Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

Clinical Cancer Research

Worth Adapting? Revisiting the Usefulness of Outcome-Adaptive Randomization

J. Jack Lee, Nan Chen and Guosheng Yin

Clin Cancer Res 2012;18:4498-4507. Published OnlineFirst July 2, 2012.

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

Worth Adapting? Revisiting the Usefulness of Outcome-Adaptive Randomization

J. Jack Lee[1], Nan Chen[1], and Guosheng Yin[1,2]

Abstract

Outcome-adaptive randomization allocates more patients to the better treatments as the information accumulates in the trial. Is it worth it to apply outcome-adaptive randomization 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 requires a smaller sample size and yields a smaller number of nonresponders than adaptive randomization by controlling type I and type II errors. Conversely, adaptive randomization produces a higher overall response rate than equal randomization with or without expanding the trial to the same maximum sample size. When there are substantial treatment differences, adaptive randomization can yield a higher overall response rate as well as a lower average sample size and a smaller number of nonresponders. Similar results are found for the survival endpoint. The differences between adaptive randomization and equal randomization quickly diminish with early stopping of a trial due to efficacy or futility. In summary, equal randomization 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 results are positive, equal randomization may lead to early approval of the treatment. Adaptive randomization focuses on treating patients best in the trial. Adaptive randomization may be preferred when the difference in efficacy between treatments is large or when the number of patients available is limited. *Clin Cancer Res; 18(17); 4498–507. ©2012 AACR.*

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 1920s (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 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 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-adaptive randomization is conceptually appealing. At the beginning of a study, not much is known about the difference in the treatment effect; hence, equal randomization is reasonable because of clinical equipoise.

Authors' Affiliations: [1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas; and [2]Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

J.J. Lee is the senior author of the manuscript.

Corresponding Author: J. Jack Lee, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Unit 1411, Houston, Texas 77030. Phone: 713-794-4158; Fax: 713-563-4243; E-mail: jjlee@mdanderson.org

doi: 10.1158/1078-0432.CCR-11-2555

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

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-adaptive randomization, 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 equal randomization 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, equal randomization is the optimal design. Conversely, adaptive randomization 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 adaptive randomization and equal randomization on designing clinical trials can be found in the literature (10, 18, 19).

Two recent publications have reinvigorated the debate about the use of outcome-adaptive randomization versus fixed randomization methods in clinical trials (20, 21). Korn and Freidlin describe adaptive randomization 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 recommend the use of equal randomization or 2:1 fixed randomization when assigning more patients to the presumably better arm would increase the study's accrual rate. While acknowledging the added complexity of adaptive randomization, Berry contends that the benefits are limited but real in 2-arm trials, that these benefits can be more evident in trials with more than 2 arms, and that adaptive randomization can shorten the time of cancer drug development and better identify responding patient populations (21). Additional letters to the editor provide further support for equal randomization over adaptive randomization (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-adaptive randomization 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 adaptive randomization and fixed randomization 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 2-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 $\Pr(p_2 > p_1|D) > \theta_T$, where D denotes the observed data and θ_T 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 adaptive randomization, we take the probability for randomizing patients to arm 2 (the experimental arm) as

$$\Pr(p_2 > p_1|D)^c / \{\Pr(p_2 > p_1|D)^c + \Pr(p_1 > p_2|D)^c\}, \quad \text{(A)}$$

where c is the tuning parameter controlling the degree of imbalance (25). A value of $c = 0$ corresponds to equal randomization; $c = \infty$ corresponds to the deterministic "play-the-winner" assignment (14). Thall and Wathen (25) recommend 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 adaptive randomization versus equal randomization, 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 increases to 1: $c = 0.87$, 0.93, and 0.97 correspond to 25%, 50%, and 75% of the patients in the trial, respectively. The adaptive randomization 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 nonresponders, 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 nonresponders. 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 3-arm randomized trial, we compare 2 experimental treatments (arms 2 and 3) and 1 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 $\Pr(p_k > p_1|D) > \theta_T$, where $k = 2$ or 3. The sample size and cutoff value θ_T 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

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

2-arm trials, we take Beta (1,1) to be the prior distribution for the response rates of all arms.

In the Bayesian adaptive randomization, we compare the response rate of each treatment with the average response rate of the 3 arms, that is, we compute $\Pr(p_k > \bar{p}|D)$ where $\bar{p} = (p_1 + p_2 + p_3)/3$. To obtain the posterior distribution of $\bar{p}$, we sample p_k for each arm k and compute the average of the 3 arms using 2,000 posterior samples. The probability of assigning a patient to arm k is

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

which reduces to equation A for a 2-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, $\mathrm{Exp}(-t/\mu)$, with mean μ. We take a conjugate prior of an inverse-gamma distribution with parameters (0.01, 0.01), and thus the posterior distribution of μ also follows an inverse-gamma distribution.

