Some Caveats for Outcome Adaptive Randomization in Clinical Trials

Peter F. Thall[1], Patricia S. Fox[1], and J. Kyle Wathen[2]

[1]*Department of Biostatistics, U.T. M.D. Anderson Cancer Center*

Houston, Texas, USA

[2] *Model Based Drug Development, Janssen Research & Development*

Titusville, NJ, USA

KEYWORDS: Adaptive design; Adaptive randomization; Bayesian design; Clinical trial

June 27, 2014

CONTENTS

1.1 Introduction

"*Intuition is a poor guide when facing probabilistic evidence.*"

Dennis V. Lindley

Outcome adaptive randomization (AR) in a clinical trial uses both the assigned treatments and observed outcomes of previous patients to compute randomization probabilities for newly accrued patients. In this chapter, we will focus on two types of AR. The scientific goal of a randomized comparative trial (RCT) of two treatments is to decide whether one treatment is substantively better than the other. Fair (also called 50:50, 1:1, or coin flip) randomization in a RCT fixes the treatment assignment probabilities at .50 throughout in order to obtain data that provide unbiased estimators of the parameters used for this comparison. While fair randomization serves the needs of future patients, flipping a coin to decide a patient's treatment looks strange to many non-statisticians, and may seem at odds with maximizing benefit to the patients in the trial and hence ethically undesirable. It also may appear to imply that the patient's physician is unduly ignorant. Many physicians refuse to participate in trials with fair randomization because they have strong beliefs about which treatment is superior. At the other extreme is a "greedy" algorithm wherein each new patient simply is given the treatment having the currently larger empirical success rate or mean survival time. It is well-known that greedy sequential decision algorithms that always choose the next action to maximize a given optimality criterion are 'sticky,' in that they have a non-trivial risk of getting stuck at a locally optimal action that is globally suboptimal. See, for example, Sutton and Barto (1998). In RCTs, a competitor to fair randomization is AR, which intentionally unbalances the sample sizes by interimly assigning patients to the empirically superior arm with higher probability (cf. Cornfield, et al., 1969; Berry and Eick, 1995; Hu and Rosenberger, 2006). For RCTs, AR provides a compromise between greedy and fairly randomized treatment assignment. For treatment assignment in a RCT with binary outcomes, Thompson (1933) first conceived AR using a Bayesian framework.

The second type of AR that we will discuss is used in early phase trials where treatment

regimes, which usually are doses, doses pairs, or dose-schedule combinations, are chosen sequentially for successive cohorts of patients. This is done, rather than randomizing fairly among the regimes, since little is known about their safety, and the fear of increasing risk of regimen-related toxicity or death with increasing dose makes fair randomization unethical. The goal is to optimize the treatment regime. Such trials usually are known as phase I if they are based on toxicity, and as phase I-II if based on both efficacy and toxicity. Recently, AR has been used in several early-phase trial settings as an additional design component to reduce stickiness (Azriel, Mandel, and Rinott, 2011; Thall and Nguyen, 2012; Thall, et al. 2013). Discussions of stickiness in the phase I setting are provided by Oron and Hoff (2013) and Carlin, et al. (2013).

There are many ways to do AR in RCTs. We will discuss several Bayesian methods. Similarly to Thall and Wathen (2007), our focus will be a trial with up to $N = 200$ patients with the goal to compare treatments A and B based on a binary response with probabilities θ_A and θ_B. Denote the achieved interim samples sizes by n_A, n_B, and $n = n_A + n_B$, and the final achieved sample sizes by N_A, N_B, and $N = N_A + N_B$. We assume that θ_A and θ_B follow independent beta(.25, .75) priors, and use the posterior probabilities $p_{A<B} = \Pr(\theta_A < \theta_B \mid data_n)$ and $p_{B<A} = 1$ - $p_{A<B}$ to compute AR probabilities. It is important to keep in mind the important fact that $p_{B<A}$ is a function of the data, hence is a statistic. This will play a critical role in the methods that we will discuss, and is true in general for any data-dependent adaptive treatment assignment rule. We denote Bayesian AR wherein patients are randomized to A with probability $r_A = (p_{B<A})^c/\{(p_{B<A})^c + (p_{A<B})^c\}$ and to B with probability $r_B = 1 - r_A$ for $c > 0$ by AR(c). We will consider AR(1), which was introduced by Thompson (1933), the commonly used AR(1/2), and AR($n/2N$). For each AR method, the trial is stopped with A declared superior to B if $p_{B<A} > .99$ and B declared superior to A if $p_{A<B} > .99$, with these rules applied continuously.

As a first comparator to AR, we include a design with fair randomization that also makes decisions continuously, and for this we used the same cut-off $p_{A<B} > .99$ or $< .01$ as used for the AR methods. We will refer to this as "Fair Contin." As a second comparator that is

closer to what would be used in practice, we consider the group sequential Bayesian design that randomizes patients fairly between A and B with the rule, applied at $n = 50$, 100, 150, and 200 patients, that stops and concludes B is superior to A if $\Pr(\theta_A + .20 < \theta_B \mid data_n) > a - b(n/200)$, that A is superior to B if $\Pr(\theta_B + .20 < \theta_A \mid data_n) > a - b(n/200)$, and otherwise continues to $N = 200$ and makes a final decision. The cut-off parameters $a = .95$ and b = .80 were used to ensure overall type I error probability $\leq .05$. We will refer to this design as "Fair GS." For both comparators, we randomize patients fairly in blocks of size 8 to avoid interim sample size imbalance.

The motivation for using AR in place of fair randomization is to obtain N_A and N_B unbalanced in favor of the arm having larger true success probability, i.e. to obtain N_B substantively larger than N_A if θ_B^{true} is substantively larger than θ_A^{true}. An important practical consideration is that, like many adaptive statistical procedures, AR sometimes behaves in ways that are not intuitively obvious. For application in actual RCTs, it is important to be aware that AR may have some undesirable properties, which we will illustrate and discuss in turn. These include :

1) introduction of substantial variability into the distributions of N_A and N_B far beyond that introduced by fair randomization;

2) the fact that there may be a surprisingly high probability of a sample size imbalance in the wrong direction, with a much larger number of patients assigned to the inferior treatment arm, so that AR has an effect that is the opposite of what was intended;

3) introduction of substantial bias, resulting from any continuous outcome-adaptive treatment comparison, that with AR is worsened if there is parameter drift due to either improving or worsening patient prognosis over time;

4) the practical problem that, for covariate-adjusted AR, is is not unlikely that patient baseline covariates may be recorded incorrectly, which potentially can have disastrous consequences.

