A RESPONSE-ADAPTIVE COVARIATE-BALANCED RANDOMIZATION FOR MULTI-ARM CLINICAL TRIALS

By

Cassandra M. Ballou

A THESIS

Presented to the Department of Public Health and Preventive Medicine

and the Oregon Health & Science University School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Science

December 2015

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank the members of my committee, Dr. Yiyi Chen, Dr. Rochelle Fu, and Dr. Barry Oken for their input on this thesis, and, particularly, Yiyi for her wonderful mentorship.

I would also like to thank my long-suffering family for their love and support and putting up with me over many sleepless nights.

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 best 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 two-stage randomization for multi-arm trials which marries response-adaptative 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. The proposed design also compared favorably with a Bayesian approach to response-adaptive covariate-balanced design, providing equal or better power and covariate-balance.

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 allow the probability of assigning a new patient to a particular arm of a trial to vary over the course

of the trial in response to the outcomes observed for previously enrolled patients 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. A properly implemented response-adaptive randomization provides higher power compared to a static unequal allocation reducing 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 than in an equally randomized trial.

The idea of response-adaptive design dates back as far as 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 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]; thus far, the Bayesian approach has predominated in terms of implementation. In particular, the M.D. Anderson Cancer Center has pioneered this area with a number of significant response-adaptively randomized studies such as I-SPY 2 and BATTLE [5,6].

Among response-adaptive randomization designs using frequentist methods, urn models have predominated in the literature [1]. They are based on a simple, intuitive model in which assignment to each of the k treatment 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 updated 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 [7,8], 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 [9,10], in which balls of type *k* are removed from the urn in response to a failure being observed for a patient on that treatment arm. The "drop-the-loser" strategy has been shown to be superior in terms of having lower randomization procedure variability (variability in the proportion of subjects assigned to each treatment arm) and consequently higher power [9], since power is a decreasing function of randomization procedure variability [10].

A potential flaw of these response-adaptive randomization procedures is that many have not considered imbalance in baseline covariates believed to be prognostic [12]. One criticism in particular of response-adaptive designs has been that they introduce bias due to population drift, in which the subject population changes over time. Covariate-balancing would help to correct for this phenomenon [13]. Particularly for trials with small to medium sample sizes, randomization alone may be inadequate to ensure that 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 [14]. Covariate imbalances, should they occur, may introduce bias into a trial's estimates of treatment success [15]. For example, if older subjects have a lower probability of treatment success regardless of treatment than younger subjects and a substantially larger proportion of older subjects are assigned to Treatment A, then Treatment 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-balanced randomization seeks to assign patients in such a way that baseline covariates are distributed 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 with a certain baseline covariate profile to the best treatment for them based on the responses of previously randomized subjects with a similar baseline covariate profile [15]. It should become apparent that if there are indeed significant differences in treatment response 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 subjects with genetic marker A, while Treatment B has a higher true probability of success for subjects 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 [12]. 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 [16]. 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 [17] and Wei [18] 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 [19]. 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 [19].

Compared to the minimization method, the prognostic score approach has two major advantages. Firstly, the prognostic score approach allows balancing of continuous covariates directly, while 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.

The goal of this current work is to provide a Frequentist covariate-balanced response-adaptive randomization for three or more arm superiority trials equivalent to what is currently available only in the Bayesian literature. Possible examples of a trial which could utilize the proposed randomization include a phase II trial with two or more novel pharmaceutical agents and a placebo or current standard care control arm or a behavioral intervention with two control arms, an active and a passive. We will consider both stratification and a prognostic score approach based on the logistic regression model as proposed by Yuan [12], who utilized this approach in combination with a Bayesian method of response-adaptive randomization.

The remainder of this thesis is laid out as follows. In the Methods section, we propose four novel designs which combine a response-adaptive and covariate-balanced approach to randomization. In the Results section, we evaluate the operating characteristics of our proposed designs via simulation and compare them with previously available randomization options including equal randomization, response-adaptive randomization without covariate balancing, and a Bayesian response-adaptive, covariate-balanced design. In the Discussion section, we discuss implications and limitations of the current work and propose future work. We conclude with a statement of the major findings in the Summary and Conclusions section.

METHODS

Four novel designs consisting of two approaches to response-adaptive randomization already present in the literature each combined with two approaches to covariate-balanced randomization present in the literature were considered. Results presented are for a three arm study with a binary outcome. All simulation studies were conducted with 1,000 repetitions using R software.

Methods of Response-Adaptive Randomization

The two methods of response-adaptive randomization considered were the generalized drop-the-loser (GDL) Urn model as presented by Sun et al. [10] and the Ridit scoring based method presented in Bandyopadhyay & De [20]. For both the Ridit and GDL Urn models, suppose there are three treatment arms, *A*, *B*, and *C*. Probabilities of treatment success, $\hat{p}_k$, where *k* is *A*, *B*, or *C*, were estimated as follows,

$$\hat{p}_k = \frac{S_k+0.5}{N_k+1},\ where\ k = A, B, C \quad (1)$$

where N_k is the number patients assigned to treatment arm *k* and S_k is the number of successes observed among those N_k subjects. At the beginning of the trial a small number of subjects, n_0, are equally assigned to the treatment arms to help stabilize the estimates of $\hat{p}_k$ prior to response-adaptive randomization. The probabilities of treatment failure, $\hat{q}_k$, are simply the complements of $\hat{p}_k$.

$$\hat{q}_k = 1 - \hat{p}_k,\ where\ k = A, B, C \quad (2)$$

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

$$R_A = \frac{1}{3} + \frac{1}{6}(2\hat{p}_A - \hat{p}_B - \hat{p}_C) + \frac{\hat{q}_B}{6}(\hat{p}_A - \hat{p}_C) + \frac{\hat{q}_C}{6}(\hat{p}_A - \hat{p}_B)$$
$$R_B = \frac{1}{3} + \frac{1}{6}(2\hat{p}_B - \hat{p}_A - \hat{p}_C) + \frac{\hat{q}_A}{6}(\hat{p}_B - \hat{p}_C) + \frac{\hat{q}_C}{6}(\hat{p}_B - \hat{p}_A) \quad (3)$$
$$R_C = \frac{1}{3} + \frac{1}{6}(2\hat{p}_C - \hat{p}_B - \hat{p}_A) + \frac{\hat{q}_B}{6}(\hat{p}_C - \hat{p}_A) + \frac{\hat{q}_A}{6}(\hat{p}_C - \hat{p}_B)$$

where R_k is the probability of assigning a new patient to treatment arm k at a given point in the trial, $\hat{p}_k$ and $\hat{q}_k$ are defined as in Equations 1 and 2.