We apply an adaptive randomization scheme that is similar to that used in the case of a binary endpoint. The probability of randomizing a patient to arm 2 is $\Pr(\mu_2 > \mu_1|D)^c/\{\Pr(\mu_2 > \mu_1|D)^c + \Pr(\mu_1 > \mu_2|D)^c\}$. We specify a threshold for the ratio of the mean survival times between 2 arms, τ (τ is set at 1.2), and a threshold value, θ_T. After the trial is completed (trial duration = 5 years), if $\Pr(\mu_1/\mu_2 > \tau|D) > \theta_T$ or $\Pr(\mu_2/\mu_1 > \tau|D) > \theta_T$, we reject the null hypothesis. The accrual rate is 60 patients per year, and the cutoff value of θ_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 2-arm trial with binary endpoints for the 1:1 and 1:2 fixed randomization designs and for the AR1 with $c = n/(2N)$ and AR2 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 equal randomization ($N = 134$) and the largest for adaptive randomization 2 ($N = 184$). In terms of the number of nonresponders for the given sample size, equal randomization has the least when $p_2 = 0.05$ or 0.2, but AR1 does the best for $p_2 \geq 0.4$. AR2 yields the highest overall response rate for all p_2 with or without expanding to the same sample size compared with the other 3 designs. The increase in the overall response rate for AR2 is more evident when the difference between p_1 and p_2 increases. The relative gains in the overall response rates for AR2 over equal randomization are 18%, 12%, 20%, and 24% for $p_2 =$ 0.05, 0.4, 0.6, and 0.8, respectively.

The 1:2 fixed randomization yields poor results when the experimental arm is worse than the control arm. In all settings, including the setting in which the experimental treatment is better than the control, AR1 performs better than the 1:2 fixed randomization. One desirable feature of adaptive randomization 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 AR1 and 19.3%, 50%, 80.6%, 85.9% for AR2, respectively

Table 1. Performance of fixed ratio (1:1 and 1:2) and adaptive randomization trial designs without early stopping

		Fixed ratio (1:1 randomization)				Fixed ratio (1:2 randomization)			AR1 $c = n/(2N)$			AR $c = (n/N)^{0.1}$	
True response rate		$N = 134$		Expd to $N = 140$	Expd to $N = 184$	$N = 153$		Expd to $N = 184$	$N = 140$		Expd to $N = 184$	$N = 184$	
Cntl (p_1)	Exp (p_2)	Nonresp, n	Overall resp %	Overall resp %	Overall resp %	Nonresp, n	Overall resp %	Overall resp %	Nonresp, n	Overall resp %	Overall resp %	Nonresp, n	Overall resp %
0.2	0.05	117.2	12.5	12.8	14.5	137.7	10.0	11.7	118.8	15.1	16.3	152.5	17.1
0.2	0.2	107.2	20.0	20.0	20.0	122.4	20.0	20.0	112.0	20.0	20.0	147.2	20.0
0.2	0.4	93.8	30.0	30.3	32.2	102.1	33.3	34.1	93.1	33.5	34.6	117.5	36.1
0.2	0.6	80.4	40.0	40.9	45.4	81.5	46.7	48.9	69.3	50.5	52.7	84.0	54.4
0.2	0.8	67.0	50.0	51.3	58.2	61.2	60.0	63.3	44.8	68.0	70.9	51.1	72.2
0.2	0.95	57.0	57.5	59.1	67.7	45.9	70.0	74.2	26.1	81.4	84.6	26.5	85.6

NOTE: 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, $\theta_T =$ 0.9 for AR1, and $\theta_T = 0.905$ for AR2 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).
Abbreviations: Cntl, control arm; Exp, experimental arm; Expd, expanded; nonresp, nonresponse; resp, response.

