A Graphical Method for Comparing Response-Adaptive Randomization Procedures

Nancy Flournoy[1], Linda M. Haines[2], and William F. Rosenberger[3,*]

[1]*Department of Statistics, University of Missouri, Columbia, MO 65211, U. S. A.*

[2]*Department of Statistical Sciences, University of Cape Town, Rondebosch 7700, South Africa*

[3]*Department of Statistics, George Mason University, 4400 University Drive, MS 4A7, Fairfax, VA 22030, U.S.A.*

[*]wrosenbe@gmu.edu

Abstract

Response-adaptive randomization procedures have a dual goal of estimating the treatment effect and randomizing patients with a higher probability of receiving the superior treatment. These are competing objectives, and no procedure in the literature is "perfect" with respect to both objectives. For clinical trials of two treatments, we discuss metrics for comparing response-adaptive randomization procedures that can be represented graphically to compare designs. These metrics involve the simulated distribution of the set of jointly sufficient statistics for estimating functions of the unknown parameters. We explore the binary response and normal cases, and compare numerous procedures found in the literature. We distinguish between metrics of efficiency and metrics that measure ethical cost. Each of these is a function of the joint sufficient statistics. When graphed against each other, we can gauge competing designs in obtaining these competing objectives. We find that, contrary to asymptotic results, tuning parameters that affect the variability of the procedure do not have much impact in the finite case. We also find that procedures that target an optimal allocation based on ethical and efficiency considerations generally provide a better compromise design than procedures that do not.

Keywords: Adaptive designs; Binary responses; Ethics; Efficiency; Multi-objective designs; Normal response; Sufficient statistics.

1 Introduction

The literature is replete with response-adaptive randomization procedures for clinical trials; many are described in [1]. In essence, the objectives of such procedures often balance tenuously between the desire to assign more patients to the better treatment and to efficiently estimate the treatment effect. In this paper, we provide a simulation-based graphical technique to compare procedures with respect to their success in achieving these competing objectives.

The paper is organized as follows. In Section 2 we introduce the clinical trial setting used in our study, that of two treatments with binary and normal responses, together with the attendant notation. In Section 3, we describe the response-adaptive randomization procedures for which we compare treatment differences. Section 3.1 assumes binary responses, while Section 3.2 assumes normally distributed responses. In Section 4, we describe the graphical method and use it to compare the procedures described in Section 3 under various scenarios. Comments on different measures of ethical performance are made in Section 5 and we draw conclusions in Section 6.

2 Background

Reference [1] describes response-adaptive randomization procedures for two treatments, A and B, that target an optimal allocation to achieve some objective. Examples include maximizing power, minimizing expected treatment failures, minimizing total mean response or hazard, or some combination. The sample treatment allocations that attain such objectives typically are unknown functions of the parameters θ; letting $\rho(\theta)$ denote the proportion of subjects randomized to treatment A when the objective is obtained, we call it the target allocation. The procedures we discuss converge almost surely to a target allocation $\rho(\theta)$. As will be seen, randomization procedures based on ad hoc rules rather than optimality properties, such as urn models, may be expected to place more subjects on the better treatment, but do not necessarily have other optimal properties.

Let $T_1, ..., T_n$ be a randomization sequence for fixed n, where $T_j = 1$ indicates the allocation of treatment A to patient j and $T_j = 0$ the allocation of treatment B to that patient. Let $N_A(n) = \sum_{j=1}^{n} T_j$ be the (random) number of assignments to treatment A, with the number $N_B(n) = n - N_A$ thus assigned to treatment B. Let $X_1, ..., X_n$ be responses of subjects to treatment, which could be binary, ordinal, or continuous. Let $\mathcal{F}_n = \sigma(T_1, ..., T_n, X_1, ..., X_n)$, be the history of all treatment assignments and responses through n subjects. Then a response-adaptive randomization procedure is given by the probability of randomizing the jth subject to treatment A for $j \geq 1$, i.e., $\phi_j = E(T_j|\mathcal{F}_{j-1})$. Note that $1 - \phi_j$ is the probability of randomizing the jth subject to treatment B.

Let $\mathcal{L}(\theta)$ be the likelihood of the data. In response-adaptive randomization, $N_A(n)$ must belong to the set of jointly sufficient statistics, since its distribution depends on θ. Most inference procedures and efficiency measures would condition on $N_A(n)$ as fixed. Because this results in a loss of information on θ [2], it is more appropriate to conduct unconditional inference, over the joint distribution of the set of sufficient statistics [3].

Consider the binary response model, where $X_j|(T_j = 1)$ and $X_j|(T_j = 0)$ are Bernoulli distributed with parameters p_A and p_B, respectively. Denote total response on treatments A and B by $S_A = \sum_{j=1}^{n} X_j T_j$ and $S_B = \sum_{j=1}^{n} X_j(1 - T_j)$, respectively. Let p_A and p_B denote the probability of success on treatments A and B, respectively, with $q_A = 1 - p_A$ and $q_B = 1 - p_B$. For brevity, let $N_A \equiv N_A(n)$. Even though they are adaptive, the designs we consider are ancillary to the treatment outcomes [4], so the likelihood is

$$\mathcal{L}(\theta) \propto p_A^{S_A} q_A^{N_A - S_A} p_B^{S_B} q_B^{n - N_A - S_B}.$$

Note that (S_A, S_B, N_A) are jointly sufficient for estimating $\theta = (p_A, p_B)$; see [3, p.193].

Assume now that $p_A > p_B$. Then, from an ethical perspective, we may wish as many patients as possible to be assigned to treatment A, since the underlying probability of success on A is higher. Many metrics have been described in the literature to gauge the ethical objectives of a response-adaptive randomization procedure. These include the proportion of patients assigned to the better treatment, $E(N_A/n)$, and the proportion

of successes, $E((S_A + S_B)/n)$, which are often used interchangeably as a metric of achieving ethical objectives in spite of their involving completely different elements from the set of sufficient statistics. While they are correlated, they also follow very different distributions. Another metric is given by $P(S_B > S_A)$, the probability of being assigned to the inferior treatment [5].

The primary goal of a clinical trial is to estimate a treatment effect. In the binary case, we can estimate the treatment effect as

$$\hat{\eta} \equiv \hat{p}_A - \hat{p}_B = \frac{S_A}{N_A} - \frac{S_B}{n - N_A},$$

which is a function of the sufficient statistics and a natural estimator of $p_A - p_B$. Note that the distribution of the statistic $\hat{\eta}$ must be computed with respect to the joint distribution of the sufficient statistics (S_A, S_B, N_A), which has enormous complexity induced by the response-adaptive randomization procedure. In many cases, the joint distribution is asymptotically normal (e.g., [1]); however, our interest will be in the moderate sample size range, and therefore the distribution of $\hat{\eta}$ must be simulated by replicating the design under a particular set of parameters (p_A, p_B). A metric can be obtained that captures the efficiency of estimation by computing the mean squared error (MSE) of $\hat{\eta}$ in estimating η:

$$\left[E\left(\left(\frac{S_A}{N_A} - \frac{S_B}{n - N_A}\right) - (p_A - p_B)\right)\right]^2 + \text{Var}\left(\frac{S_A}{N_A} - \frac{S_B}{n - N_A}\right). \tag{1}$$

Similarly the MSE can be computed for the relative risk or odds ratio, or the logarithms of either.

Consider now the normal response model, where $X_j|(T_j = 1) \sim N(\mu_A, \sigma_A^2)$ and $X_j|(T_j = 0) \sim N(\mu_B, \sigma_B^2)$. Let $S_A = \sum_{j=1}^n X_j T_j, S_B = \sum_{j=1}^n X_j(1 - T_j), SS_A = \sum_{j=1}^n X_j^2 T_j$ and $SS_B = \sum_{j=1}^n X_j^2(1 - T_j)$ denote the sums and sums of squares of the responses for the two treatments A and B in an obvious way. The log-likelihood can be written as

$$\ln \mathcal{L}(\theta) \propto -\frac{N_A}{2} \ln \sigma_A^2 - \frac{SS_A - 2\mu_A S_A + N_A \mu_A^2}{2\sigma_A^2} - \frac{N_B}{2} \ln \sigma_B^2 - \frac{SS_B - 2\mu_B S_B + N_B \mu_B^2}{2\sigma_B^2}.$$

For this model, for fixed n, the jointly sufficient statistics for estimating $\theta = (\mu_A, \mu_B, \sigma_A^2, \sigma_B^2)$ are given by the set $(S_A, S_B, SS_A, SS_B, N_A)$.

