Sankhyā : The Indian Journal of Statistics
Special Issue on Biostatistics
2000, Volume 62, Series B, Pt. 1, pp. 43–69

USING URN MODELS FOR THE DESIGN OF CLINICAL TRIALS

By A. GREGORY DIRIENZO
Brown University, Providence

SUMMARY. When conducting a clinical trial to compare $K \geq 2$ treatments, it is essential to randomize patients to treatment in order to reduce experimental bias. Typically, an equal number of patients are assigned to each treatment group in order to increase the precision of inference concerning treatment effect. However, if the treatment difference is large and the endpoint potentially fatal, it seems unethical in an unmasked trial to sacrifice a large number of study patients for the purpose of maintaining balance in the number assigned to each treatment group. In this article, we survey those treatment randomization rules that are designed for use in each of the above two experimental situations, and have the empirical advantage of being well explained in terms of an urn model.

1. Introduction

The essential feature of controlled clinical trials is the random assignment of subjects to two or more treatment groups under investigation. The principle of randomization provides a protection against hidden biases and thus increases the validity of the trial's findings. Another fundamental aspect of clinical trials is the incorporation into the design and/or analysis those prognostic factors (e.g. gender, age, study center) which are known or thought to affect a patient's response to treatment. Ignoring such explanatory information may result in the trial yielding wrong conclusions, either by falsely claiming one treatment superior, or by obscuring true treatment differences.

Typically, eligible subjects arrive at an experimental site sequentially and must be treated immediately. The complete randomization scheme, which, for two treatment arms, determines treatment membership by tossing a fair coin independently each time a patient arrives, reduces or eliminates different kinds of experimental bias, and also provides a basis for statistical inference (Lehmann, 1975). However, in small-sized experiments, (less than 200 patients) complete randomization may result in a severe imbalance in the final number of patients assigned to each treatment group (Lachin, 1988b). As a result, inference about treatment effect may suffer from a loss in precision. Assigning an equal number of patients to each treatment group increases the accuracy of the inference concerning no treatment difference, result-

AMS (1991) *subject classification.* Primary: 62L05; Secondary: 60K30.
Key words and phrases. Clinical Trial, randomization, urn design, adaptive biased-coin design, play-the-winner design, response-adaptive design.

ing in the test having higher power. Another possible shortcoming of complete randomization is that a chance run of treatment assignments to one of the treatment groups can easily occur. If the baseline characteristics of the entering patients change with time, complete randomization could generate differences among treatment arms in the distribution of patients' baseline characteristics.

The permuted-block randomization design, which randomly allocates each treatment an equal number of times within each (of more than one) block, achieves periodic and (if the last block is filled) final balance in the number of patients assigned to each treatment arm. However, this design may easily introduce bias into an unmasked experiment by way of investigators selecting patients based on their expected response and the observed assignments made thus far in the block. In such situations, the Type I error probability may be substantially inflated. That is, under the null hypothesis of no treatment effect, one arm might exhibit a significantly better response to treatment and the null hypothesis rejected, solely because this treatment arm was composed of 'healthier' patients compared to the other arms. This type of bias is referred to as 'selection bias' (Blackwell and Hodges, 1957).

It is therefore desirable to implement a randomization scheme that is a compromise between a completely randomized one and one that achieves perfect balance. Efron (1971) introduced a biased coin design that operates in exactly the same manner as the complete randomization scheme, except that for each assignment, the coin is biased in favor of the treatment group that has thus far been assigned least frequently. However, the biased coin design is unrealistic in that it is independent of both the actual value of the difference in the number of patients currently assigned to each treatment group (apart from the sign of this difference) and the total number of subjects enrolled in the study so far.

In section 2, we describe the working character, how to implement and how to analyze a class of designs that are a generalization of the biased coin design of Efron (1971). Unlike Efron's biased coin design, these designs take into account the current degree of imbalance in the number of patients assigned to each treatment group before allocating the next patient to treatment. Furthermore, these designs have the advantage of forcing periodic and (in small-sized trials) final balance while asymptotically tending toward the complete randomization scheme, for which experimental bias is eliminated. The adaptive-biased coin designs that we describe have the advantage of being well explained in terms of an urn model and thus are in principle easily implemented in practice.

A completely different set of qualities of a randomization scheme are required when the response of the patient is potentially fatal and the treatment difference large. In these situations, a randomization scheme that attempts to force equality in the numbers assigned to each treatment group might sacrifice a large number of study patients. Instead, it would be attractive to randomize more patients to the better treatment while at the same time providing sound information about the relative merit of the treatment regimes. Such designs are referred to as 'response-adaptive designs' and by nature are unbalanced, since they deliberately attempt to randomize more patients to the superior treatment arm. As a consequence, these designs may not be as efficient as a perfectly balanced one. In section 3, we describe

a class of response-adaptive randomization rules that may also be explained in terms of an urn model. We discuss the properties of these designs, how to implement them, and present methods to analyze the results. Several real-life examples are provided.

Turning to the analysis of study results, the assumption that patient responses are independent and identically distributed (i.i.d.) observations from some well-defined homogeneous population provides a formal sampling basis for statistical inference to be developed from a *population model*. However, this assumption is sometimes questionable in clinical trials. As a result, unless otherwise stated, we do not impose such requirements. Instead, we adopt as the basis for inference that which is provided by the use of randomization, called the *randomization model*, from which permutation tests are derived. In a randomization model, a popular way in which information from prognostic factors is incorporated into the study is through the separate application of the treatment assignment rule within each prognostic category of patients. An overall permutation test is then easily constructed, since prospectively stratified permutation tests are mutually independent for any randomization rule. On the other hand, although such prognostic information may be accounted for in an analysis by stratifying patients at trial completion, a post-stratified permutation test may or may not be tractable, depending on the treatment rule employed.

2. Urn models for adaptive-biased coin designs

In this section, we overview a class of adaptive-biased coin designs that are a generalization of the biased coin design introduced by Efron (1971).

For ease of discussion, suppose that there are two treatment groups, denoted by A and B. An explanation will be given when the extension to $K \geq 2$ treatment groups is not straightforward. During the execution of the trial, eligible patients arrive at a study center sequentially and must be immediately assigned to therapy. Suppose that the size of the trial, either overall, within a stratum or at a stopping time, is not known beforehand. What is therefore faced is a design problem of a sequential clinical trial whose size cannot be predetermined.

The class of adaptive-biased coin designs (Smith, 1984) that we entertain may best be explained after an arbitrary number of assignments, say $n > 1$. Since this is a symmetric design, the first treatment assignment is determined by tossing a fair coin. Let n_A and n_B, $n = n_A + n_B$, denote the number of previous assignments to treatment group A and B, respectively. Define the present imbalance in the number of patients assigned to each treatment group by $D_n = n_A - n_B$ and denote by $\varphi(D_n, n)$ a nonincreasing function in D_n/n with range from zero to one. The function $\varphi(.)$ is assumed to satisfy $\varphi(D_n, n) + \varphi(-D_n, n) = 1$. The $(n+1)$st subject is assigned to treatment group A with probability $\varphi(D_n, n)$ and to treatment group B with probability $1 - \varphi(D_n, n)$. Consider the sub-class of designs with $\varphi(D_n, n) = \{1 - g(D_n, n)\}/2$, which, for a particular transformation $g(.)$, have been shown to generate a satisfactory design that may be well explained in terms of an urn model (Wei, 1977). These designs thus are in principle easily implemented and satisfactory for practical use.

A generalized Friedman's urn model (Friedman, 1949) is used to illustrate the working character of the sub-class of adaptive-biased coin designs with

$$\varphi(D_n, n) = \frac{1}{2}\left(1 - \frac{\beta D_n}{2\alpha + \beta n}\right), \qquad \alpha, \beta \geq 0, \tag{1}$$

with at least one of α, β positive. Note that $\varphi(D_n, n)$ is monotonically decreasing in D_n for fixed n and tends toward $1/2$ as n becomes large for fixed D_n. Formation of the treatment assignment sequence using (1) may be explained as follows. An urn initially contains α white and α red balls. When a patient is available for assignment, a ball is randomly drawn with replacement from the urn and its color noted. If it is a white ball, the assignment is to treatment A, and β red balls are added to the urn; otherwise if it is a red ball, the assignment is to treatment B and β white balls are added to the urn. This design we denote by UD(α, β). For the case when $\alpha = 0, \beta > 0$, the first assignment is determined by tossing a fair coin, and then adding to the urn β balls of the corresponding color of the treatment not assigned, commencing the draws from the urn when the next eligible patient arrives. Note that a larger value of α with respect to β introduces more randomness in the beginning of the trial, otherwise more balance is enforced.

The UD$(0, \beta), \beta > 0$ is independent of the value of β and is similar to the random allocation design (permuted-block design with one block), which is perfectly balanced. Notice that the UD$(\alpha, 0), \alpha > 0$ is the complete randomization scheme.

The function $\varphi(.)$ corresponding to the biased coin design [BCD(p)] of Efron (1971) is

$$\varphi(D_n, n) = \begin{cases} 1 - p & D_n/n > 0 \\ 1/2 & D_n/n = 0 \\ p & D_n/n < 0 \end{cases}$$

where $1/2 \leq p \leq 1$. Other choices of $\varphi(.)$ are given in Smith (1984).

For the case when there are $K > 2$ treatment groups, the allocation of patients to therapy with the UD(α, β) proceeds in exactly the same manner as described above, except that now, β balls of the corresponding color of each treatment other than the one chosen are added to the urn after each assignment. Application and properties of the urn design when multiple treatment groups are being compared are described in Wei (1978) and Wei, Smythe and Smith (1986), respectively.