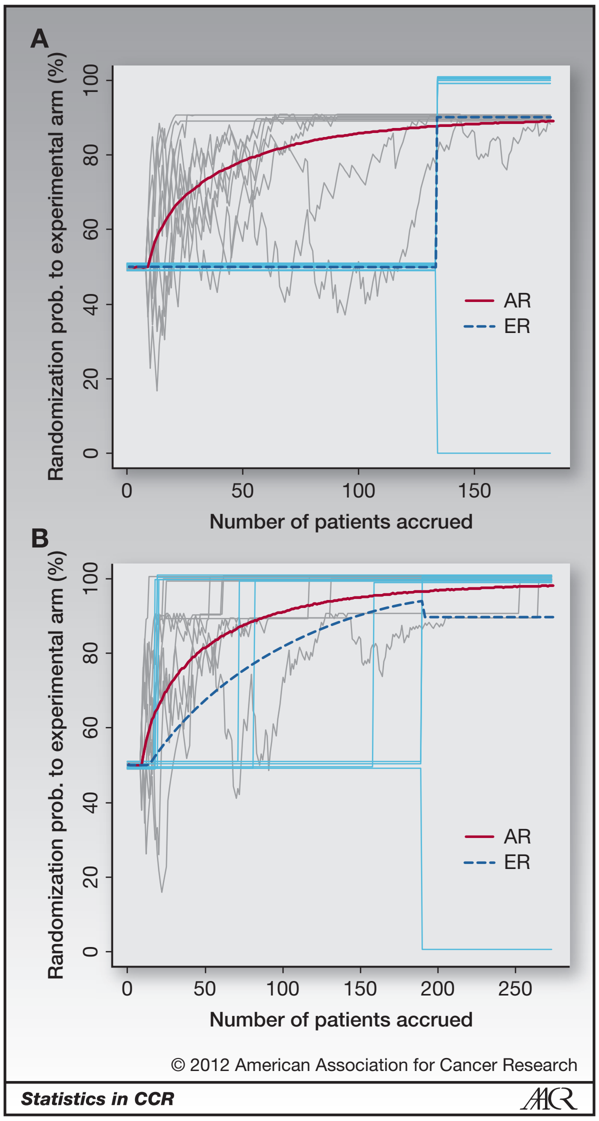
Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

(Supplementary Table S1). The results illustrate a "learning" feature of adaptive randomization: the larger the response rate for the experimental arm compared with 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 equal randomization and AR2 when $p_1 = 0.2$ and $p_2 = 0.4$. In particular, we show individual trial results for 10 randomly selected trials (gray for adaptive randomization and light blue for equal randomization), as well as the averages (maroon for adaptive randomization and dark blue for equal randomization) over the entire 500,000 simulations. Without early stopping, Fig. 1A shows that r_2 remains at 50% for $n = 0$ to 134, then increases to 100% in 9 trials (as the null hypothesis is rejected) and decreases to 0% for one trial (as the null hypothesis is not rejected) under equal randomization. Under adaptive randomization, 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 increases from 50% to 90% as the trial progresses. Because of the stochastic nature of adaptive randomization, 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 equal randomization, AR1, and AR2 are 190, 208, and 274, respectively. The early efficacy and futility stopping rates are similar between equal randomization and adaptive randomization designs (Supplementary Table S2). Because of 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 AR2 design (N = 134.7, 237.5, 110.0, and 39.2 in these 4 respective settings) are considerably larger than those for the equal randomization design ($N = 85.5$, 162.8, 84.0, and 35.5). In contrast, the average sample size for the AR1 design is similar than that of the equal randomization 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 AR1 and AR2, respectively (Supplementary Table S2).

For the number of nonresponders, equal randomization produces the smallest for p_2 = 0.05, 0.2, and 0.4 but AR2 does the best for $p_2 \geq 0.6$. The overall response rate for the AR2 design is higher than that of the AR1 design, which is higher than the equal randomization design in all settings. The differences among the designs, however, are smaller in settings with early stopping than in 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 adaptive randomization could be seen. In this case, the role of adaptive randomization is substantially mitigated by early stopping. When the sample size is expanded to 274, the relative gains in the overall response rate for adaptive randomization over equal randomization are reduced to less than 5% in all settings. Note that in an extreme setting of $p_1 = 0.2$ and $p_2 = 0.95$, AR2 has a smaller average sample size and a smaller number of nonresponders than equal randomization. Figure 1B is



Statistics in CCR

Figure 1. Randomization probability to the experimental treatment versus the number of patients accrued for 2-arm trials with binary endpoints. The results of adaptive randomization (AR) and equal randomization (ER) are compared. For AR, the AR2 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 solid blue lines for ER. The averages of 500,000 trials are also shown in the maroon line for AR and in the blue dashed line for ER. Response rates are $p_1 = 0.2$ and $p_2 = 0.4$. A, without early stopping. The sample size of the AR2 design sample size is 184 and that of the ER design is 134, which is expanded to 184 after the trial is completed. B, with early stopping. The maximum sample size for the AR2 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. prob., probability.