Many response-adaptive randomization procedures have the property that $N_A(n)$ is asymptotically distributed as $N(n\rho(\theta), nv)$, where the variance v is derived for many different procedures in [1]. A lower bound for v was found in [6]. They call a procedure *asymptotically best* among all asymptotically normal procedures with mean $n\rho(\theta)$ if it attains the lower bound. This is one metric of the efficiency of a procedure, but it only makes sense asymptotically. Procedures with asymptotically normal limits may be very slow to attain them, and we found that, in some procedures, $n = 500$ was required to attain v. Ignoring a possibly larger (than asymptotic) variance for finite $N_A(n)$ may give unfounded confidence on the precision of estimators such as $\hat{p}_A \equiv S_A/N_A(n)$, depending on the variance of S_A and the covariance between S_A and N_A.

3 Response-Adaptive Randomization Procedures

Most response-adaptive designs in the literature are for binary random variables. We discuss them first. For each procedure, we specify the target allocation proportion on treatment A, that is, $\rho(\theta)$, and the probability of assigning the next subject to treatment A, that is, ϕ_j.

3.1 Binary Responses

One goal of response-adaptive randomization procedures is to target some optimal allocation $\rho(\theta)$, where $\theta = (p_A, p_B)$ (e.g., [7]). In order to maximize power for fixed sample size, for example, *Neyman allocation* is given by

$$\rho(\theta) = \frac{\sqrt{p_A q_A}}{\sqrt{p_A q_A} + \sqrt{p_B q_B}}.$$

For fixed power, the following allocation minimizes the expected number of treatment failures $N_A q_A + N_B q_B$:

$$\rho(\theta) = \frac{\sqrt{p_A}}{\sqrt{p_A} + \sqrt{p_B}}$$

and is termed the RSIHR allocation, using an acronym based on the surnames of the authors [7]. These target allocations depend on the unknown parameters θ and must

be sequentially estimated by substituting $\rho(\hat{\theta})$, where $\hat{\theta}$ is the the usual estimator of (p_A, p_B), that is, $\hat{p}_A = S_A/N_A, \hat{p}_B = S_B/N_B$.

One procedure with favorable operating characteristics and a known asymptotic distribution theory is Hu and Zhang's [8] version of Eisele's [9] *doubly-adaptive biased coin design* (DBCD) for which the probability of allocating the jth subject to treatment A is

$$\phi_j = \widehat{E}\left(T_j \middle| \mathcal{F}_{j-1}\right) = \frac{\rho\left(\widehat{\theta}_{j-1}\right)\left(\frac{\rho\left(\widehat{\theta}_{j-1}\right)}{N_A(j-1)/(j-1)}\right)^{\gamma}}{\rho\left(\widehat{\theta}_{j-1}\right)\left(\frac{\rho\left(\widehat{\theta}_{j-1}\right)}{N_A(j-1)/(j-1)}\right)^{\gamma} + \left(1-\rho\left(\widehat{\theta}_{j-1}\right)\right)\left(\frac{1-\rho\left(\widehat{\theta}_{j-1}\right)}{N_B(j-1)/(j-1)}\right)^{\gamma}}, \quad (2)$$

where $N_A(j-1)$ and $N_B(j-1) = (j-1) - N_A(j-1)$ are the numbers of subjects allocated to treatments A and B, respectively, after a total of $j-1$ subjects have been randomized; γ is a tuning parameter that affects the variability of the procedure. This procedure reduces to $\phi_j = \rho\left(\widehat{\theta}_{j-1}\right)$ if $\gamma = 0$, which is a procedure first investigated by Melfi and Page [10]. The procedure is highly variable, but choosing γ between 2 and 5 will reduce the variability substantially. Note that if $\gamma = \infty$, the procedure is deterministic unless $N_A(j-1)/(j-1) = \rho\left(\widehat{\theta}_{j-1}\right)$. Hu and Zhang's procedure can target any $\rho(\theta) \in (0,1)$, and has an asymptotically normal limit. However, it is not asymptotically best in the sense of [6].

Reference [11] describes a randomization procedure that can target any allocation $\rho(\theta)$, and is asymptotically best in the sense of [6]. It can be described as follows:

$$\phi_j = \begin{cases} \alpha\rho(\widehat{\theta}), & \text{if } N_A(j-1)/(j-1) > \rho(\widehat{\theta}); \\ \rho(\widehat{\theta}), & \text{if } N_A(j-1)/(j-1) = \rho(\widehat{\theta}); \\ 1-\alpha(1-\rho(\widehat{\theta})), & \text{if } N_A(j-1)/(j-1) < \rho(\widehat{\theta}). \end{cases} \quad (3)$$

where the parameter α controls the amount of variability and is termed an *efficient response-adaptive design*, with acronym ERADE. When $\rho(\theta) = 1/2$, the procedure reduces to a class of restricted randomization procedures described by [12] and when, in addition, $\alpha = 2/3$ to Efron's biased coin design [13]. When $\alpha = 0$, we have complete randomization.

Urn models. Urn models have been proposed as a method to generate response-adaptive randomization sequences, but, unlike the doubly adaptive biased coin design, most urn models converge to a specific target that cannot be changed according to the objectives of the trial.

The randomized play-the-winner rule [14] is a simple model that randomizes with a higher probability to treatment A if there have been more successes on A or failures on B. Specifically,

$$\phi_j = \frac{S_A(j-1) + n - N_A(j-1) - S_B(j-1)}{j-1}.$$

For the randomized play-the-winner rule, N_A/n has an asymptotically normal distribution, but it targets $\rho(\theta) = q_B/(q_A + q_B)$, which is often too skewed to the better treatment to offer efficient estimation. It can have a large variability, so it is not the asymptotically best procedure targeting $\rho(\theta) = q_B/(q_A + q_B)$

The drop-the-loser rule [15] is a more complicated rule in which there are balls of three types in an urn. The first two types, if drawn, assign the patient to either treatment A or B. The ball is replaced only if there is a success. In addition to two treatments, there is an external element, called "immigration balls", that when "drawn", do not result in a treatment assignment, but replenish the urn through adding a type A and type B ball; the immigration ball is replaced. Because of the immigration component, it is not possible to write ϕ_j in a nice form. Reference [14] shows that N_A/n is asymptotically normal and also targets $q_B/(q_A + q_B)$, but with lower variability than the randomized play-the-winner rule. The variance attains the lower bound and therefore the procedure is asymptotically best for targeting $\rho(\theta) = q_B/(q_A + q_B)$.

Reference [16] introduced a modification of Wei and Durham's [14] randomized play-the-winner rule which has

$$\phi_j = \frac{S_A(j-1) + \alpha}{S_A(j-1) + S_B(j-1) + \alpha + \beta},$$

where α and β are positive numbers that, although they do not have to be integers, can be conceptualized as the initial number of type A and B balls, respectively, in an urn. Reference [17] shows that N_A/n converges almost surely to 1 if $p_A > p_B$. Reference

[18] studies this procedure in a dose–finding context. Note that only successes change the probability of assignment. So this urn procedure is constructed to only target an ethical objective and is expected to have low power for tests of treatment effects, although one can test a treatment effect using a modified t–test based on the data accrued on both treatments prior to the asymptotic limit [19].

Other procedures. Other procedures have been proposed for response-adaptive randomization with binary responses. Reference [20] described a Bayesian procedure to compute the probability that one treatment is better than another with binary responses. Under a uniform prior distribution, the procedure yields the following formula. Given sufficient statistics (S_A, S_B, N_A) after $j-1$ patients,

$$\hat{P}(p_A > p_B) = \frac{\sum_{a=0}^{S_A} \begin{pmatrix} S_A + S_B - a \\ S_B \end{pmatrix} \begin{pmatrix} N_A - S_A + n - N_A - S_B + a \\ n - N_A - S_B \end{pmatrix}}{\begin{pmatrix} n+2 \\ n - N_A + 1 \end{pmatrix}}.$$

Reference [21] reformulated this procedure by proposing the following response-adaptive procedure:

$$\phi_j = \frac{[\hat{P}(p_A > p_B)]^{1/2}}{[\hat{P}(p_A > p_B)]^{1/2} + [1 - \hat{P}(p_A > p_B)]^{1/2}}.$$

3.2 Normal Responses

One can apply the doubly-adaptive biased coin design to any $\rho(\theta)$ in the case of continuous outcomes. If responses are normal with $\theta = (\mu_A, \mu_B, \sigma_A^2, \sigma_B^2)$, one can target allocations that involve only the variances, in order to improve efficiency of estimation. These include Neyman allocation, given by $\rho(\theta) = \sigma_A/(\sigma_A + \sigma_B)$, and the E-optimal allocation of [2], given by $\rho(\theta) = \sigma_A^2/(\sigma_A^2 + \sigma_B^2)$. Suppose we assume that $\mu_A > 0$ and $\mu_B > 0$, and that a higher mean response is undesirable, that is $\mu_A < \mu_B$. If one wishes, for fixed power, to minimize the mean total response, the optimal allocation is given by [22]:

$$\rho(\theta) = \frac{\sqrt{\mu_B}\sigma_A}{\sqrt{\mu_B}\sigma_A + \sqrt{\mu_A}\sigma_B}.$$

However, the relative magnitude of μ_B to σ_A and μ_A to σ_B may be such that $\phi_j = \rho(\hat{\theta})$ assigns more patients to the inferior treatment. The authors therefore recommend constraining ϕ_j so that it never assigns more probability to the treatment performing worse thus far.

