A Response-Adaptive Covariate-Balanced Randomization for Multi-Arm Clinical Trials

Cassandra Ballou[a] and Yiyi Chen[a]
[a] OHSU-PSU School of Public Health, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, , Portland, OR 97239-3098, USA

Abstract
Randomization is a key characteristic of clinical trials which makes them the gold standard for determining treatment effectiveness. Response-adaptive randomization is desirable because it allows more patients to receive the winning treatment; however, compared to traditional equal randomization response-adaptive randomization is more likely to allow imbalance in prognostic baseline covariates. We propose a simple yet flexible randomization for multi-arm trials which marries response-adaptation and covariate-balancing designs. The operating characteristics of the proposed methods were assessed via simulation for a variety of scenarios in which values of treatment success probability and patient response delay time were varied. The newly proposed methods consistently outperformed equal randomization in terms of reducing the proportion of treatment failures for subjects and compared favorably to response-adaptive only randomization while significantly improving the balance of prognostic covariates between treatment arms.

Key Words: Clinical Trials, Multi-arm, Randomization, Response-adaptive, Covariate-balanced

1. Introduction

In trials with human subjects and particularly when treatment failure may mean serious morbidity or mortality there is a strong ethical imperative to treat subjects with the most promising treatment available. Response-adaptive randomization designs, which fall under the broader category adaptive design, allow the probability of assigning a new patient to a particular arm of a trial to be varied over the course of the trial in response to the outcomes observed for previously enrolled patients as they become available in a systematic manner which does not compromise the validity of the results of the trial [1].

Interest in response-adaptive randomization stems not only from its ethicality, but also from more logistical advantages. Also, a properly implemented response adaptive randomization can be expected to provide higher power compared to a static unequal allocation allowing a reduction in sample size. It has been suggested that this advantage is more pronounced for trials having three or more arms [2]. In addition, recruitment may be easier if patients are more willing to enroll knowing their chances of receiving the best treatment are higher.

The idea of response-adaptive design dates back as far as Thompson in 1933 [3]; however, early attempts suffered from being deterministic. For example, the “play-the-winner” rule allocated the next patient to the same treatment if the previous patient’s outcome had been

Author for Correspondence: Yiyi Chen, OHSU-PSU School of Public Health, Oregon Health & Science University, 3181 SW Sam Jackson Park Road, Mail Code: CR-145, Portland, OR 97239-3098, USA
Email: chenyiy@ohsu.edu

success and to the other treatment if the previous patient had experienced treatment failure [4]. More recently, numerous randomized response-adaptive designs, both Bayesian and Frequentist, have been proposed in the literature [1].

Among response-adaptive randomization designs using frequentist methods, urn models have predominated [1]. They are based on a simple, intuitive model in which assignment to each of the k arms is represented by k types of balls contained in an urn. A ball is drawn at random from the urn for each subject as they enroll and they are assigned to the arm corresponding to the type of ball which was drawn. The composition of balls in the urn is varied over time depending on observed successes and failures for previously assigned patients. Notable variations on this basic urn model include the randomized "play-the-winner" strategy [5,6], in which additional balls of type k are added to the urn in response to success being observed for a patient on that arm, and the "drop-the-loser" strategy [7,8], in which balls of type k are removed from the urn in response to a failure being observed for a patient on that arm. The "drop-the-loser" strategy has been shown to be superior in terms of having lower variability and, by extension higher power [7], since power is a decreasing function of randomization procedure variability [9].

A potential flaw of these response-adaptive randomization procedures is that they many have not considered imbalance in baseline covariates believed to be prognostic [10]. Particularly for trials with small to medium sample sizes, randomization alone may be inadequate to ensure important covariates are balanced across multiple treatment arms. Campbell and McPherson found that for a two arm trial as many as 1000 subjects may be required before simple randomization provides adequate covariate balance [11]. Covariate imbalances, should they occur, may introduce bias into a trial's estimates of treatment success [12]. For example, if older persons have a lower probability of treatment success regardless of treatment than younger persons and a substantially larger proportion of older persons are assigned to Treatment A, then A might wrongfully be concluded to be inferior. Imbalanced prognostic covariates can, and should, be adjusted for in a post-hoc manner at the analysis stage; however, a covariate-balanced design will improve the efficiency of the trial.

It is important to distinguish covariate-balanced randomization from another adaptive approach involving baseline prognostic factors: covariate-adjusted randomization. Covariate-balancing seeks to assign patients with a certain value of some baseline covariate more evenly across all treatment arms for the purpose of reducing bias in the results of the trial. In contrast, the purpose of covariate-adjusted randomization is to assign more subjects to the best treatment for them by increasing the probability of assigning a new subject to a given treatment arm in response to subjects with similar baseline covariate profiles previously assigned to that treatment achieving treatment success [12]. It should become apparent that if there are indeed significant differences in which treatment has the best probability of success based on a given covariate then covariate-adjusted randomization will result in a greater amount of imbalance between treatment arms in regard to that covariate. Covariate-adjusted randomization is appropriate for prognostic factors where an interaction between treatment and covariate is expected. For example, treatment A has a higher true probability of success for persons with genetic marker A, while treatment B has a higher true probability of success for subject with genetic marker B. In contrast, covariate-balanced randomization is appropriate for prognostic factors where the effect of the covariate would be expected to be consistent across treatments. Our current discussion will be restricted to covariate balancing.

Simple stratification has been the traditional approach to covariate balancing; however, prognostic score based randomization offers a more versatile approach because it allows for balancing on continuous covariates and a larger total number of covariates [10]. Pre-stratification is adequate when only a small number of binary and/or categorical baseline variables (resulting in only a few strata) are of interest, but if balance across many categorical variables or continuous variables is desired achieving marginal balance on each covariate between treatment arms becomes impractical if not impossible. Provided that achieving balance on the baseline covariates is only of interest in so far as they are predictive of the primary outcome, a potential alternative to pre-stratification based methods is to balance on a prognostic score, a linear combination of the covariates predictive of the outcome.