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

Table 2. Performance of equal and adaptive randomization designs with both futility and efficacy early stopping

True response rate		Equal randomization					AR1 $c = n/(2N)$					AR2 c $= (n/N)^{0.1}$			
		$N_{max} = 190$			Expd to $N = 208$	Expd to $N = 274$	$N_{max} = 208$			Expd to $N = 208$	Expd to $N = 274$	$N_{max} = 274$			Expd to $N = 274$
Cntl (p_1)	Exp (p_2)	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %	Overall resp %	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %	Overall resp %	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %
0.2	0.05	74.8	85.5	12.5	16.9	17.7	77.3	89.4	14.3	17.2	17.9	112.5	134.7	17.3	18.3
0.2	0.2	130.5	162.8	20.0	20.0	20.0	142.2	177.7	20.3	20.0	20.0	190.0	237.5	20.4	20.0
0.2	0.4	59.4	84.0	30.0	35.6	36.2	59.9	87.4	31.6	36.3	36.7	71.5	110.0	34.9	37.8
0.2	0.6	21.4	35.5	40.0	56.5	57.4	21.0	35.6	41.3	56.7	57.5	20.3	39.2	47.4	58.3
0.2	0.8	11.3	22.5	50.0	76.7	77.5	11.1	22.5	50.7	76.8	77.6	9.2	22.7	58.7	78.3
0.2	0.95	7.7	18.2	57.5	91.7	92.5	7.6	18.2	57.8	91.8	92.6	5.8	17.8	66.8	93.2

NOTE: 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 AR1, $\theta_H = \theta_T = 0.98$ for AR2, and 0.9835 for ER to yield 10% type I error rate at $p_1 = p_2 = p_3 = 0.2$ and 90% power at $p_1 = 0.2, p_2 = p_3 = 0.4$ (Shaded cells).

Abbreviations: Avg, average; Cntl, control arm; Exp, experimental arm; Expd, expanded; N_{max}, maximum sample size; Resp, response.

similar to Fig. 1A but includes early stopping rules. With early stopping, the average r_2 for the adaptive randomization design is consistently higher than that of the equal randomization design across the entire accrual period.

We also compare the results of equal randomization, AR1, and AR2 in a 3-arm clinical trial that incorporates early stopping rules for both futility and efficacy (Table 3). The equal randomization design requires the enrollment of up to 231 patients, the AR1 and AR2 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 equal randomization, AR1, and AR2, 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 equal randomization, AR1, and AR2 when $p_1 = p_2 = 0.2$ and $p_3 = 0.4$. The

Table 3. Equal randomization design for three treatment arms, with both futility and efficacy early stopping

True response rate			Equal randomization					AR1 $c = n/(2N)$					AR2 $c = (n/N)^{0.1}$			
			$N_{max} = 231$			Expd to $N = 255$	Expd to $N = 321$	$N_{max} = 255$			Expd to $N = 255$	Expd to $N = 321$	$N_{max} = 321$			Expd to $N = 321$
Cntl (p_1)	Exp 1 (p_2)	Exp 2 (p_3)	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %	Overall resp %	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %	Overall resp %	Nonresp, *n*	Avg sample size	Overall resp %	Overall resp %
0.2	0.05	0.05	93.8	104.3	10.8	15.9	16.7	91.4	103.1	11.1	16.1	16.9	91.8	104.2	12.7	17.4
0.2	0.2	0.2	161.2	201.4	20.2	20.0	20.0	176.7	221.0	20.2	20.0	20.0	221.5	276.9	20.2	20.0
0.2	0.4	0.4	69.6	104.5	33.7	37.0	37.2	72.5	110.1	34.3	37.4	37.5	84.3	131.9	36.3	38.3
0.2	0.6	0.6	23.6	44.3	46.8	57.7	58.2	23.6	44.7	47.3	57.8	58.2	23.9	49.3	51.2	58.7
0.2	0.8	0.8	11.5	28.7	59.8	77.8	78.2	11.4	28.8	60.1	77.8	78.2	10.0	29.5	65.7	78.7
0.2	0.3	0.5	53.2	82.8	33.8	43.4	44.3	54.9	83.3	34.3	43.7	44.6	53.6	85.2	37.3	45.7
0.2	0.4	0.6	32.9	54.8	40.3	53.6	54.4	32.8	55.2	40.7	53.8	54.5	31.8	57.2	44.3	55.5
0.2	0.4	0.8	18.0	33.8	46.7	74.2	75.1	17.9	33.7	47.0	74.3	75.1	15.5	32.6	52.3	76.1
0.2	0.1	0.6	40.1	57.4	30.5	53.1	54.5	38.2	55.5	31.3	53.6	54.9	31.2	50.0	37.3	56.4
0.2	0.2	0.4	96.3	131.4	27.1	32.2	32.8	97.1	134.2	27.7	33.0	33.6	97.0	138.6	30.2	35.3
0.2	0.2	0.6	38.2	57.3	33.8	53.7	54.9	37.1	56.2	34.3	54.0	55.1	31.7	52.1	39.2	56.4
0.2	0.2	0.8	20.4	34.0	40.2	74.5	75.6	20.1	33.8	40.6	74.6	75.7	16.5	31.4	47.3	76.7