Reference [23] proposed the following procedure that depends only on the means:

$$\rho(\theta) = \Phi\left(\frac{\mu_A - \mu_B}{T}\right),$$

where Φ is the probit function and T is a positive constant. Reference [24] generalized the binary optimal allocation for normal responses in terms of failures. This amounts to minimizing the total number of patients with response greater than a constant c. The corresponding allocation rule is

$$\phi_j = \frac{\sqrt{\Phi\left(\frac{\hat{\mu}_B - c}{\hat{\sigma}_B}\right)}\hat{\sigma}_A}{\sqrt{\Phi\left(\frac{\hat{\mu}_B - c}{\hat{\sigma}_B}\right)}\hat{\sigma}_A + \sqrt{\Phi\left(\frac{\hat{\mu}_A - c}{\hat{\sigma}_A}\right)}\hat{\sigma}_B}.$$

Reference [16] introduced a very general class of randomized treatment allocation rules for nonnegative responses. A special case of these has come to be called the randomized reinforcement urn (RRU); see [25] for more complete information. We consider a special case for two treatments with continuous random variables in which the jth subject is allocated as

$$\phi_j = \frac{N_A^*(j-1) + a}{N_A^*(j-1) + N_B^*(j-1) + a + b} \tag{4}$$

where $N_A^*(j-1)$ and $N_B^*(j-1)$ are the "numbers" of balls which have been added to the urn after $j-1$ subjects have been treated and $a > 0$ and $b > 0$ represent the initial "numbers" in the urn for treatments A and B respectively. Updating of the urn is based on the response x_j of the jth patient and more specifically, in order to avoid negative probabilities of allocation, on a positive, monotonic function $g(x_j)$ of that response. Thus, for example, if the jth subject is allocated treatment A, then $N_A^*(j) = N_A^*(j-1) + g(x_j)$ and $N_B^*(j)$ remains unchanged as $N_B^*(j-1)$. Note that for $\mu_A < \mu_B$ with treatment A prefered, $g(x_j)$ will be a monotonic decreasing

function of x_j. References [26] and [27] showed, by different methods, that ϕ_j converges to 1 as j approaches infinity and thus that the target allocation is $\rho(\theta) = 1$. The procedure is sensitive to the choice of the initial numbers a and b and the function $g(\cdot)$ as demonstrated in the simulation study of [28], but sensible recommendations are available.

4 Comparing Procedures Graphically

To summarize the ethical and estimation performance for each procedure we study, we compute the expected proportion of subjects on the inferior treatment, $E(N_B/n)$, and the "root MSE" for estimating the treatment difference, $\sqrt{MSE\,(\hat{p}_A - \hat{p}_B)}$ or $\sqrt{MSE\,(\hat{\mu}_A - \hat{\mu}_B)}$, from 1,000,000 simulated replications under different parameterizations using the programming language Gauss [29]. The number of simulations is chosen to ensure that the Monte Carlo error, that is the between-simulation variation, is such that all quantities of interest are estimated to within at least ± 0.0015. Simulations involving 10,000 replications are in fact adequate for the graphics but it is wise to be cognizant of the fact that Monte Carlo error can be high, as emphasized recently in [30]. All graphs were produced using the software R [31]. Another natural measure of ethical performance for binary responses is the total proportion of successes $(S_A + S_B)/n$. We discuss this alternative briefly in Section 5.

4.1 Binary Responses

We implement procedures targeting Neyman and RSIHR allocations with the DBCD [8] and $\gamma = 0$ and 2, denoted NM0, NM2, and RSIHR0 and RSIHR2, and procedures targeting the RSIHR allocation with the ERADE and $\alpha = 0.4, 0.5$ and 0.7, denoted ERADE4, ERADE5 and ERADE7. Recall that the DBCD with $\gamma = 0$ reduces to $\phi_j = \rho(\hat{\theta}_{j-1})$. We also implement the randomized play-the-winner rule, the drop-the-loser rule with 1 immigration ball, the randomized reinforced urn with $a = b = 3$ and $a = b = 5$ following reference [28] for normal responses ([32] used $a = b = 1$

and $a = b = 3$) and the Thall and Walthen [21] procedure denoted RPW, DL, RRU3, RRU5, and TW, respectively. In addition, for comparison purposes, we simulate selected *restricted randomization procedures*, specifically Efron's biased coin design [13], denoted BCD, and the generalized biased coin designs using:

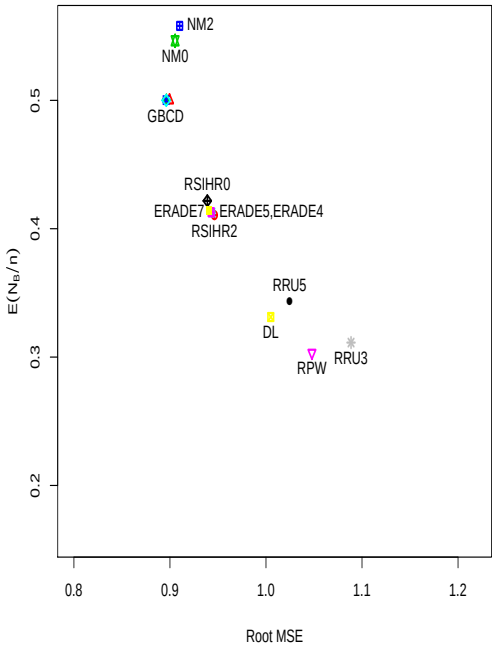
$$\phi_j = \frac{[N_B(j-1)]^\gamma}{[N_A(j-1)]^\gamma + [N_B(j-1)]^\gamma}$$

with $\gamma = 0$ (completely randomized), $\gamma = 1$ [33], $\gamma = 2$ [34] and $\gamma = 5$ [12], denoted GBDC0, GBCD1, GBCD2 and GBCD5 respectively. All these BCDs target $\rho(\theta) = 1/2$.

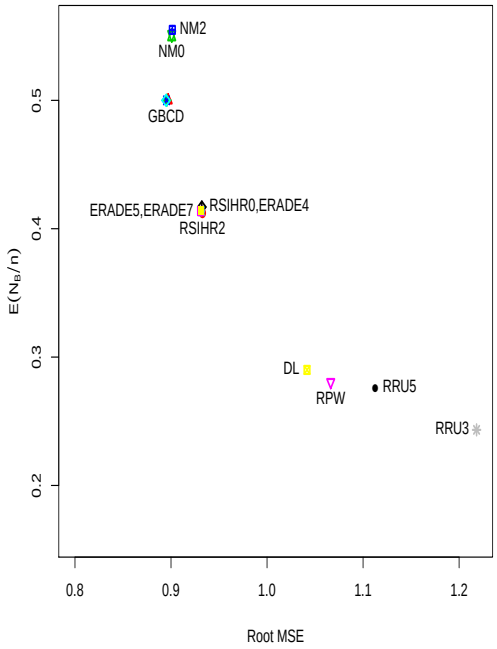
Each procedure has a "start-up" rule that assigns subjects with probability 1/2 until S_A/N_A and S_B/N_B are not equal to 0 or 1. The average number of subjects in the start-up phase is approximately 13 (see the Appendix). All procedures, with the exception of the RRU, are then begun with values of S_A, S_B, N_A and N_B obtained from the start-up phase. The RRU rule, after start-up, is implemented with the initial urn allocations $a = b = 3$ and $a = b = 5$ as specified above.