Covariate-balanced randomization was first proposed by Taves in 1974 [13]. Taves minimization method, so-named for its intent to minimize differences between groups in regard to important baseline covariates, suffered from the same short-coming of determinism as early attempts at response-adaptive randomization; however, randomized versions of the minimization method from Pocock and Simon [14] and Wei [15] soon followed. Although they have been known for some time, Scott et al. found in their review of the literature that minimization methods of covariate balance are still rarely employed with only 4% of randomized trials published in the *Lancet* and the *New England Journal of Medicine* in 2001 reporting use of this method [16]. The authors cite the perception of additional administrative burden and uncertainty about the proper analysis techniques to employ in evaluating the results of a trial randomized in this way as major barriers to wider use [16].

Compared to the minimization method, the prognostic score approach has two major advantages. Firstly, like stratification methods, minimization requires the categorization of continuous variables which poses a challenge if optimal cutoff values are unknown. Secondly, minimization methods fail in the presence of interactions between covariates introducing larger alpha errors, while the prognostic score approach can easily accommodate interaction terms in the logistic regression model. The major disadvantage to the prognostic score approach; however, is that due to being model-based it may be less robust.

We will consider both stratification and a prognostic score approach based on the logistic regression model as proposed by Yuan [10] to covariate balancing for a response-adaptive clinical trial with a binary outcome. The goal of this current work is to provide a method of response-adaptive covariate-balanced randomization suitable for a three arm superiority trial. Possible examples might include two pharmaceutical agents with a placebo control arm or a behavioral intervention with two control arms, an active and a passive.

The remainder of this article is laid out as follows. In Section 2, we propose several novel designs which combine a response-adaptive and covariate-balanced approach to randomization. In Section 3, we evaluate the operating characteristics of our proposed designs via simulation. In Section 4, we conclude with a brief discussion.

2. Methods

Two methods of response-adaptive randomization were considered: the generalized drop-the-loser rule as presented by Sun et al. [8] and the Ridit scoring based method presented in Bandyopadhyay & De [17] and their performance in combination with simple

stratification or prognostic scoring for covariate balance was assessed. Three criteria were used in assessing the performance of a given method: the proportion of total subjects assigned to the best treatment with a higher proportion being superior, the proportion of treatment failures experienced by subjects in the trial with a low value being superior, and the imbalance in prognostic scores between the treatment arms at the conclusion of the trial which was measured using Kolmogorov–Smirnov (KS) statistics for which smaller values indicate better balance between treatment arms.

For both the Ridit and GDL Urn models, probabilities of treatment success, $\hat{p}_k$, were estimated as follows,

$$\hat{p}_k = \frac{S_k+0.5}{N_k+1},\ k = A, B, C \tag{1}$$

where N_K is the number patients assigned to one of the three treatment arms, A, B, or C, and S_K is the number of successes observed among those N_K subjects. The probabilities of treatment failure, $\hat{q}_k$, are simply the complements of $\hat{p}_k$.

$$\hat{q}_k = 1 - \hat{p}_k,\ k = A, B, C \tag{2}$$

The algorithm for randomization using the Ridit method for three treatment arms is as follows,

$$\begin{aligned} R_A &= \frac{1}{3} + \frac{1}{6}(2p_A - p_B - p_C) + \frac{q_B}{6}(p_A - p_C) + \frac{q_C}{6}(p_A - p_B) \\ R_B &= \frac{1}{3} + \frac{1}{6}(2p_B - p_A - p_C) + \frac{q_A}{6}(p_B - p_C) + \frac{q_C}{6}(p_B - p_A) \\ R_C &= \frac{1}{3} + \frac{1}{6}(2p_C - p_B - p_A) + \frac{q_B}{6}(p_C - p_A) + \frac{q_A}{6}(p_C - p_B) \end{aligned} \tag{3}$$

where R_k is the probability of assigning a new patient to treatment arm k at a given point in the trial, p_k and q_k are defined as in Equations 1 and 2.

The GDL urn model proposed by Sun et al. [8] utilizes k+1 types of balls present in the urn. The additional type are termed immigration balls and when an immigration ball is randomly drawn the composition of balls, X_k, in the urn is updated as follows prior to the new subject being randomized,

$$\begin{aligned} X_I &= X_{I0},\ where\ X_{I0} > 0 \\ X_A &= X_{A0} + \frac{\frac{1}{q_A}}{\frac{1}{q_A}+\frac{1}{q_B}+\frac{1}{q_C}} \\ X_B &= X_{B0} + \frac{\frac{1}{q_B}}{\frac{1}{q_A}+\frac{1}{q_B}+\frac{1}{q_C}} \\ X_C &= X_{C0} + \frac{\frac{1}{q_C}}{\frac{1}{q_A}+\frac{1}{q_B}+\frac{1}{q_C}} \end{aligned} \tag{4}$$

where q_K is defined as in Equation 2 and X_{K0} is the number of balls of type k present in the urn prior to the most recent immigration ball draw and X_{I0} is the number of immigration balls which is a constant over of the course of the trial. This makes the equivalent expressions to R_K for the GDL Urn model, U_K, the probability of assignment to a given treatment arm at a given time, as follows,

$$U_A = \frac{X_A}{X_A+X_B+X_C}$$
$$U_B = \frac{X_B}{X_A+X_B+X_C} \quad (5)$$
$$U_C = \frac{X_C}{X_A+X_B+X_C}$$

When prognostic scoring was used as the method of covariate-balancing, the probability of assignment to a given treatment arm, R_K or U_K, was weighted by a factor of π_K as follows,

$$\pi_k = \begin{cases} \phi & if\ assigment\ to\ k\ minimizes\ \sum KS \\ \frac{1-\phi}{k-1} & if\ assigment\ to\ k\ does\ not\ minimize \sum KS \end{cases} \quad (6)$$

where ϕ is a constant satisfying $1/k < \phi \leq 1$.