The GDL urn model proposed by Sun et al. [10] 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,

$$X_i = X_{i0},\ where\ X_{i0} > 0$$
$$X_k = X_{k0} + \frac{\frac{1}{\hat{q}_k}}{\frac{1}{\hat{q}_A}+\frac{1}{\hat{q}_B}+\frac{1}{\hat{q}_C}},\ \ where\ k = A, B, C \quad (4)$$

where $\hat{q}_k$ is defined as in (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. A value of 3 was chosen for X_{i0}. Results were found to be insensitive to the numbers of treatment balls in excess of that value. Then, the probability of assignment to a given treatment arm at a given time, U_k, as follows,

$$U_k = \frac{X_k}{X_A+X_B+X_C},\ where\ k = A, B, C \quad (5)$$

Methods of Covariate-Balanced Randomization

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.

Results involving the prognostic scoring method assume a standard normal distribution of prognostic scores with no assumptions made about the number or distributions of individual prognostic covariates. 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 [21].

Novel Response-Adaptive Covariate-Balanced Designs

In summary, the novel models considered are: Ridit with prognostic scoring, in which the probability of assignment to a given treatment arm *k*, *pr(k)*, is given by,

$$pr(k) = \pi_k R_k \quad (7)$$

where π_k is defined as in (6) and R_k is defined as in (3).

GDL Urn with prognostic scoring, in which the probability of assignment to a given treatment arm *k*, *pr(k)*, is given by,

$$pr(k) = \pi_k U_k \quad (8)$$

where π_k is defined as in (6) and U_k is defined as in (5).

Stratified Ridit in which the probability of assignment to a given treatment arm *k*, *pr(k)*, is given by,

$$pr(k) = R_{k_s} \quad (9)$$

where R_{k_s} is R_k as defined in (3) for only patients in strata *s*.

Stratified GDL Urn in which the probability of assignment to a given treatment arm *k*, *pr(k)*, is given by,

$$pr(k) = U_{k_s} \quad (10)$$

where U_{k_s} is U_k as defined in (5) for only patients in strata *s*.

Bayesian Method

The novel methods were compared to the Bayesian approach taken by Yuan et al. [12] in which probability of assignment to a given treatment arm is based on the posterior probability that treatment *k* is superior to all others, S_k,

$$\begin{aligned} S_A &= pr(p_A = \max\{p_k, 1 \le k \le 3\}|data) \\ S_B &= pr(p_B = \max\{p_k, 1 \le k \le 3\}|data) \\ S_C &= pr(p_C = \max\{p_k, 1 \le k \le 3\}|data) \end{aligned} \quad (11)$$

Noninformative priors (Beta (1, 1)) were used for the parameters p_A, p_B, and p_C to facilitate a more direct comparison with the novel frequentist methods. The posterior probability that treatment *k* is superior to all others is stabilized by taking the square root transformation as follows and this transformation was used to determine the allocation probability to each treatment arm, T_k,

$$T_k = \frac{\sqrt{S_k}}{\sqrt{S_A}+\sqrt{S_B}+\sqrt{S_C}} \quad (12)$$

This value is then weighted by the prognostic scoring factor, π_k, to give a probability of assignment to a given treatment arm *k*, *pr(k)*, for the Bayesian design of,

$$pr(k) = \pi_k T_k \quad (13)$$

Equal Randomization Method

The novel methods were also compared to simple equal randomization in which the probability of assignment to a given treatment arm *k*, *pr(k)*, is fixed,

$$pr(k) = \frac{1}{k} \quad (14)$$

Operating Characteristics

Four criteria were used to assess the performance of the methods: the proportion of subjects assigned to the best treatment with a higher proportion being superior, the proportion of treatment failures experienced by subjects with a low value being superior, the average statistical power archived for a fixed total number of subjects in the trial with values above 80% considered acceptable, and the imbalance in prognostic scores between the treatment arms at the conclusion of the trial as measured using Kolmogorov–Smirnov (KS) statistics for which smaller values indicate better balance between treatment arms.

Scenarios

Without loss of generality, Treatment A was fixed as the best treatment in terms of true probability of treatment success. Three potential scenarios about the true probabilities of success for the three treatment arms were considered. In the first scenario, 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 varied from 0.3 to 0.9 in increments of 0.1. In the second scenario, there was a small difference between the true probability of success for Treatment B (0.4) and that for Treatment C (0.3) and the probability of success for Treatment A varied from 0.5 to 0.9 in increments of 0.1. In the third scenario, there was a larger difference between the true probability of success for Treatment B (0.5) and that for Treatment C (0.3) and the probability of success for Treatment A varied from 0.6 to 0.9 in increments of 0.1.

Both the novel designs and the Bayesian design used for comparison are two-stage with 15 patients being equally randomized to provide adequate initial estimates of $\hat{p}_k$ in the first stage and 50 patients adaptively randomized in the second stage for a moderate total sample size of 65. The sample size for equal randomization was fixed at 65 subjects for comparison. Differing times to availability of patient treatment outcome was simulated by updating the estimates of $\hat{p}_k$ after the randomization of every 1 subject for instantaneous, every 10 subjects for short delay, and every 25 subjects for long delay.

RESULTS

Comparison with Equal Randomization

All four novel designs, stratified Ridit, Ridit with prognostic scoring, stratified GDL Urn, and GDL Urn with prognostic scoring, significantly outperformed equal randomization with lower the proportions of total treatment failures (see Figure 1).

