Properties of the Urn Randomization in Clinical Trials

L.J. Wei, PhD, and John M. Lachin, ScD

Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan (L.J.W.) and the George Washington University, Department of Statistics/ Computer and Information Systems, The Biostatistics Center, Rockville, Maryland (J.M.L.)

ABSTRACT: In this article we review the important statistical properties of the urn randomization (design) for assigning patients to treatment groups in a clinical trial. The urn design is the most widely studied member of the family of adaptive biased-coin designs. Such designs are a compromise between designs that yield perfect balance in treatment assignments and complete randomization which eliminates experimental bias. The urn design forces a small-sized trial to be balanced but approaches complete randomization as the size of the trial (n) increases. Thus, the urn design is not as vulnerable to experimental bias as are other restricted randomization procedures.

In a clinical trial it may be difficult to postulate that the study subjects constitute a random sample from a well-defined homogeneous population. In this case, a randomization model provides a preferred basis for statistical inference. We describe the large-sample permutational null distributions of linear rank statistics for testing the equality of treatment groups based on the urn design. In general, these permutation tests may be different from those based on the population model, which is equivalent to assuming complete randomization.

Poststratified subgroup analyses can also be performed on the basis of the urn design permutational distribution. This provides a basis for analyzing the subset of patients with observed responses when some patients' responses can be assumed to be missing-at-random. For multiple mutually exclusive strata, these tests are correlated. For this case, a combined covariate-adjusted test of treatment effect is described.

Finally, we show how to generalize the urn design to a prospectively stratified trial with a fairly large number of strata.

KEY WORDS: *Randomization, urn design, adaptive biased-coin design, permutation tests, stratified analysis, missing data, selection bias, accidental bias*

INTRODUCTION

One of the fundamental statistical issues in comparative clinical trials is the determination of the method of assigning patients to treatment groups, say a and b. In most trials, eligible patients become available sequentially for study and must be assigned shortly thereafter to receive either treatment a or b. Although the target size of the trial can be determined beforehand, the

Address reprint requests to: John M. Lachin, ScD, The George Washington University, Department of Statistics/Computer and Information Systems, The Biostatistics Center, 6110 Executive Boulevard, Suite 750, Rockville, MD 20852.

Received April 29, 1987; revised June 13, 1988.

Controlled Clinical Trials 9:345–364 (1988)

655 Avenue of the Americas, New York, New York 10010

0197-2456/1988/$3.50

actual number of patients entered into the trial may not be known in advance. For example, in a prospectively stratified trial, patients are often grouped into several strata based on some important demographic or clinical factors, almost always by clinic. If the number of strata is not too large, one generally treats each stratum as a separate independent subtrial. In this case, the actual numbers of patients who fall in each stratum are impossible to know in the beginning of the study. Therefore, in a staggered-entry clinical trial, it is preferable to have a treatment assignment rule that maintains some degree of balance between the numbers of patients assigned to a and b at any stage of the trial.

On the other hand, randomization plays a vital role in the control of bias. Complete randomization [1], analogous to tossing a fair coin, reduces or eliminates bias in treatment comparisons. However, in a small- or moderate-sized trial, complete randomization may result in a severe imbalance between the numbers of patients assigned to the two groups. In fact, it has been recommended [1,2] that one should not use complete randomization for a single-stratum trial with a target sample size under 200.

Several restricted randomization rules have been proposed, including the permuted-block design [3], the biased-coin design [4], the urn design [5], and the adaptive biased-coin design [6]. In particular, the urn design forces a small trial to be balanced but behaves like complete randomization as the size of the trial increases. As a result, the treatment assignments within a sequence generated by the urn design are not as predictable as those of other restricted randomization procedures, and the vulnerability to bias is likewise reduced.

In this article, we review the important statistical properties of the urn design along the lines described in ref [7]. We describe the balancing and bias reduction properties of the urn design. We then describe permutation tests of the equality of two groups based on the urn randomization distribution for the family of linear rank tests. We also discuss how to perform a post-stratified permutation test or subgroup analysis based on the urn design. Finally, we describe the generalization of the urn design to a prospectively stratified study for which the number of factors is so large that it is not feasible to employ an independent randomization within each possible stratum.