2.1. *Balancing property of the urn design.* The UD(α, β) design achieves periodic balance in the number of patients assigned to each treatment arm and as a result forces small sized trials to be balanced with respect to the final number of patients assigned to group A and B, denoted by N_A and N_B, respectively. The final total sample size is denoted by $N = N_A + N_B$. Table 1, taken from Wei (1977), displays, for several different randomization rules, the probability that a small trial will be perfectly balanced. Notice that the UD$(0, \beta)$ forces the trial to be more balanced than the other designs for the small values of n. This feature of the UD$(0, \beta)$ design is very attractive when the total size of the trial is unknown beforehand.

For moderate to large n, the probability of an imbalance with the UD(α, β) design may be assessed by noting that $P(|n_A - n_B| > d)$ may be approximated by

$2\Phi(-Z_d)$, where

$$Z_d = (d+0.5) \Big/ \left\{ \frac{n(\alpha+\beta)}{3\beta+\alpha} \right\}^{1/2}$$

and $\Phi(.)$ is the standard normal distribution function. For the UD$(0,1)$ design, Z_d reduces to $Z_d = (d+0.5)/\sqrt{n/3}$. For complete randomization, the probability of an imbalance, $P(|n_A - n_B| > d)$, may be approximated by $2\Phi\{-(d+0.5)/\sqrt{n}\}$. Stated in terms of proportions, define $q = \max(n_A, n_B)/n$. The probability of an imbalance is now written as $P(q > r)$ and is approximately $2\Phi\{-2(r-0.5)\sqrt{3n}\}$ for the UD(0,1) and $2\Phi\{-2(r-0.5)\sqrt{n}\}$ for complete randomization. The corresponding percentile of the standard normal distribution decreases on the order of $\sqrt{3n}$ for the UD(0,1) and on the order of $\sqrt{n}$ for complete randomization. As a result, as n increases, imbalances are far less likely with the UD(α, β) than with complete randomization.

Table 1. PROBABILITY THAT A TRIAL IS EXACTLY BALANCED AFTER n TREATMENT ALLOCATIONS.

	n				
Design	2	4	6	8	10
permuted-block with length 10	0.56	0.48	0.48	0.56	1.00
UD$(0,\beta)$	1.00	0.67	0.55	0.48	0.43
BCD$(\frac{2}{3})$	0.67	0.59	0.56	0.54	0.53
UD$(\alpha,0)$	0.50	0.38	0.31	0.27	0.25

The efficiency of the UD$(0,\beta)$ design may be shown to compare with that of the random allocation design which yields a perfectly balanced design. Denote by Y_A and Y_B the response of a patient treated by strategy A and B, which have mean-value denoted by μ_A and μ_B, respectively. Suppose that the responses have a common variance denoted by σ^2. Consider as an estimate of $\mu_A - \mu_B$ the difference between the sample mean response in treatment A and the sample mean response in treatment B, $\overline{Y}_A - \overline{Y}_B$, which has variance $\sigma^2(1/N_A + 1/N_B)$. Prespecifying the total sample size as $N = 2m$ allows a random allocation rule to be invoked, and as a consequence minimizes the quantity $1/N_A + 1/N_B$ at $2/m$.

If N is not known beforehand, an interesting question is how many more additional observations would the UD$(0,\beta)$ design need in order for

$$\frac{1}{N_A} + \frac{1}{N_B} \leq \frac{2}{m}. \tag{2}$$

If we define the random variable $U = N_A + N_B - 2m$ to be the number of extra patients needed to satisfy (2), then for large m and any $u > 0$, with the UD$(0,\beta)$ design

$$P(U \leq u) \doteq \Phi(\sqrt{3u}) - \Phi(-\sqrt{3u})$$

(Wei, 1978b). For example, when m is large, $P(U \leq 4)$ is approximately 0.9995. It is essentially guaranteed that for large m, when using the UD$(0,\beta)$ design, at most four additional observations would be needed to achieve at least as precise inference

concerning treatment effect as that if the allocation had been perfectly balanced. Incidentally, Blackwell and Hodges (1957) show that for complete randomization, $P(U \leq u) \doteq \Phi(\sqrt{u}) - \Phi(-\sqrt{u})$ with $P(U \leq 4) \doteq 0.95$.

More generally, when both the mean and variance of the outcome depend on treatment group, Eisele (1994) considers a doubly adaptive design that takes into account the current number of subjects assigned to each treatment along with current estimates of treatment effect to estimate the desired allocation probability. For instance, for the fixed width interval estimation of the difference between two population means, the design of Eisele (1994) will at each stage estimate the correct allocation probability for minimizing total sample size, while retaining pre-assigned coverage probability and interval width.

Table 2. EXPECTED DEVIATION OF THE NUMBER OF CORRECT GUESSES FROM $N/3$.

	N						
Design	3	6	9	12	21	27	36
permuted-block with length N	0.83	1.67	2.50	3.33	5.83	7.50	10.00
UD$(1,1)$	0.16	0.37	0.54	0.69	1.06	1.26	1.53

Each entry for UD$(1,1)$ is based on 5,000 simulations.

2.2. *Randomization property of the urn design.* A natural measure for the amount of selection bias of a randomization rule is the expected number of correct guesses of treatment assignments in excess of that possible by chance alone that an experimenter can make if they guess optimally (Blackwell and Hodges, 1957). This has been called the 'expected bias factor' by Lachin (1988a). For complete randomization, the expected bias factor is equal to zero and there is no selection bias.

For the case of the UD(α, β), the optimal strategy is to guess the treatment which has been thus far assigned least frequently, with no preference if there is a tie. After n treatment assignments, the probability of guessing the $(n+1)$st treatment assignment correctly when using the UD(α, β) rule is

$$\frac{1}{2} + \frac{\beta E|n_A - n_B|}{2(2\alpha + \beta n)}$$

(Wei, 1977), which is asymptotically equal to 1/2. Therefore, as the size of the trial becomes large, selection bias is gradually eliminated with the UD(α, β). For the BCD(p), the expected number of excess correct guesses in $2m$ assignments is asymptotically $(\eta - 1)m/(2\eta)$, where $\eta = p/(1-p)$ (Efron, 1971). Table 2 compares, for $K = 3$ treatments, the permuted-block design to the UD$(1,1)$ in terms of the expected deviation of the number of correct guesses from $N/3$. Compared to the permuted-block design, the UD$(1,1)$ greatly reduces selection bias.

Another type of experimental bias is accidental bias (Efron, 1971), which may be caused by an imbalance between treatment groups with respect to one or more prognostic factors, either known or unknown to the investigator. With the UD(α, β),

as $n \to \infty$, the sequence of treatment assignments becomes uncorrelated. Thus, in a large scale trial, the UD(α, β) gradually behaves like complete randomization and is asymptotically free of accidental bias.

2.3. *Randomization-based analysis of study results.* There is a fundamental difference between the population model and the randomization model as a basis for constructing a statistical test of the null hypothesis H_0 of no difference in the distribution of the outcome variable between treatment groups. The population model assumes that the patients that are enrolled in the trial are a random sample from a well defined homogeneous population, in which case their responses may be considered to be i.i.d. under the null hypothesis. This convention provides a vehicle to derive the null distribution of the test statistic. However, the population model is sometimes questionable in clinical trials, since those selected, eligible, consenting patients enrolled are recruited from various sources by a non-random selection of clinics.

The use of randomization in a sequentially controlled clinical trial provides a basis for an assumption-free statistical test of the hypothesis of equal efficacy of the two treatments on response among the patients entered in the trial (Lehmann, 1975). With a randomization model, the sequence of responses are considered fixed constants and the sequence of treatment assignments considered random. Thus for any given sequence of treatment assignments, the value and associated probability of the statistic employed for testing H_0 is entirely defined by the observed sequence of responses and treatments, and the particular randomization scheme invoked, respectively. It is therefore possible to enumerate all possible sequences of treatment assignments (and test statistic values), along with their associated probability of selection as determined by the particular design implemented, to generate the exact null distribution of the test-statistic and provide a decision on H_0. With complete randomization, all possible treatment sequences have the same probability, equal to the inverse of the total number of sequences, $1/2^N$. For the urn design, however, not all treatment sequences have the same probability of occurrence and some sequences actually have probability equal to 0.

We now describe how to perform randomization-based analysis of the results of a single trial whose treatment sequence was generated from the UD(α, β). Note that the trial may be a sub-trial corresponding to a stratified randomization, for which the analysis would be performed separately within each stratum.

Consider the family of linear rank test-statistics. Let Y_i be equal to 0 or 1, according to whether treatment A or B is assigned to the ith eligible patient, respectively, $i = 1, \ldots, N$, and denote the sequence of observed responses by $bx = x_1, \ldots, x_N$. Denote the corresponding scores of bx, which may be the rank of x_i among bx, by $a_{1N}, \ldots, a_{NN}$, with overall mean $\overline{a}_N$. The linear-rank test statistic is written as

$$T_N = \sum_{i=1}^{N} (a_{iN} - \overline{a}_N) \left(Y_i - \frac{1}{2} \right).$$

As given by Lachin (1988a), under a population model or a simple (complete or random allocation) randomization model, the scores a_{iN} may be selected so as

to result in T_N being algebraically equivalent to, for example, the chi-square test for 2×2 tables, the Wilcoxon rank sum test, the Log-rank or Peto-Peto-Prentice-Wilcoxon tests, depending on the nature of the responses.

