse than the control, or when the experimental treatment is overwhelmingly better than the control, AR may reach a smaller average sample size, a smaller number of non-responders, and a higher overall response rate than ER at the same time. Similar conclusions hold in three-arm trials. Note that the extremely large efficacy differences are infrequently observed in clinical trials. However, such large differences could happen in certain settings with matching treatments and biomarkers for targeted therapies.

In practice, because we do not know the relative efficacy between treatment arms (Were we to know, we would not need to conduct the trial!), it is sensible to allow the randomization ratio to depend on the observed data rather than having
it pre-set throughout the trial. AR is an adaptive learning process. It does not need to speculate a priori as to which treatment arm is better: During the course of a trial, AR adapts the randomization probabilities automatically and continuously based on the observed data.

There are pros and cons to all designs. ER designs emphasize on maximizing the statistical power and is favored from a global, population-based view. On the other hand, AR designs stress more on individual benefit by assigning more patients to the putative better treatments during the trial based on the available data. The imbalance in patient allocation between treatments causes a loss of statistical power and requires an increased sample size to achieve the same target power. The allocation ratio can be changed by varying the tuning parameter to be more or less imbalanced. ER designs have the advantage of reaching a conclusion earlier. Hence, if the trial is positive, the result can be announced sooner or the drug can be approved earlier to benefit future patients in the population. ER designs also have a smaller sample size under the null hypothesis. The potential benefit can be large if the population size is bigger than the trial size. On the other hand, in rare diseases (such as certain pediatric cancers), there is only a limited population. After the trial result is known, future patients arrive at the same rate as before and will receive the better treatment. This setting can be mimicked by expanding the trials of the ER design to the same sample size as the AR design. With the expanded sample size, the AR design always yields higher overall success compared to the ER design. The
benefit of AR is more prominent when one treatment is substantially better than the others. AR focuses on how to best treat patients in the trial while ER emphasizes on making the right decision early in comparing the treatment effect. Hence, when there is a treatment difference, ER is preferred if the population size is much bigger than the trial size and AR is preferred otherwise. When there is no treatment difference, ER is preferred over AR because the required sample size to reach a conclusion is smaller. More detailed comparison of the relative merit of ER and AR with respect to the population size and trial size is given in the Appendix. Our results also show that the difference between ER and AR designs quickly diminishes when early stopping rules for efficacy and futility are implemented.

AR takes the “learn as we go” approach to adjusting the randomization ratio based on the observed data. AR assigns more patients to better treatments and, as a result, these treatments can be better studied with larger sample sizes. Yet, statistical power suffers from imbalanced samples. AR 1 is commonly used in practice while AR 2 serves as an example to illustrate more extreme imbalance. How much imbalance one would like to reach depends on the specific objectives. The “optimal” adaptive randomization ratio can be derived based upon the choice of optimization criteria or the utility function. For example, we may maximize the efficacy in the patient horizon (i.e., the total patient population available in the whole society) or minimize the loss function using Bayesian decision theoretic approaches (26,27).
From the trial conduct point of view, trials using outcome-AR require more effort in planning and implementation. A robust infrastructure should be set up to ensure that AR can be carried out properly throughout the trial. The outcome-AR is more applicable to trials with short-term endpoints. To ensure that AR works as it is supposed to, the primary endpoint needs to be recorded accurately and timely. For example, an integrated web-based database system for patient registration, eligibility checking, randomization, and follow up can be developed to facilitate the conduct of the AR trials. A scheduling module and an e-mail notification module can be included to ensure that the primary endpoints are collected and reported timely. To enhance the accuracy of the endpoint determination, clear criteria should be established and consistently followed. Endpoint review should be conducted blinded to the treatment assignment. One caveat is that AR is more prone to the danger of population drift. When the patient characteristics change over time, AR is more likely to result in a biased estimate of the treatment difference than ER (28). One solution is to lay out well-defined eligibility criteria such that a homogeneous population of patients can be enrolled throughout the trial. Another approach is to set a minimal portion of the patients to be assigned to the control arm to ensure a fair comparison. Block outcome-AR can be considered to reduce the bias caused by the population drift (29, 30). Covariate adjusted regression methods may also be applied to reduce the bias. Another limitation is that the current discussions only focus on the efficacy outcome of the treatment. In the real trial setting, treatment toxicities
should be monitored concurrently. The decision of treatment allocation should consider both efficacy and toxicity outcomes.