In the simple stratification approach to covariate-balancing $\hat{p}_k$ was calculated separately for each strata with the estimate being based only on data from subjects belonging to the same stratum as the new subject to be randomized.

3. Results

Treatment A was fixed as the best treatment in terms of true probability of treatment success without loss of generality. Three potential relationships between the true probabilities of success for the three treatment arms were considered. In the first case, the true probabilities of success for both Treatment B and Treatment C were set to be equal and low (both 0.3) and the probability of success for Treatment A was varied from 0.3 to 0.9 in increments of 0.1. In the second case, there was a small difference between the true probability of success for Treatment B and that for Treatment C (0.4 and 0.3 respectively) with B being superior to C and the probability of success for Treatment A was varied from 0.5 to 0.9 in increments of 0.1. In the third case, there was a larger difference between the true probability of success for Treatment B and that for Treatment C (0.5 and 0.3 respectively) with B still being superior to C and the probability of success for Treatment A was varied from 0.6 to 0.9 in increments of 0.1. In the figures which follow, these three scenarios are presented left to right. Results presented are for a three arm study with a binary outcome and a moderate sample size of 65 patients with 15 of those being equally randomized to provide adequate initial estimates of $\hat{p}_T$ and 50 being adaptively randomized. Results involving the prognostic scoring method assume a standard normal distribution of prognostic scores. Results involving stratification assume a single equally distributed binary prognostic covariate. This can be thought of as equivalent to dichotomizing the prognostic score variable and stratifying based on prognostic score [18]. All simulation studies were conducted with 1,000 repetitions using R software.

All four designs, Stratified GDL Urn model, GDL Urn model with prognostic scoring, Stratified Ridit, and Ridit with prognostic scoring, significantly outperformed equal randomization in terms of lowering the proportion of total treatment failures (see Figure 1).

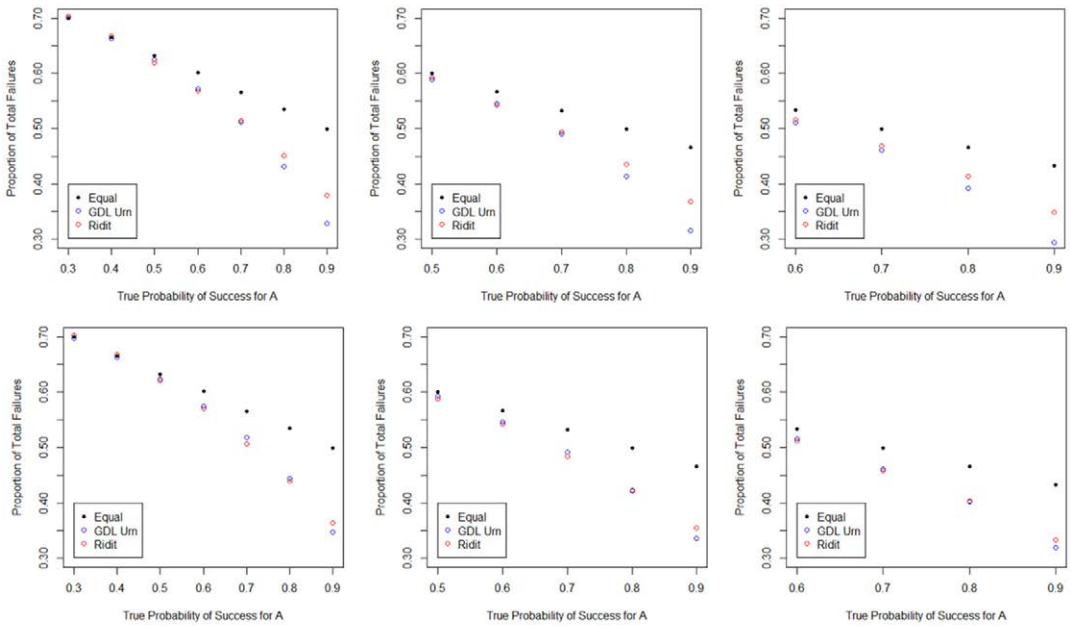

Figure 1: Comparison of proportions of total treatment failures for equal randomization versus Ridit and GDL Urn with prognostic scoring (top row) and stratification (bottom row).

There was also little or no increase in the proportion of total treatment failures observed as a result of incorporating a covariate-balancing component compared to response-adaptive randomization alone regardless of whether the baseline covariates chosen to be balanced were predictive of the outcome (see Figure 2.)