NOTE: Cutoff value for futility stopping is $\theta_L = 0.06$. Cutoff value for efficacy stopping and final decision is $\theta_H = \theta_T = 0.9904$ for ER, $\theta_H = \theta_T = 0.9904$ for AR1, $\theta_H = \theta_T = 0.988$ for AR2 to yield 10% type I error rate at $p_1 = p_2 = p_3 = 0.2$ and 90% power at $p_1 = 0.2, p_2 = p_3 = 0.4$ (shaded cells).

Abbreviations: Avg, average; Cntl, control arm; Exp, experimental arm; Expd, expanded; N_{max}, maximum sample size; Nonresp, nonresponse. Resp, response.

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

Table 4. Performance of equal and adaptive randomization designs for survival analysis with both futility and efficacy early stopping

True median survival rime		Parameters of exponential distribution		Equal randomization			AR1 $c = n/(\mathbf{2}/N)$			AR2 $c = (n/N)^{0.1}$		
				$N_{\max} = 170$		Expd to $N = 218$	$N_{\max} = 180$		Expd to $N = 218$	$N_{\max} = 218$		Expd to $N = 218$
Cntl	Exp	Cntl (μ_1)	Exp (μ_2)	Avg sample size	Median survival time	Median survival time	Avg sample size	Median survival time	Median survival time	Avg sample size	Median survival time	Median survival time
0.69	0.35	1.00	0.50	106.0	0.54	0.61	108.1	0.56	0.63	127.2	0.62	0.65
0.69	0.52	1.00	0.75	157.7	0.62	0.64	165.4	0.63	0.64	195.0	0.64	0.65
0.69	0.69	1.00	1.00	163.3	0.70	0.70	169.8	0.70	0.70	193.3	0.70	0.70
0.69	0.87	1.00	1.25	162.0	0.79	0.79	169.2	0.80	0.79	198.4	0.81	0.81
0.69	1.04	1.00	1.50	152.7	0.88	0.91	160.3	0.91	0.93	189.6	0.95	0.96
0.69	1.21	1.00	1.75	138.4	0.98	1.06	143.4	1.02	1.08	166.7	1.09	1.11
0.69	1.39	1.00	2.00	123.2	1.07	1.20	127.0	1.13	1.23	142.9	1.22	1.28
0.69	2.08	1.00	3.00	90.6	1.45	1.81	91.4	1.54	1.85	97.7	1.74	1.93

NOTE: The cutoff values for declaring significant results are $\theta_T = 0.721$, 0.721, and 0.742 for ER, AR1, and AR2, 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).

Abbreviations: Avg, average; Cntl, control arm; Exp, experimental arm; Expd, expanded; $N_{\max}$, maximum sample size.

early stopping probabilities are comparable for the 3 designs in all other scenarios (Supplementary Table S3).

A desirable design should have the smallest average sample size and the smallest number of nonresponders. These 2 features generally go together. Among the 3 designs, equal randomization is the best in 6 of the 11 scenarios in Table 3. The exceptions are (i) when both experimental arms are worse $p_2 = p_3 = 0.5$, AR1 is the best and (ii) when a large difference is seen across treatments (in 4 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 $p_2 = 0.2$; $p_3 = 0.8$), AR2 is the best. Similar to the 2-sample scenarios, in all the alternative cases, the overall response rate under AR2 is always higher than those under AR1 and equal randomization, 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 AR2 over equal randomization 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 an HR of 0.5 to 3. To achieve the same 10% type I error rate and 80% power, equal randomization requires up to 170 patients, whereas the AR1 and AR2 schemes require up to 180 and 218 patients, respectively. We compare the performance of the 3 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 equal randomization, followed by AR1, and AR2 is the largest. Conversely, in a comparison of the median survival time, AR 2 outperforms AR1 followed by equal randomization, with or without expanding to $N = 218$. For HRs of 0.5 and 3, the gains for AR2 in the median survival time are 15% and 20%, respectively, over equal randomization without expanding the sample size. When the sample size is expanded to 218 for equal randomization, the advantage of AR2 remains but the relative gain in the median survival time is reduced by 7% or less. Similar as before, for adaptive randomization, more patients are randomized to the better arm in all cases. For HRs 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 AR2 than in those in equal randomization (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 equal randomization and adaptive randomization in their totality and determine their relative strengths and weaknesses.