The severity of each of these problems depends on the particular AR method used and the

specific trial design parameters. These problems may be mitigated by adopting some sort of fix-up strategy, such as stratifying or blocking (Cornfield, 1969; Jennison and Turnbull, 2000; Karrison, Huo and Chappell, 2003), doing an initial 'burn-in' using fair randomization before applying AR, modifying the AR probabilities to shrink them toward .50 as AR(c) does, or correcting for bias in some *post hoc* manner. With such fix-up methods, the resulting gain in the number of patients treated with the superior treatment over the inferior treatment, if in fact θ_B^{true} and θ_A^{true} differ, often becomes much smaller. Consequently, it is worthwhile to consider the actual gain with AR in light of the noted problems and the increased risk of an incorrect conclusion, that in turn may lead to giving an inferior treatment to a large number of future patients.

1.2 Achieved Sample Size Distributions

While it is well known that any form of random treatment assignment produces random achieved sample sizes, it is important to consider the particular forms of the distributions of N_B and N_A for a given AR method. Simulation-based properties of AR methods often are reported in terms of the means, $E(N_B)$ and $E(N_A)$. The difference $E(N_B - N_A)$ often is used, *per se*, to quantify a putative advantage of an AR method over fair randomization, which has $E(N_B - N_A) = 0$, for specific $\theta_B^{true} > \theta_A^{true}$. Mean sample sizes alone may be misleading, however, due to the fact that AR produces much more disperse sample size distributions compared to fair randomization.

For all simulations, each combination of trial conduct method and case was replicated 10,000 times. Table 1 gives simulation results for a trial run using each of the three AR and two fair designs. We write "$A > B$" in place of "$\theta_A > \theta_B$" for brevity. These simulations are similar to those reported by Thall and Wathen (2007). We note some important typos in that paper: (i) The prior of θ_A and θ_B actually used for the simulations was not beta(.50, .50) as reported, but rather was beta(.25, .75) as used here; (ii) in the left-hand column on page 862 'A' and 'B' are incorrectly reversed; (iii) in the header of Table 1, $\Pr(N_A > N_B + .20)$ should be $\Pr(N_A > N_B + 20)$.

The simulation results reported here in Table 1 show that, while the mean of $N_B - N_A$ is

quite large for each AR method when $\theta_B^{true} = .35$ or .45 and $\theta_A^{true} = .25$, the distributions of this difference and of the total sample size $N = N_A + N_B$ both are much more disperse with AR compared to fair randomization. Figure 1 illustrates this important point by giving the empirical distributions of $N_B - N_A$ for each of the three AR designs in the case where $\theta_A^{true} = .25$ and $\theta_B^{true} = .35$. The densities of $N_B - N_A$ for the two fair randomization designs are not included because they are very concentrated near 0 thus would distort the figure. While $\mathrm{E}(N_B - N_A) > 0$ for each AR method, Figure 1 shows that the distribution of $N_B - N_A$ has a very long left tail for AR(1) and AR(1/2). This is because because these AR methods carry a nontrivial risk of getting "stuck" with a larger sample for the inferior arm A early in the trial, due to the low sample size, and hence high variability in the posterior of arm B. When this occurs, many patients are treated on the inferior arm A, the stopping boundary for correctly declaring B superior is not crossed, and consequently the distribution $N_B - N_A$ ends up with a long left tail. The second possibility is that a small to modest number of patients are treated with A and a larger number are treated with B, which is the goal of the AR procedures, and thus the distribution of $N_B - N_A$ is shifted to the right. The third possibility is that, due to the play of chance, the AR method does not favor either arm substantially and the distribution of $N_B - N_A$ is centered near 0, as with fair randomization.

The long left tail of the distribution of $N_B - N_A$ has a very undesirable consequence. For example, in the case where $\theta_A^{true} = .25$ and $\theta_B^{true} = .35$, while AR(1) gives $E(N_B - N_A) = 66$, due to the long left tail of the distribution of $N_B - N_A$, which has 2.5^{th} percentile -164, AR(1) has the ethically very undesirable property that $\pi_{20} = \Pr[N_A > N_B + 20] = .14$. *That is, in this case, AR(1) has a 14% chance of producing a sample size imbalance of 20 patients or more in the wrong direction.* AR(1/2) reduces this probability to $\pi_{20} = .069$, and AR(n/2N) has $\pi_{20} = .03$, while $\pi_{20} = 0$ for both designs with fair randomization. These values of π_{20} for the three AR methods may be considered as trade-offs for the corresponding values $\mathrm{E}(N_B - N_A) = 66$, 37, and 21. This example illustrates two important points. First, since N_B and N_A are highly disperse with AR, using mean achieved sample sizes is not enough to adequately describe a given method's behavior. Second, the general claim that any AR

method is ethically more desirable than fair randomization is false.

Another interesting result shown by Table 1 is that, for the cases $\theta_B^{true} = .35$ or .45 where B is superior to A, all four methods that make decisions continuously have much larger bias in the final estimate of $\theta_B - \theta_A$ compared to fair GS. That is, most of the bias appears to be due to continuous treatment comparison, rather than AR *per se*. In general, the magnitude of this bias increases with the frequency of comparative decision making and $\theta_B^{true} - \theta_A^{true}$. The Fair Contin method exhibits as much or slightly more bias than the AR designs, while the Fair GS design reduces the bias by 66% to 71% compared to all the other designs.