The exhaustive enumeration of all possible sequences of treatment assignments can become unwieldy even for moderately sized trials. As a result, the finite-sample null distribution of the linear-rank test statistic may usually be approximated by a normal distribution, with variance depending on the randomization rule used. When the UD(α, β) rule is used to assign patients to treatment groups, a sufficient condition on the scores for the (null) asymptotic normality of the test statistic T_N is, as $N \to \infty$

$$\frac{\max\limits_{1 \le i \le N} (a_{iN} - \overline{a}_N)^2}{\sum_{i=1}^{N} (a_{iN} - ova_N)^2} \to 0 \tag{3}$$

(Smythe and Wei, 1983). It is demonstrated in Smythe and Wei (1983) that this condition is general enough for practical use. For instance, (3) is trivially satisfied if a_{iN} is the rank of x_i among $\mathbf{x}$, in which case T_N corresponds to the Wilcoxon statistic. Smythe and Wei (1983) also provide a consistent estimate of the asymptotic variance of T_N (note that under H_0, $E(T_N) = 0$) which takes the form

$$\widehat{v}(T_N) = \frac{1}{4} \sum_{i=1}^{N} b_{iN}^2,$$

where

$$\begin{aligned} b_{iN} &= (a_{iN} - \overline{a}_N) - \sum_{j=i+1}^{N} \frac{\{2\alpha + \beta(i-1)\}\beta(a_{jN} - \overline{a}_N)}{\{2\alpha + \beta(j-1)\}\{2\alpha + \beta(j-2)\}}, \quad 1 \le i \le (N-1) \\ b_{NN} &= a_{NN} - \overline{a}_N. \end{aligned}$$

For the special case of the UD$(0, \beta)$ design, the form of b_{iN} above reduces to

$$b_{iN} = (a_{iN} - \overline{a}_N) - \sum_{j=i+1}^{N} \frac{(a_{jN} - \overline{a}_N)(i-1)}{(j-1)(j-2)}, \quad 1 \le i \le (N-1).$$

It should be noted that for complete randomization, $b_{iN} = a_{iN} - \overline{a}_N$ and the variance estimate of T_N is $\widehat{v}(T_N) = \sum (a_{iN} - \overline{a}_N)^2/4$, with the sum from $i = 1, \ldots, N$. In either case, $T_N / \sqrt{\widehat{v}(T_N)}$ is asymptotically standard normal.

In the case when there are more than two treatment arms, a large-sample approximation to the null permutation distribution of the linear rank test-statistic for a class of adaptive-biased coin designs is given by Wei *et al.* (1986).

For smaller sized experiments using restricted randomization, Cox (1982) suggests that it is more appealing to use a conditional randomization test, whereby the null permutational distribution of the test-statistic is taken not over all possible

treatment sequences, but only those that result in nearly the same final balance, e.g. $\delta_N = N_A - N_B$. This is similar to conditioning on ancillary statistics in a population model; for restricted randomization designs, N_A and N_B provide no information regarding the true treatment difference. For several adaptive-biased coin designs, including the UD (α, β) and permuted-block design, Mehta, Patel and Wei (1988) provide an efficient recursive algorithm which generates the exact permutational distribution for linear rank test-statistics, given the final difference between the number of patients assigned to each treatment arm, δ_N.

For a class of adaptive-biased coin designs, Wei *et al.* (1986) provide the asymptotic null permutational distribution of T_N conditional on δ_N, which is justified later by Smythe (1988). For the UD$(0, \beta)$ design, the conditional distribution of T_N given δ_N can by asymptotically approximated by a normal distribution with mean

$$\widehat{\mu} = \left(\delta_N \sum_{i=1}^{N} b_{iN}\bar{b}_{iN}\right) \Bigg/ \left(2\sqrt{N}\sum_{i=1}^{N}\bar{b}_{iN}^2\right)$$

and variance

$$\widehat{v} = \frac{1}{4}\sum_{i=1}^{N} b_{iN}^2 \left[1 - \left\{\left(\sum_{i=1}^{N} b_{iN}\bar{b}_{iN}\right)^2 \Bigg/ \left(\sum_{i=1}^{N}\bar{b}_{iN}^2\right)\right\}\right],$$

where b_{iN} is defined above and $\bar{b}_{iN}$ is defined as b_{iN} with $n^{-1/2}$ replacing $(a_{iN} - \bar{a}_N)$. Mehta *et al.* (1988) suggest that this large sample approximation is satisfactory by the time the total trial size reaches around 30 patients.

It is illustrated in Mehta *et al.* (1988), that in general, one cannot ignore the actual treatment assignment rule employed when calculating the permutational distribution of the linear rank test-statistic. Different randomization rules can generate vastly different permutational distributions of the linear rank test-statistic when the responses are not i.i.d., for example when there is a time trend in the observations. As a consequence, assuming a different randomization scheme than one implemented for an analysis may lead to the resulting permutational test having either supra-nominal or sub-nominal size (Efron, 1971). On the other hand, it should be noted that, if for various randomization rules, the permutational distribution of the linear rank test-statistic does not differ considerably from that corresponding to complete randomization, then it is plausible to assume that the responses are a random sample from some well-defined homogeneous population, and the analysis may, for simplicity, assume that complete randomization generated the treatment sequence.

As previously described, under a randomization model information from prognostic factors thought or known to influence response is properly accounted for by treating each of S stratum as an independent sub-trial, with the randomization and analysis performed separately within each one (if the number of patients that fall within each stratum is not too small). In such a prospectively stratified trial, the overall linear rank test-statistic is a linear combination of the S independent linear rank test-statistics. That is, if $T_N^{(s)}$ is the stratum-specific rank test with null

expectation $\mu^{(s)}$ and variance $v^{(s)}$, $s = 1, \ldots, S$, then the overall test of H_0 is given by

$$\frac{\sum_{s=1}^{S} w_s (T_N^{(s)} - \mu^{(s)})}{(\sum_{s=1}^{S} w_s^2 v^{(s)})^{1/2}},$$

which is asymptotically standard normal under H_0 for any weight function $\{w_s\}$. Methods for how to choose the weight function are given in Lachin (1988a). Mehta *et al.* (1988) describe how to extend their recursive algorithm in order to obtain the exact conditional permutational distribution of such an overall test-statistic generated from the UD(α, β) rule.

2.4. *Post-stratified analyses.* Suppose that a randomization model is the basis for inference, and at the conclusion of the trial, it is of interest to conduct a post-stratified or subgroup analysis based on, say, demographic, or clinical factors that were not considered at the onset of the trial for use in executing a stratified randomization. Although certain forms of post-stratified analysis have been cautioned in the past (Peto *et al.*, 1977), they may provide useful information for the conduct of future trials. For instance, the investigator may wish to know if there are treatment effects within specific sub-classes of patients, corresponding to say, age, gender, prior treatment, etc. Additionally, it may be of interest to know if there is interaction between treatment effect and the levels of a particular factor. Post-stratified analyses may also provide a covariate adjusted test of treatment effect by combining the separate tests over all strata considered.

When randomization (of any type) is prospectively stratified, each stratum is considered an independent sub-trial and the corresponding S linear rank test-statistics calculated within each are treated as being mutually independent. This is also the case when S post-stratified analyses are performed when the treatment sequence was generated from complete randomization, or when a population model is the basis for inference. In both cases, multivariate inference is straightforward, in that only S variance components and not $S(S+1)/2$ variance-covariance components need to be estimated. However, with restricted randomization rules, such as the UD(α, β), post-stratified linear rank test-statistics possess a non-identity joint correlation structure.

For the urn design, Davis (1989) has investigated the asymptotic theory for two-sample post-stratified permutation tests of the (global) null hypothesis and other hypotheses. Here we will describe the unconditional test corresponding to the UD$(0, 1)$. For simplicity, assume that no prospectively stratified randomization was conducted, the extension to such cases is trivial. Consider the two sample linear rank test-statistics $\{T_N^{(s)}\}$, calculated within each post-stratified stratum as tests of the stratum-specific null hypotheses $\{H_s\}$, $s = 1, \ldots, S$. The test-statistics take the form

$$T_N^{(s)} = \sum_{i=1}^{N} \eta_{is} (a_{iN} - \overline{a}_{sN}) \{Y_i - E(Y_i - 0.5)\},$$

where

$$\overline{a}_{sN} = \frac{\sum_{i=1}^{N} \eta_{is} a_{iN}}{\sum_{i=1}^{N} \eta_{is}}, \qquad s = 1, \ldots, S.$$

Here, the score a_{iN} may now be the rank of x_i among those in the stratum it belongs to, η_{is} is one if subject i belongs to stratum s, zero otherwise, $i = 1, \ldots, N$, $s = 1, \ldots, S$, and $\sum \eta_{is} = 1$ where the summation is from $s = 1, \ldots, S$. Unconditional on the final imbalance δ_N, Davis (1989) justifies, for mild conditions on the scores $\{a_{iN}\}$ equivalent to (3), the convergence of the joint null permutational distribution of $\{T_N^{(s)}\}$ to a multivariate normal with mean-vector $\mathbf{0}$ and (nonstochastic) $S \times S$ variance-covariance matrix for which

$$\widehat{\mathbf{\Psi}} = \frac{1}{4} \sum_{i=1}^{N} D_i' D_i,$$

with $D_i = (d_{i1}, \ldots, d_{iS})$ and

$$d_{is} = \eta_{is}(a_{iN} - \overline{a}_{sN}) - \sum_{j=i+1}^{N} \frac{\eta_{is}(a_{iN} - \overline{a}_{sN})(i-1)}{(j-1)(j-2)}, \qquad i = 1, \ldots, N.$$

Note that the $\{d_{is}\}$ may be generalized to correspond to the UD(α, β). Unfortunately, the post-stratified permutation test has not been developed to condition on the final within stratum treatment imbalance when treatment assignments were made using the urn design.

The condition on the scores is satisfied if for example, both the score a_{iN} is defined as the rank of x_i within its respective stratum, and $\sum \eta_{is} = O(N)$, for each $s = 1, \ldots, S$, where the sum is from $i = 1, \ldots, N$. If the censoring mechanism acts the same on both treatment groups, then both the log-rank and Peto-Peto-Prentice generalized Wilcoxon test satisfy this condition. It should be noted that this distributional result does not apply to Efron's (1971) biased coin design, since the associated probability function $\varphi(.)$ is not (in general) continuous at zero (Davis, 1989).