Our extensive simulation studies show that for a binary endpoint, equal randomization typically results in a smaller sample size and a smaller number of nonresponders than adaptive randomization to control the type I and type II errors. Equal randomization has a smaller average sample size than adaptive randomization when there is no difference in the response rates. Adaptive randomization consistently achieves a higher overall response rate by allocating more patients to more effective treatments during the course

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

of the trial. By expanding the sample size to the same number across all designs, adaptive randomization yields a higher overall response rate (a lower number of nonresponders as well) than equal randomization. In particular, when the experimental treatment is unexpectedly worse than the control, or when the experimental treatment is overwhelmingly better than the control, adaptive randomization may reach a smaller average sample size, a smaller number of nonresponders, and a higher overall response rate than equal randomization at the same time. Similar conclusions hold in 3-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 preset throughout the trial. Adaptive randomization 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; adaptive randomization adapts the randomization probabilities automatically and continuously based on the observed data.

There are pros and cons to all designs. Equal randomization designs emphasize maximizing the statistical power and are favored from a global, population-based view. On the other hand, adaptive randomization designs put more emphasis 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. Equal randomization 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. Equal randomization 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 equal randomization design to the same sample size as the adaptive randomization design. With the expanded sample size, the adaptive randomization design always yields higher overall success than the equal randomization design. The benefit of adaptive randomization is more prominent when one treatment is substantially better than the others. Adaptive randomization focuses on how to best treat patients in the trial, whereas equal randomization emphasizes making the right decision early in comparing the treatment effect. Hence, when there is a treatment difference, equal randomization is preferred if the population size is much bigger than the trial size and adaptive randomization is preferred otherwise. When there is no treatment difference, equal randomization is preferred over adaptive randomization because the required sample size to reach a conclusion is smaller. More detailed comparison of the relative merit of equal randomization and adaptive randomization with respect to the population size and trial size is given in the Appendix. Our results also show that the difference between equal randomization and adaptive randomization designs quickly diminishes when early stopping rules for efficacy and futility are implemented.

Adaptive randomization takes the "learn as we go" approach to adjusting the randomization ratio based on the observed data. Adaptive randomization assigns more patients to better treatments and, as a result, these treatments can be better studied with larger sample sizes. Yet, statistical power may suffer from imbalanced samples. AR1 is commonly used in practice whereas AR2 serves as an example to illustrate more extreme imbalance. How much imbalance one would like to reach depends on the specific objectives. The "optimal" randomization ratio can be derived based on the design settings and 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). In fact, by allowing the randomization ratio to vary, equal randomization can be considered as a special case of adaptive randomization. The best randomization ratio, which can be equal or unequal, depends on the design parameters and the optimization criteria.

From the trial conduct point of view, trials using outcome-adaptive randomization require more effort in planning and implementation. A robust infrastructure should be set up to ensure that adaptive randomization can be carried out properly throughout the trial. The outcome-adaptive randomization is more applicable to trials with short-term endpoints. To ensure that adaptive randomization 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 adaptive randomization trials. A scheduling module and an e-mail notification module can be included to ensure that the primary endpoints are collected and reported in a timely manner. 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 adaptive randomization is more prone to the danger of population drift. When the patient characteristics change over time, adaptive randomization is more likely to result in a biased estimate of the treatment difference than equal randomization (28). One solution is to lay out well-defined eligibility criteria such that a homogeneous population of patients can be

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

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-adaptive randomization 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 2 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.

Appendix: Comparison of Equal and Adaptive Randomization by the Number of Nonresponders and the Equivalence Ratio of the Population Size versus Trial Size

To provide further comparison between equal randomization and adaptive randomization, we plot the additional number of nonresponders versus the number of patients accrued for 2-arm trials with binary endpoints. The computations are based on the average of 500,000 trials. The number of additional nonresponders is defined as the excess number of nonresponders for the respective designs compared with the case that all patients had received the better treatment.