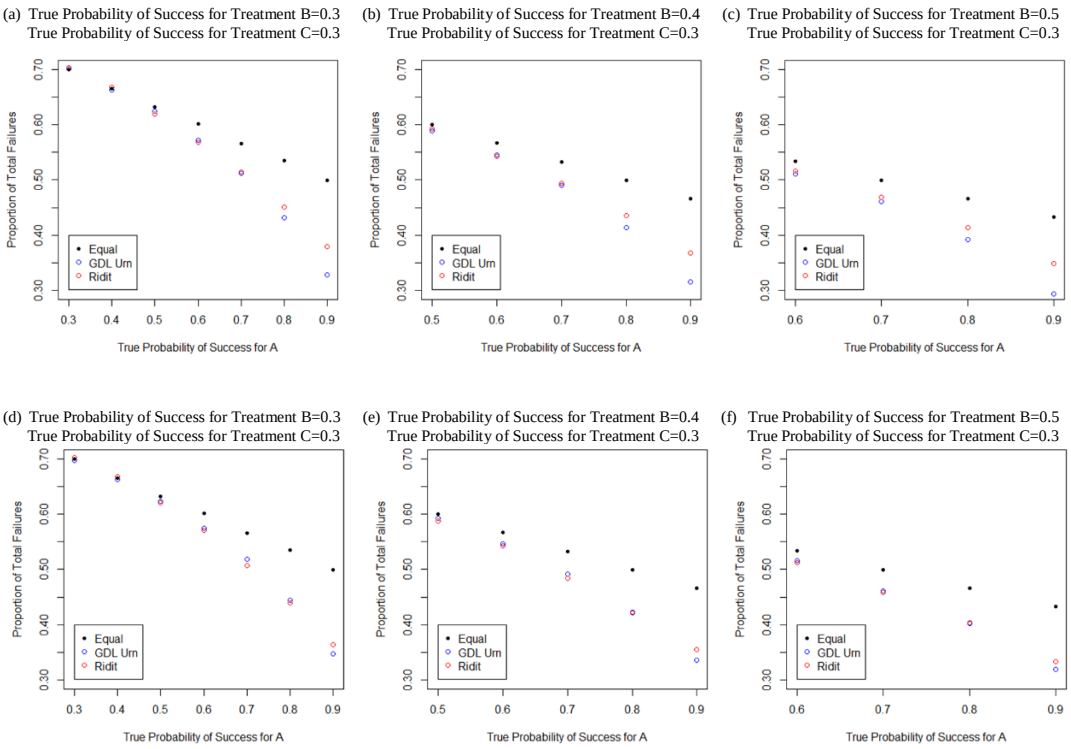

Figure 1: Comparison of proportions of total treatment failures for equal randomization versus Ridit and GDL Urn with prognostic scoring (a-c) and stratification (d-f). Results are also available in table form in the appendices (See also Tables S1 & S2).

Comparison with Response-Adaptive Randomization

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 prognostic score was predictive of the outcome (see Figure 2).

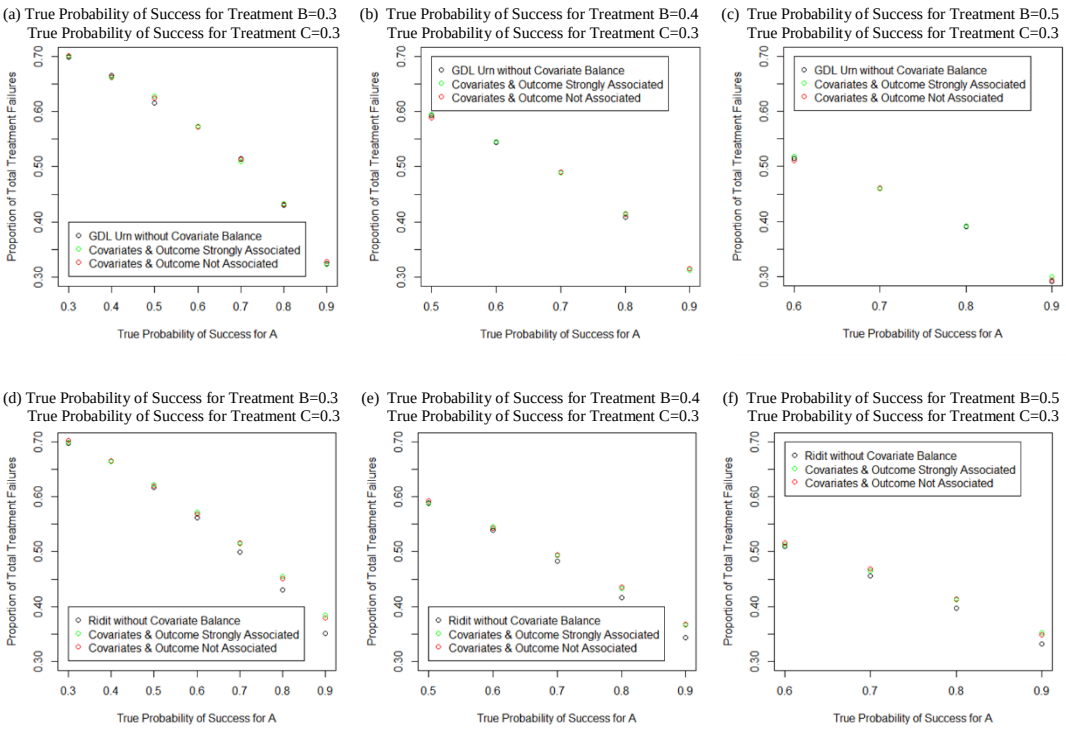

Figure 2: Comparison of proportions of total treatment failures for GDL Urn (a-c) and Ridit (d-e) response-adaptive randomization alone versus response-adaptive randomization with prognostic scoring. (See also Table S3)

Comparison with Bayesian Randomization

When compared to the Bayesian design proposed by Yuan et al. [12], the novel designs had lower proportions of subjects assigned to Treatment A and higher proportions of total treatment failures (see Figure 3).

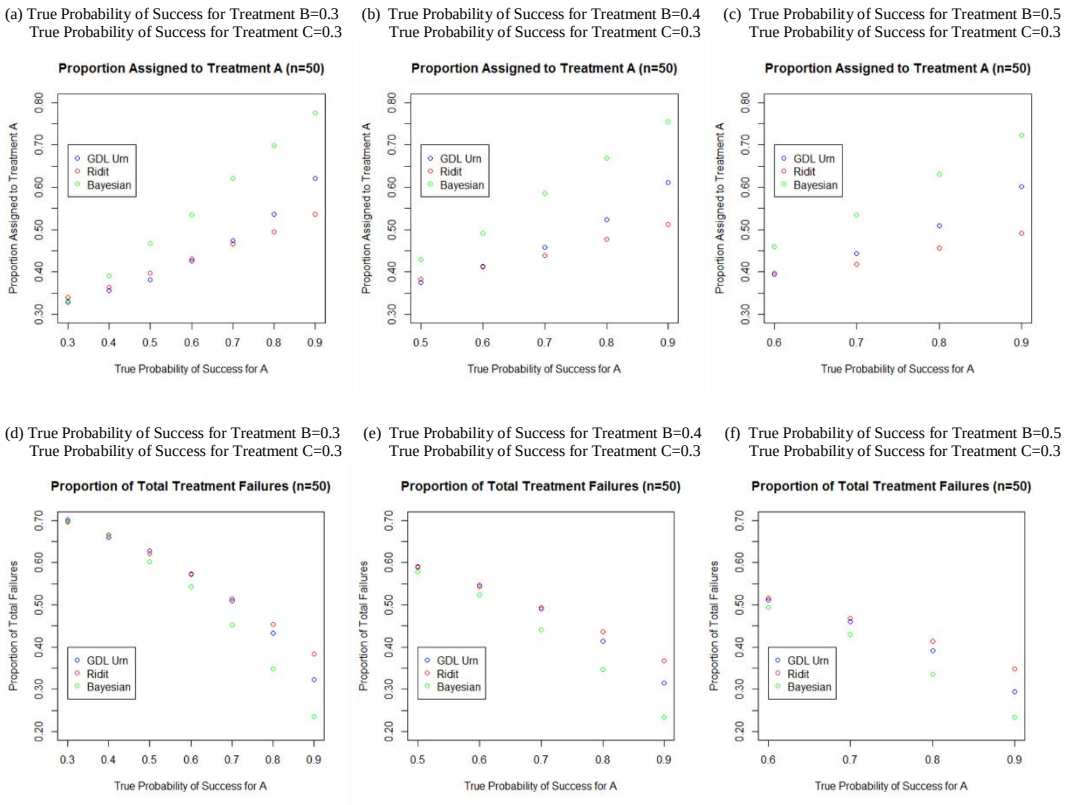

Figure 3: Comparison of novel designs with Bayesian design for proportion of subjects assigned to the best treatment (a-c) and proportion of total treatment failures (d-f). (See also Table S1)

However, the novel designs outperformed the Bayesian design in terms of covariate balance and also power (see Figure 4) particularly at high probabilities of success for Treatment A.