Figure 2 gives the distributions of the posterior mean of Δ for the Fair GS, AR(n/2N), and AR(1) designs for the three cases where $\Delta^{true} = 0$, .10, or .20. The density plots of the estimates provide graphical illustrations of the nature of the bias and dispersion of the posterior means obtained from the competing methods given in Table 2. For all three designs, the distributions are multi-modal in all three cases, including the null, with a large mode at Δ^{true} and smaller modes located quite far from the null. In the two alternative cases where $\Delta^{true} = .10$ or .20, all three methods have a mode to the right of Δ^{true}, and the two AR methods have much more probability mass to the right of Δ^{true} compared to Fair GS. These plots illustrate why the AR methods have larger bias compared to Fair GS. The general messages are that all adaptive designs introduce bias, as is well known, but that the bias is much larger when AR methods are used rather than Fair GS, and that AR(1) produces by far the largest bias.

It also is worth mentioning that, for Bayesian methods that make decisions continuously, the seemingly trivial difference between assuming a beta(.50, .50) prior or a beta(.25, .75) prior can have non-trivial effects on the design's properties. A beta (.50, .50) prior decreases the probability of correctly concluding that $B > A$ for AR(1) or AR(1/2). For example, when $\theta_B^{true} = .35$, AR(1) and AR(1/2) have respective power figures .20 and .35 compared to .30 and .40 with a beta(.25, .75) prior. This effect is due to the great variability in the statistic $p_{A<B}$ for small samples early in the trial. AR(n/2n) does not suffer from this problem since its AR probabilities are very close to .50 for small n, i.e. it behaves similarly to Fair Contin

early in the trial.

A modification of AR methods that commonly is used to mitigate these problems arising with small early sample sizes is to begin the trial with a "burn-in" using fair randomization, and start using AR once the data from the burn-in are obtained. If a burn-in with 20 fairly randomized patients is used for each of the three AR methods, there is virtually no effect on power but, as might be expected, the sample size dispersion is reduced. In the case where $\theta_A^{true} = .25$ and $\theta_B^{true} = .35$ in Table 1, this in turn reduces the π_{20} values .138, .069, .028 for AR(1), AR(1/2), AR(n/2N) to .084, .050, .024, respectively. Starting the trial with a burn-in also slightly decreases the bias, but also reduces the values of $N_B - N_A$.

1.3 Drift and Bias

One of the most prominent arguments against the use of AR is that it can lead to biased estimates in the presence of parameter drift. This occurs, for example, if θ_A and θ_B both increase over time due to improving prognosis of patients enrolled over the course of the trial, but the comparative effect $\theta_B - \theta_A$ remains constant. Karrison, Kuo and Chappell (2003) recommended a blocking method to mitigate the biasing effects of drift when using AR, and provided a detailed simulation study. To examine the effects of drift on the behavior of the designs in Table 1, we re-simulated each case with θ_A and θ_B each increasing linearly from its initial value to that value plus .20 at the end of the trial. In each case, θ_A took on the true values $\theta_A(n)^{true} = .20 + .20(n/200)$ when n patients had been accrued while, for example in each case where nominally $\theta_B = .35$, it took on the true values $\theta_B(n)^{true} = .35 + .20(n/200)$. Thus, the treatment effect $\theta_B(n) - \theta_A(n)$ remained constant throughout the trial, while each probability increased.

The results of these simulations are summarized in Table 2. While drift had little effect on the sample size distributions, it substantively increased the false positive rate of AR(1) from .18 to .36 and of AR(1/2) from .24 to .32. Drift also increased the bias of all three AR methods for $\theta_B^{true} = .35$ or .45, with larger increases seen for AR(1) and AR(1/2). It is important to note that the bias for the Fair GS design changed trivially when drift was introduced, with $\widehat{bias}(\theta_B^{true} = .35) = .017$ and $\widehat{bias}(\theta_B^{true} = .45) = .032$ without drift (Table

1) compared to $\widehat{bias}(\theta_B^{true} = .35) = .019$ and $\widehat{bias}(\theta_B^{true} = .45) = .033$ with drift (Table 2). It thus appears that, while the Fair GS design does introduce some bias, this is not made substantively worse by drift. In contrast with the results in Table 1, in the presence of drift the Fair Contin method produces slightly less bias than the AR designs, although the Fair GS design still greatly reduces the bias, by at least 67%. It also should be noted that the estimates of θ_A and θ_B in Table 2 do not really correspond to the tabled θ_A^{true} and θ_B^{true} values, but rather $\theta_A(n)^{true}$ and $\theta_B(n)^{true}$ changed over time.

1.4 Adaptive Use of Patient Covariates

1.4.1 A Highly Structured Design

Thall and Wathen (2005) described a Bayesian adaptive design for a randomized comparative trial of two chemotherapy regimes, T = G+D (gemcitabine+ docetaxel) and T = G (gemcitabine alone) for advanced or metastatic unresectable soft tissue sarcoma. The design was quite ambitious. It accounted for up to four stages of therapy, each stage lasting six weeks, with stage k having ordinal outcome, Y_k, taking on the three possible values $\{R, F, S\}$ where R = complete or partial remission (response), F = progressive disease or death (failure), and S = neither R nor F (stable disease). It included a within-patient outcome-adaptive decision rule for k =1, 2, 3, reflecting actual physician behavior, that terminated therapy if $Y_k = F$ or R and continued if $Y_k = S$. Thus, outcome was not a univariate binary variable, but rather was a longitudinally observed vector $(Y_1, \cdots, Y_k)$ of random length k =2, 3, or 4. In addition to including stage-specific treatment effects, the underlying model and method also accounted for two binary baseline covariates, $\mathbf{Z} = (Z_1, Z_2)$, indicating whether the patient had received previous pelvic radiotherapy (PPR) and whether the patient's disease was a leimeiosarcoma (LMS) or some other sarcoma subtype. A generalized logistic conditional probability model for Y_k in stage k = 1,2,3,4 accounted for T, $\mathbf{Z}$, treatment-covariate interactions, and previous outcome history, $(Y_1, \cdots, Y_{k-1})$, giving a total of nine distinct possible likelihood contributions for each $(T, \mathbf{Z})$. The AR was based on the posteriors of the probabilities $\xi_T(R, \mathbf{Z}, \boldsymbol{\theta})$ and $\xi_T(F, \mathbf{Z}, \boldsymbol{\theta})$ of overall treatment success and failure, respectively, by the end of the patient's

regime. Using weights elicited from the physicians to quantify the relative importance of overall success and overall failure, the AR treatment assignment probabilities were $r_T(\mathbf{Z}, \boldsymbol{\theta})$ $= 0.435\,\xi_T(R, \mathbf{Z}, \boldsymbol{\theta}) + 0.565\,\{1 - \xi_T(F, \mathbf{Z}, \boldsymbol{\theta})\}$, for $T = G, G + D$. This AR method was applied after a burn-in of fair randomization with the first 30 patients. This is an example of a dynamic treatment regime (cf. Moodie, et al., 2007) in that, for each patient, the medical decision at each stage $k > 1$ was based on the patient's history of previous treatments and outcomes.