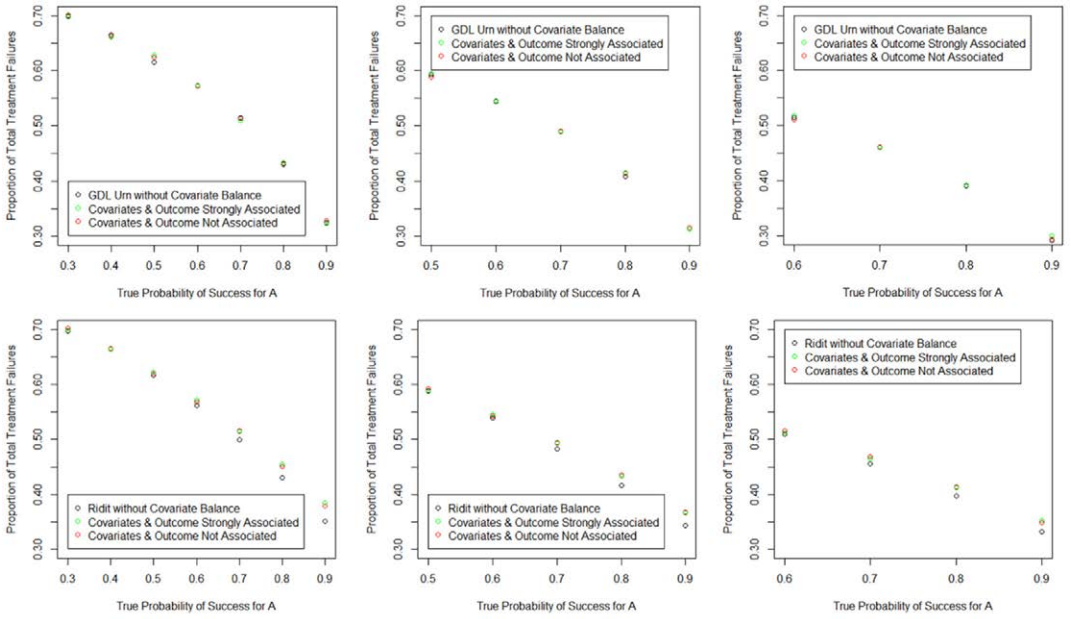

Figure 2: Comparison of proportions of total treatment failures for GDL Urn (top row) and Ridit (bottom row) response-adaptive randomization alone versus response-adaptive randomization with prognostic scoring

In terms of proportion of subjects assigned to the best treatment, when prognostic scoring was used as the method of covariate balancing, the Ridit method performed as well or better than the GDL Urn model when the probability of treatment success for the best treatment was below 0.7; however, at higher values of probability of treatment success for the best treatment the GDL Urn model dominated; however when stratification was used there was no advantage to the GDL urn model until the probability of treatment success for the best treatment reached 0.9 (see Figure 3).

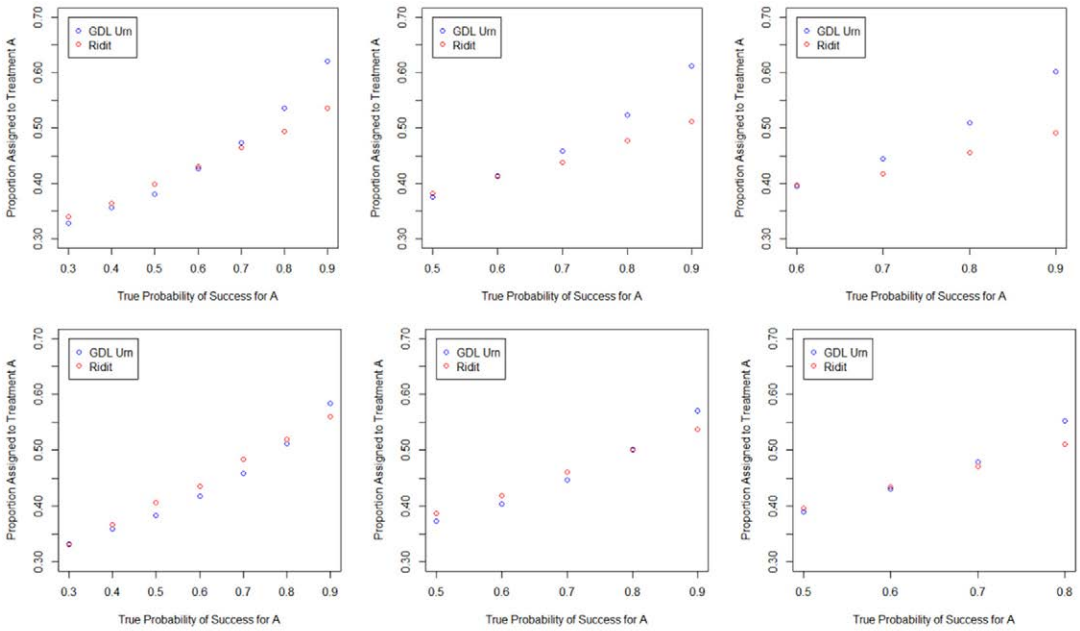

Figure 3: Comparison of proportions of subjects assigned to the treatment with the highest true probability of success for Ridit versus GDL Urn with Prognostic Scoring (top row) and Stratification (bottom row)

It is important to note that due to the way the composition of balls (and; therefore, the probability of assignment to each treatment arm) in the urn model is updated there is a built-in delay between ascertainment of the outcome for a patient and the incorporation of that information into the assignment of new subjects entering the trial. This has the important implications that the GDL urn model would be expected to perform less well compared to the Ridit method as the rate of subject recruitment or the delay in obtaining patient outcomes increases. Although, in general, the benefit of any response-adaptive randomization strategy will be reduced if relatively few patient outcomes will become available before the conclusion of recruitment, this effect may be compounded for the GDL Urn model. This intuition was found to be correct. When there was a significant delay in knowledge of patient outcomes (results shown are for a delay of 25 patients), the advantage to the GDL Urn model at high values of probability of treatment success for the best treatment observed with prognostic scoring was reduced in magnitude and the Ridit method more consistently outperformed the GDL Urn at values below 0.7 (see Figure 4).