Figures 1 to 3 are plots of $E(N_B/n)$ versus root MSE, for the six parameter settings and with $n = 100$ and 200. The TW procedure was found to have substantially higher root MSE than do the other procedures. Specifically, values for the root MSE for TW ranged from 0.973 to 1.802 and for $E(N_B/n)$ from 6.3% to 28.7%. The results for TW are therefore excluded from the graphs, since otherwise the remaining procedures appear indistinguishable. In fact the TW procedure performs the best of all procedures in terms of ethics and the worst in terms of estimation. Note that the points corresponding to the BCD, GBCD0, GBCD1, GBCD2 and GBCD5 procedures almost overlap on the graphs and, for the sake of visual clarity, are therefore labeled generically as GBCD.

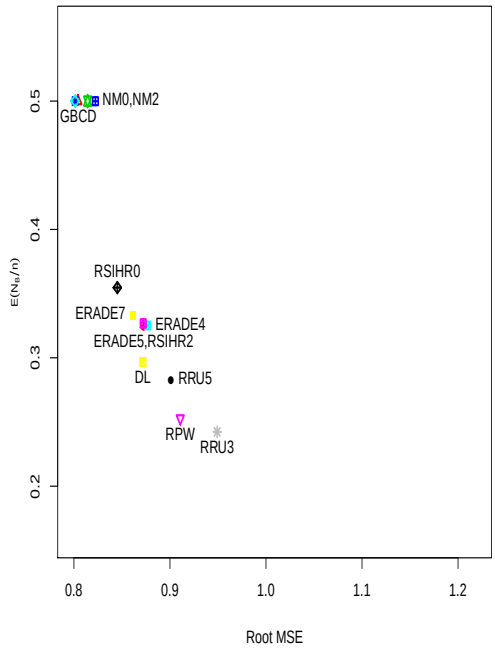
For the GBCD procedure, $E(N_B/n)$ stays right at 0.50 for all parameter combinations considered. The other restricted randomization designs all cluster nearby with $E(N_B/n)$ slightly lower in plots with $p_B = 0.20$ and higher when p_B reaches 0.60; all restricted randomization designs have relatively small root MSE. The ERADE, doubly adapted biased coin design and RSIHR procedures cluster tightly together with relatively better ethical performance and yet very similar root MSE.



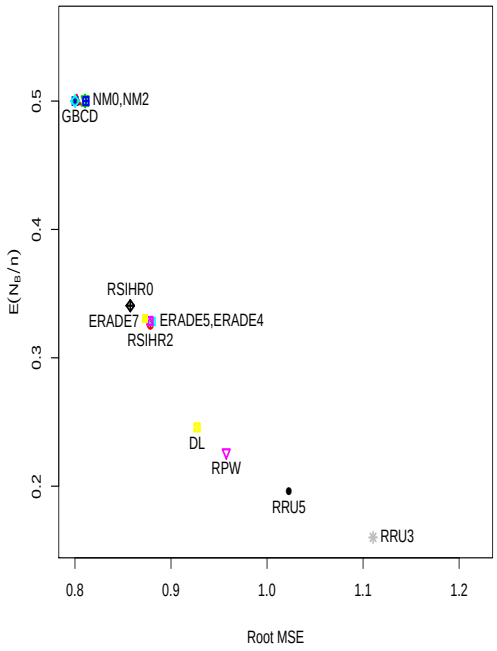
(a) $p_A = 0.8, p_B = 0.4,\ n = 100$



(b) $p_A = 0.8, p_B = 0.4,\ n = 200$

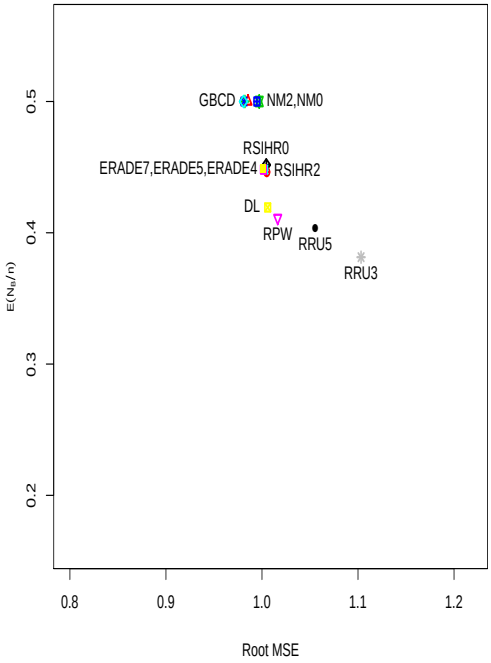


(c) $p_A = 0.8, p_B = 0.2,\ n = 100$

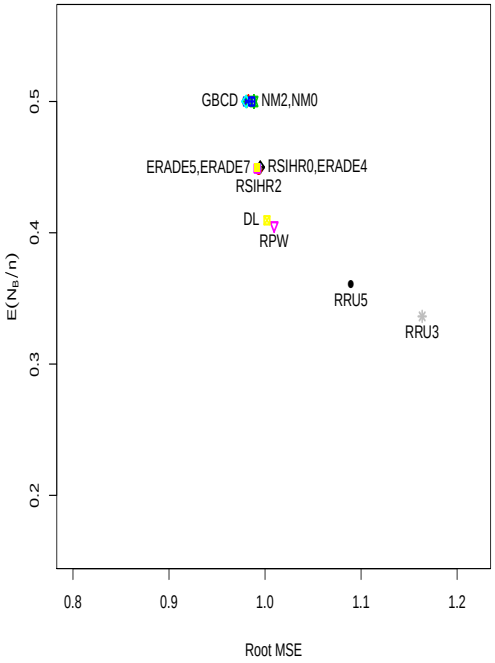


(d) $p_A = 0.8, p_B = 0.2,\ n = 200$

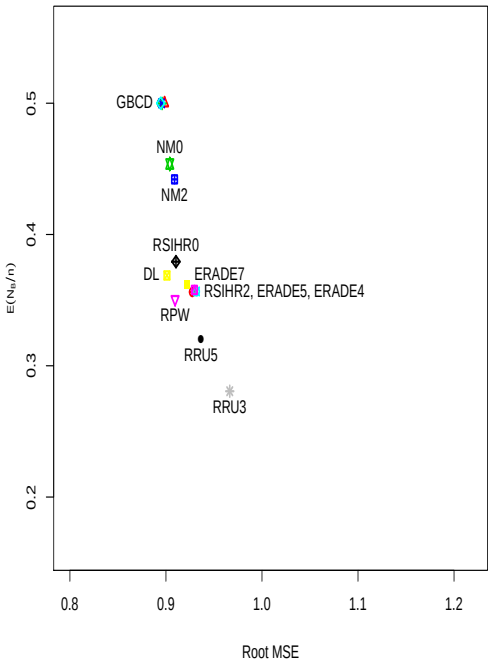
Figure 1: Ethics versus estimation with binary responses