A website-based user interface for real-time data entry was constructed, written in ASP.NET, with physicians and research nurses at all participating clinical centers each given a user name and a password for access. The website had modules for both training and trial conduct, including detailed, explicit instructions, to allow each user to become familiarized with the program before enrolling patients.

1.4.2 What Actually Happened in the Sarcoma Trial

Most clinical trials are not conducted exactly as designed, for a variety of reasons, including the fact that medical oncology practice is inherently complex. Conduct of a trial with sequential outcome-adaptive rules that rely on accumulating data requires an extra step in the clinic to enter the necessary patient outcome information into a database. This may be delayed or forgotten because physicians and nurses are busy treating patients. Moreover, while a research nurse may have scored a patient's outcome, the attending physician must sign off on it before it can be recorded officially.

In the sarcoma trial, the covariates $\mathbf{Z}$ were to be determined at patient entry and entered into the database. Unfortunately, people make mistakes, and this occurred many times in this trial. Table 3 summarizes the final actual and website-based failure count data at the end of the trial, and the corresponding AR probabilities, for each covariate subgroup. The table shows that incorrect $\mathbf{Z}$ values were entered into the website for many patients, and that this had substantial effects on the AR probabilities $r_{G+D}(\mathbf{Z}, \boldsymbol{\theta})$ in two of the four subgroups. While all incorrect AR probabilities were unbalanced in the right direction, this was due to sheer luck, since it turned out that $G + D$ was superior to G in all four subgroups (Maki, et

al., 2007).

The take-away message from this trial is that, while accounting for patient covariates when making interim adaptive decisions may seem like a good idea from an ideal perspective, there is a non-trivial probability that covariates will be entered incorrectly, or not at all. As seen in the sarcoma trial, this may have a severe detrimental effect on a design's actual behavior. While there is always a risk that people may make mistakes when entering variables into database, if these variables are used as the basis for real-time adaptive treatment decisions then such mistakes can have very severe consequences in terms of actual treatment assignments. The likelihood of these types of errors can be reduced by including second-party verification of all data that is used in adaptive decision making.

A similar logistical problem arises for biomarker covariates, such as gene or protein signatures, which may require a non-trivial amount of time to be evaluated using a blood or tissue sample. This delays computation of any covariate-specific adaptive treatment assignment rule, and thus delays the actual time when the patient's treatment may begin. Consequently, such biomarker-adaptive rules may be ignored by physicians or nurses in the clinic in order to treat patients in a timely fashion.

1.5 Adaptive Randomization to Reduce Stickiness

AR in RCTs should not be confused with the use of AR in sequentially adaptive early phase trials, where the goal is to select an optimal regime (dose, dose pair, schedule, or dose-schedule combination). In such trials, using AR is motivated by the desire to reduce the probability of getting stuck at a suboptimal regime due to the sequential use of a greedy algorithm such as the CRM or posterior mean utility optimization. For trials based on toxicity alone, this usually is done using closeness to a targeted probability, as done by the continual reassesment method (CRM, O'Quigley, et al.; Cheung, 2011). Trials using both efficacy and toxicity employ a criterion that quantifies the trade-off between toxicity and efficacy (cf. Braun, 2002; Thall and Cook, 2004) or their joint utility (cf. Thall and Nguyen, 2012). In this setting, most methods use a greedy algorithm that optimizes the criterion based on the current data to select a regime for each cohort.

In general, for bivariate ordinal outcome $\mathbf{Y}$ taking on values $\mathbf{y} = (y_E, y_T)$, let $U(\mathbf{y})$ denote elicited joint consensus utilities, d = dose, and $\boldsymbol{\theta}$ the parameter vector for the probability $\pi(\mathbf{y}|d,\boldsymbol{\theta}) = \Pr(\mathbf{Y} = \mathbf{y}|d,\boldsymbol{\theta})$. A greedy utility-based algorithm chooses each new cohort's dose to maximize the posterior mean utility

$$\phi(d, data) = \mathrm{E}\Big[\mathrm{E}\big\{U(Y_E, Y_T)|d,\boldsymbol{\theta}\big\}|data\Big] = \int_{\boldsymbol{\theta}} \sum_{\mathbf{y}} U(\mathbf{y})\pi(\mathbf{y}|d,\boldsymbol{\theta})p(\boldsymbol{\theta}|data)d\boldsymbol{\theta},$$

where $p(\boldsymbol{\theta}|data)$ is the posterior computed using the current data. An alternative AR method assigns dose d with probability

$$r_d(data) = \frac{\phi(d, data)}{\sum_{x=1}^{m} \phi(x, data)}, \quad \text{for } d = 1, \cdots, m.$$

In practice, the trial begins using the greedy algorithm, constrained by safety requirements, and AR is begun after a specified amount of initial data are obtained. A frequentist version of this would use a plug-in estimator $\hat{\boldsymbol{\theta}}$ instead of computing the posterior expectation, and

$$\phi(d, \hat{\boldsymbol{\theta}}) = \mathrm{E}\big\{U(Y_E, Y_T)|d,\hat{\boldsymbol{\theta}}\big\} = \sum_{\mathbf{y}} U(\mathbf{y})\pi(\mathbf{y}|d,\hat{\boldsymbol{\theta}})$$

would be used to define r_d.