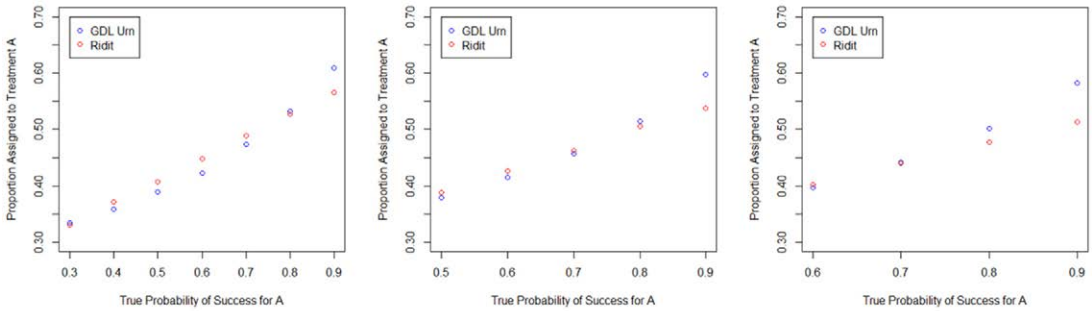

Figure 4: Comparison of proportions of subjects assigned to the treatment with the highest true probability of success for Ridit versus GDL Urn with prognostic scoring with a delay of 25 subjects between randomization and knowledge of subject outcome

When simple stratification was used as the method of covariate-balancing, the Ridit method outperformed the GDL Urn method in terms of achieving better covariate balance while covariate-balancing via prognostic scoring favored the GDL Urn method over the Ridit in terms of covariate balance (see Figure 5).

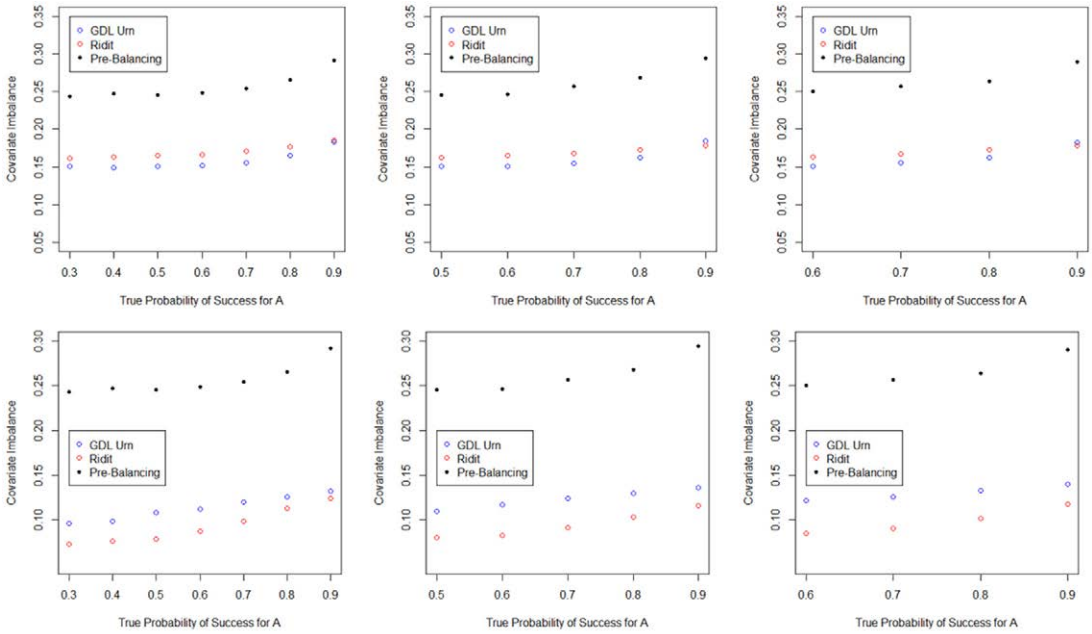

Figure 5: Comparison of average KS statistic (a measure of baseline covariate imbalance) for Ridit versus GDL Urn with prognostic scoring (top row) and stratification (bottom row)

4. Discussion

It was found that, on average, all four designs, Stratified GDL Urn model, GDL Urn model with prognostic scoring, Stratified Ridit, and Ridit with prognostic scoring, significantly outperformed equal randomization in terms of lowering the proportion of total treatment failures. In addition, there was little or no increase in the proportion of total treatment failures as a result of incorporating a covariate-balancing component compared to response-adaptive randomization alone regardless of whether the baseline covariates chosen to be balanced were predictive of the outcome while achieving significant improvement in covariate balance. The choice of the Ridit or GDL Urn model as the better

method of response-adaptive randomization was found to depend on the choice of covariate-balancing method, the true probability of success for the most successful treatment, and the delay between patient randomization and knowledge of the treatment outcome. Prognostic scoring, true probability of success for the most successful treatment above 0.6, and short delays favored the GDL Urn. Stratification, true probability of success for the most successful treatment at or below 0.6, and long delays favored the Ridit method (see Table 1).

Table 1: Factors Influencing Choice of Ridit or GDL Urn for Response-Adaptive Design

	True Probability of Success for Most Successful Treatment	Covariate Balancing Method	Delay in Availability of Treatment Outcomes[1]
Favors GDL Urn	>0.6	Prognostic Scoring	≤1/5
Favors Ridit	≤0.6	Stratification	>1/5

1 Defined as the faction of the total number of patients in the trial whose outcome is still unknown when the last patient is randomized

Because, in reality, only one study can be performed, the variability in, as well as the average values of, the three performance criteria (proportion of total subjects assigned to the best treatment, the proportion of treatment failures, and the covariate imbalance between treatment arms) is of interest. When stratification was used as the method of covariate balancing, the GDL Urn showered lower variability than the Ridit method; however, when prognostic scoring was used, the Ridit method showed lower variability for all three measures when the true success probability of the best treatment was 0.7 or higher and the GDL Urn showed lower variability when the true success probability for the best treatment was lower (see Table S1 in the supplemental materials). Differences between the two methods of response-adaptive randomization can be attributed to the inherent differences in the algorithms by which U_k and R_k values are calculated. Of particular note is the fact that U_k updates at random intervals in response to the drawing immigration balls while R_k is consistently updated after each patient outcome is observed.