In summary, outcome-adaptive randomization is an active research area in both medicine and statistics (31-35). It has been implemented successfully in two recent trials, namely, BATTLE and I-SPY2 (36-37). Instead of always applying equal randomization in clinical trials, we advocate challenging the status quo by considering adaptive randomization. The final verdict of the relative advantages and disadvantages for various designs should ultimately be based upon results from real trials and the benefit of the entire population.
REFERENCES


35. Lei X, Yuan Y, Yin G. Bayesian phase II adaptive randomization by jointly modeling time-to-event efficacy and binary toxicity. Lifetime Data Anal 2011;17:156–74.


Appendix: Comparison of Equal and Adaptive Randomization by the Number of Non-responders and the Equivalence Ratio of the Population size versus Trial size

To provide further comparison between equal randomization (ER) and adaptive randomization (AR), we plot the additional number of non-responders versus the number of patients accrued for two-arm trials with binary endpoints. The computations are based on the average of 500,000 trials. The number of additional non-responders is defined as the excess number of non-responders for the respective designs compared to the case that all patients had received the better treatment.

Supplementary Figure 1A shows the additional number of non-responders over time for the ER and AR 2 designs compared to the case that all patients had received the better treatment. The blue solid line shows that the ER design stopped at N=134 with 13.4 more non-responders than the theoretically best case where all patients are assigned to Arm 2. The blue dashed line shows the additional number of non-responders in the expansion cohort after ER while the black line represented the straight ER throughout. The red line indicates the “excess” number of non-responders for AR 2 compared to the theoretically best case scenario. By 184 patients, the “excess” numbers of non-responders for straight ER, ER + expansion, and AR 2 are 18.4, 14.4, and 7.1, respectively. Throughout the trial, AR 2 is better than ER by yielding a smaller number of non-responders when a total of 184 patients are treated.
Similarly, Figure 1B plots the “excess” number of non-responders over the best case scenario when the designs implement early stopping rules. The horizontal location of the green dots indicate the average sample size while the blue dot and the red dot show the maximal sample size for ER and AR 2, respectively. When expanding to 274 patients, the “excess” number of non-responders is 5.9 for the AR 2 design which is smaller than 10.5 of the ER design.

The construct of the expansion cohort shown above resembles the rare disease setting, in which the total disease population is small and all patients participate the trial. At the conclusion of the trial, the information learned from the trial is applied to treat future patients with the best treatment. Future patients arrive at the same rate as the enrollment rate in the trial. We consider the cases with different population sizes with respect to the trial size. The performance of ER and AR is compared by computing the number of non-responders in the entire patient population with similar conditions and could be affected by the similar treatments – both current patients enrolled in the trial as well as future patients in the population outside the trial. Taking the example of comparing two treatments with the response rates $p_1 = 0.2$ and $p_2 = 0.4$ with early stopping, to achieve 90% power with a 10% type I error rate, the ER design needs an average of 84 patients with 59.4 non-responders. In contrast, the AR 2 design (with the tuning parameter $c = (n / N)^{0.1}$) requires 110 patients with 71.5 non-responders. One major difference between ER and AR 2 is that ER reaches the conclusion earlier by $110 - 84 = 26$ patients. Supposing that there are $x$ number of patients available
outside the trial between the time of the end of ER and the end of AR, we compute the expected number of non-responders as follows.