An example is given in Thall and Nguyen (2012), who present a Bayesian utility-based phase I-II design for a trial to optimize palliative radiation therapy dose for pediatric brain tumors. Each patient's outcome is characterized as $\mathbf{Y} = (Y_E, Y_T)$ where Y_R is the number of improvements in clinical symptoms, radiographic appearance of the tumor, or quality of life, hence has possible values $y_E = 0$, 1, 2, or 3, and Y_T is toxicity severity, categorized as y_T = low, moderate, high, or severe, so there are 16 possible elementary outcomes. Application of AR in this type of setting, as an alternative to a pure greedy algorithm, is useful because it reduces the probability of getting stuck at a suboptimal dose. Compared to the greedy utility-based method, AR gives slightly smaller probabilities of selecting the best regime in some settings, but much larger correct selection probabilities in settings where the greedy algorithm has very poor performance. Thus, in early phase trials, AR may be considered insurance against disaster in some cases, with the price being slightly worse performance on

other cases. This methodology also may be applied in more complex phase I-II settings, where the goal is to jointly optimize dose and schedule (Thall, et al., 2013).

1.6 Discussion

"*If it ain't broke, don't fix it.*"

Colloquial advice.

We have described problems with AR in RCTs, in the context of both a simple toy trial and a complex trial involving multi-stage regimes that actually was conducted. Our general aim has been to provide simple illustrations of important practical problems with AR that have both ethical and scientific consequences. In many actual applications of AR methods, it appears that the practitioners do not fully understand the properties of their methodologies. While the first author of this chapter certainly has been guilty of this failing in the past, it seems worthwhile to learn from one's mistakes in order to avoid repeating them.

The use of AR methods in place of fair randomization in clinical trials remains controversial. In the setting where treatment A is standard therapy and B is experimental, Korn and Freidlin (2011) compared BAR(n/2N) to unbalanced randomization with fixed probabilities in the proportions r_A:r_B = 1:1, 2:1, 4:1, or 9:1. They modified the Thall and Wathen (2007) design by capping the AR probabilities at .80, not including any early stopping rules, and setting $N = 140$ rather than 200. They concluded that a trial with AR should use block-randomization and a block-stratified analysis, as given in Jennison and Turnbull (2000) and discussed by Karrison and Chappell (2003), and conclude "Adaptive randomization is inferior to 1:1 randomization . . ." in terms of benefit to future patients, and that it offers "modest to no benefit" to the patients in the trial. Two aspects of the particular AR design that Korn and Freidlin studied are worth noting. First, in general all elements of a clinical trial design affect its operating characteristics, and effects on design performance of the interplay between different elements, such as a particular AR method and various other design components and parameters, often are non-trivial and cannot be anticipated based on intuition. Second, con-

ducting a 2-arm RCT without any early stopping rule is at odds with actual practice, and may be considered unethical. Their article triggered a letter by Yuan and Yin (2012), who noted that the optimal AR that minimizes the mean number of nonresponders has allocation ratio $\sqrt{(}\theta_B) : \sqrt{(}\theta_A)$, but that in many settings the actual gain of such an optimal AR with $r_A = \sqrt{(}\hat{\theta}_A)/\{\sqrt{(}\hat{\theta}_A) + \sqrt{(}\hat{\theta}_B)\}$ using the continuously updated empirical estimates $\hat{\theta}_A$ and $\hat{\theta}_B$ compared to using 2:1 fixed probability randomization is small. Yuan and Yin also state, "The gain from using AR in terms of reducing the number of nonresponders actually decreases when the response rate of the experimental arm increases." Of course, statistically, distinguishing between θ_A and θ_B that are far apart is easy and distinguishing between θ_A and θ_B that are close to each other is hard. Given these facts, and the well-known fact that balanced sample sizes provide a nearly optimal allocation for efficient (minimum variance) estimation of $\theta_A - \theta_B$, discarding this important advantage of fair randomization by using AR does not not seem to make sense in any case.

Comparison of AR(n/2N) to a simpler design with fair block randomization and a Bayesian group sequential stopping rule showed that the latter performs quite well, and does not suffer from the risks of adverse effects carried by AR. These adverse effects include (1) a nontrivial probability of greatly unbalancing sample size in the wrong direction, (2) increased bias in the final inferences due to continuous treatment comparison, which is increased if there is parameter drift, and (3) logistical difficulties during trial conduct in recording, either accurately or at all, patient covariates at accrual.

In recent years there as been an explosion of a wide variety of methods to correct for various sources of bias in observational data, or to correct for noncompliance in RCTs (cf. Robins, 1994; Hernan et al., 2000; Robins, 2005; Austin, 2009). This suggests that such methods may be used to analyze data from trials conduct using AR. These techniques essentially attempt to correct for bias by using the available data to create a dataset as close as possible to what would have been obtained with fair randomization, and they rely on the assumption that no unobserved confounders are at play. Of course, all of this can be avoided by using fair randomization in the first place.

Our conclusions and recommendations are as follows. First, in sequentially adaptive early phase trials, such as phase I-II dose-finding, where the goal is to select a best treatment regime from several candidates, adding suitably calibrated AR to a greedy algorithm can improve both the design's reliability and its ethical desirability. One *caveat* is that, in some cases, the use of additional dose acceptability rules may reduce or obviate the advantage of adding AR. For AR in RCTs, where the goal is unbiased comparison, there are several central issues. AR introduces greater bias, and this is worsened by drift. Additionally, AR produces much greater variability of $N_B - N_A$ and $N_B + N_A$, which carries the risks of unbalancing the sample size in the wrong direction and reducing the reliability of the estimates of $\theta_B - \theta_A$. Certainly, some of these problems are due to continuous decision making, which is associated with any continuously adaptive procedure, and can be mitigated by the use of a group sequential decision scheme or blocking. Similarly, the simple but disastrous problem with incorrect values of patient covariates being input that occurred in the sarcoma trial could arise in any trial that uses covariates adaptively in real time, and is not limited to AR. For AR, important fix-ups include some sort of shrinkage of the AR probabilities toward .50, blocking, and the use of a burn-in with fair randomization. Inclusion of such design elements, however, requires careful preliminary simulation study of the design's behavior and calibration of its parameters on that basis. Consequently, for RCTs, given the severity and consequences of the problems with AR that we have discussed, broadly recommending the use of AR with available computer software is not unlike giving a loaded gun to a child with the instruction, "Shoot all the bad guys." Based on our simulation results, the simulation results of many others, and our experiences with AR in real trials, it is clear that AR often introduces more problems than it solves. For RCTs where treatment comparison is the primary scientific goal, it appears that in most cases designs with fixed randomization probabilities and group sequential decision rules are preferable to AR scientifically, ethically, and logistically.