Ideal values for the parameters n_0 and ϕ were also investigated. Values of 3 or 5 were considered for n_0 (see Table S2 in the supplemental materials.). Although, in most cases, estimates of the true treatment success probabilities were found to be stable with only 3 patients equally randomized to each treatment arm, we have chosen and generally recommend the more conservative value of 5. If limiting sample size is a major concern and delays in outcome availability are expected to be short, such as in an emergency medicine setting [19], 3 will be adequate (see Table S2 in the supplemental materials). For the prognostic scoring parameter, ϕ, values considered were 1/2 and 2/3 with 2/3 being selected as optimal because it yielded a significant improvement in covariate balance with only minor losses in terms of proportion of patients assigned to the best treatment and proportion of total treatment failures compared to 1/2 (see Table S3 in the supplemental materials). If covariate-balance is of only secondary concern 1/2 is the superior choice. An example for which this might be the case is that of a study with a larger total number of subjects than that considered here, as covariate imbalance decreases with increasing sample size even for a response-adaptive design with no covariate-balancing component [12].

In conclusion, a response-adaptive covariate-balanced randomization provides a significant gain in ethicality over an equal randomization and also a significant gain in

efficiency over a response-adaptive only randomization while sacrificing little in terms of ethicality compared to the response-adaptive only randomization.

References

1. Rosenberger, W. F., Sverdlov, O., & Hu, F. Adaptive randomization for clinical trials. *Journal of Biopharmaceutical Statistics.* 2012; 22: 719–736.
2. Berry, D.A. Adaptive clinical trials: the promise and the caution. *Journal of Clinical Oncology.* 2010; 21: 606–9.
3. Thompson, W. R. On the likelihood that one unknown probability exceeds another in the view of the evidence of the two samples. *Biometrika*. 1933; 25: 275–294.
4. Zelen, M. Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association.* 1969; 64: 131–146.
5. Wei, L. J., Durham, S. The randomized play-the-winner rule in medical trial. *Journal of the American Statistical Association.* 1978; 73: 840–843.
6. Andersen, J., Faries, D. Tamura, R. A randomized Play-the-Winner design for multi-arm clinical trials. *Communications in Statistics Theory and Methods.* 1994; 23: 309–323.
7. Ivanova, A. A play-the-winner type urn model with reduced variability. *Metrika.* 2003; 58: 1–13.
8. Sun R, Cheung AH, Zhang L-X. A generalized drop-the-loser rule for multi-treatment clinical trials. *Journal of Statistical Planning and Inference.* 2007; 137: 2011–2023.
9. Hu, F. Rosenberger, W. F. Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association.* 2003; 98: 671–678.
10. Yuan Y, Huang X, Liu S. A Bayesian response-adaptive covariate-balanced randomization design with application to a leukemia clinical trial. *Statistics in medicine.* 2011; 30: 1218–1229.
11. Campbell, M. K., McPherson, G. C., Simple randomisation or minimisation: the impact of trial size. *Controlled Clinical Trials.* 2001; 22-87.
12. Rosenberger, W. F., Sverdlov, O. Handling Covariates in the Design of Clinical Trials. *Statistical Science.* 2008; 23: 404–419.
13. Taves, D. R. Minimization: A new method of assigning patients to treatment and control groups. *Journal of Clinical Pharmacology Therapy.* 1974; 15:443–453.
14. Pocock S. J., Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics.* 1975; 31:103–115.
15. Wei, L. J. An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association.* 1978; 72:382–386.
16. Scott, N. W., McPherson, G. C., Ramsay, C. R., Campbell, M. K. The method of minimization for allocation to clinical trials: a review. *Controlled Clinical Trials.* 2002; 23:662-674.
17. Bandyopadhyay, U. & De, S. On Multi-Treatment Adaptive Allocation Design for Dichotomous Response. *Communication in Statistics – Theory and Methods.* 2011; 40: 4104–4124.
18. Magyar, A. F.. Zhou, J., Jenkins, B., Haag-Molkenteller, C. A Profile-Based Stratified Randomization and Its Application to a Double-Blind, Placebo-Controlled Clinical Trial. *JSM.* August 11, 2015.
19. Flight, L., Julious, S. A., Goodacre, S. Assessing the Current and Potential Use of Adaptive Study Designs in Emergency Medicine Clinical Trials. *JSM*. August 10, 2015.

Supplemental Materials

Table S1: Standard deviations of operating characteristics (i-iii) for Ridit versus GDL Urn
(i) S.D. of Proportion of Subjects Assigned to Treatment A
(ii) S.D. of Proportion of Total Treatment Failures
(iii) S.D. of Measure of Prognostic Score Imbalance Between Arms