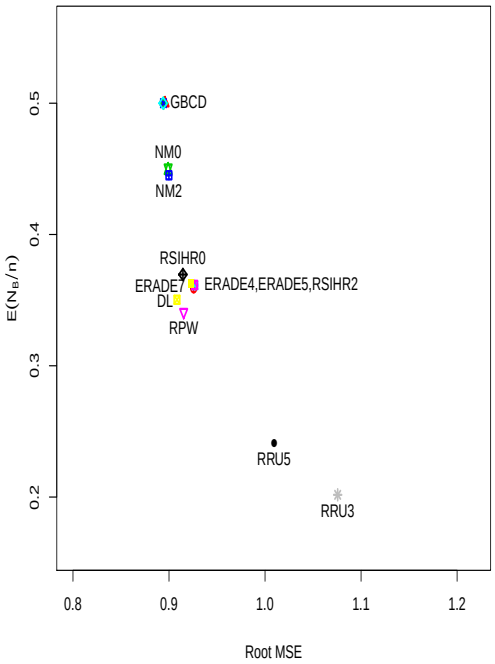
(a) $p_A = 0.6, p_B = 0.4,\ n = 100$



(b) $p_A = 0.6, p_B = 0.4,\ n = 200$

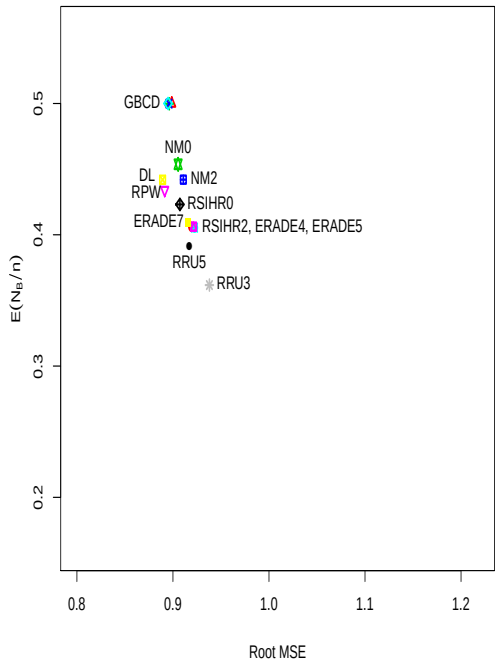


(c) $p_A = 0.6, p_B = 0.2,\ n = 100$

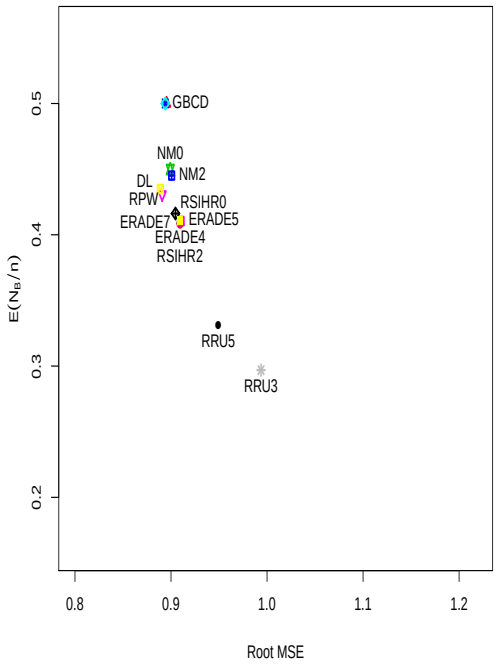


(d) $p_A = 0.6, p_B = 0.2,\ n = 200$

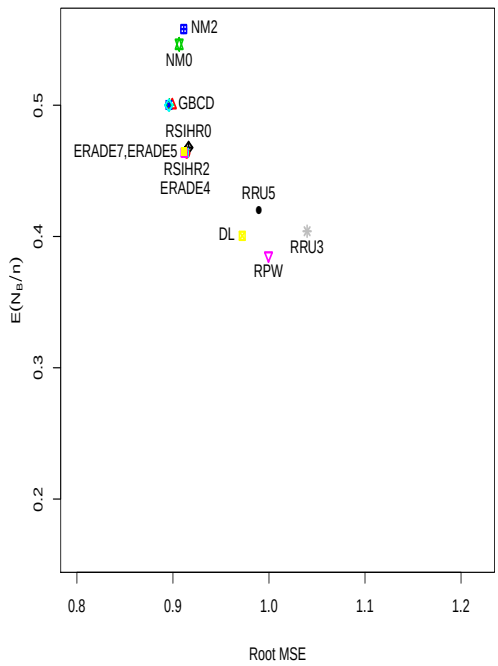
Figure 2: Ethics versus estimation with binary responses



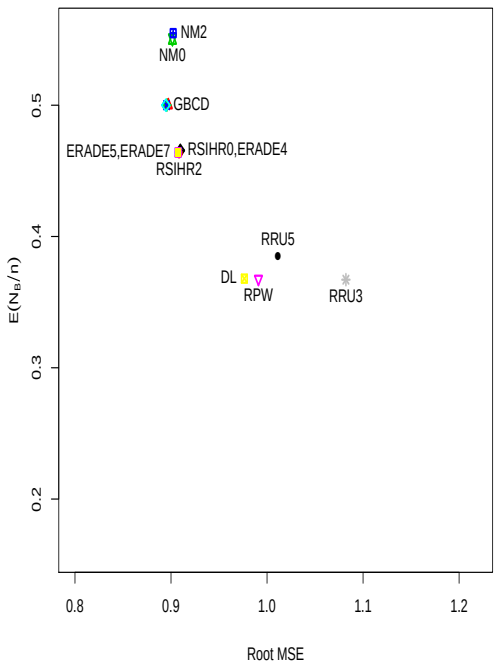
(a) $p_A = 0.4, p_B = 0.2$, $n = 100$



(b) $p_A = 0.4, p_B = 0.2$, $n = 200$



(c) $p_A = 0.8, p_B = 0.6$, $n = 100$



(d) $p_A = 0.8, p_B = 0.6$, $n = 200$

Figure 3: Ethics versus estimation with binary responses

The urn procedures tend to outperform the others in terms of ethics, and underperform in terms of estimation, with RRU3 being the extreme. RPW and DL tends to cluster closer to the ERADE procedures than the other urns in terms of estimation slightly better ethical performance, except when $p_A = 0.80$ and $p_B = 0.60$ in which case RPW and DL cluster with the RRUs. The RRUs have the lowest ethical cost and the greatest root MSE, with exceptions only in the case $p_A = 0.80$ and $p_B = 0.60$ when clustered with RPW and DL.

The ethical performance of the urn procedures improves noticeably as the sample size increases from 100 to 200, at the cost of higher MSEs, while the other procedures improve in terms of estimation and change much less in terms of ethics.

4.2 Normal Responses

We implement procedures with target allocations Neyman, [2], [22], [23] with $T = 2$ and [24] with $c = 0$, and all with $\phi_j = \rho\left(\hat{\theta}_{j-1}\right)$ and denote these NM, AG, ZR, BB and BM respectively. We follow reference [28] in implementing the RRU with an initial allocations of $a = b = 3$ and $a = b = 5$ but, since $\mu_A < \mu_B$ with treatment A preferred, we apply the reverse transformation function

$$g(x) = \begin{cases} 1 & \text{if } x < 0.1 \\ 1/x & \text{if } 0.1 \leq x \leq 10 \\ 0 & \text{if } x > 10 \end{cases}$$

We denote these procedures RRU3 and RRU5, respectively.

Each procedure has a "start–up" rule that assigns subjects to treatments with probability 1/2 until the smallest number of subjects assigned to either treatment is 2. The average number of subjects to attain this bound is 5.5 (see the Online Supplemental Material). Each response-adaptive procedure then begins with the sums and sums of squares for treatments A and B calculated using the start-up data and, in addition, the RRU begins with initial allocations $a = b = 3$ and $a = b = 5$ as indicated above.

We compare the procedures' ethical and estimation performance graphically. Figures 4 and 5 are plots of $E(N_B/n)$ versus the root MSE, $\sqrt{MSE\left(\hat{\mu}_A - \hat{\mu}_B\right)}$, for $n = 100$

and 200. We compare procedures for parameter combinations $(\mu_A, \mu_B, \sigma_A, \sigma_B) = (1, 2, 0.2, 0.2)$, $(1, 3, 0.2, 0.2)$, $(1, 3, 0.2, 0.33)$ and $(1, 3, 0.33, 0.2)$.

AG distinguishes itself as having significantly larger root MSE than the other procedures, except when $(\mu_A, \mu_B, \sigma_A, \sigma_B) = (1, 3, 0.2, 0.33)$ in which case BB's is equally large. Furthermore, for every parameter combination considered there are procedures that outperform AG in terms of ethics. BB is consistently the best procedure in terms of ethical performance.

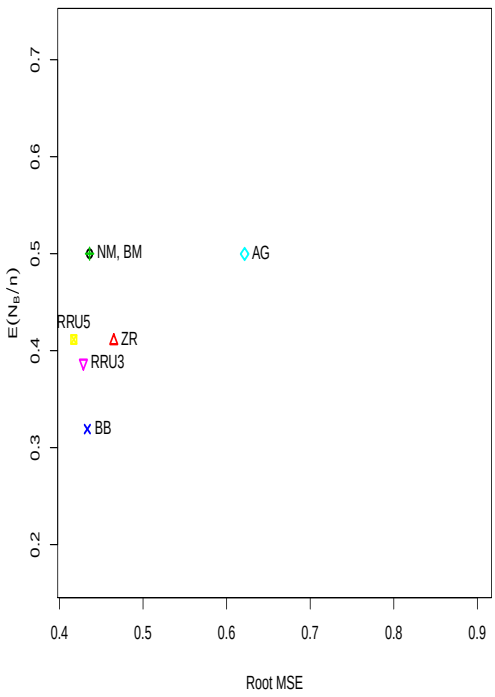
When the two groups have equal variance, $\mathrm{E}(N_B/n)$ for NM and BM is consistently at 0.5; but even with unequal variances, other procedures perform almost as well in terms of root MSE, yet much better in terms of ethics.