Acknowledgement

This research was supported by NIH/NCI grant RO1 CA 83932.

Bibliography

1. P.C. Austin. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29:21372148, 2010.

2. D. Azriel, M. Mandel. and Y. Rinott. The treatment versus experiment dilemma in dose-finding studies. *J. Statistical Planning and Inference* 141:2759-68, 2011.

3. D.A. Berry and S.G. Eick. Adaptive assignment versus balanced randomization in clinical trials: A decision analysis. *Statistics in Medicine*, 14:231246, 1995.

4. T. M. Braun. The bivariate continual reassessment method: extending the CRM to phase I trials of two competing outcomes. *Contemporary Clinical Trials* 23:240256, 2002.

5. B.P. Carlin, W. Zhong and J.S. Koopmeiners. Discussion of Small-sample behavior of novel phase I cancer trial designs by Assaf P Oron and Peter D Hoff. *Clinical Trials* 10:81-85, 2013.

6. R. Chappell and T. Karrison. Letter to the editor. *Statistics in Medicine*, 26:3046-3056, 2007.

7. Y.K. Cheung. *Dose Finding by the Continual Reassessment Method*. Chapman & Hall/CRC, 2011.

8. Y.K. Cheung, L.Y.T. Inoue, J.K. Wathen and P.F. Thall. Continuous Bayesian adaptive randomization based on event times with covariates. *Statistics in Medicine*, 25:55-70, 2006.

9. J. Cornfield, M. Halperin and S.W. Greenhouse. An adaptive procedure for sequential clinical trials. *J American Stat Association* 64:759770, 1969.

10. M.A. Hernan, B. Brumback and J.M. Robins. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology.* 5:561-570, 2000.

11. F. Hu and W. F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials.* Wiley Series in Probability and Statistics, Hoboken, 2006.

12. T.G. Karrison, D. Huo, and R. Chappell. A group sequential, response-adaptive design for randomized clinical trials, *Controlled Clinical Trials*, 24:506-22,2003.

13. E. L. Korn and B. Freidlin. Outcome-adaptive randomization: Is it useful? *J Clin Oncology* 29:771-776, 2011.

14. Lee, J.J., Chen, N and Yin, G. Worth adapting? Revisiting the usefulness of outcome-adaptive randomization. *Clinical Cancer Research* 18:4498-4507, 2012.

15. R.G. Maki, J.K. Wathen, M.L. Hensley, S.R. Patel, D.A. Priebat, S. Okuno, B. Samuels, M. Fanucchi, D. Harmon, S.M. Schuetze, D. Reinke, P.F. Thall, R.S. Benjamin, L.H. Baker. An adaptively randomized phase III study of gemcitabine and docetaxel versus gemcitabine alone in patients with metastatic soft tissue sarcomas. *J Clinical Oncology*, 25:2755-1763, 2007.

16. J. O′Quigley, M. Pepe, and L. Fisher, L. Continual reassessment method: a practical design for phase I clinical trials in cancer. *Biometrics* 46: 3348, 1990.

17. A. P. Oron and P. D. Hoff. Small-sample behavior of novel phase I cancer trial designs. *Clinical Trials* 10:63-80, 2013.

18. J.M. Robins, J. M. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics, Theory and Methods* 23:23792412, 1994.

19. W.R. Rosenberger and J.M. Lachin . The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials* 14:471-84, 1993.

20. R.S. Sutton and A.G. Barto, A.G. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press, 1998.

21. P.F. Thall and J.D. Cook. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 60:684-693, 2004.

22. P.F. Thall and H.Q. Nguyen. Adaptive randomization to improve utility-based dose-finding with bivariate ordinal outcomes. *J Biopharmaceutical Statistics* 22:785-801, 2012.

23. P.F. Thall, H.Q. Nguyen, T.M. Braun and M. Qazilbash. Using joint utilities of the times to response and toxicity to adaptively optimize schedule-dose regimes. *Biometrics* 69:673-682, 2013.

24. P.F. Thall and J.K. Wathen. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Statistics in Medicine*, 24:1947-1964, 2005.

25. P.F. Thall and J.K. Wathen. Practical Bayesian adaptive randomization in clinical trials. *European J Cancer*. 43:860-867, 2007.

26. W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of the two samples. *Biometrika* 25:285294, 1933.

27. L.J. Wei and S. Durham. The randomized play-the-winner rule in medical trials. *J American Stat Assoc* 1978;73:840843.

28. Y. Yuan and G. Yin. On the usefulness of outcome-adaptive randomization. *J Clin Oncology* 29:390-392-776, 2011.

29. Zelen M (1979). A new design for randomized clinical trials, *New England J Med* 300:1242-6.

Table 1: Operating characteristics of three Bayesian adaptively randomized designs, and two fairly randomized Bayesian designs with either continuous or group sequential (GS) comparisons. The fair randomizations were done in blocks of size 8 to obtain balanced samples. We denote $\pi_{20} = \Pr(N_A > N_B + 20)$, the probability of a large imbalance in the wrong direction, and bias = ($\hat{\theta}_B$- $\hat{\theta}_A$) - ($\theta_B^{true} - \theta_A^{true}$). In all cases, maximum $N = 200$ and $\theta_A^{true} = .25$.