(p_A, p_B, p_C)		GDL Urn With Prognostic Scoring $n_0=5$	Ridit With Prognostic Scoring $n_0=5$	GDL Urn With Stratification $n_0=5$	Ridit With Stratification $n_0=5$
(0.3, 0.3, 0.3)	(i)	0.0497	0.0720	0.0409	0.0703
	(ii)	0.0546	0.0565	0.0573	0.0580
	(iii)	0.0274	0.0309	0.0491	0.0425
(0.4, 0.3, 0.3)	(i)	0.0588	0.0761	0.0472	0.0756
	(ii)	0.0584	0.0605	0.0575	0.0593
	(iii)	0.0254	0.0302	0.0527	0.0450
(0.5, 0.3, 0.3)	(i)	0.0629	0.0775	0.0525	0.0754
	(ii)	0.0603	0.0643	0.0606	0.0628
	(iii)	0.0266	0.0316	0.0579	0.0464
(0.6, 0.3, 0.3)	(i)	0.0754	0.0785	0.0591	0.0813
	(ii)	0.0676	0.0628	0.0635	0.0661
	(iii)	0.0283	0.0311	0.0597	0.0512
(0.7, 0.3, 0.3)	(i)	0.0835	0.0780	0.0652	0.0790
	(ii)	0.0710	0.0642	0.0669	0.0646
	(iii)	0.0287	0.0324	0.0618	0.0578
(0.8, 0.3, 0.3)	(i)	0.0893	0.0788	0.0656	0.0780
	(ii)	0.0777	0.0621	0.0683	0.0628
	(iii)	0.0341	0.0353	0.0651	0.0586
(0.9, 0.3, 0.3)	(i)	0.0896	0.0791	0.0637	0.0750
	(ii)	0.0764	0.0599	0.0648	0.0570
	(iii)	0.0478	0.0401	0.0667	0.0656
(0.5, 0.4, 0.3)	(i)	0.0654	0.0745	0.0519	0.0778
	(ii)	0.0613	0.0595	0.0647	0.0615
	(iii)	0.0247	0.0307	0.0567	0.0451
(0.6, 0.4, 0.3)	(i)	0.0750	0.0774	0.0592	0.0795
	(ii)	0.0648	0.0629	0.0628	0.0641
	(iii)	0.0258	0.0313	0.0619	0.0488
(0.7, 0.4, 0.3)	(i)	0.0852	0.0796	0.0657	0.0777
	(ii)	0.0710	0.0631	0.0635	0.0601
	(iii)	0.0274	0.0310	0.0646	0.0539
(0.8, 0.4, 0.3)	(i)	0.0928	0.0760	0.0721	0.0791
	(ii)	0.0742	0.0620	0.0660	0.0626
	(iii)	0.0378	0.0340	0.0683	0.0602
(0.9, 0.4, 0.3)	(i)	0.0947	0.0766	0.0679	0.0728
	(ii)	0.0764	0.0550	0.0613	0.0544
	(iii)	0.0451	0.0380	0.0698	0.0658
(0.6, 0.5, 0.3)	(i)	0.0780	0.0762	0.0577	0.0784

	(ii)	0.0643	0.0623	0.0607	0.0652
	(iii)	0.0271	0.0306	0.0642	0.0527
(0.7, 0.5, 0.3)	(i)	0.0896	0.0803	0.0698	0.0794
	(ii)	0.0670	0.0619	0.0644	0.0652
	(iii)	0.0283	0.0323	0.0693	0.0545
(0.8, 0.5, 0.3)	(i)	0.0977	0.0753	0.0753	0.0772
	(ii)	0.0726	0.0630	0.0630	0.0613
	(iii)	0.0298	0.0337	0.0700	0.0606
(0.9, 0.5, 0.3)	(i)	0.0964	0.0790	0.0723	0.0764
	(ii)	0.0707	0.0540	0.0625	0.0559
	(iii)	0.0423	0.0400	0.0710	0.0674

Table S2: Averages of operating characteristics (i-iii) for Ridit versus GDL Urn with increasing delay in knowledge of subject outcomes (delay is measure in terms of number of subjects who will be randomized between when a given subject is randomized and when that subject's treatment outcome is known)

(i) Proportion of Subjects Assigned to Treatment A

(ii) Proportion of Total Treatment Failures

(iii) Measure of Prognostic Score Imbalance Between Arms

(p_A, p_B, p_C)		GDL Urn n_0=3 Delay=0	Ridit n_0=5 Delay=0	GDL Urn n_0=3 Delay=10	Ridit n_0=5 Delay=10	GDL Urn n_0=3 Delay=25	Ridit n_0=5 Delay=25
(0.3, 0.3, 0.3)	(i)	0.3306	0.3356	0.3316	0.3373	0.3326	0.3304
	(ii)	0.6999	0.6973	0.7001	0.7031	0.6999	0.7015
	(iii)	0.2626	0.2520	0.2675	0.2500	0.2742	0.2533
(0.4, 0.3, 0.3)	(i)	0.3799	0.3752	0.3774	0.3709	0.3758	0.3717
	(ii)	0.6606	0.6646	0.6642	0.6649	0.6614	0.6624
	(iii)	0.2665	0.2531	0.2685	0.2520	0.2701	0.2544
(0.5, 0.3, 0.3)	(i)	0.4192	0.4131	0.4163	0.4135	0.4093	0.4069
	(ii)	0.6145	0.6167	0.6180	0.6175	0.6177	0.6160
	(iii)	0.2706	0.2568	0.2717	0.2568	0.2754	0.2571
(0.6, 0.3, 0.3)	(i)	0.4605	0.4587	0.4625	0.4549	0.4571	0.4488
	(ii)	0.5640	0.5618	0.5617	0.5627	0.5624	0.5666
	(iii)	0.2742	0.2628	0.2760	0.2611	0.2787	0.2620
(0.7, 0.3, 0.3)	(i)	0.5122	0.4920	0.5016	0.4936	0.4965	0.4890
	(ii)	0.4954	0.4996	0.5028	0.5014	0.5002	0.5025
	(iii)	0.2889	0.2674	0.2873	0.2685	0.2891	0.2641
(0.8, 0.3, 0.3)	(i)	0.5551	0.5375	0.5479	0.5357	0.5306	0.5275
	(ii)	0.4239	0.4306	0.4272	0.4332	0.4352	0.4376
	(iii)	0.2963	0.2762	0.2923	0.2786	0.2944	0.2760
(0.9, 0.3, 0.3)	(i)	0.5976	0.5798	0.5955	0.5738	0.5769	0.5654
	(ii)	0.3446	0.3514	0.3412	0.3540	0.3524	0.3610
	(iii)	0.3104	0.2876	0.3099	0.2871	0.3096	0.2883
(0.5, 0.4, 0.3)	(i)	0.3976	0.3946	0.3908	0.3905	0.3863	0.3881
	(ii)	0.5881	0.5880	0.5901	0.5896	0.5865	0.5891
	(iii)	0.2698	0.2561	0.2722	0.2544	0.2754	0.2586
(0.6, 0.4, 0.3)	(i)	0.4409	0.4276	0.4408	0.4245	0.4297	0.4266
	(ii)	0.5346	0.5391	0.5411	0.5425	0.5425	0.5377