One may test the null hypothesis H_s of no treatment effect on the distribution of the defined outcome variable for those individuals characterized by membership to stratum $s = 1, \ldots, S$, by the normal deviate $T_N^{(s)}/(\widehat{\psi}_{ss})^{1/2}$, $\widehat{\psi}_{ss}$ the sth diagonal element of $\widehat{\mathbf{\Psi}}$. However, since the S hypotheses are related, a simultaneous inference scheme is probably more appropriate. That is, one should test the global null hypothesis $H = \cap H_s$. The following are some commonly used tests of H: First, when all test-statistics point in the same direction, one may calculate a standardized weighted average (weights depending upon $\widehat{\mathbf{\Psi}}$) of the S test statistics, which takes the form

$$\frac{\mathbf{w}\boldsymbol{T}'_N}{(\mathbf{w}\widehat{\mathbf{\Psi}}\mathbf{w}')^{1/2}},$$

where $\mathbf{w} = (w_1, \ldots, w_s)$ are the weights and $\boldsymbol{T}_N = (T_N^{(1)}, \ldots, T_N^{(S)})$. Next, an

omnibus test statistic may be calculated, that takes the quadratic form

$$\boldsymbol{T}_N \widehat{\boldsymbol{\Psi}}^{-1} \boldsymbol{T}'_N,$$

which is asymptotically chi-square on S d.f. under H. Finally, in order to identify treatment differences within individual strata while simultaneously maintaining the global type-I error level, one may implement a sequential multiple testing method (Marcus, Peritz and Gabriel, 1976; Holm, 1979). Each of these methods is well explained in Wei and Glidden (1997) and the references therein. An excellent example that illustrates these methods on a prostatic cancer study and demonstrates that the specific randomization rule cannot in general be ignored when making randomization-based, post-stratified inference about treatment effect is given in Davis (1989).

Note that when some patients' responses are missing at random (independent of treatment group), a valid permutation test of treatment effect can be justified by considering the single stratum of patients with observed responses as a post-stratified subgroup.

2.5. *Accommodating large numbers of strata.* A standard feature of the controlled clinical trial is to take into account in either the design and/or analysis various prognostic factors that are known or thought to influence a patients response to treatment. Suppose that there are M prognostic factors, each one having L_i levels, $i = 1, \ldots, M$. There are thus $S = \prod L_i$ strata and $\sum L_i$ levels in the trial, where the product and sum is over $i = 1, \ldots, M$. Under a randomization model, the proper design is to perform the experiment separately within each stratum. When the number of strata is so large that too few subjects fall within each one to allow a separate randomization to be performed within each stratum (e.g. Zelen, 1974), an overall treatment assignment rule should be considered (Pocock and Simon, 1975; Freedman and White, 1976).

Consider the case when there is both no interaction among prognostic factors and no interaction between treatment and prognostic factors. In such situations it is not necessary to balance the treatment group numbers within each stratum, rather it is satisfactory to maintain balance only within each level of each factor. In this section, we show how to implement the UD(α, β) in order to achieve treatment balance simultaneously across all factor levels.

For each level j of factor i, let there be a corresponding urn U_{ij}, which initially consists of α_i white and red balls each, $j = 1, \ldots, L_i$, $i = 1, \ldots, M$. Consider some arbitrary stage of the sequential trial. At this time, suppose that there are n_{Aij} and n_{Bij} white and red balls, respectively, in each urn U_{ij}, with $n_{Aij} + n_{Bij} = n_{ij}$ denoting the total number of balls in each urn thus far. Suppose that the next patient to enter the study has a characteristic that is given by the levels $\ell_1, \ldots, \ell_M$ of the M corresponding prognostic factors. To obtain balance in the numbers of patients assigned to each treatment group within each level of each prognostic factor we do the following. Select from the set of M urns $\{U_{i\ell_i} : i = 1, \ldots, M\}$ with largest probability the one with the greatest imbalance in the numbers of patients randomized to each treatment group so far and in turn use this urn to generate the

present treatment assignment. Since imbalance in the number of previously assigned patients to each treatment group in each factor level ℓ_i is evidenced by the variability of $\{n_{Ai\ell_i}/n_{i\ell_i}, n_{Bi\ell_i}/n_{i\ell_i}\}$, $i = 1, \ldots, M$, rank the urns from smallest to largest (with a random ordering in case of ties) in terms of the value $\nu_i = \sigma(n_{Ai\ell_i}/n_{i\ell_i}, n_{Bi\ell_i}/n_{i\ell_i})$, $i = 1, \ldots, M$, for $\sigma(.)$ some function that measures variation such as the range or variance. Each of the M ranked urns is selected for generating the treatment to be assigned with corresponding probability denoted by $p_{(i)}$, $i = 1, \ldots, M$, where $p_{(1)} \leq \ldots \leq p_{(M)}$ and $p_{(i)}$ is some constant [e.g. $p_{(M)} = 1, p_{(i)} = 0; 1 \leq i \leq (M-1)$] or a function of $\{\nu_i\}$, with $\sum p_{(i)} = 1$, where the sum is from $i = 1, \ldots, M$. Thus the urn with the most imbalance (largest ν) is selected with the highest probability. Suppose that the urn $U_{m\ell_m}$ is chosen. A ball is drawn at random and replaced. If it is a white ball, treatment A is assigned and β_i red balls are added to each urn $\{U_{i\ell_i}\}$, $i = 1, \ldots, M$. If it is a red ball, treatment B is assigned and β_i white balls are added to each urn $\{U_{i\ell_i}\}$, $i = 1, \ldots, M$. The procedure is repeated when the next eligible patient is ready for treatment assignment.

As an illustration, consider the case when two treatments are being compared in an AIDS clinical trial. Suppose that the following three prognostic factors are known to affect a patients response: institution (total of 10), duration of prior zidovudine use (< 6 weeks, ≥ 6 weeks), and baseline CD4 cell count (< 250 cells/mm^3, ≥ 250 cells/mm^3). Let the function $\sigma(.)$ be the range, i.e. $\sigma(x, y) = x - y$. For this example take $\alpha_i = 1$, $\beta_i = 1$, and $p_{(3)} = 1$, where $i = 1, 2, 3$.

Since there are 10 institutes in the study, very few patients are expected to fall within each stratum. Suppose that the next eligible patient to enter the trial appears at institution #3, has been taking zidovudine for more than 6 weeks and has a CD4 count of greater than 250 cells/mm^3. Further suppose that the distribution of prior treatment assignments in the corresponding factor levels is as given in Table 3.

Table 3. DISTRIBUTION OF PRIOR TREATMENT ASSIGNMENTS.

	Factor 1	Factor 2	Factor 3
Treatment	INSTITUTION #3	ZIDOVUDINE $\geq$ 6 WEEKS	CD4$\geq$ 250
A	2	5	8
B	0	6	11
TOTAL	2	11	19

Table 4. DISTRIBUTION OF NUMBERS OF BALLS IN URNS.

	Urn		
Ball color	U_{13}	U_{22}	U_{32}
A (white)	1	7	12
B (red)	3	6	9
TOTAL	4	13	21

Even though the greatest difference in the number of patients assigned to the two treatment groups occurs in the CD4 stratum, the imbalance in the institution stratum is the most serious. Treatment B should be assigned to this patient with higher probability than treatment A.

The distribution of the numbers of balls in each urn is given in Table 4. It follows that $\nu_1 = 1/2$, $\nu_2 = 1/13$ and $\nu_3 = 1/7$, therefore the urn U_{13} would be used to determine which treatment this patient should receive. The probability that this patient would be assigned to treatment B is 3/4.

3. **Urn models for response-adaptive designs**

Typically, when conducting a randomized, controlled sequential clinical trial to compare two treatment strategies, a conventional 50/50 randomization rule, such as the permuted-block or urn design, is used to allocate treatment to patients. Such conventional rules randomly assign patients sequentially, but also keep a certain degree of balance between the number of patients assigned to each group throughout the trial. Inference concerning treatment effect tends to more efficient when there is an equal number of patients assigned to each treatment group. However, when performing the trial, it is not only desirable to derive reliable information about the relative efficacy of the treatments. We are ethically obliged to treat each patient in the best way possible. This is especially true when the response is potentially fatal and the treatment difference large. In these situations, it would be attractive to be able to randomize more patients to the better treatment, while simultaneously providing sound information about the relative efficacy of the two treatment strategies.

A response-adaptive design is one that sequentially uses accumulating information about the treatment difference, as evidenced by the outcome data already observed, to make it more likely for incoming patients to be assigned to the treatment group which is performing better thus far (although not conclusively). For example, a particular response-adaptive design, called the randomized play-the-winner design, was implemented in a recent prospective controlled randomized trial for the use of extracorporeal membrane oxygenation (ECMO) to treat newborns with persistent pulmonary hypertension (Bartlett *et al.*, 1985; Cornell, Landenberger and Bartlett, 1986). ECMO is an artificial heart-lung machine that recycles blood through a membrane exposed to a high concentration of oxygen. The control treatment was the conventional mechanical ventilation therapy and historically had an associated probability of death of at least 0.8. The ECMO therapy, however, entailed a surgical procedure with potential life-threatening complications to the newborns. The response to treatment, either death or lung recovery, could be observed within a few days after treatment. It was determined by the investigators that a response-adaptive design should be used to randomize treatment to patients in order to best accommodate the needs of the patients while simultaneously providing an experiment capable of both a valid and fruitful treatment comparison.