		Pr(Conclusion)		Mean (2.5^{th}, 97.5^{th})		Estimates			
θ_B^{true}	Design	$B > A$	$A > B$	$N_B - N_A$	N	$\hat{\pi}_{20}$	$\hat{\theta}_A$	$\hat{\theta}_B$	$\widehat{bias}$
.25	AR(1)	.09	.09	0(-186,186)	180(17,200)	.431	.222	.222	.000
	AR(1/2)	.12	.12	0(-100,100)	164(12,200)	.327	.237	.237	.000
	AR(n/2N)	.12	.13	0(-60,60)	161(6,200)	.220	.254	.250	-.004
	Fair Contin	.12	.12	0(-1,1)	160(5,200)	.000	.258	.258	.000
	Fair GS	.024	.026	0(0,0)	196(100,200)	.000	.251	.250	-.001
.35	AR(1)	.30	.03	66(-164,188)	162(12,200)	.138	.196	.348	.052
	AR(1/2)	.41	.04	37(-50,116)	140(8,200)	.069	.208	.366	.058
	AR(n/2N)	.44	.04	21(-22,74)	135(5,200)	.028	.220	.383	.063
	Fair Contin	.46	.04	0(-2,2)	134(5,200)	.000	.223	.387	.064
	Fair GS	.34	.00	0(-2,2)	180(50,200)	.000	.243	.360	.017
.45	AR(1)	.59	.01	80(-62,184)	128(6,200)	.048	.182	.469	.087
	AR(1/2)	.80	.01	38(-7,116)	94(5,200)	.010	.188	.483	.095
	AR(n/2N)	.84	.01	16(-7,70)	85(5,200)	.002	.202	.498	.096
	Fair Contin	.86	.02	0(-2,2)	82(5,200)	.000	.202	.498	.096
	Fair GS	.86	.00	0(-2,2)	130(50,200)	.000	.235	.467	.032

Table 2: Comparison of the designs in Table 1 when $(\theta_A^{true}(n), \theta_B^{true}(n))$ drift upward together over time, with maximum drift .20 at the end of the trial, corresponding to improving prognosis of enrolled patients over time. All other parameters are as in Table 1.

		% Conclusion		Mean (2.5^{th}, 97.5^{th})		Estimates			
$\theta_B^{true}(n)$	Design	$B > A$	$A > B$	$N_B - N_A$	N	$\hat{\pi}_{20}$	$\hat{\theta}_A$	$\hat{\theta}_B$	$\widehat{bias}$
.25	AR(1)	.18	.18	0(-174,176)	165(17,200)	.426	.292	.292	.000
	AR(1/2)	.16	.16	0(-98,96)	157(11,200)	.331	.311	.312	.001
	AR(n/2N)	.14	.14	-1(-60,58)	159(6,200)	.216	.329	.331	.002
	Fair Contin	.12	.13	0(-1,1)	160(5,200)	.000	.336	.335	-.001
	Fair GS	.04	.04	0(0,0)	195(100,200)	.000	.347	.347	.000
.35	AR(1)	.47	.05	56(-142,178)	147(12,200)	.145	.244	.424	.080
	AR(1/2)	.51	.04	33(-52,110)	133(8,200)	.075	.261	.436	.075
	AR(n/2N)	.50	.04	19(-26,70)	134(5,200)	.032	.280	.453	.073
	Fair Contin	.43	.04	0(-2,2)	136(5,200)	.000	.292	.450	.058
	Fair GS	.36	.00	0(-2,2)	177(50,200)	.000	.330	.449	.019
.45	AR(1)	.74	.02	66(-72,180)	113(8,200)	.060	.214	.526	.112
	AR(1/2)	.85	.02	33(-8,110)	87(5,200)	.011	.221	.527	.106
	AR(n/2N)	.88	.01	14(-8,64)	83(5,200)	.002	.237	.541	.104
	Fair Contin	.84	.02	0(-2,2)	85(5,200)	.000	.242	.543	.101
	Fair GS	.87	.00	0(-2,2)	129(50,200)	.000	.297	.530	.033

Table 3: Comparison of the Website versus Actual treatment failure count data from the completed trial of G = gemcitabine versus G + D = gemcitabine + docetaxel for unresectable soft tissue sarcoma. Patient covariates are Z_1 = I(Leimyosarcoma) and Z_2 = I(prior pelvic radiation). For each combination of covariate pair (Z_1, Z_2), Data = Website or Actual, and chemotherapy arm = G+D or G,the tabled values are Number of failures / Number of patients.

Z_1	Z_2	Data	G+D	G	$r_{G+D}(\mathbf{Z}, \boldsymbol{\theta})$
1	1	Website	5/10	5/6	.90
		Actual	5/10	2/3	.91
1	0	Website	13/24	12/12	.96
		Actual	9/19	6/6	.52
0	1	Website	7/10	5/7	.66
		Actual	6/8	8/8	.97
0	0	Website	14/29	17/24	.71
		Actual	19/36	23/32	.79
Totals			39/73	39/49	