For ZR and the urns, root MSE is similar and only noticeably larger than for NM and BM with unequal variances. In terms of ethics, On the other hand, ZR and the urns perform consistently better than NM and BM in terms of ethics. The relative ranking of ZR and the urns in terms of ethics is inconsistent; and sometimes they cluster together and sometimes they are quite disparate.

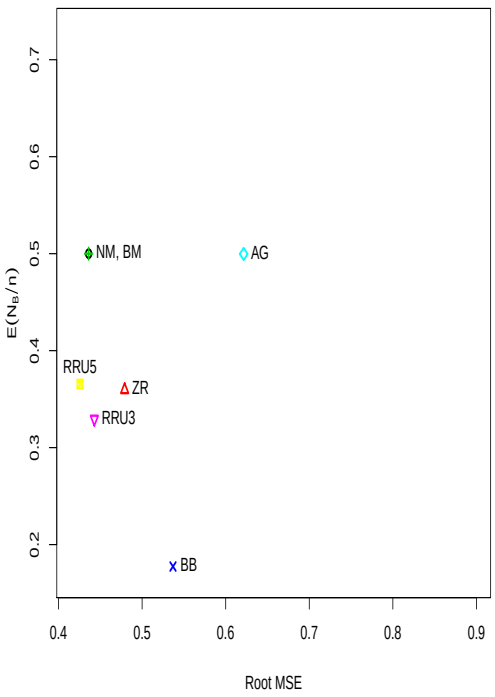
5 Measures of Ethical Performance

With binary responses, two natural measures of ethical performance are the overall proportion of successes and the proportion of subjects on the best treatment. The comparison of procedures is virtually unchanged if, versus root MSE, we plot the expected proportion of successes instead of the expected proportion of subjects on the best treatment. To see this Figure 5 plots the $E((S_A + S_B)/n)$ versus $E(N_B/n)$ for $p_A = 0.8$, $p_B = 0.4$ and $n = 100$ for each procedure. The points lie almost on a straight line. Plots for other parameters show similar high correlations. So for comparing procedures, these measures are interchangeable.

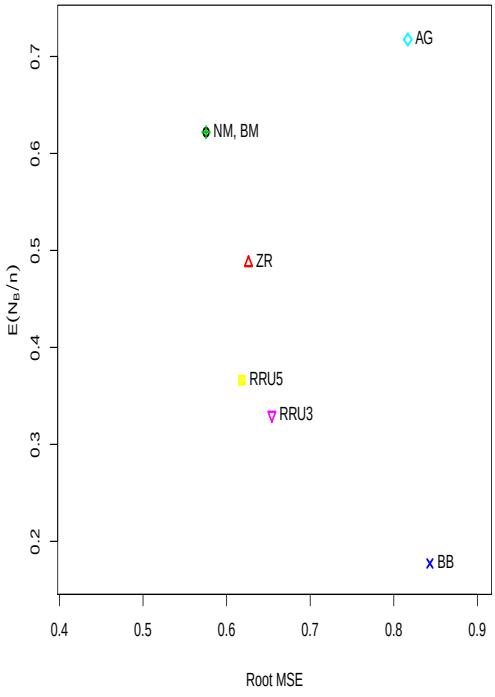
We emphasize that the similarity between the two ethics measures is shown regarding their mean values. Correlations of the two measures within procedures, instead of between procedures, are highly variable in the binary response case, with correlations ranging from -0.328 for the NM2 procedure with $n = 100$ through 0.026 for the BCD



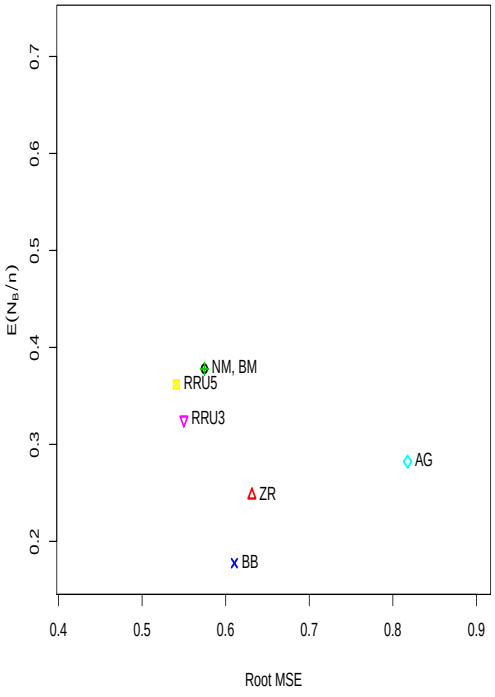
(a) $\mu_A = 1.0, \mu_B = 2.0, \sigma_A = 0.20, \sigma_B = 0.20$



(b) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.20, \sigma_B = 0.20$

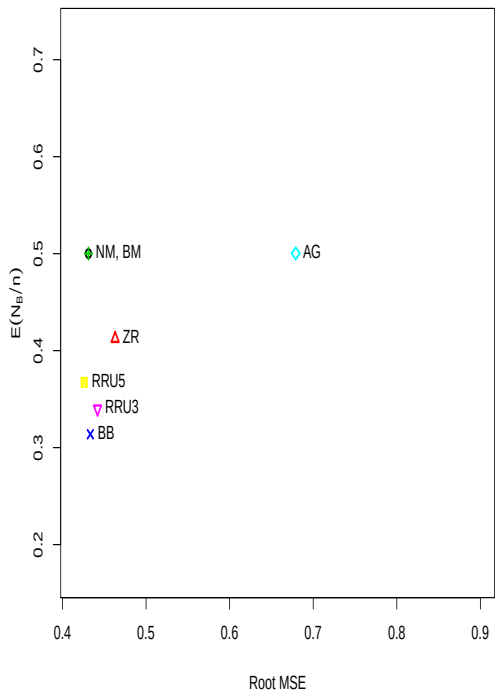


(c) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.20, \sigma_B = 0.33$

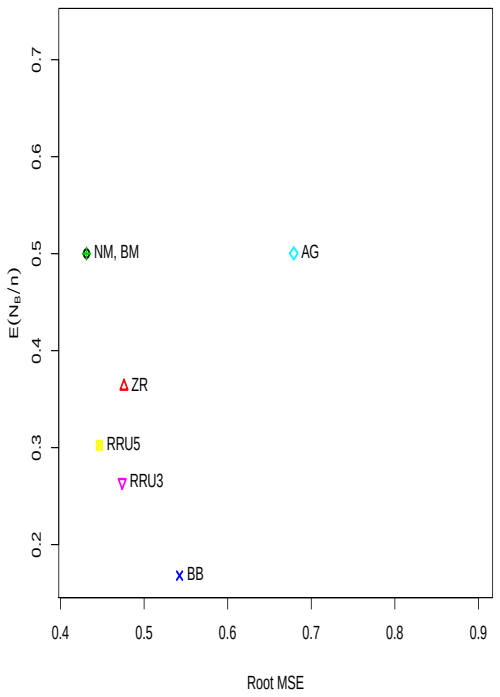


(d) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.33, \sigma_B = 0.20$

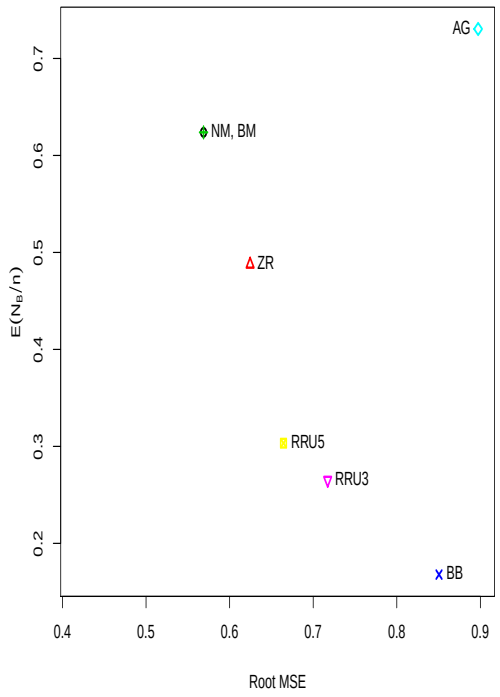
Figure 4: Ethics versus estimation with normal responses, $n = 100$