Supplementary Figure S1A shows the additional number of nonresponders over time for the equal randomization and AR2 designs compared with the reference case that all patients had received the better treatment. The blue solid line shows that the equal randomization design stopped at $N = 134$ with 13.4 more nonresponders than the theoretically best case where all patients are assigned to arm 2. The blue dashed line shows the additional number of nonresponders in the expansion cohort after equal randomization, whereas the black line represents the straight equal randomization throughout. The red line indicates the "excess" number of nonresponders for AR2 compared with the theoretically best-case scenario. By 184 patients, the "excess" numbers of nonresponders for straight equal randomization, equal randomization + expansion, and AR2 are 18.4, 14.4, and 7.1, respectively. Throughout the trial, AR2 is better than equal randomization by yielding a smaller number of nonresponders when a total of 184 patients are treated.

Similarly, Supplementary Fig. S1B plots the "excess" number of nonresponders over the best-case scenario when the designs implement early stopping rules. The horizontal location of the green dots indicates the average sample size, whereas the blue and the red dots show the maximal sample size for equal randomization and AR2, respectively. When expanding to 274 patients, the "excess" number of nonresponders is 5.9 for the AR2 design which is smaller than 10.5 of the equal randomization 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 in 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. In addition, we consider the cases with different population sizes with respect to the trial size. The performance of equal randomization and adaptive randomization is compared by computing the number of nonresponders in the entire patient population with similar conditions and could be affected by the similar treatments—both current patients enrolled in the trial and future patients in the population outside the trial. Taking the example of comparing 2 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 equal randomization design needs an average of 84 patients with 59.4 nonresponders. In contrast, the AR2 design (with the tuning parameter $c = (n/N)^{0.1}$ requires 110 patients with 71.5 nonresponders. One major difference between equal randomization and AR2 is that equal randomization 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 equal randomization and the end of adaptive randomization, we compute the expected number of nonresponders as follows.

1 Using the equal randomization 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 nonresponders including patients treated during and after the trial is $59.4 + (0.9 \times 0.6 + 0.1 + 0.8)\, x$.
2 Using the AR2 design, the trial ends at 110 patients, which is 26 patients more than using the equal randomization design. During this period of time, all x number of patients available outside the trial are treated on the control arm. The total expected number of nonresponders 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 2 equations. Taking the ratio of 48.3 and 26, we get 1.86. The ratio can be considered as the "equivalence ratio" of equal randomization and adaptive randomization in terms of yielding the same number of nonresponders. 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, equal randomization is better. Otherwise, AR2 is better. Similar calculations can

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

be applied to other settings. For example, the equivalence ratios for the settings in Table 1 with 2 arms without early stopping are also about 1.8 to 1.9. The equivalence ratios for Table 2 with 2 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 3-arm trials shown in Table 3, the equivalence ratios are between 1.5 and 4.6 except for the 4 cases mentioned in the text with larger differences between the experimental arms and the control arm. In those cases, AR2 yields a smaller sample size and has a smaller number of nonresponders. Generally speaking, equal randomization is preferred when the patient population outside the trial is large (e.g., more than twice the trial size) because the trial result can be reported earlier to benefit the entire population. On the other hand, adaptive randomization is preferred if there are not many patients available outside the trial, such as in the rare disease setting and when effective treatments are available. Adaptive randomization has the benefit of minimizing the number of nonresponders 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 nonresponders for the equal randomization design are (n_1, m_1) and those for the adaptive randomization design are (n_1, m_1), respectively. Also assume that the power for the test is w, it can be shown that the solution of x for the 2-arm trial is $[(1 - p_1)(n_2 - n_1) - (m_2 - m_1)]/[w(p_2 - p_1)]$. The equivalence ratio is $x/(n_2 - n_1)$, which depends on the true response rates, differences of the trial sample size, and the number of nonresponders between the 2 arms, as well as statistical power. The equivalence ratios can be calculated for different adaptive randomization methods with different tuning parameters that determine the degree of imbalance. For example, the equivalence ratio changes to 3.63 if we compare AR1 and equal randomization 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 nonresponders. The variation of the number of nonresponders tends to be larger in adaptive randomization than in equal randomization.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: J.J. Lee, G. Yin
Development of methodology: J.J. Lee, N. Chen, G. Yin
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): J.J. Lee, N. Chen, G. Yin
Writing, review, and/or revision of the manuscript: J.J. Lee, N. Chen, G. Yin
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): G. Yin
Study supervision: J.J. Lee

Acknowledgments

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

Grant Support

The study was supported, in part, by NCI grants CA16672 (J.J. Lee), CA97007 (J.J. Lee and N. Chen), and the Hong Kong RGC grant (G. Yin).

Received October 4, 2011; revised May 19, 2012; accepted June 10, 2012; published OnlineFirst July 2, 2012.