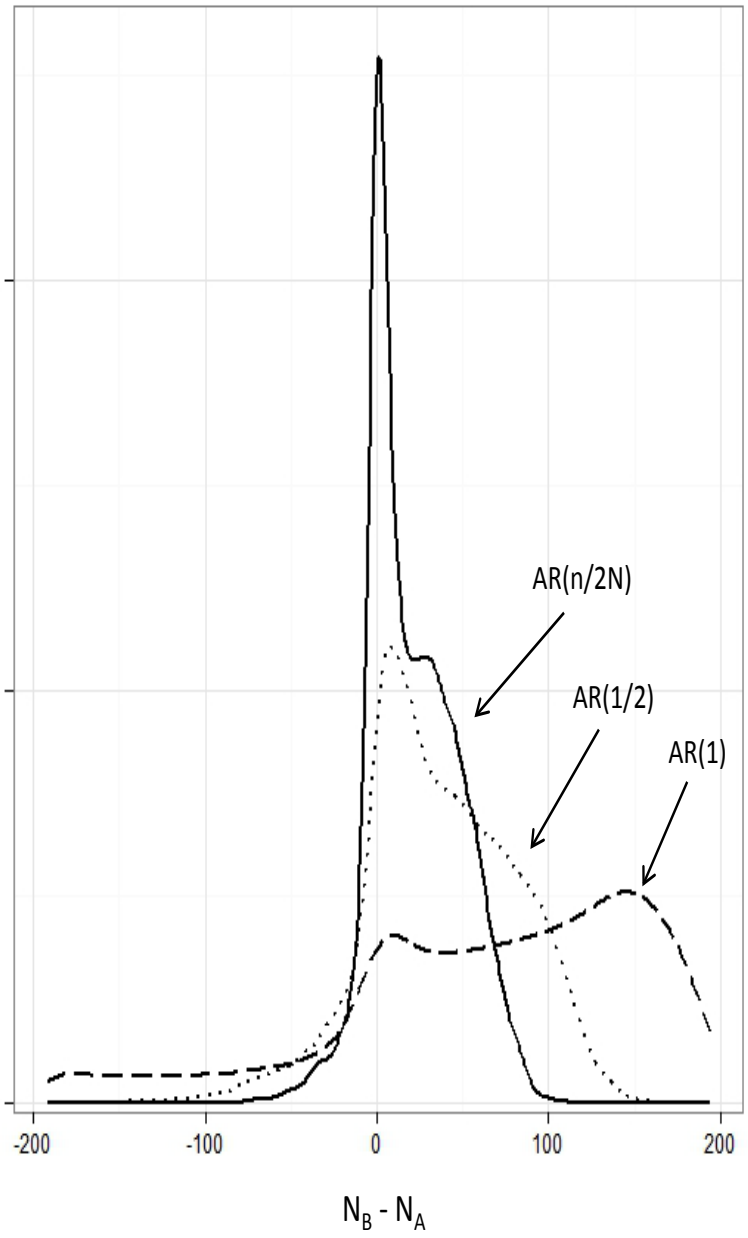

Figure 1: Distributions of the sample size difference, $N(B) - N(A)$, for AR(1), AR(1/2), and AR(n/2N) when $\theta_A^{true} = .25$ and $\theta_B^{true} = .35$, for a 200-patient trial.

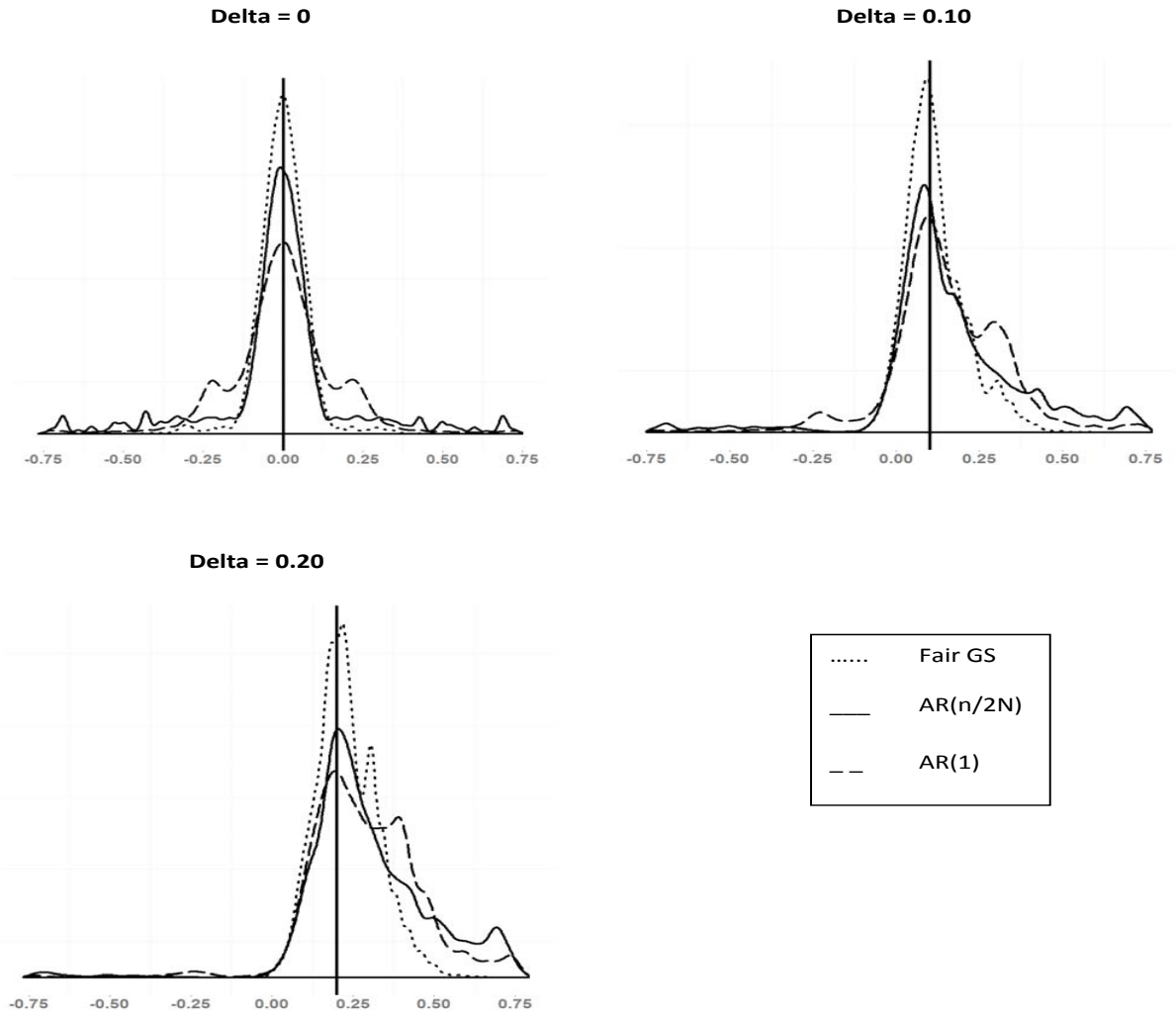

Figure 2: Distributions of the posterior means of $\Delta = \theta_B - \theta_A$ for a 200-patient trial conducted using the Fair GS, AR(n/2N), or AR(1) methods when $\theta_A^{true} = .25$ and $\theta_B^{true} = .25$, .35, or .45, corresponding to $\Delta_A^{true} = 0$, .10, or .20.