(1) Using the ER design, the trial ends at 84 patients. All subsequent \( x \) patients are treated with the better treatment with a probability of 0.9. The total expected number of non-responders including patients treated during and after the trial is
\[
59.4 + (0.9 \times 0.6 + 0.1 \times 0.8) x.
\]

(2) Using the AR 2 design, the trial ends at 110 patients which is 26 patients more than using the ER design. During this period of time, all \( x \) number of patients available outside the trial are treated with the control arm. The total expected number of non-responders including patients treated inside and outside the trial is
\[
71.5 + 0.8(x - 26).
\]

We can solve \( x = 48.3 \) by equating the above two equations. Taking the ratio of 48.3 and 26, we get 1.86. The ratio can be considered as the “equivalence ratio” of ER and AR in terms of yielding the same number of non-responders. The result suggests that if the outside trial patients are available to be treated at the rate of 1.86 or higher than the rate of patients enrolled in the trial, ER is better. Otherwise, AR 2 is better. Similar calculations can be applied to other settings. For example, the equivalence ratios for the settings in Table 1 with two arms without early stopping are also about 1.8 to 1.9. The equivalence ratios for Table 2 with two arms and early stopping are 2.7 for \( p_1 = 0.2 \) and \( p_2 = 0.6 \) and 18.8 for \( p_1 = 0.2 \) and \( p_2 = 0.8 \). For the three-arm trials shown in Table 3, the equivalence ratios are between 1.5 to 4.6 except for the four cases mentioned in the text with larger differences between the experimental arms and the control arm. In those
cases, AR 2 yields a smaller sample size and has a smaller number of non-responders. Generally speaking, ER is preferred when the patient population outside the trial is large (e.g., more than twice of the trial size) because the trial result can be reported earlier to benefit the entire population. On the other hand, AR is preferred if there are not many patients available outside the trial such as in the rare disease setting. AR has the benefit of minimizing the number of non-responders in the trial by assigning more patients to better treatments in the entire course of the trial.

Assume that the sample size and the number of non-responders for the ER design are \((n_1, m_1)\) and those for the AR design are \((n_2, m_2)\), respectively. Also assume that the power for the test is \(w\). It can be shown that the solution of \(x\) for the two-arm trial is \(\frac{((1-p_1)(n_2-n_1)-(m_2-m_1))}{w(p_2-p_1)}\). The equivalence ratio is \(\frac{x}{n_2-n_1}\), which depends on the true response rates, differences of the trial sample size and the number of non-responders between the two arms, as well as statistical power. The equivalence ratios can be calculated for different AR methods with different tuning parameters which determine the degree of imbalance. For example, the equivalence ratio changes to 3.63 if we compare AR 1 and ER when \(p_1 = 0.2\) and \(p_2 = 0.4\) with early stopping. Note that the above calculation is focused only on comparing the mean number of non-responders. The variation of the number of non-responders tends to be larger in AR than in ER.
Figure Legends

**Figure 1**: Randomization probability to the experimental treatment versus the number of patients accrued for two-arm trials with binary endpoints. The result of adaptive randomization (AR) and equal randomization (ER) are compared. For AR, the AR 2 design with the tuning parameter $c = (n / N)^{0.1}$ was applied. Performances of 10 randomly selected trials are shown in gray lines for AR and in light blue lines for ER. The averages of 500,000 trials are also shown in the maroon line for AR and in the dark blue dashed line for ER. Response rates are $p_1 = 0.2$ and $p_2 = 0.4$. **Panel (A)**: Without early stopping. The sample size of the AR 2 design sample size is 184; that of the ER design is 134, which is expanded to 184 after the trial is completed. **Panel (B)**: With early stopping. The maximum sample size for the AR 2 design is 274 and for the ER design is 190. If a trial is stopped early (or completed for ER), additional patients are added to reach a total of 274 patients. Additional patients are allocated to the better treatment if the null hypothesis is rejected, or to the control arm if the null hypothesis is not rejected.
Table 1: Performance of fixed-ratio (1:1 and 1:2) and adaptive randomization trial designs without early stopping.