References

1. Peirce CS, Jastrow J. On small differences in sensation. Memoirs Natl Acad Sci 1885;3:73–83. Available from: http://psychclassics.yorku.ca/Peirce/small-diffs.htm.
2. Hacking I. Telepathy: origins of randomization in experimental design. Isis 1988;79:427–51.
3. Stigler SM. Mathematical statistics in the early states. Ann Stat 1978;6:239–65.
4. Stigler SM. A historical view of statistical concepts in psychology and educational research. Am J Educ 1992;101:60–70.
5. Fisher RA. Statistical methods for research workers. London (UK): Oliver and Boyd; 1925.
6. Fisher RA. Design of experiments. London, UK: Oliver and Boyd; 1935.
7. Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council investigation. Br Med J 1948;2:769–82.
8. Lachin JM. Statistical properties of randomization in clinical trials. Control Clin Trials 1988;9:289–311.
9. Schulz KF, Grimes DA. Generation of allocation sequences in randomised trials: chance, not choice. Lancet 2002;359:515–9.
10. Rosenberger WF, Lachin JM. Randomization in clinical trials: theory and practice. New York: John Wiley & Sons; 2002.
11. Friedman LM, Furberg CD, DeMets DL. Fundamentals of clinical trials. 4th ed. New York: Springer; 2010.
12. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics 1975;31:103–15.
13. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. Biometrika 1933;25:285–94.
14. Zelen M. Play the winner rule and the controlled clinical trial. J Am Stat Assoc 1969;64:131–46.
15. Wei LJ, Durham SD. The randomized play-the-winner rule in medical trials. J Am Stat Assoc 1978;85:156–62.
16. Hu F, Rosenberger WF. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. J Am Stat Assoc 2003;98:671–8.
17. Hu F, Rosenberger WF. The theory of response-adaptive randomization in clinical trials. Hoboken (NJ): John Wiley & Sons; 2006.
18. Thall PF. Ethical issues in oncology biostatistics. Stat Methods Med Res 2002;11:429–48.
19. Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. Stat Sci 2004;19:175–87.
20. Korn EL, Freidlin B. Outcome-adaptive randomization: is it useful? J Clin Oncol 2010;21:100–20.
21. Berry DA. Adaptive clinical trials: the promise and the caution. J Clin Oncol 2010;21:606–9.
22. Yuan Y, Yin G. On the usefulness of outcome-adaptive randomization. J Clin Oncol 2011;29:e390–2.
23. Korn EL, Freidlin B. Reply to Y. Yuan et al. J Clin Oncol 2011;29:e393.
24. Berry SM, Carlin BP, Lee JJ, Müller P. Bayesian adaptive methods for clinical trials. Boca Raton (FL): Chapman & Hall; 2010.
25. Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. Eur J Cancer 2007;43:859–66.
26. Berry DA, Eick SG. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. Stat Med 1995;14: 231–46.
27. Cheng Y, Berry DA. Optimal adaptive randomized designs for clinical trials. Biometrika 2007;94:673–87.

Published OnlineFirst July 2, 2012; DOI: 10.1158/1078-0432.CCR-11-2555

28. Karrison TG, Huo D, Chappell R. A group sequential, response-adaptive design for randomized clinical trials. Control Clin Trials 2003;24: 506–22.

29. Jennison C, Turnbull BW. A multi-stage adaptive design with time trend. Group sequential methods with applications to clinical trials. New York: Chapman & Hall; 2000. p. 331–3.

30. Magirr D. Block response-adaptive randomization in clinical trials with binary endpoints. Pharm Stat 2011;10:341–6.

31. Zhou X, Liu S, Kim ES, Herbst RS, Lee JJ. Bayesian adaptive design for targeted therapy development in lung cancer: a step toward personalized medicine. Clin Trials 2008;5:181–93.

32. Ji Y, Bekele BN. Adaptive randomization for multiarm comparative clinical trials based on joint efficacy/toxicity outcomes. Biometrics 2009;65:876–84.

33. Eickhoff JC, Kim K, Beach J, Kolesar JM, Gee JR. A Bayesian adaptive design with biomarkers for targeted therapies. Clin Trials 2010;7: 546–56.

34. Lee JJ, Gu X, Liu S. Bayesian adaptive randomization designs for targeted agent development. Clin Trials 2010;7:584–96.

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.

36. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, et al. The BATTLE Trial: personalizing therapy for lung cancer. Cancer Discov 2011;1:44–53.

37. Berry DA. Adaptive trials in oncology. Nat Rev Clin Oncol 2011;9: 199–207.