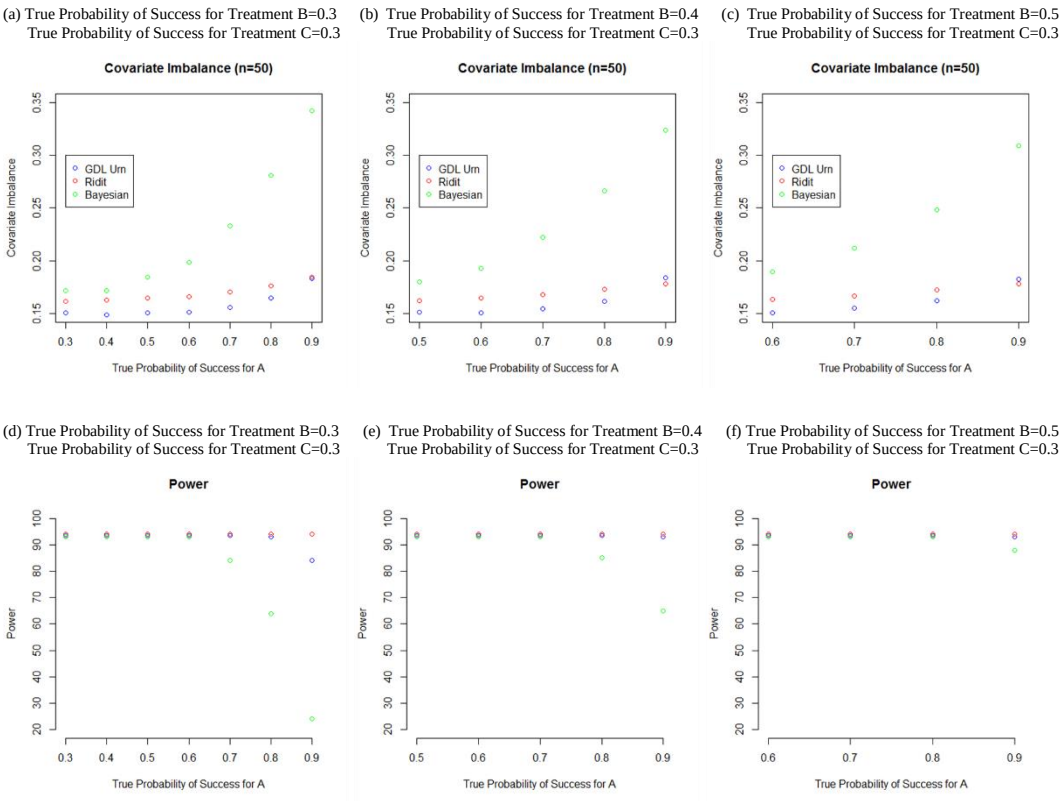

Figure 4: Comparison of novel designs with Bayesian design for covariate imbalance as measured by KS statistic (a-c) and statistical power (d-f) (See also Table S1).

These results indicate that there is a maximum threshold past which higher proportions of subjects assigned to the best treatment become undesirable as very small sample sizes for the other two treatment arms both make it impossible to provide adequate prognostic covariate balance and inflate the type II error rate unacceptably, compromising the results of the trial.

Comparison of Ridit and GDL Urn

Although all novel approaches performed well, several noteworthy differences between the Ridit and GDL Urn methods when combined with covariate balancing methods were

observed. 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. 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 5).

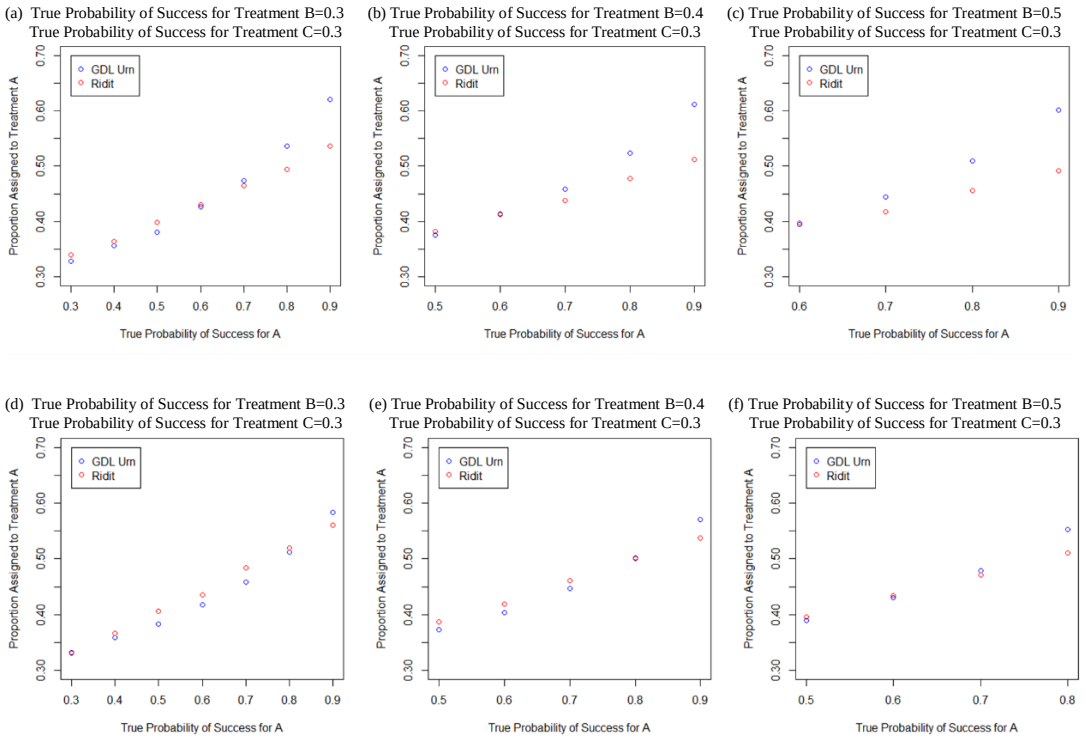

Figure 5: Comparison of proportions of subjects assigned to the treatment with the highest true probability of success for Ridit versus GDL Urn with Prognostic Scoring (a-c) and Stratification (d-f). (See also Table S1 & S2)

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 in terms of assigning subjects to the best treatment and reducing treatment failures 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. 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 6).