The randomized play-the-winner design (Wei, 1978), is well explained in terms of an urn model, and thus in principle, easily implemented in practice. An urn initially contains $u \geq 0$ balls of each color white and red. When a patient is available for assignment, a ball is drawn at random from the urn and replaced. If it is a white ball the assignment is to treatment A, if it is a red ball, the assignment

is to treatment B. When the response of a previous patient on treatment $k = A, B$ becomes available, the structure of the urn is changed as follows. If the response was a 'success', then β balls of type k are added to the urn and α balls of type j are added to the urn. On the other hand, if the response on treatment k was a 'failure', then β balls of type j and α balls of type k are added to the urn. Here $\beta \geq \alpha \geq 0$ and $k \neq j$. The design is denoted by RPW(u, α, β). In the case when the urn is empty, treatment group assignment is decided by the independent toss of a fair coin. Note that the response has been assumed to be binary. Extensions to continuous response variables are discussed below.

Previously, Zelen (1969) introduced the play-the-winner rule, which is implemented in exactly the same way as the RPW $(0, 0, 1)$, except that for each treatment assignment, the ball is drawn *without* replacement from the urn. Thus, if on average, the time to response is much longer than the time between patient entries into the trial, the play-the-winner rule assigns most treatments by the complete randomization rule, which doesn't tend to put more patients on the better arm. If, however, a patients' response to treatment is known before the succeeding patient is available for assignment, the play-the-winner rule specifies that after each successful response to treatment, we assign the same treatment and after each failed response, we assign the opposite treatment. Zelen (1969) termed this the modified play-the-winner rule, which does assign more patients to the better treatment arm, but is deterministic and may easily induce experimental bias. The randomized play-the-winner design assigns more patients to the better treatment, yet is not deterministic and is less vulnerable to experimental bias than the modified play-the-winner rule, while at the same time permits a delayed patient response.

The randomized play-the-winner rule is applicable when $K \geq 2$ treatments are under study, with explanation facilitated by an urn model as well, see Wei (1979). The two-arm RPW$(u, 0, \beta)$ may be extended to explain a special case of this multi-arm design. An urn initially contains u balls for each of K different colors. When an eligible patient arrives at a site a ball is drawn at random from the urn, its color k noted and then replaced. The treatment corresponding to k is assigned. When a response from a patient assigned to treatment k is available and is a 'success', $K - 1$ balls of color k are added to the urn; otherwise if the response on k was a 'failure', one ball of each color $j = 1, \ldots, K$, $j \neq k$ is added to the urn. Andersen, Faries and Tamura (1994) have modified the multi-arm design of Wei (1979) in the following way. When a response to treatment k is available and is a 'failure', instead of placing in the urn the same number of balls (one) of every color other than k [$(K - 1)$ total balls are added], partition the $K - 1$ balls according to the existing proportions of balls other than color k in the urn. This procedure will make the ratios of any two non-k balls independent of whether the response on treatment k were a success or failure.

Suppose that one of the $K \geq 2$ treatments is performing relatively poorly. Some would argue that it is unethical to add balls corresponding to this least favorable treatment as a result of another treatment's failure. Li (1995) considers a modified randomized-play-the-winner design where a success on a particular treatment rewards this treatment by adding β balls of its corresponding color to the urn,

while a failure on this treatment leaves the urn composition unchanged, i.e. other treatments are not rewarded based on a particular ones failure.

Although the randomized-play-the-winner design does not formally require responses to be instantaneous, its capability to assign more patients to the better treatment will be diminished if relatively few outcomes are observed before the last patient is enrolled. In such situations, an alternative to relying on the (delayed) primary outcome to evidence treatment efficacy would be to employ information provided by a surrogate for response that may be observed with less waiting time. An example of a clinical trial that used the randomized play-the-winner design and incorporated information from a surrogate response to update the urn is given by Tamura *et al.* (1994).

Any clinical trial demands close cooperation of the physicians, statisticians, data managers and patients; this becomes an increasingly difficult task when there are many centers involved. Thus mistakes in executing the trial, including errors in treatment assignments are not unlikely with a very simple design. Response-adaptive designs such as the randomized-play-the-winner rule, although conceptually simple, can be logistically complex and thus increase the chance of errors in the conduct of the trial. Some actual examples are given in Tamura *et al.* (1994) and mentioned later in this article. For a review of logistical considerations with response-adaptive designs, see Rosenberger and Lachin (1993).

3.1. *Properties of the randomized play-the-winner rule.* Suppose that an individuals response to treatment is known before the succeeding patient is available for assignment. It is interesting to note the following results. If n_A and n_B are the number of patients assigned to treatments A and B, respectively, after $n = n_A + n_B$ treatment assignments, then for the RPW$(u, 0, \beta)$ rule, as $n \to \infty$, $E(n_A/n)$ converges to the asymptotic proportion of patients treated by arm A for the (modified) play-the-winner rule (Wei and Durham, 1978). Thus for a large trial, the RPW$(u, 0, \beta)$ rule is as good as the (modified) play-the-winner rule for assigning patients to the better treatment. The probability of correctly guessing the $(n+1)$st assignment in the RPW$(u, 0, \beta)$ design converges to $1/2 + \psi$ as the trial size gets large, where the positive term ψ is increasing in the value of β. The ratio β/α in the RPW(u, α, β) design ($\alpha > 0$) governs to what extent we 'play the winner'. As β/α becomes large, more patients are assigned to the better arm; however, if β/α is too large, the design becomes deterministic and is subject to experimental bias. More properties of the randomized play-the-winner rule are given in Matthews and Rosenberger (1997) and Rosenberger (1996), the latter considering a more general urn design.

Since at any point during the trial, there is deliberate imbalance in the number of patients assigned to each treatment group, inference procedures concerning treatment effect are less efficient (than a balanced experiment). Thus, it is possible that at an interim analysis, there is not enough power to detect a treatment difference when there is actually one present. This may result in not being able to stop the trial as early as if a balanced randomization rule had been implemented. As a consequence, even though fewer study patients are sacrificed, the benefit to the greater population is delayed. Even though it is possible that the loss in efficiency may be

negligible when the treatment difference is large, this issue needs to be seriously addressed before it is decided to implement a response-adaptive design. Note that if the treatment difference is small or moderate, the RPW(u, α, β) design would put roughly half of the study patients in each treatment group.

One way around the possibility of not being able to detect an early treatment difference with response-adaptive designs is to implement a proper stopping rule. For illustration, assume that the response of every previously assigned patient is available before the current assignment. Wei and Durham (1978) consider an inverse stopping rule for the RPW$(0, 0, 1)$ design that observes patients until either $S_A(n) + F_B(n) = r$ or $S_B(n) + F_A(n) = r$, where $S_k(n)$ and $F_k(n)$ are the number of successes and failures on treatment $k = A, B$, after n assignments, respectively, and r is some preselected positive integer. For example, if $S_A(n) + F_B(n) = r$, the trial would be terminated and treatment A would prevail as the better treatment. Note that at most $2r - 1$ observations are needed before the trial is terminated. As $r \to \infty$, the probability of correctly selecting the superior treatment strategy approaches one (Wei and Durham, 1978). Furthermore, the inverse stopping rule has the same power to select the better treatment as does a trial of fixed size $2r-1$, however, the average sample size needed to terminate the trial is smaller (for the inverse stopping rule) (Wei and Durham, 1978). In a numerical comparison of the probability of correct selection and the average sample size needed, Wei and Durham (1978) conclude that the RPW$(0, 0, 1)$ design is not inferior to the play-the-winner rule of Zelen (1969) for practical use.

Throughout the remainder of the paper, each assignment rule that we consider is meant to be applied separately within each (baseline) category of patients, and therefore each category is treated as an independent sub-trial. Thus, we drop all reference to separate categories below.

3.2. *Analysis of study results.* Unlike the play-the-winner rule of Zelen (1969), the randomized play-the-winner rule is never deterministic, and as a result, a permutation test of the null hypothesis can be generated from this design under a randomization model. Wei (1988) provides an efficient network algorithm to construct exact permutation tests of equal efficacy of the two treatments on response with the RPW$(u, 0, \beta)$ rule. The RPW$(u, 0, \beta)$ rule has, for several reasons, including practicality, been a popular selection from the class of RPW(u, α, β) designs. The setting from which the test is derived is as follows. Again, we denote by $Y_1, \ldots, Y_N$ the sequence of (random) binary treatment assignments, where $Y_i = 1$ if patient i was assigned treatment A, 0 otherwise, and by $x_1, \ldots, x_N$ the (fixed) sequence of (binary) responses, where $x_i = 1$ if the ith response was a success and 0 otherwise, $i = 1, \ldots, N$. Consider the test-statistic $G_N = \sum x_i Y_i$, the number of successes of those assigned to treatment A, where the sum is over $i = 1, \ldots, N$. If the realization of treatment indicators $y_1, \ldots, y_N$ generates an extreme value of G_N, then the null hypothesis H_0 of no treatment effect is rejected.

Wei (1988) applied this exact testing procedure to the ECMO study described above, which actually implemented the RPW$(1, 0, 1)$ design, and obtained a one-sided $p-$value of 0.051 in favor of the ECMO therapy. Wei (1988) also showed that if the analysis of the ECMO trial assumed that complete randomization generated the

treatment sequence, then an exaggerated one-sided $p-$value of 0.001 is obtained. However, the ECMO trial has been the subject of intense criticism from the scientific community (Begg, 1990). The RPW$(1,0,1)$ design randomized 11 patients to ECMO therapy and only one to the conventional treatment, thus the major aspersion with this trial was that one observation was not sufficient to reflect the survival experience of the conventional therapy.