<table>
<thead>
<tr>
<th>True Response Rate</th>
<th>Fixed Ratio 1:1 Randomization</th>
<th>Fixed Ratio 1:2 Randomization</th>
<th>Adaptive Randomization 1</th>
<th>Adaptive Randomization 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N=134</td>
<td>Expd. to N=140</td>
<td>Expd. to N=184</td>
<td>N=153</td>
</tr>
<tr>
<td></td>
<td>Overall Resp %</td>
<td>Overall Resp %</td>
<td>Overall Resp %</td>
<td>Overall Resp %</td>
</tr>
<tr>
<td>Cntl</td>
<td>Exp (p₁)</td>
<td>Exp (p₂)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.05</td>
<td>117.2</td>
<td>12.5</td>
<td>12.8</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>107.2</td>
<td>20.0</td>
<td>20.0</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>93.8</td>
<td>30.0</td>
<td>30.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.6</td>
<td>80.4</td>
<td>40.0</td>
<td>40.9</td>
</tr>
<tr>
<td>0.2</td>
<td>0.8</td>
<td>67.0</td>
<td>50.0</td>
<td>51.3</td>
</tr>
<tr>
<td>0.2</td>
<td>0.95</td>
<td>57.0</td>
<td>57.5</td>
<td>59.1</td>
</tr>
</tbody>
</table>

Cntl: Control Arm; Exp: Experimental Arm; ER: Equal Randomization; AR: Adaptive Randomization; Expd.: Expanded; Resp: Response.

Cutoff values for declaring significant results are chosen as \( \theta_T = 0.892 \) for 1:2 randomization, \( \theta_T = 0.9 \) for 1:1 randomization, \( \theta_T = 0.900 \) for AR 1 and \( \theta_T = 0.905 \) for AR 2 to yield 10% type I error rate at \( p_1 = p_2 = 0.2 \) and 90% power at \( p_1 = 0.2, p_2 = 0.4 \) (Shaded cells).
Table 2: Performance of equal and adaptive randomization designs with both futility and efficacy early stopping.

<table>
<thead>
<tr>
<th>True Response Rate</th>
<th>Equal Randomization</th>
<th>Adaptive Randomization 1 (c = n / (2N))</th>
<th>Adaptive Randomization 2 (c = (n / N)^{\gamma})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cntl ((p_1)) Exp ((p_2))</td>
<td>N(_{\text{max}})=190</td>
<td>Expd. to N=208</td>
<td>Expd. to N=274</td>
</tr>
<tr>
<td>0.2 0.05</td>
<td>74.8 85.5</td>
<td>12.5 16.9</td>
<td>17.7</td>
</tr>
<tr>
<td>0.2 0.2</td>
<td>130.5 162.8</td>
<td>20.0 20.0</td>
<td>20.0</td>
</tr>
<tr>
<td>0.2 0.4</td>
<td>59.4 84.0</td>
<td>30.0 35.6</td>
<td>36.2</td>
</tr>
<tr>
<td>0.2 0.6</td>
<td>21.4 35.5</td>
<td>40.0 56.5</td>
<td>57.4</td>
</tr>
<tr>
<td>0.2 0.8</td>
<td>11.3 22.5</td>
<td>50.0 76.7</td>
<td>77.5</td>
</tr>
<tr>
<td>0.2 0.95</td>
<td>7.7 18.2</td>
<td>57.5 91.7</td>
<td>92.5</td>
</tr>
</tbody>
</table>

Cntl: Control Arm; Exp: Experimental Arm; ER: Equal Randomization; AR: Adaptive Randomization; Expd.: Expanded; Resp: Response; Avg: Average; N\(_{\text{max}}\): Maximum sample size.

Cutoff value for futility stopping is \(\theta_L=0.02\). Cutoff value for efficacy stopping and final decision is \(\theta_H=\theta_T=0.983\) for AR 1, \(\theta_H=\theta_T=0.98\) for AR 2, and 0.9835 for ER to yield 10% type I error rate at \(p_1=p_2=0.2\) and 90% power at \(p_1=0.2, p_2=0.4\) (Shaded cells).
Table 3: Equal randomization design for three treatment arms, with both futility and efficacy early stopping.

<table>
<thead>
<tr>
<th>True Response Rate</th>
<th>Cntl (p₁)</th>
<th>Exp 1 (p₂)</th>
<th>Exp 2 (p₃)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cutoff value for futility stopping is θₐ=0.06. Cutoff value for efficacy stopping and final decision is θₐ=θᵢ=0.9904 for ER, θₐ=θᵢ=0.9904 for AR 1, θₐ=θᵢ=0.988 for AR 2 to yield 10% type I error rate at p₁=0.2 and 90% power at pᵢ=0.2, pᵢ=0.4 (Shaded cells).</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 4: Performance of equal and adaptive randomization designs for survival analysis with both futility and efficacy early stopping.