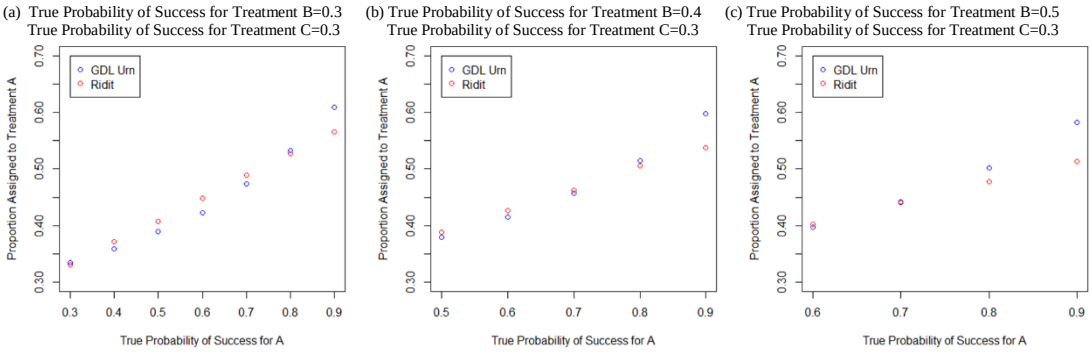

Figure 6: Comparison of proportions of subjects assigned to the treatment with the highest true probability of success for Ridit versus GDL Urn with prognostic scoring with a delay of 25 subjects between randomization and knowledge of subject outcome. (See also Table S5)

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 7).

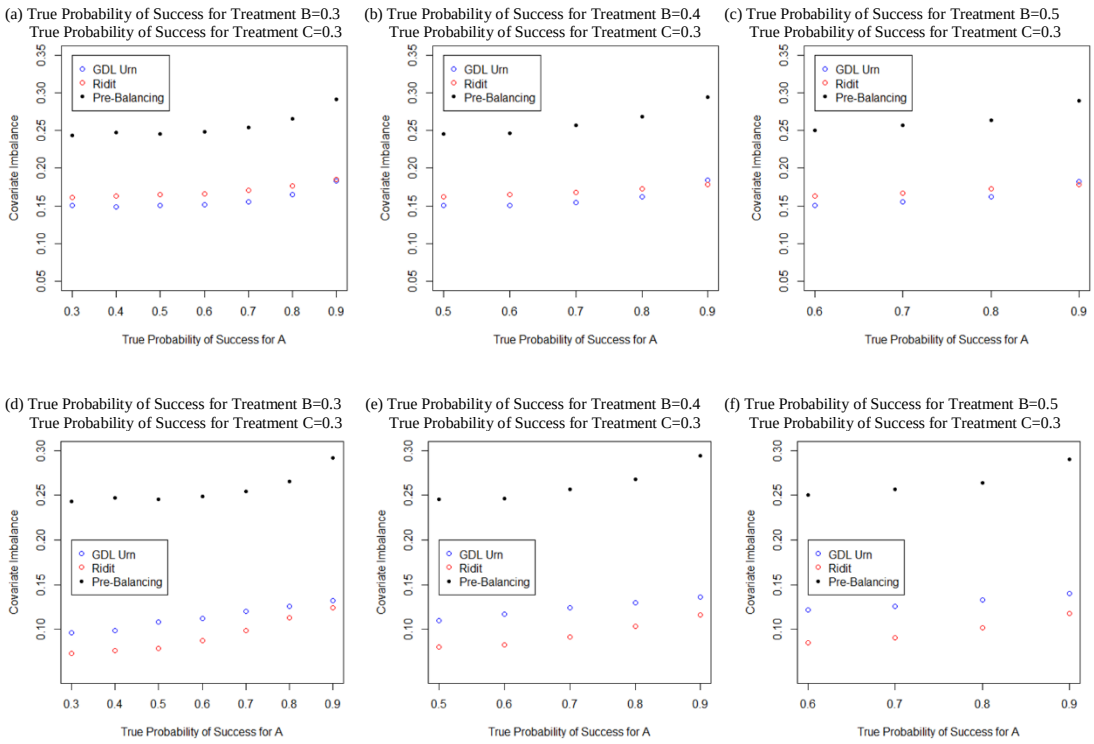

Figure 7: Comparison of average KS statistic (a measure of baseline covariate imbalance) for Ridit versus GDL Urn with prognostic scoring (a-c) and stratification (d-f). See also (Table S7).

Variability

Because, in reality, only one study can be performed, the variability in, as well as the average values of, the 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 S4 in the appendices).

Optimization of Parameters

Ideal values for the parameters n_0 and ϕ were also investigated. Values of 3 or 5 were considered for n_0 (see Table S5 in the appendices.). Although, in most cases, estimates of the true treatment success probabilities were found to be sufficiently stable for the purposes of response-adaptive randomization with only 3 patients equally randomized to each treatment arm in the first stage of the design, we have chosen and generally recommend the more conservative value of 5 to reduce variability. 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 [22], 3 will be adequate (see Table S5 in the appendices). 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 S6 in the appendices). 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 the 65 considered here, as covariate imbalance decreases with increasing sample size even for a response-adaptive design with no covariate-balancing component [15].

DISCUSSION

It was found that, on average, all four proposed novel designs, stratified Ridit, Ridit with prognostic scoring, stratified GDL Urn, and GDL Urn 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. When compared to the Bayesian response-adaptive covariate-balanced design proposed by Yuan et al. [12], the novel designs had higher proportions of total treatment failures; however, in order to achieve lower proportions of total treatment failures the Bayesian design makes sacrifices in other operating characteristics, having higher covariate-imbalance and lower power, particularly at high values of true probability of success for the best treatment (above 0.6), due to the very high proportion of subjects assigned to one treatment which produces small sample sizes for the other two groups. The proposed methods maintain good covariate balance and acceptable power (greater than 80%) regardless of true probabilities of success for the three treatment arms. It is important to note that Bayesian designs incorporate the flexibility of specifying different priors. A "skeptical" prior, representing a strong belief that all three treatments will be ineffective, and equally so, could reduce the proportion of subjects assigned to the best treatment, potentially bringing the results for covariate-balance and power from the Bayesian method more in line with those observed for the novel frequentist methods than the noninformative prior presented here.

The choice of the Ridit or GDL Urn model as the better method for response-adaptive covariate-balanced 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

Differences between these 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.

In general, the potential limitations of this work are similar to those of any response-adaptive randomization method. One example is that, in our simulations, equal sample sizes have been compared, while, in practice, an equal randomization will require fewer subjects to achieve the same power as a response-adaptive randomization. Each method of covariate balancing has its own limitations. Covariate balancing by stratification limits

the type and number of prognostic covariates which may be balanced; while prognostic scoring is less robust, being susceptible to errors in the logistic regression model. The better choice is highly study dependent and for that reason, results for both have been presented here. A further limitation is that early stopping rules for either futility or efficacy, which may improve the performance of equal randomization in terms of minimizing the proportion of treatment failures, were not considered [13]. It would be of interest to explore the potential effects of stopping rules, for example, expanding the design to three-stages with poor performing treatment arms being dropped at an interim analysis. Although this work has presented results for three arm studies with a binary outcome, either the GDL Urn or Ridit methods can be extended to four or more arms and modification to accommodate a continuous primary outcome is possible, perhaps by using a method of calculating $\hat{p}_k$ which weights patient outcomes based on magnitude of the observation. In addition, more scenarios involving different true probabilities of success for the three treatment arms could be considered. This work remains to be done in the future.