THE URN DESIGN

The urn design is a generalization of the class of biased-coin designs (BCDs) that were introduced by Efron [4] and which can be described in the following manner. Suppose that after n assignments, n_a as and n_b bs have been assigned. We then let d_n be some function of n_a and n_b such that $d_n = 0$ if $n_a = n_b$. Efron suggested $d_n = n_a - n_b$. Then the following rule is used: if $d = 0$, assign the patient to either treatment with probability 1/2; if $d < 0$ (excess of bs), assign the patient to a with probability p; if $d > 0$ (excess of as), assign the patient to b with probability p. A value for p is used such that $p \geq 1/2$. Thus, the BCD(p) forces the trial to tend to be balanced. However, it may not be satisfactory in some situations because the probability of assignment (p) is constant regardless of the magnitude of the imbalance. To alleviate this problem, Wei [5,6] proposed the family of adaptive biased-coin designs in

which p fluctuates as a function of the degree of imbalance. The urn design is the most widely studied member of this family of designs.

A generalized Friedman's urn model [8] can be used to explain the urn design. An urn contains α white and α red balls originally. For a treatment assignment a ball is drawn at random and replaced. If the ball is white, treatment a is assigned; if red then b is assigned. Furthermore, β additional balls of the opposite color of the ball chosen are added to the urn. Here α and β can be any reasonable nonnegative numbers. This drawing procedure is repeated for each assignment. This urn design is designated by UD(α, β). Note that the UD(α, 0) is simply complete randomization.

For the case $\alpha = 0$ and $\beta > 0$, namely UD(0, β), either treatment a or b will be chosen with probability 1/2 for the first assignment. This particular design has the following interesting property. Again let n_a and n_b be the number of prior assignments to a and b after n assignments. Then the $(n + 1)$th patient will be assigned to a with probability $p_{n+1} = n_b/n$ (or to b with probability $1 - p_{n+1} = n_a/n$), regardless of the value of β. That is, the probability of having treatment a on the next assignment equals the proportion so far assigned to b. This particular design can easily be implemented and has an efficiency similar to the random allocation rule [1], which has been claimed to be a nearly optimum design [9,10] in the case where the size of each treatment group is predetermined. This design is also a member of the class of adaptive restricted randomization designs studied by Wei [6,11], Smith [12], and Wei, Smythe, and Smith [13].

BALANCING PROPERTY OF THE URN DESIGN

Consider the general UD(α, β) design. Let D_n be the absolute value of the difference between the numbers in the two treatment groups after the nth assignment. Then D_n forms a stochastic process with possible values $d \in \{0, 1, 2, \ldots, n\}$. Initially $D_0 = 0$. The $(n + 1)$ stage transition probabilities are

$$\begin{aligned} &\Pr(D_{n+1} = d - 1 \mid D_n = d) = 1/2 + \beta d/[2(2\alpha + \beta n)] = P(d, n) \\ &\Pr(D_{n+1} = d + 1 \mid D_n = d) = 1/2 - \beta d/[2(2\alpha + \beta n)], \\ &\Pr(D_{n+1} = 1 \mid D_n = 0) = 1, \end{aligned} \tag{1}$$

where $1 \leq d \leq \mathbf{n}$ [5]. We note that $P(d, n)$ is monotonically increasing with respect to d, monotonically decreasing with respect to n, and tends toward to 1/2 as n increases for fixed $d > 0$. Therefore, the UD(α, β) forces the trial to be more balanced when severe imbalance occurs and also forces a small-sized experiment to be balanced. However, as n increases, the UD(α, β) behaves like the complete randomization design.

The transition probabilities in eq. (1) can be used recursively to calculate the probability of an imbalance of degree d at any stage of the trial as

$$\begin{aligned} \Pr(D_{n+1} = d) = {} & \Pr(D_{n+1} = d \mid D_n = d - 1)\Pr(D_n = d - 1) \\ & + \Pr(D_{n+1} = d \mid D_n = d)\Pr(D_n = d) \end{aligned} \tag{2}$$

for $0 \leq d \leq n + 1$. A comparison among the UD(0, β), the BCD(2/3), the permuted block design with length 10, and the complete randomization UD(α,

Table 1 Probability[a] That a Trial Is Exactly Balanced After n Allocations for Small n

Design	n				
	2	4	6	8	10
UD (0,β)	1.00	0.667	0.550	0.479	0.430
BCD (2/3)	0.667	0.593	0.560	0.541	0.530
Permutated blocks with length 10	0.556	0.476	0.476	0.555	1.000
Complete randomization	0.500	0.375	0.313	0.273	0.246

[a]From refs. 4 and 5.

0) is given in Table 1. Table 1 presents the probabilities that a small trial ($n \leq 10$) will be exactly balanced. Notice here that the UD(0, β) forces the experiment to be much more balanced than the other designs.