<table>
<thead>
<tr>
<th>True Median Survival Time</th>
<th>Parameters of Exponential Distribution</th>
<th>Equal Randomization</th>
<th>Adaptive Randomization 1 ( c = n / (2N) )</th>
<th>Adaptive Randomization 2 ( c = (n / N)^{\alpha} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cntl Exp (( \mu_1 ))</td>
<td>Exp (( \mu_2 ))</td>
<td>( N_{\text{max}}=170 ) Expd. to N=218</td>
<td>( N_{\text{max}}=180 ) Expd. to N=218</td>
<td>( N_{\text{max}}=218 ) Expd. to N=218</td>
</tr>
<tr>
<td>Cntl Exp (( \mu_1 ))</td>
<td>Exp (( \mu_2 ))</td>
<td>Avg. Sample Size</td>
<td>Median Survival Time</td>
<td>Median Survival Time</td>
</tr>
<tr>
<td>0.69 0.35 1.00 0.50</td>
<td>106.0 0.54 0.61</td>
<td>108.1 0.56 0.63</td>
<td>127.2 0.62 0.65</td>
<td></td>
</tr>
<tr>
<td>0.69 0.52 1.00 0.75</td>
<td>157.7 0.62 0.64</td>
<td>165.4 0.63 0.64</td>
<td>195.0 0.64 0.65</td>
<td></td>
</tr>
<tr>
<td>0.69 0.69 1.00 1.00</td>
<td>157.7 0.62 0.64</td>
<td>165.4 0.63 0.64</td>
<td>195.0 0.64 0.65</td>
<td></td>
</tr>
<tr>
<td>0.69 0.87 1.00 1.25</td>
<td>162.0 0.79 0.79</td>
<td>169.2 0.80 0.79</td>
<td>198.4 0.81 0.81</td>
<td></td>
</tr>
<tr>
<td>0.69 1.04 1.00 1.50</td>
<td>152.7 0.88 0.91</td>
<td>160.3 0.91 0.93</td>
<td>189.6 0.95 0.96</td>
<td></td>
</tr>
<tr>
<td>0.69 1.21 1.00 1.75</td>
<td>138.4 0.98 1.06</td>
<td>143.4 1.02 1.08</td>
<td>166.7 1.09 1.11</td>
<td></td>
</tr>
<tr>
<td>0.69 1.39 1.00 2.00</td>
<td>123.2 1.07 1.20</td>
<td>127.0 1.13 1.23</td>
<td>142.9 1.22 1.28</td>
<td></td>
</tr>
<tr>
<td>0.69 2.08 1.00 3.00</td>
<td>90.6 1.45 1.81</td>
<td>91.4 1.54 1.85</td>
<td>97.7 1.74 1.93</td>
<td></td>
</tr>
</tbody>
</table>

Cntl: Control Arm; Exp: Experimental Arm; Expd.: ER: Equal Randomization; AR: Adaptive Randomization; Expanded; \( N_{\text{max}} \): Maximum sample size. The cutoff values for declaring significant results are \( \theta_T = 0.721, 0.721, \) and 0.742 for ER, AR 1 and AR 2, respectively. Early stopping cutoff values for claiming efficacy stopping and futility stopping are 0.99 and 0.2, respectively, to yield 10% type I error rate at \( \mu_1 = \mu_2 = 1.00 \) and 80% power at \( \mu_1 = 1.00, \mu_2 = 1.50 \) (Shaded cells).
Clinical Cancer Research

WORTH ADAPTING? REVISITING THE USEFULNESS OF OUTCOME-ADAPTIVE RANDOMIZATION

J. Jack Lee, Nan Chen and Guosheng Yin

Clin Cancer Res Published OnlineFirst July 2, 2012.

Updated version
Access the most recent version of this article at:
doi:10.1158/1078-0432.CCR-11-2555

Supplementary Material
Access the most recent supplemental material at:
http://clincancerres.aacrjournals.org/content/suppl/2012/07/02/1078-0432.CCR-11-2555.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.