SUMMARY AND CONCLUSIONS

In conclusion, the novel response-adaptive covariate-balanced randomizations presented here provide a significant gain in ethicality over an equal randomization by increasing the proportion of subjects who experience treatment success. They also offer 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. They provide a viable non-Bayesian alternative which was not previously available in the literature.

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. Zhou, X, Liu, S, Kim, E, Herbst, R, Lee, J. Bayesian adaptive design for targeted therapy development in lung cancer: A step toward personalized medicine. *Clinical Trials.* 2008; 5: 181–193.
6. Barker, A, Sigman, C, Kelloff G, Hylton, N, Berry, D, Esserman, L. I-SPY 2: An adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy. *Clinical Pharmacology & Therapeutics*. 2009; 86: 97–100.
7. Wei, L. J., Durham, S. The randomized play-the-winner rule in medical trial. *Journal of the American Statistical Association.* 1978; 73: 840–843.
8. 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.
9. Ivanova, A. A play-the-winner type urn model with reduced variability. *Metrika.* 2003; 58: 1–13.
10. 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.

11. 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.

12. 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.

13. Lee, J, Chen, N, Yin, G. Worth Adapting? Revisiting the Usefulness of Outcome-Adaptive Randomization. *Clinical Cancer Research.* 2012; 18: 4498-4507.

14. Campbell, M. K., McPherson, G. C., Simple randomisation or minimisation: the impact of trial size. *Controlled Clinical Trials.* 2001; 22-87.

15. Rosenberger, W. F., Sverdlov, O. Handling Covariates in the Design of Clinical Trials. *Statistical Science.* 2008; 23: 404–419.

16. Taves, D. R. Minimization: A new method of assigning patients to treatment and control groups. *Journal of Clinical Pharmacology Therapy.* 1974; 15:443–453.

17. Pocock S. J., Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics.* 1975; 31:103–115.

18. 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.

19. 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.

20. Bandyopadhyay, U. & De, S. On Multi-Treatment Adaptive Allocation Design for Dichotomous Response. *Communication in Statistics – Theory and Methods.* 2011; 40: 4104–4124.

21. 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.

22. 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.

APPENDICES

Table S1: Averages of operating characteristics (i-iv) with prognostic scoring as the method of covariate balancing.

(i) Proportion of Subjects Assigned to Treatment A

(ii) Proportion of Total Treatment Failures

(iii) Measure of Prognostic Score Imbalance Between Arms

(iv) Statistical Power

(p_A, p_B, p_C)		Ridit With Prognostic Scoring $n_0=5$	GDL Urn With Prognostic Scoring $n_0=5$	Bayesian With Prognostic Scoring $n_0=5$
(0.3, 0.3, 0.3)	(i)	0.3400	0.3282	0.3319
	(ii)	0.7026	0.7013	0.6986
	(iii)	0.1616	0.1507	0.1718
	(iv)	94%	94%	94%
(0.4, 0.3, 0.3)	(i)	0.3641	0.3567	0.3907
	(ii)	0.6650	0.6628	0.6625
	(iii)	0.1631	0.1485	0.1718
	(iv)	94%	94%	94%
(0.5, 0.3, 0.3)	(i)	0.3980	0.3809	0.4675
	(ii)	0.6188	0.6241	0.6024
	(iii)	0.1645	0.1504	0.1844
	(iv)	94%	94%	94%
(0.6, 0.3, 0.3)	(i)	0.4302	0.4268	0.5352
	(ii)	0.5688	0.5722	0.5433
	(iii)	0.1659	0.1513	0.1988
	(iv)	94%	94%	93%
(0.7, 0.3, 0.3)	(i)	0.4653	0.4742	0.6213
	(ii)	0.5148	0.5129	0.4516
	(iii)	0.1703	0.1556	0.2328
	(iv)	94%	94%	84%
(0.8, 0.3, 0.3)	(i)	0.4944	0.5359	0.6995
	(ii)	0.4510	0.4313	0.3491
	(iii)	0.1760	0.1649	0.2812
	(iv)	94%	93%	64%
(0.9, 0.3, 0.3)	(i)	0.5360	0.6210	0.7758
	(ii)	0.3788	0.3283	0.2357
	(iii)	0.1842	0.1842	0.3425
	(iv)	94%	84%	24%
(0.5, 0.4, 0.3)	(i)	0.3823	0.3757	0.4289

	(ii)	0.5918	0.5890	0.5784
	(iii)	0.1623	0.1510	0.1803
	(iv)	94%	94%	94%
(0.6, 0.4, 0.3)	(i)	0.4121	0.4138	0.4920
	(ii)	0.5431	0.5457	0.5244
	(iii)	0.1646	0.1508	0.1928
	(iv)	94%	94%	94%
(0.7, 0.4, 0.3)	(i)	0.4384	0.4585	0.5854
	(ii)	0.4939	0.4905	0.4408
	(iii)	0.1678	0.1544	0.2221
	(iv)	94%	94%	85%
(0.8, 0.4, 0.3)	(i)	0.4769	0.5239	0.6691
	(ii)	0.4356	0.4132	0.3468
	(iii)	0.1729	0.1618	0.2660
	(iv)	94%	93%	65%
(0.9, 0.4, 0.3)	(i)	0.5122	0.6113	0.7542
	(ii)	0.3675	0.3148	0.2329
	(iii)	0.1783	0.1838	0.3236
	(iv)	94%	94%	94%
(0.6, 0.5, 0.3)	(i)	0.3970	0.3940	0.4595
	(ii)	0.5162	0.5111	0.4941
	(iii)	0.1634	0.1509	0.1894
	(iv)	94%	94%	94%
(0.7, 0.5, 0.3)	(i)	0.4179	0.4439	0.5340
	(ii)	0.4683	0.4605	0.4301
	(iii)	0.1669	0.1550	0.2119
	(iv)	94%	94%	94%
(0.8, 0.5, 0.3)	(i)	0.4559	0.5089	0.6311
	(ii)	0.4133	0.3916	0.3360
	(iii)	0.1725	0.1624	0.2486
	(iv)	94%	94%	94%
(0.9, 0.5, 0.3)	(i)	0.4921	0.6011	0.7229
	(ii)	0.3486	0.2939	0.2330
	(iii)	0.1781	0.1825	0.3087
	(iv)	94%	93%	88%

Table S2: Averages of operating characteristics (i-iii) with stratification as the method of covariate balancing.