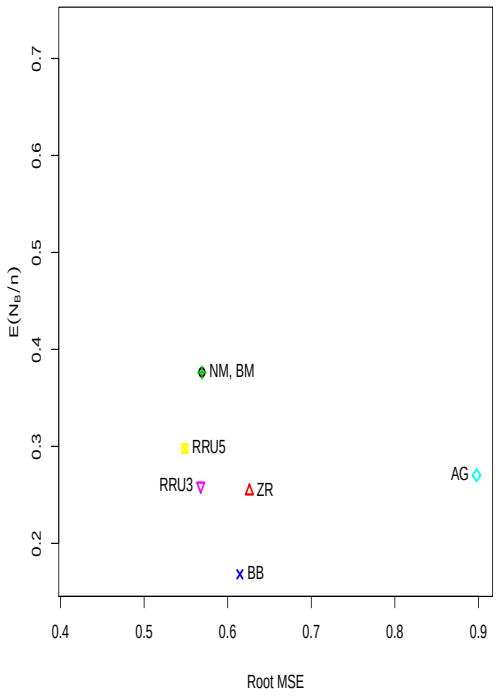
(a) $\mu_A = 1.0, \mu_B = 2.0, \sigma_A = 0.20, \sigma_B = 0.20$



(b) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.20, \sigma_B = 0.20$



(c) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.20, \sigma_B = 0.33$



(d) $\mu_A = 1.0, \mu_B = 3.0, \sigma_A = 0.33, \sigma_B = 0.20$

Figure 5: Ethics versus estimation with normal responses, $n = 200$

with $n = 200$ to 0.818 for the RRU3 with $n = 200$.

Correlations decrease with sample size for the BCD and GBCD0, whereas they increase with sample size for DL, RRU3, RRU5. For other procedures, the relationship of correlation with sample size is more complicated. The point is that within procedure, the two ethics variables are not interchangeable. Indeed, even in the best case with RRU3, 24% (i.e., $1 - 0.76^2$) of the variation in the proportion of successes is left unexplained by the proportion treated on the best treatment. So for evaluating ethics within procedures, both measures need to be considered.

6 Discussion

We have presented simple plots that compare dozens of procedures simultaneously with respect to the tradeoffs between ethics and efficiency. The plots offer immediate easy-to-view comparison as opposed to tedious tables scattered in the literature. For a fixed sample size, an investigator can determine an appropriate procedure based on the competing goals of the clinical trial.

The plots are based on metrics computed as functions of the jointly sufficient statistics. While countless simulation papers have been published on randomization procedures computing many different measures, the only measures that need to be computed and saved from simulations are the sufficient statistics. In the case of binary responses, these are (N_A, S_A, S_B), and for normal response, these are $(N_A, S_A, S_B, SS_A, SS_B)$. Once these are computed for a fixed sample size, all ethical metrics and efficiency metrics can be computed from them. We chose to illustrate only two on our graphs, but it should be clear from the previous sections that many more can be explored.

Because of the different scaling of efficiency and ethical measures, one needs to exercise caution in interpreting a 45 degree line as a measure of how well a procedure balances competing goals. In our plots, "closer to the origin" on both axes must be considered.

We have noted several important global properties of these procedures that have heretofore been examined only specifically. First, many papers in the literature have

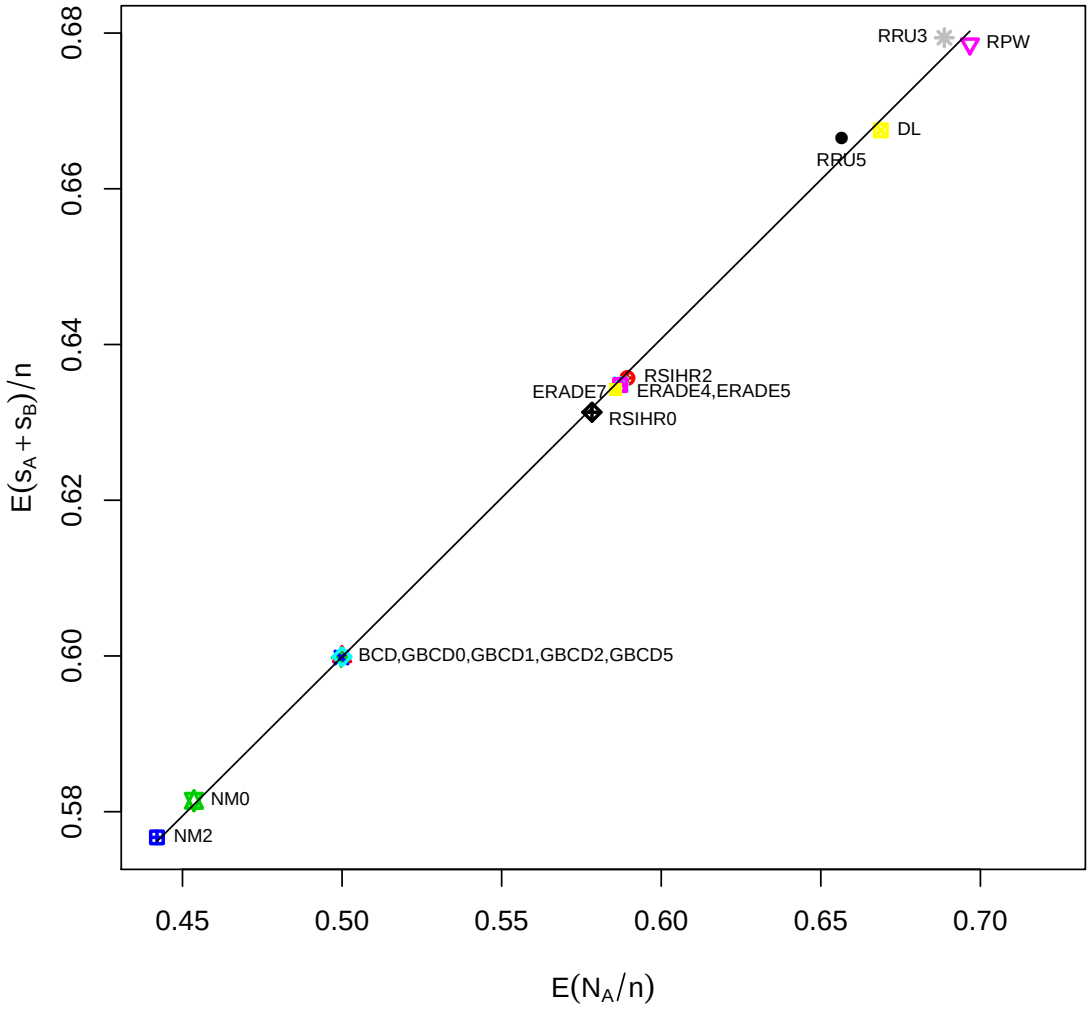


Figure 6: Measures of ethical performance for binary responses with $n = 100, p_A = 0.8$ and $p_B = 0.4$

addressed important asymptotic properties of procedures, and we have examined these procedures for realistic sample sizes found in clinical trials. While the theory of randomization has been enriched by asymptotic results, and many procedures have been described that have optimal asymptotic properties, we find that some of the theoretical results become less important in the finite sample case. For example, many procedures, such as the generalized biased coin designs, the doubly adaptive biased coin designs, and the ERADE have "tuning parameters" that affect the variability of the procedure. While these parameters tend to be important asymptotically, our plots show that many of these procedures cluster on top of each other for different parameter values, and, in fact, some plots do not follow the natural ordering of the parameter values. The ERADE, although asymptotically best, is not globally much better than the doubly adaptive biased coin design targeting RSIHR allocation in finite samples.

Second, while urn models and other types of response-adaptive randomization procedures that do not have optimal "targets" may sometimes outperform procedures that are based on optimal targets, in general the latter types of procedures capture the trade-offs in competing objectives better across all values of θ. This is a global observation that is seen throughout our graphics. Procedures based on Hu and Zhang's function (including the Zhang and Rosenberger procedure in the normal case) and the ERADE seem to balance the trade-off best. Of the procedures that do not target an optimal allocation in the binary case, the drop-the-loser rule seems to be globally better than others.

If $E(N_A/n)$, rather than $E(N_B/n)$, is plotted against root MSE then the graphs can be regarded as representing risk-return and are clearly the same, at least in principle, as the risk-return plots that are ubiquitous in finance. It is therefore tempting to take this analogy further and to consider constructing an efficient frontier for our settings by plotting the smallest root MSE that can be attained for a fixed $E(N_A/n)$ and then allowing $E(N_A/n)$ to vary. The calculation of such a frontier is however undoubtedly challenging and is left as an open problem for future research.