For a moderate or large n, the probability of imbalance, $\Pr(D_n > d)$, for the UD(α, β) can be approximated by $2\Phi[-Z_d]$, where

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

[14] and where $\Phi(\cdot)$ is the distribution function of the standard normal variate. For the UD(0, 1), $Z_d = (d + 0.5)/\sqrt{n/3}$. With complete randomization $\Pr(D_n > d)$ can be approximated by $2\Phi[-(d + 0.5)/\sqrt{n}]$. In terms of the proportions of assignments, let $q_u = \max(n_a, n_b)/n$. It then follows that the probability of an imbalance $\Pr[q_u > r]$ 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. Since the standardized deviate increases on the order of $\sqrt{3n}$ with UD(0, 1) versus $\sqrt{n}$ with complete randomization, as n increases imbalances are increasingly far less likely with UD(α, β) than with complete randomization.

It can also be shown that the efficiency of the UD(0, β) compares favorably to that of the random allocation rule which yields perfect balance. Suppose that Y_a and Y_b are responses of patients treated by a and b with a common variance σ^2 and means μ_a and μ_b, respectively. At some stage, let n_a and n_b be the numbers assigned to a and b. Then the variance of $\bar{Y}_b - \bar{Y}_a$, an estimator of $\mu_b - \mu_a$, is $\sigma^2[1/n_a + 1/n_b]$, where $\bar{Y}_a$ and $\bar{Y}_b$ are sample means for a and b, respectively. If the total sample size $n = 2m$ is prespecified, a perfectly balanced design with $n_a = n_b = m$ can be obtained with a random allocation rule [1] for which the quantity $\eta = [1/n_a + 1/n_b]$ is minimized at $\eta = 2/m$. Now, if n is not known beforehand, it is interesting to know how many extra observations are needed for the UD(0, β) to reduce η to be less than or equal to $2/m$. That is, we continue taking observations until n_a and n_b satisfy

$$\frac{1}{n_a} + \frac{1}{n_b} \leq \frac{2}{m}. \tag{4}$$

If we write $n_a + n_b = 2m + U$, then U is the number of additional observations

required by the UD(0, β) to satisfy this condition. It follows from Wei [6] that for any given u and large m,

$$\Pr(U \leq u) \cong \Phi\left[(3u)^{\frac{1}{2}}\right] - \Phi\left[-(3u)^{\frac{1}{2}}\right]. \tag{5}$$

For large m, $\Pr[U \leq 4]$ is approximately 0.9995, and thus the UD(0, β) needs at most *four* extra observations to satisfy the above inequality, that is, yield the same efficiency as the perfectly balanced random allocation rule. By way of comparison, for complete randomization, Blackwell and Hodges [15] show that $\Pr(U \leq u) \cong \Phi(u^{\frac{1}{2}}) - \Phi(-u^{\frac{1}{2}})$. In this case, $\Pr[U \leq 4] \cong 0.95$.

REDUCTION OF EXPERIMENTAL BIASES

As reviewed by Lachin [7], a measure of the selection bias of a treatment assignment rule is the expected number of correct guesses of treatment assignments in excess of that possible by chance alone which the investigator can make if he guesses optimally. This was termed the expected bias factor, designated as $E(F)$ [7].

Consider the case of even sample sizes, $n = 2m$. For complete randomization $E(F) = 0$ and there is no expected selection bias. With respect to the BCD(p), the expected number of excess correct guesses in $2m$ assignments asymptotically approaches $(\gamma - 1)m/2\gamma$, where $\gamma = p/(1 - p)$ [4]. For the UD(α, β), the probability of guessing correctly on the $(n + 1)$th assignment is

$$g_{n+1} = 1/2 + \frac{E(D_n)\beta}{2(2\alpha + \beta n)} \tag{6}$$

[5], where $E(D_n) = \sum_{d=0}^{n} d \Pr(D_n = d)$ can be obtained by the recursive relationship in eq. (2). Therefore, the expected number of excess correct guesses after $2m$ assignments for UD(α, β) is $E(F) = \sum_{i=1}^{2m} g_i - m$.

Figure 1 shows a comparison among the UD(0, β), the random allocation design with length $2m$, the permuted block design of block length 10, and BCD(2/3). The expected bias factor for the random allocation rule is described in ref [1] and for the permuted-block design in ref. 3. For the BCD(p), it follows from ref. 4 that $E(F) = (r - 1)m/2r$, where $r = p/(1 - p)$. Each of these has greater potential for selection bias than the urn design. Also, since g_{n+1} in eq. (6) converges to 1/2 as n increases, the UD(α, β) again tends to behave like complete randomization as n increases, thus gradually eliminating selection bias for future assignments.