Rosenberger (1993) developed a large-sample analog of the exact permutation test of Wei (1988) for the RPW$(u,0,1)$ rule. Re-define the fixed sequence of binary outcomes $x_1,\ldots,x_N$ as $x_j=1$ for a success and $x_j=-1$ otherwise, $j=1,\ldots,N$. The test statistic

$$T_N=\frac{2\sum_{j=1}^{N}x_j(Y_j-\frac{1}{2})}{\left(\sum_{j=1}^{N}a_{jN}^2\right)^{1/2}}$$

where

$$a_{NN}=1,\qquad a_{jN}=\prod_{k=j+1}^{N}\left(1+\frac{x_k}{2u+k-1}\right),j=1,\ldots,N-1,$$

has an asymptotic standard normal distribution under H_0, provided that

$$\frac{\max\limits_{1\le j\le N}a_{jN}^2}{\sum_{i=j}^{N}a_{jN}^2}\to 0\ (N\to\infty)$$

and

$$\frac{\sum_{j=1}^{N}\left\{x_j(Y_j-\frac{1}{2})+\frac{1}{2}\right\}}{N}\to 0 \text{ in probability } (N\to\infty).$$

Both of these conditions hold if the responses $x_1,\ldots,x_N$ are an independent centered Bernoulli sequence with $P(X_j=1)=p<0.75$. Simulations have shown that the test is conservative for $u=1$, with a better approximation to normality for $u=5$ (Rosenberger, 1993).

Other tests of the null hypothesis H_0 of no treatment effect for the RPW$(u,0,\beta)$ design are also available. Let the probability of success on treatment A and B be respectively given by $p_A=P(X_i=1|Y_i=1)$ and $p_B=P(X_i=1|Y_i=0)$, for $i=1,\ldots,N$. For testing H_0, Wei *et al.* (1990) provide for the RPW$(u,0,\beta)$ rule, exact conditional, exact unconditional, and for other response-adaptive designs satisfying mild conditions (including the RPW$(u,0,\beta)$ design), asymptotic confidence intervals for the odds ratio $\theta=p_A(1-p_B)/\{(1-p_A)p_B\}$ and difference in proportions $\Delta=p_A-p_B$. The results are developed from a frequentist point of view, i.e. the dichotomous responses $X_1,\ldots,X_N$ are now treated as i.i.d. observations under H_0. The asymptotic confidence intervals are derived from a likelihood-ratio statistic and have been shown to perform well in finite samples of size $N>50$ (see Wei *et al.*, 1990). A thorough application of the above methods to the ECMO study are provided in Wei *et al.* (1990).

Unlike traditional 50/50 treatment allocation rules, such as the permuted-block design, the final number of patients assigned to each treatment group in a response-adaptive designs, N_A and N_B, carry subsequent information about H_0. Thus, performing an analysis conditional on, say, $\delta_N = N_A - N_B$ may not be very efficient. In fact, in an extensive simulation study, Wei *et al.* (1990) showed that for the RPW$(1, 0, 1)$ rule, although exact unconditional confidence intervals tended to be conservative (more than nominal coverage), they were shown to be more powerful than the exact conditional confidence intervals. In fact, if the odds ratio is not too far from 1, then it was shown that the exact unconditional tests are better than the randomized version of the conditional test (Lehmann, 1986; p.74), in that the corresponding type I and II error probabilities are smaller.

As an illustration, consider an example in Wei *et al.* (1990), where, for testing $H_0 : \Delta = 0$ against $H_1 : p_A = 0.8, p_B = 0.2$, with type I error level 0.05 and 77% power, the unconditional exact inference procedure with a RPW$(3, 0, 1)$ rule required roughly half of the sample size needed for a corresponding design initially using a permuted-block randomization rule. Therefore, in contrast to conventional randomization schemes (that attempt to balance treatment allocation), with a response-adaptive design, the efficiency of tests of H_0 are dramatically increased if their $p-$values are not computed conditionally on the final imbalance in the numbers of patients allocated to each treatment group.

It is interesting to investigate whether or not the RPW(u, α, β) design can in general be ignored in the analysis of study results. That is, do tests based on assuming that complete randomization generated the treatment assignments have correct coverage probabilities? Wei *et al.* (1990) conducted an extensive simulation study to investigate this. For various values of p_A and p_B, observations were generated by the RPW$(1, 0, 1)$ rule, but the analysis assumed that complete randomization generated the treatment sequence. The two-sided exact unconditional confidence intervals for Δ had sub-nominal coverage in several cases.

If a patient's response occurs relatively quickly, for instance, research on emergency medicine, it has been argued that for ethical reasons, response-adaptive designs should be implemented more frequently in practice (Bather, 1985). However, investigators planning to use response-adaptive designs should ensure sufficient concurrent experience with each treatment strategy. In the RPW$(u, 0, \beta)$ design, this may be accomplished by taking u to be somewhat large relative to β, e.g. RPW$(3, 0, 1)$.

3.3. *Applications of the randomized play-the-winner rule.* In this section, we use perhaps the most pivotal pediatric clinical trial conducted by the AIDS Clinical Trials Group (ACTG) to date, ACTG 076, in order to illustrate some of the properties of the RPW(u, α, β) design. ACTG 076 investigated the effectiveness of the drug zidovudine (AZT) for reducing the risk of vertical transmission of HIV from infected mothers to their newborns. The endpoint was the HIV status of the infant, with virus positivity being able to be ascertained at approximately 2 months of age. Ascertainment of virus negativity requires approximately 6 months. A stratified (with respect to study center) permuted-block design was implemented and assigned 238 women and their newborns to each treatment arm, AZT and placebo. The women in

the AZT arm received AZT during pregnancy through delivery and the infants took AZT orally. The trial was stopped early by the Data Safety and Monitoring Board (DSMB) and the AZT treatment strategy was declared superior.

A short time after the trial had been stopped, the data were updated from that which was presented at the DSMB meeting and of the 476 women, 442 had given birth to live infants, 220 on the AZT arm and 222 on the placebo arm. In the AZT group, 16 newborns had an HIV-positive culture, while 52 newborns had a HIV-positive culture in the placebo group. The test statistic for testing the null hypothesis H_0 of no treatment effect was the difference between the treatment arm-specific Kaplan-Meier estimates of the probability that the time to first positive HIV culture was greater than 18 months. The corresponding values of these probabilities were 0.928 for the AZT group and 0.752 for the placebo group, with 2×10^{-7} being the p-value of the test that their difference is zero.

The findings of ACTG 076 benefit the newborns of HIV positive women in the general population, and this is a remarkable advancement. However, a substantial number of newborns in the placebo group were detrimentally infected with HIV. An obvious question is whether or not a response-adaptive design, such as the RPW(u, α, β), would have led to fewer assignments to the placebo group, resulting in fewer HIV infections, while simultaneously retaining enough power to stop the trial just as quickly so as to not delay the benefit to the general population. Response-adaptive designs by nature are unbalanced so one has to ensure that the loss in power from implementing such designs is minimal.

Yao and Wei (1996) conducted an extensive simulation study to investigate the benefit of using the RPW$(1, 0, 1)$ design instead of the permuted-block design actually implemented in ACTG 076. The simulation used the actual entry dates and birth dates of the women and associated newborns in ACTG 076. To explain the simulation procedure, consider an arbitrary stage of the trial and suppose that a woman who was just enrolled was assigned to the AZT group using the RPW$(1, 0, 1)$ rule. Her infant would then have a probability of (1-0.928) 0.072 of being infected with HIV, where as if she were assigned to the placebo group, the infants probability of being infected would be (1-0.752) 0.248. These are the transmission rates observed in ACTG 076. For each mother, the response of the child, either HIV positive or negative, was then generated as a Bernoulli random variable with probability of HIV positivity depending on treatment group, 0.072 and 0.248 for AZT and placebo, respectively. If the child's response was HIV positive, then the time to first HIV positive culture was generated according to the treatment arm-specific Kaplan-Meier estimates of time to first positive culture. However, since the two Kaplan-Meier curves are flat after 24 weeks, which indicates that the majority of HIV transmissions were detected before 24 weeks, when the generated time to first positive culture was more than 24 weeks, the response was changed to be HIV negative. If the child's response was HIV negative, then the time to this event was defined to be 6 months, the time needed to reliably ascertain HIV negative status. The time to obtain a response for a particular mother-infant pair with the RPW (u, α, β) design was the duration between study entry of the mother and when the HIV status of the infant was determined. The urn was updated at the ascertainment time of, and according to, the infants HIV status.

This process was completed for all 476 women. The above described simulated trial was independently generated a total of 500 times, and for each one, the total number of women assigned to each group and the total number of HIV positive infants were recorded. A similar simulation study was also conducted that used the permuted-block randomization rule with block length 4 to assign treatment to mothers.

From the 500 simulations, an average of 300 out of 476 women were assigned to the AZT arm with the RPW$(1,0,1)$ rule, compared to 238 out of 476 for the permuted-block design. This resulted in the permuted-block design generating an average of 11 more HIV infected infants per trial. Table 5 displays the results.

Table 5. RESPONSE-ADAPTIVE VERSUS PERMUTED-BLOCK DESIGNS IN ACTG 076.

Design	Sample size (AZT, placebo)	Number HIV+	Power
permuted-block	(238, 238)	76	0.92
RPW(1,0,1)	(300, 176)	65	0.88
MS(8,4,4,4,4,4,4)	(287, 189)	67	0.90
MS(8,8,8,8)	(288, 188)	67	0.89
MS(14,6,6,6)	(273, 203)	70	0.91

NOTE: For the multi-stage design, the numbers $x \ldots y$ in MS$(x, \ldots, y)$ are the time in weeks between updates of the urn.