Acknowledgements

Professor Rosenberger is supported by a grant from the National Science Foundation under the 2009 American Reinvestment and Recovery Act. Professor Haines is supported by a grant from the National Research Foundation of South Africa and the University of Cape Town. All three authors were supported as Visiting Fellows by the Isaac Newton Institute for Mathematical Sciences at the University of Cambridge during the Programme on Design and Analysis of Experiments. They thank Professor Rosemary Bailey and the organizing committee for the invitation. Professors Flournoy and Rosenberger completed some of the research while invited to the Institute for Mathematical Sciences at the National University of Singapore. They thank Professor Louis Chen for his hospitality.

Appendix: Average Number of Trials in the Start-Up Procedures

Binary Responses

Observe that

$$\begin{aligned} P(S_A, S_B, N_A|n) &= P(S_A, S_B|N_A, n)P(N_A|n) \\ &= P(S_A|N_A, n)P(S_B|(n-N_A)|N_A, n)P(N_A|n) \\ &= \binom{n_A}{s_A} p_A^{s_A}(1-p_A)^{n_A-s_A}\binom{n-n_A}{s_B} p_B^{s_B}(1-p_B)^{n-n_B-s_B}\binom{n}{n_A}\left(\frac{1}{2}\right)^n. \end{aligned}$$

Thus, for N_{stop} the (random) number of trials in the start-up, it follows that

$$P(N_{stop} \leq n) = \sum_{n_A=2}^{n-2}\sum_{s_B=1}^{n_B-1}\sum_{s_A=1}^{n_A-1} P(S_A, S_B, N_A|n)$$

and hence, by some straightforward but tedious algebra, that

$$\begin{aligned} P(N_{stop}=n) &= P(N_{stop}\le n)-P(N_{stop}\le (n-1)) \\ &= \frac{1}{2^n}\left[(1+p_A)(1-p_A)^{n-1}+(1+p_B)(1-p_B)^{n-1}+(2-p_A)p_A^{n-1}+(2-p_B)p_B^{n-1}\right. \\ &\quad +p_A(2-p_A)^{n-1}+p_B(2-p_B)^{n-1}+(1-p_A)(1+p_A)^{n-1}+(1-p_B)(1+p_B)^{n-1} \\ &\quad -(1-p_A+p_B)(1+p_A-p_B)^{n-1}-(1+p_A-p_B)(1-p_A+p_B)^{n-1} \\ &\quad \left.-(p_A+p_B)(2-p_A-p_B)^{n-1}-(2-p_A-p_B)(p_A+p_B)^{n-1}-2\right], \end{aligned}$$

for $n=4,5,\ldots$ Thus for $(p_A,p_B)=(0.4,0.2),(0.6,0.2),(0.8,0.4)$ and $(0.8,0.6)$, $E(N_{stop})=12.3611$, for $(p_A,p_B)=(0.8,0.2)$ $E(N_{stop})=15.3056$ and for $(p_A,p_B)=(0.6,0.4)$ $E(N_{stop})=8.8571$, giving an average expected number of 12.2679.

Normal Responses

Observe that

$$P(N_A=1 \text{ and } N_B\ge 2|n-1 \text{ trials }) = \binom{n-1}{1}\left(\frac{1}{2}\right)^{n-1} \text{ for } n\ge 3$$

and vice versa. Hence, by conditioning arguments,

$$P(\min(N_A,N_B)=2|n \text{ trials }) = \binom{n-1}{1}\left(\frac{1}{2}\right)^{n-1} = (n-1)\left(\frac{1}{2}\right)^{n-1} \text{ for } n\ge 4$$

and thus the expected number of trials in the start-up procedure is 11/2.

REFERENCES

1. Hu F, Rosenberger WF. *The Theory of Response-Adaptive Randomization in Clinical Trials*. Wiley, Inc.: New York, 2006.
2. Baldi Antognini A, Giovagnoli A. On the large sample optimality of sequential designs for comparing two or more treatments. *Sequential Analysis* 2005; **24**:205–217.
3. Rosenberger WF, Lachin JM. *Randomization in Clinical Trials: Theory and Practice*. Wiley: New York, 2002.
4. Rosenberger WF, Flournoy N, Durham SD. Asymptotic normality of maximum likelihood estimators from multiparameter response-driven designs. *Journal of Statistical Planning and Inference*. 1997; **60**:69–76.
5. Robinson D. A comparison of sequential allocation rules. *Biometrika* 1983; **70**:492–495.

6. Hu F, Rosenberger WF, and Zhang L-X. Asymptotically best response-adaptive randomization procedures. *Journal of Statistical Planning and Inference* 2007; **136**:1911–1922.

7. Rosenberger WF, Stallard N, Ivanova AV, Harper CN, Ricks ML. Optimal adaptive designs for binary response trials. *Biometrics* 2001; **57**:909–913.

8. Hu F, Zhang L-X. Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Annals of Statistics* 2004; **32**:268–301.

9. Eisele JR. The doubly adaptive biased coin design for sequential clinical trials. *Journal of Statistical Planning and Inference* 1994; **38**:249–262.

10. Melfi V, Page C. Estimation after adaptive allocation. *Journal of Statistical Planning and Inference* 2000; **29**:107–116.

11. Hu F, Zhang L-X, He X. Efficient randomized-adaptive designs. *Annals of Statistics* 2009; **37**:2543–2560.

12. Smith RL. Properties of biased coin designs in sequential clinical trials. *Annals of Statistics* 1984; **12**:1018–1034.

13. Efron B. Forcing a sequential experiment to be balanced. *Biometrika* 1971; **62**:403–417.

14. Wei LJ, Durham S. The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association* 1978; **73**:840–843.

15. Ivanova A. A play-the-winner type urn design with reduced variability. *Metrika* 2003; **58**:1–13.

16. Durham SD, Yu KF. Randomized play-the leader rules for sequential sampling from two populations. *Probability in Engineering and Information Science* 1990; **26**:355–367.

17. Li W, Durham SD, Flournoy N. Randomized Polya urn designs. *Proceedings of the Biometric Section of the American Statistical Association* 1996; 166–170.

18. Durham SD, Flournoy N, Li W. Sequential designs for maximizing the probability of a favorable response. *Canadian Journal of Statistics* 1998; **3**:479–495.

19. May C, Flournoy N. Asymptotics in response-adaptive designs generated by a two-color, randomly reinforced urn. *Annals of Statistics* 2009; **37**:1058–1078.

20. Thompson WR. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 1933; **25**:285–294.

21. Thall PF, Wathen JK. Practical Bayesian adaptive randomization in clinical trials. *European Journal of Cancer* 2007; **43**:859–866.

22. Zhang L, Rosenberger WF. Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics* 2006; **62**:562–569.

23. Bandyopadhyay U, Biswas A. Adaptive designs for normal responses with prognostic factors. *Biometrika* 2001; **88**:409–419.

24. Biswas A, Mandal S. Optimal adaptive designs in phase III clinical trials for continuous responses

with covariates. In (Di Bucchianico A, Lauter H, Wynn HP, eds.). *mODa7 – Advances in Model Oriented Design and Analysis.* Physica-Verlag: Heidelberg, 2004; 51–60.

25. Flournoy N, May C, Secchi P. Asymptotically optimal response-adaptive designs for allocating the best treatment: an overview *International Statistical Review*; 2012, in press.
26. Beggs AW. On the convergence of reinforcement learning. *Journal of Economic Theory* 2005; **122**:1–36.
27. Muliere P, Paganoni A, Secchi P. A randomly reinforced urn. *Journal of Statistical Planning and Inference* 2006; **136**:1853–1874.
28. Paganoni AM, Secchi P. A numerical study for comparing two response-adaptive designs for continuous treatment effects. *Statistical Methods and Applications* 2007; **16**:321–346.
29. GAUSS Programming Language. Aptech Systems, Inc, 2011.
30. Koehler E, Brown E, Haneuse J-P A. On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician* 2009; **63**:155–162.
31. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/, 2011.
32. Flournoy N, May C, Moler, JA, and Plo, F. On testing hypotheses in response-adaptive designs targeting the best treatment. In (Giovagnoli A, Atkinson, AC, Torsney B, May C, eds.). *mODa9 – Advances in Model Oriented Design and Analysis.* Physica-Verlag: Heidelberg, 2010; 81–88.
33. Wei LJ. The adaptive biased coin design for sequential experiments. *Annals of Statistics* 1978; **6**:92–100.
34. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika* 2008; **69**:61–67.