	(iii)	0.2726	0.2570	0.2758	0.2594	0.2778	0.2595
(0.7, 0.4, 0.3)	(i)	0.4816	0.4730	0.4793	0.4685	0.4743	0.4624
	(ii)	0.4760	0.4824	0.4799	0.4837	0.4864	0.4885
	(iii)	0.2782	0.2643	0.2816	0.2650	0.2812	0.2654
(0.8, 0.4, 0.3)	(i)	0.5314	0.5140	0.5218	0.5098	0.5100	0.5055
	(ii)	0.4068	0.4163	0.4161	0.4204	0.4200	0.4210
	(iii)	0.2945	0.2722	0.2921	0.2730	0.2893	0.2697
(0.9, 0.4, 0.3)	(i)	0.5756	0.5537	0.5661	0.5517	0.5529	0.5380
	(ii)	0.3309	0.3438	0.3366	0.3472	0.3440	0.3521
	(iii)	0.3025	0.2799	0.3066	0.2822	0.3028	0.2808
(0.6, 0.5, 0.3)	(i)	0.4227	0.4073	0.4096	0.4066	0.4083	0.4021
	(ii)	0.5045	0.5095	0.5074	0.5092	0.5062	0.5132
	(iii)	0.2714	0.2592	0.2725	0.2554	0.2760	0.2581
(0.7, 0.5, 0.3)	(i)	0.4557	0.4470	0.4592	0.4473	0.4529	0.4401
	(ii)	0.4529	0.4559	0.4521	0.4608	0.4530	0.4564
	(iii)	0.2774	0.2654	0.2796	0.2656	0.2856	0.2610
(0.8, 0.5, 0.3)	(i)	0.5050	0.4843	0.5022	0.4862	0.4865	0.4780
	(ii)	0.3880	0.3970	0.3917	0.3971	0.3964	0.3994
	(iii)	0.2931	0.2683	0.2891	0.2687	0.2932	0.2718
(0.9, 0.5, 0.3)	(i)	0.5418	0.5276	0.5344	0.5270	0.5256	0.5129
	(ii)	0.3198	0.3322	0.3235	0.3315	0.3298	0.3351
	(iii)	0.2970	0.2808	0.3021	0.2789	0.2968	0.2801

Table S3: Averages of operating characteristics (i-iii) for different values of ϕ

(i) Proportion of Total Treatment Failures

(ii) Measure of Prognostic Score Imbalance Between Arms

(p_A, p_B, p_C)		GDL Urn $n_0=5$ $\phi=2/3$	GDL Urn $n_0=5$ $\phi=1/2$
(0.3, 0.3, 0.3)	(i)	0.3301	0.3347
	(ii)	0.6991	0.6964
	(iii)	0.1513	0.1809
(0.4, 0.3, 0.3)	(i)	0.3537	0.3628
	(ii)	0.6618	0.6649
	(iii)	0.1499	0.1811
(0.5, 0.3, 0.3)	(i)	0.3834	0.3935
	(ii)	0.6276	0.6183
	(iii)	0.1511	0.1813
(0.6, 0.3, 0.3)	(i)	0.4234	0.4293
	(ii)	0.5739	0.5722
	(iii)	0.1526	0.1851
(0.7, 0.3, 0.3)	(i)	0.4740	0.4825
	(ii)	0.5099	0.5051
	(iii)	0.1557	0.1904
(0.8, 0.3, 0.3)	(i)	0.5354	0.5478
	(ii)	0.4331	0.4250
	(iii)	0.1632	0.1991
(0.9, 0.3, 0.3)	(i)	0.6278	0.6386
	(ii)	0.3233	0.3164

	(iii)	0.1867	0.2215
(0.5, 0.4, 0.3)	(i)	0.3712	0.3764
	(ii)	0.5944	0.5919
	(iii)	0.1494	0.1800
(0.6, 0.4, 0.3)	(i)	0.4128	0.4172
	(ii)	0.5451	0.5435
	(iii)	0.1505	0.1840
(0.7, 0.4, 0.3)	(i)	0.4562	0.4675
	(ii)	0.4885	0.4843
	(iii)	0.1540	0.1869
(0.8, 0.4, 0.3)	(i)	0.5203	0.5373
	(ii)	0.4146	0.4061
	(iii)	0.1612	0.1976
(0.9, 0.4, 0.3)	(i)	0.6112	0.6276
	(ii)	0.3132	0.3030
	(iii)	0.1799	0.2200
(0.6, 0.5, 0.3)	(i)	0.3984	0.3973
	(ii)	0.5180	0.5124
	(iii)	0.1523	0.1841
(0.7, 0.5, 0.3)	(i)	0.4434	0.4511
	(ii)	0.4597	0.4595
	(iii)	0.1548	0.1871
(0.8, 0.5, 0.3)	(i)	0.5060	0.5123
	(ii)	0.3920	0.3888
	(iii)	0.1617	0.1942
(0.9, 0.5, 0.3)	(i)	0.5927	0.6040
	(ii)	0.2995	0.2931
	(iii)	0.1811	0.2172