The RPW$(1,0,1)$ rule is obviously preferable to the permuted-block design from the point of view of study patients, since fewer HIV infections occur than if a perfectly balanced design were used. However, the unbalanced nature of the RPW(u,α,β) rule results in a reduction of the power of tests of treatment effectiveness and as a consequence, might result in AZT not being declared superior as quickly as if a balanced design were implemented. In this case, the benefit to the study patients comes at an expense to the greater population by delaying their access to AZT. To estimate if a trial like ACTG 076 would have the same probability of being stopped at the first interim analysis with the RPW$(1,0,1)$ rule as it would if it employed the permuted-block randomization, the proportion of times (out of 500) that the test statistic used in ACTG 076 had a $p-$value of less than the stopping boundary used for the first interim analysis (0.0005) was calculated. Based on 500 simulations, the chance that a trial using the RPW$(1,0,1)$ design would be stopped at the first interim analysis is approximately 0.88. As for the permuted-block design with length 4, based on 500 simulations, the chance that a trial would be stopped at the first interim analysis is approximately 0.92. Therefore, in terms of power, the RPW(u,α,β) rule appears to be almost as good as the perfectly balanced design in those clinical trial settings like that of ACTG 076.

A trial besides the ECMO study that actually implemented the randomized play-the-winner rule was a double-blind, stratified, placebo controlled trial of out-patients suffering from depressive disorder (Tamura *et al.*, 1994). The treatment was fluoxetine hydrochloride and was thought to possibly effect patients differently, depending if their time between sleep onset and first rapid eye movement (REML) was short or normal. Patients with a short REML were believed to possibly exhibit a better response to treatment. At baseline, patients were stratified into two groups corresponding to short and normal REML. One primary endpoint was binary, defined as

a 50% or greater reduction from baseline to final visit in the Hamilton Depression Scale after a minimum of 3 weeks of therapy. Patients that achieved this endpoint were termed responders and were said to have a 'successful' response. Based on this definition of a successful response, it was desired to execute the RPW$(1, 0, 1)$ rule separately for each stratum of patients.

The time from randomization to final visit was approximately eight weeks with a rapid patient accrual. This relative delay in observing a response would cause the RPW$(1, 0, 1)$ rule to not absorb sufficient information regarding treatment efficacy to bias the treatment allocation probability in favor of the leading treatment. A surrogate response that could be observed more quickly was considered. A surrogate response was defined as a 50% or greater drop in the Hamilton Depression Scale in two consecutive visits after a minimum of 3 weeks of therapy. A surrogate nonresponder was one who completed at least 3 weeks of therapy but could not be classified as a surrogate responder prior to the final visit. Patients not completing at least 3 weeks of therapy did not yield a surrogate response.

The first six patients within each stratum were randomized to treatment via independent permuted block designs; subsequent patients were assigned to treatment separately within each stratum according to independent RPW$(1, 0, 1)$ rules based on the surrogate responses. The purpose of the permuted block design was to ensure sufficient concurrent experience with each treatment regime. This may also have been accomplished by adding more balls of both treatment types to the urn at baseline, i.e. taking u to equal 2 or 3.

A total of 89 patients were randomized in the trial, 46 to fluoxetine and 43 to placebo. Both randomization and Bayesian methods were used to analyze the data. For the randomization test, both the standardized difference in the proportion of responders and a t-test for the differences in the change in the Hamilton Depression Scale were considered. The null permutational distribution of each statistic was approximated based on 500,000 simulations. Both randomization and Bayesian analyses found that in the shortened REML stratum, fluoxetine was significantly associated with a drop in the Hamilton Depression Scale.

Forty-five patients were randomized in the shortened REML stratum, 23 to fluoxetine. The balanced allocation observed in this stratum was investigated by Tamura *et al.*, (1994), and they concluded that such a balanced allocation would not be unusual for this stratum, even if there was nearly a two-fold increase in the difference between the observed response rates.

The response-adaptive study design and double-blind nature undoubtedly strained logistical aspects and complicated implementation. Indeed, various protocol violations were encountered. For instance, 3 patients were incorrectly stratified based on their REML status; for two of these patients the wrong urn was used for randomization and this urn was updated based on the patients' surrogate response. Also, the surrogate response of one patient was mistakenly used to update the urn two times instead of only once.

The response-adaptive design implemented in this study was not successful in assigning more patients to the better treatment. Ways in which this study may have made their increased effort worthwhile with respect to this objective would

have been to increase the number of balls added to the urn after each surrogate response, i.e. set β to equal 2 or 3.

3.4. *A multi-stage randomized play-the-winner rule.* In practice, it may not be very feasible to change the treatment allocation rule after each patients response during the trial. This is especially true in a multi-center trial, when response time is relatively short. In these cases, executing the randomized play-the-winner rule would induce quite an administrative burden.

The multi-stage RPW$(1,0,1)$ design updates the urn not at each response time, but only at the defined study-time points $\{\xi_i\}, i = 1, \ldots, m$, where $0 = \xi_0 < \xi_1 < \cdots < \xi_m < \xi_{m+1} = \tau$, and τ is the study time corresponding to the end of follow-up. It is implemented as follows. Let $S_k^{(i)}$ and $F_k^{(i)}$ denote the number of success and failures, respectively, on treatment $k = A, B$ within the time interval $(\xi_{i-1}, \xi_i]$, $i = 1, \ldots, m$. At each study time ξ_i, $S_A^{(i)} + F_B^{(i)}$ white balls and $S_B^{(i)} + F_A^{(i)}$ red balls are added to the urn, $i = 1, \ldots, m$. For each treatment assignment during the next study-time interval $(\xi_i, \xi_{i+1}]$, $i = 1, \ldots, m$, a ball is drawn from the urn with replacement, if it is white, treatment A is assigned, otherwise treatment B is assigned. Note that the choice of the time points $\{\xi_i\}$ may depend upon post-randomization variables, such as the accrual rate, and hence may be empirically adjusted in progress.

In order to investigate the performance of the multi-stage RPW$(1,0,1)$ design in settings like that of ACTG 076, Yao and Wei (1996) conducted simulations with the same set-up as those described in the previous section, except that the multi-stage RPW$(1,0,1)$ design was implemented instead of the RPW$(1,0,1)$ design. Table 5 displays, out of 500 simulations, the average number of HIV infected infants and the proportion of times that the trial was stopped at the first interim analysis, for several multi-stage RPW$(1,0,1)$ designs. Substantially more infants are spared HIV infection with the multi-stage RPW$(1,0,1)$ design than with the permuted-block design. Furthermore, the power loss from using these response-adaptive designs appears to be slight.

3.5. *Continuous responses.* The RPW(u,α,β) design proposed so far required that the response of each patient be dichotomous, i.e. either a success or failure. In many clinical trials, however, the patients defined response is continuous, e.g. time from randomization to failure. We make use of the previously defined multi-stage randomized play-the-winner design to describe how to accommodate such continuous response data in a response-adaptive design.

The procedure is as follows. At each time point ξ_i in the multi-stage design, $i = 1, \ldots, m$, compute a summary statistic $G_N^{(i)}$ which reflects the treatment difference with respect to (continuous) patient response thus far. For example, with failure-time data, the statistic may be the standardized two-sample Gehan statistic. Suppose that a positive value of $G_N^{(i)}$ corresponds to preference for treatment A, and a negative value preference for treatment B. Then the probability of assigning a patient to treatment A during the next interval of study-time $(\xi_i, \xi_{i+1}]$ becomes, for instance, $\varphi_i(G_N^{(i)}, r) = 0.5 + rG_N^{(i)}$, where $0.1 < \varphi_i(G_N^{(i)}, r) < 0.9$, $i = 1, \ldots, m$ and $r > 0$ is a pre-selected constant reflecting how much one would

like to 'play-the-winner' for the next interval of time.

For complete continuous responses, Rosenberger (1993) developed a response-adaptive design based on a linear-rank statistic. We consider the multi-stage design for illustration. At each time point ξ_j, $j = 1, \ldots, m$ suppose that there are k of N outcomes available. Let $r_{ik} (i \leq k)$ denote the rank of the ith patient based on the outcome variable, where a larger rank corresponds to a better response. Define a_{ik} to be some score function of the r_{ik}, with $\sum a_{ik} = 0$, the sum from $i = 1, \ldots, k$. In the succeeding interval of study time $(\xi_j, \xi_{j+1}]$, each patient is randomized to treatment A with a probability that is a function of the rank statistic computed at study time ξ_j, for instance with probability

$$\frac{1}{2}\left\{1 + \frac{\sum_{i=1}^{k} a_{ik}(Y_i - \frac{1}{2})}{\sum_{i=1}^{k} a_{ik} I(a_{ik} > 0)}\right\}.$$

A positive value of the considered rank statistic favors treatment A while a negative value favors treatment B. The better the responses on treatment A relative to B, the larger the value of the above probability of assignment to A in the next study time interval. The exact form of the rank statistic is given in Rosenberger (1993) along with conditions for its asymptotic normality.

As an illustration of how to implement the multi-stage response-adaptive design with survival time data, consider a long term prostatic cancer trial conducted by the Veterans Administration Cooperative Urological Research Group (VACURG). The trial is described in Slud and Wei (1982). Patients were enrolled between 1960 and 1967 and randomized to either prostatectomy and oestrogen (group A) or prostatectomy and placebo (group B). In total, 43 and 46 patients were enrolled in treatment group A and B, respectively, with all except 27 of the 89 patients being observed to die. The outcome was time between study entry and death. The median survival times in group A and B, were 5.7 and 9.2 years, respectively, with the Gehan test yielding a one-sided $p-$value of 0.021. It is interesting to know whether a response-adaptive design would have assigned more patients to treatment B in the VACURG trial. Furthermore, would such a response-adaptive design, with its unbalanced nature, only induce a mild loss in the power of the test-statistic of H_0.

Yao and Wei (1996) conducted a simulation study that used the actual entry dates of the patients in the VACURG trial and for each one, randomized them to either group A or B using the response-adaptive multi-stage design described above. Survival times were generated based on the corresponding treatment arm-specific Kaplan-Meier estimate of time to death calculated from the actual VACURG trial. The censoring time for each patient was simulated from the Kaplan-Meier estimate of time to censoring, also calculated from the VACURG data. Each year from 1963 through 1966, the standardized Gehan statistic $G_N^{(i)}$ was calculated and the probability of assignment to treatment group A from thereafter until the next year became $\varphi(G_N^{(i)}, r) = 0.5 + rG_N^{(i)}$, for $r = 0.1, 0.15$, with $0.1 < \varphi(G_N^{(i)}, r) < 0.9$.