(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)		Ridit	GDL Urn

		with Stratification n_0=5	with Stratification n_0=5
(0.3, 0.3, 0.3)	(i)	0.3306	0.3326
	(ii)	0.7000	0.6981
	(iii)	0.0731	0.0963
(0.4, 0.3, 0.3)	(i)	0.3667	0.3593
	(ii)	0.6678	0.6634
	(iii)	0.0767	0.0986
(0.5, 0.3, 0.3)	(i)	0.4055	0.3829
	(ii)	0.6199	0.6228
	(iii)	0.0791	0.1080
(0.6, 0.3, 0.3)	(i)	0.4357	0.4175
	(ii)	0.5711	0.5748
	(iii)	0.0874	0.1122
(0.7, 0.3, 0.3)	(i)	0.4835	0.4590
	(ii)	0.5070	0.5189
	(iii)	0.0990	0.1204
(0.8, 0.3, 0.3)	(i)	0.5199	0.5118
	(ii)	0.4395	0.4439
	(iii)	0.1129	0.1255
(0.9, 0.3, 0.3)	(i)	0.5600	0.5838
	(ii)	0.3645	0.3479
	(iii)	0.1239	0.1320
(0.5, 0.4, 0.3)	(i)	0.3866	0.3724
	(ii)	0.5873	0.5922
	(iii)	0.0806	0.1102
(0.6, 0.4, 0.3)	(i)	0.4186	0.4040
	(ii)	0.5428	0.5466
	(iii)	0.0827	0.1172
(0.7, 0.4, 0.3)	(i)	0.4608	0.4463
	(ii)	0.4834	0.4910
	(iii)	0.0915	0.1240
(0.8, 0.4, 0.3)	(i)	0.5004	0.5019
	(ii)	0.4213	0.4229
	(iii)	0.1033	0.1297
(0.9, 0.4, 0.3)	(i)	0.5380	0.5706
	(ii)	0.3550	0.3354
	(iii)	0.1162	0.1359
(0.6, 0.5, 0.3)	(i)	0.3955	0.3900
	(ii)	0.5123	0.5152
	(iii)	0.0849	0.1219
(0.7, 0.5, 0.3)	(i)	0.4340	0.4307
	(ii)	0.4595	0.4597
	(iii)	0.0905	0.1255
(0.8, 0.5, 0.3)	(i)	0.4709	0.4791
	(ii)	0.4036	0.4022
	(iii)	0.1017	0.1327
(0.9, 0.5, 0.3)	(i)	0.5113	0.5528

	(ii)	0.3336	0.3190
	(iii)	0.1179	0.1401

Table S3: Averages of operating characteristics (i-iii) for comparison between no covariate-balancing method, covariate-balancing with strong association between balanced covariate and outcome and no association between balanced covariates and outcome

(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 without Covariate-Balancing n_0=5	GDL Urn with Prognostic Scoring and Strong Association n_0=5	GDL Urn With Prognostic Scoring and No Association n_0=5
(0.3, 0.3, 0.3)	(i)	0.3340	0.3301	0.3282
	(ii)	0.6989	0.6991	0.7013
	(iii)	0.2433	0.1513	0.1507
(0.4, 0.3, 0.3)	(i)	0.3586	0.3537	0.3567
	(ii)	0.6654	0.6618	0.6628
	(iii)	0.2471	0.1499	0.1485
(0.5, 0.3, 0.3)	(i)	0.3899	0.3834	0.3809
	(ii)	0.6159	0.6276	0.6241
	(iii)	0.2457	0.1511	0.1504
(0.6, 0.3, 0.3)	(i)	0.4283	0.4234	0.4268
	(ii)	0.5726	0.5739	0.5722
	(iii)	0.2486	0.1526	0.1513
(0.7, 0.3, 0.3)	(i)	0.4742	0.4740	0.4742
	(ii)	0.5146	0.5099	0.5129
	(iii)	0.2542	0.1557	0.1556
(0.8, 0.3, 0.3)	(i)	0.5380	0.5354	0.5359
	(ii)	0.4303	0.4331	0.4313
	(iii)	0.2657	0.1632	0.1649
(0.9, 0.3, 0.3)	(i)	0.6244	0.6278	0.6210
	(ii)	0.3240	0.3233	0.3283
	(iii)	0.2916	0.1867	0.1842
(0.5, 0.4, 0.3)	(i)	0.3799	0.3712	0.3757
	(ii)	0.5918	0.5944	0.5890

	(iii)	0.2456	0.1494	0.1510
(0.6, 0.4, 0.3)	(i)	0.4147	0.4128	0.4138
	(ii)	0.5437	0.5451	0.5457
	(iii)	0.2463	0.1505	0.1508
(0.7, 0.4, 0.3)	(i)	0.4587	0.4562	0.4585
	(ii)	0.4885	0.4885	0.4905
	(iii)	0.2566	0.1540	0.1544
(0.8, 0.4, 0.3)	(i)	0.5315	0.5203	0.5239
	(ii)	0.4088	0.4146	0.4132
	(iii)	0.2682	0.1612	0.1618
(0.9, 0.4, 0.3)	(i)	0.6186	0.6112	0.6113
	(ii)	0.3124	0.3132	0.3148
	(iii)	0.2941	0.1799	0.1838
(0.6, 0.5, 0.3)	(i)	0.3961	0.3984	0.3940
	(ii)	0.5147	0.5180	0.5111
	(iii)	0.2502	0.1523	0.1509
(0.7, 0.5, 0.3)	(i)	0.4436	0.4434	0.4439
	(ii)	0.4600	0.4597	0.4605
	(iii)	0.2566	0.1548	0.1550
(0.8, 0.5, 0.3)	(i)	0.5099	0.5060	0.5089
	(ii)	0.3905	0.3920	0.3916
	(iii)	0.2638	0.1617	0.1624
(0.9, 0.5, 0.3)	(i)	0.6074	0.5927	0.6011
	(ii)	0.2905	0.2995	0.2939
	(iii)	0.2899	0.1811	0.1825
(p_A, p_B, p_C)		Ridit without Covariate-Balancing n_0=5	Ridit with Prognostic Scoring and Strong Association n_0=5	Ridit With Prognostic Scoring and No Association n_0=5
(0.3, 0.3, 0.3)	(i)	0.3356	0.3350	0.3400
	(ii)	0.6973	0.6983	0.7026
	(iii)	0.2520	0.1630	0.1616
(0.4, 0.3, 0.3)	(i)	0.3752	0.3673	0.3641
	(ii)	0.6646	0.6640	0.6650
	(iii)	0.2531	0.1615	0.1631
(0.5, 0.3, 0.3)	(i)	0.4131	0.3959	0.3980
	(ii)	0.6167	0.6215	0.6188
	(iii)	0.2568	0.1622	0.1645
(0.6, 0.3, 0.3)	(i)	0.4587	0.4265	0.4302
	(ii)	0.5618	0.5724	0.5688
	(iii)	0.2628	0.1662	0.1659
(0.7, 0.3, 0.3)	(i)	0.4920	0.4584	0.4653
	(ii)	0.4996	0.5146	0.5148
	(iii)	0.2674	0.1691	0.1703
(0.8, 0.3, 0.3)	(i)	0.5375	0.4962	0.4944