Table 6 presents the mean number of subjects assigned to each treatment group along with the proportion of times the null hypothesis was rejected from 500 in-

dependent simulations of the above multi-stage design with $r = 0.1, 0.15$. For comparison purposes, results from 500 independent trial simulations that instead used the permuted-block design with length 4 are included as well. On average, the multi-stage response-adaptive design assigns 12 ($r = 0.1$) and 16 ($r = 0.15$) more patients to group B per trial than does the permuted-block design. Furthermore, the loss in power of the response-adaptive design due to its unbalanced nature seems to be relatively small. Finally, Table 6 displays the results from 500 simulations of the VACURG trial with the treatment allocation rule being the multi-stage RPW(u, α, β) design, treating a death as a failure to update the urn at the same time points as the multi-stage design with the Gehan statistics. It appears that in this trial setting, the multi-stage RPW (u, α, β) rule does not perform a well as the multi-stage rule with Gehan statistics.

Table 6. MULTI-STAGE RESPONSE-ADAPTIVE DESIGNS VERSUS PERMUTED-BLOCK DESIGN.

Design	Sample size (A, B)	Power
permuted-block	(45, 44)	0.76
MS($r = 0.1$)	(33, 56)	0.72
MS($r = 0.15$)	(28, 61)	0.70
RPW(1,0,1)	(38, 51)	0.74
RPW(4,0,1)	(39, 50)	0.74

3.6. *Ethical issues.* Some may argue that, when using a response-adaptive design to allocate treatment to patients, it would be ethically difficult for an investigator to assign a patient to a treatment group that (s)he believes to be inferior, just because a biased coin lands the 'wrong' way. However, consider the alternative. Suppose that a 50/50 allocation rule, such as the permuted-block or urn design, was used to assign treatment to patients. Furthermore, suppose, which is often the case, that the trial is being monitored by some sequential procedure and that at an interim analysis, one treatment appears to be superior, although not statistically significant so as to be able to stop the trial. If the treatment assignments are not masked and the results of the interim analysis are advertised to incoming patients, it would seem to be very difficult for an investigator to administer a patient to the perceived inferior treatment with probability 1/2, just to avoid imbalances with respect to treatment totals and covariates. Using a masked 50/50 allocation rule and sequential stopping protect trials from ethical criticisms.

In clinical studies on humans, there is a conflict between the interests of a study patient and the advancement of science (Bather, 1985). Nondeterministic response-adaptive designs seem to be a fair compromise between these two concerns and as a result are open to criticism from both ends. One thing is certain when considering implementing a response-adaptive design: It should be ensured that sufficient concurrent experience with each treatment arm is provided by the design. For a review of ethical issues with response adaptive designs, see Rosenberger and Lachin (1993) and Bather (1985).

Acknowledgments. Research was partially supported by grants from the National Institute of Health. The author is grateful to L.J. Wei for helpful comments and

the invitation to write the article. The author is also thankful to a reviewer for insightful comments.

References

Andersen, J., Faries, D. and Tamura, R. (1994). A randomized play-the-winner design for multi-arm clinical trials. *Communications in Statistics, Part A–Theory and Methods,* **23,** 309–323.

Bartlett, R. H., Roloff, D. W., Cornell, R. G., Andrews, A. F., Dillon, P. W. and Zwischenberger, J. B. (1985). Extracorporeal circulation in neonatal respiratory failure: a prospective randomized trial. *Pediatrics,* **76,** 479–487.

Bather, J. A. (1985). On the allocation of treatments in sequential medical trial (C/R: P25-36). *International Statistical Review,* **53,** 1–13.

Begg, C. B. (1990). On inferences from Wei's biased coin design for clinical trials (C/R: P473-484). *Biometrika,* **77,** 467–473.

Blackwell, D. and Hodges, J. (1957). Design for the control of selection bias. *Ann. Math. Statist.,* **28,** 449–460.

Cornell, R. G., Landenberger, B. D. and Bartlett, R. H. (1986). Randomized play-the-winner clinical trials. *Communications in Statistics, Part A–Theory and Methods,* **15,** 159–178.

Cox, D. R. (1982). A remark on randomization in clinical trials. *Utilitas Mathematica,* **21,** 245–252.

Davis, C. S. (1989). Two-sample post-stratified or subgroup analysis with restricted randomization rules. *Communications in Statistics, Part A–Theory and Methods,* **18,** 367–378.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika,* **58,** 403–417.

Eisele, J. R. (1994). The doubly adaptive biased coin design for sequential clinical trials. *Jour. Statistical Planning and Inference,* **38,** 249–261.

Freedman, L. S. and White, S. J. (1976). On the use of Pocock and Simon's method for balancing treatment numbers over prognostic factors in the controlled clinical trial. *Biometrics,* **32,** 691–694.

Friedman, B. (1949). A simple urn model. *Communications on Pure and Applied Mathematics,* **2,** 59–70.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics,* **6,** 65–70.

Lachin, J. M. (1988). Statistical properties of randomization in clinical trials. *Controlled Clinical Trials,* **9,** 289–311.

– – – – (1988). Properties of simple randomization in clinical trials. *Controlled Clinical Trials,* **9,** 312–326.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks,* Holden-Day, San Francisco.

Lehmann, E. L. (1982). *Testing Statistical Hypotheses (Second Edition),* Wiley, New York.

Li, W. (1995). Sequential designs for opposing failure functions. *Ph.D. dissertation, College of Arts and Sciences, American University, Washington D.C.*

Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika,* **63,** 655–660.

Matthews, P. and Rosenberger, W. F. (1997). Variance for randomized play-the-winner clinical trials. *Statistics and Probability Letters,* **35,** 233-240.

Mehta, C. R., Patel, N. R. and Wei, L. J. (1988). Constructing exact significance tests with restricted randomization rules. *Biometrika,* **75,** 295–302.

Peto, R., Pike, M. C., Armitage, P., Breslow, N. E., Cox, D. R., Howard, S., Mantel, N., McPherson, K., Peto, J. and Smith, P. (1977). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II: Analysis and examples. *British Journal of Cancer,* **35,** 1–39.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial (Corr: V32 P954-955). *Biometrics,* **31,** 103–115.

ROSENBERGER, W. F. (1993). Asymptotic inference with response-adaptive treatment allocation designs. *The Annals of Statistics,* **21,** 2098–2107.

– – ––– (1996). New directions in adaptive designs. *Statistical Science,* **11,** 137–149.

ROSENBERGER, W. F. and LACHIN, J. M. (1993). The use of response-adaptive designs in clinical trials. *Controlled Clinical Trials,* **14,** 471–484.

SLUD, E. and WEI, L. J. (1982). Two-sample repeated significance tests based on the modified Wilcoxon statistic. *Jour. Amer. Statist. Assoc.,* **77,** 862–868.

SMITH, R. L. (1984). Sequential treatment allocation using biased coin designs. *J. Roy. Statist. Soc.,* Series B, **46,** 519–543.

SMYTHE, R. T. and WEI, L. J. (1983). Significance tests with restricted randomization design. *Biometrika,* **70,** 496–500.

SMYTHE, R. T. (1988). Conditional inference for restricted randomization designs. *The Annals of Statistics,* **16,** 1155–1161.

TAMURA, R. N., FARIES, D. E., ANDERSEN, J. S. and HEILIGENSTEIN, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Jour. Amer. Statist. Assoc.,* **89,** 768–776.

WEI, L. J. (1977). A class of designs for sequential clinical trials. *Jour. Amer. Statist. Assoc.,* **72,** 382–386.

– – ––– (1978a). The adaptive biased coin design for sequential experiments. *Ann. Statist.,* **6,** 92–100.

– – ––– (1978b). An application of an urn model to the design of sequential controlled clinical trials. *Jour. Amer. Statist. Assoc.,* **73,** 559–563.

– – ––– (1979). The generalized Pólya's urn design for sequential medical trials. *Ann. Statist.,* **7,** 291–296.

– – ––– (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. *Biometrika,* **75,** 603–606.

WEI, L. J. and DURHAM, S. (1978). The randomized play-the-winner rule in medical trials. *Jour. Amer. Statist. Assoc.,* **73,** 840–843.

WEI, L. J., SMYTHE, R. T. and SMITH, R. L. (1986). K-treatment comparisons with restricted randomization rules in clinical trials. *Ann. Statist.,* **14,** 265–274.

WEI, L. J. and LACHIN, J. M. (1988). Properties of the urn randomization in clinical trials. *Controlled Clinical Trials,* **9,** 345–364.

WEI, L. J., SMYTHE, R. T., LIN, D. Y. and PARK, T. S. (1990). statistical inference with data-dependent treatment allocation rules. *Jour. Amer. Statist. Assoc.,* **85,** 156–162.

WEI, L. J. and GLIDDEN, D. V. (1997). An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine,* **16,** 833–839.

YAO, Q. and WEI, L. J. (1997). Play the winner for phase II/III clinical trials (Disc: P2455-2458). *Statistics in Medicine,* **15,** 2413–2423.

ZELEN, M. (1969). Play the winner rule and the controlled clinical trial. *Jour. Amer. Statist. Assoc.,* **64,** 131–146.

– – ––– (1974). The randomization and stratification of patients to clinical trials. *Jour. Chronic Disease,* **27,** 365–375.

A. GREGORY DIRIENZO
CENTER FOR STATISTICAL SCIENCES
BROWN UNIVERSITY
BOX G-H
PROVIDENCE, RI 02912
U.S.A.
FAX: 401-863-9182
Email: gregd@stat.brown.edu