	(ii)	0.4306	0.4540	0.4510
	(iii)	0.2762	0.1754	0.1760
(0.9, 0.3, 0.3)	(i)	0.5798	0.5297	0.5360
	(ii)	0.3514	0.3842	0.3788
	(iii)	0.2876	0.1813	0.1842
(0.5, 0.4, 0.3)	(i)	0.3946	0.3816	0.3823
	(ii)	0.5880	0.5876	0.5918
	(iii)	0.2561	0.1640	0.1623
(0.6, 0.4, 0.3)	(i)	0.4276	0.4152	0.4121
	(ii)	0.5391	0.5453	0.5431
	(iii)	0.2570	0.1632	0.1646
(0.7, 0.4, 0.3)	(i)	0.4730	0.4447	0.4384
	(ii)	0.4824	0.4925	0.4939
	(iii)	0.2643	0.1678	0.1678
(0.8, 0.4, 0.3)	(i)	0.5140	0.4752	0.4769
	(ii)	0.4163	0.4328	0.4356
	(iii)	0.2722	0.1720	0.1729
(0.9, 0.4, 0.3)	(i)	0.5537	0.5092	0.5122
	(ii)	0.3438	0.3674	0.3675
	(iii)	0.2799	0.1795	0.1783
(0.6, 0.5, 0.3)	(i)	0.4073	0.3934	0.3970
	(ii)	0.5095	0.5118	0.5162
	(iii)	0.2592	0.1655	0.1634
(0.7, 0.5, 0.3)	(i)	0.4470	0.4272	0.4179
	(ii)	0.4559	0.4648	0.4683
	(iii)	0.2654	0.1686	0.1669
(0.8, 0.5, 0.3)	(i)	0.4843	0.4569	0.4559
	(ii)	0.3970	0.4129	0.4133
	(iii)	0.2683	0.1711	0.1725
(0.9, 0.5, 0.3)	(i)	0.5276	0.4838	0.4921
	(ii)	0.3322	0.3520	0.3486
	(iii)	0.2808	0.1750	0.1781

Table S4: Standard deviations of operating characteristics (i-iii) for Ridit, GDL Urn, and Bayesian

(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 $n_0=5$	Stratified GDL Urn $n_0=5$	Stratified Ridit $n_0=5$	Bayesian With Prognostic Scoring $n_0=5$

(0.3, 0.3, 0.3)	(i)	0.0497	0.0720	0.0409	0.0703	0.0962
	(ii)	0.0546	0.0565	0.0573	0.0580	0.0571
	(iii)	0.0274	0.0309	0.0491	0.0425	0.0371
(0.4, 0.3, 0.3)	(i)	0.0588	0.0761	0.0472	0.0756	0.1041
	(ii)	0.0584	0.0605	0.0575	0.0593	0.0612
	(iii)	0.0254	0.0302	0.0527	0.0450	0.0370
(0.5, 0.3, 0.3)	(i)	0.0629	0.0775	0.0525	0.0754	0.1151
	(ii)	0.0603	0.0643	0.0606	0.0628	0.0653
	(iii)	0.0266	0.0316	0.0579	0.0464	0.0441
(0.6, 0.3, 0.3)	(i)	0.0754	0.0785	0.0591	0.0813	0.1134
	(ii)	0.0676	0.0628	0.0635	0.0661	0.0697
	(iii)	0.0283	0.0311	0.0597	0.0512	0.0504
(0.7, 0.3, 0.3)	(i)	0.0835	0.0780	0.0652	0.0790	0.1084
	(ii)	0.0710	0.0642	0.0669	0.0646	0.0723
	(iii)	0.0287	0.0324	0.0618	0.0578	0.0740
(0.8, 0.3, 0.3)	(i)	0.0893	0.0788	0.0656	0.0780	0.0909
	(ii)	0.0777	0.0621	0.0683	0.0628	0.0697
	(iii)	0.0341	0.0353	0.0651	0.0586	0.0912
(0.9, 0.3, 0.3)	(i)	0.0896	0.0791	0.0637	0.0750	0.0596
	(ii)	0.0764	0.0599	0.0648	0.0570	0.0536
	(iii)	0.0478	0.0401	0.0667	0.0656	0.0976
(0.5, 0.4, 0.3)	(i)	0.0654	0.0745	0.0519	0.0778	0.1118
	(ii)	0.0613	0.0595	0.0647	0.0615	0.0636
	(iii)	0.0247	0.0307	0.0567	0.0451	0.0419
(0.6, 0.4, 0.3)	(i)	0.0750	0.0774	0.0592	0.0795	0.1199
	(ii)	0.0648	0.0629	0.0628	0.0641	0.0672
	(iii)	0.0258	0.0313	0.0619	0.0488	0.0512
(0.7, 0.4, 0.3)	(i)	0.0852	0.0796	0.0657	0.0777	0.1164
	(ii)	0.0710	0.0631	0.0635	0.0601	0.0715
	(iii)	0.0274	0.0310	0.0646	0.0539	0.0676
(0.8, 0.4, 0.3)	(i)	0.0928	0.0760	0.0721	0.0791	0.1017
	(ii)	0.0742	0.0620	0.0660	0.0626	0.0701
	(iii)	0.0378	0.0340	0.0683	0.0602	0.0845
(0.9, 0.4, 0.3)	(i)	0.0947	0.0766	0.0679	0.0728	0.0692
	(ii)	0.0764	0.0550	0.0613	0.0544	0.0518
	(iii)	0.0451	0.0380	0.0698	0.0658	0.0936
(0.6, 0.5, 0.3)	(i)	0.0780	0.0762	0.0577	0.0784	0.1151
	(ii)	0.0643	0.0623	0.0607	0.0652	0.0639
	(iii)	0.0271	0.0306	0.0642	0.0527	0.0526
(0.7, 0.5, 0.3)	(i)	0.0896	0.0803	0.0698	0.0794	0.1258
	(ii)	0.0670	0.0619	0.0644	0.0652	0.0691
	(iii)	0.0283	0.0323	0.0693	0.0545	0.0613
(0.8, 0.5, 0.3)	(i)	0.0977	0.0753	0.0753	0.0772	0.1156
	(ii)	0.0726	0.0630	0.0630	0.0613	0.0697
	(iii)	0.0298	0.0337	0.0700	0.0606	0.0792
(0.9, 0.5, 0.3)	(i)	0.0964	0.0790	0.0723	0.0764	0.0883
	(ii)	0.0707	0.0540	0.0625	0.0559	0.0574
	(iii)	0.0423	0.0400	0.0710	0.0674	0.0964

Table S5: 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 S6: Averages of operating characteristics (i-iii) for different values of ϕ

(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 $n_0=5$ $\phi=2/3$	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