As also reviewed by Lachin [7], another kind of bias is accidental bias [4], which may be caused by an imbalance between treatment groups in the distributions of a prognostic factor, known or unknown to the investigator. One way to evaluate the vulnerability of a design to such bias is to examine $\rho_{n,k} = \text{cov}(T_n, T_{n+k})$ for all positive integers n and k, where $T_n = 1$, if the nth patient is assigned to a and -1 if to b. For the UD(α, β), it can be shown that for any given k, $\rho_{n,k} \to 0$ as $n \to \infty$ [6]. This indicates that the components of the tail of the vector of treatment assignments $T = [T_1, \ldots, T_n, \ldots, T_{n+k}]$ are almost uncorrelated for the UD(α, β). Therefore, asymptotically the urn design is free of accidental bias.

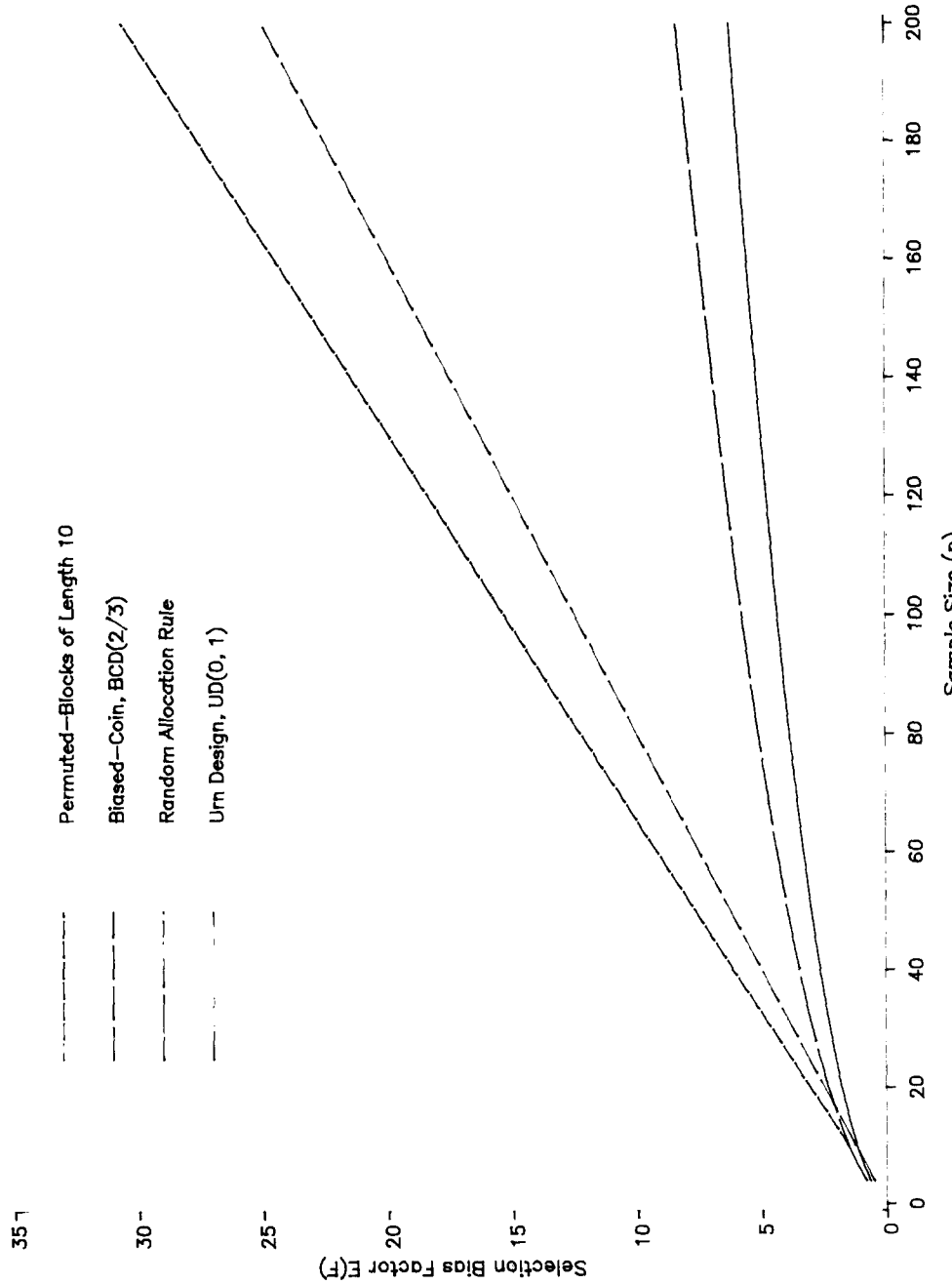

Figure 1 Selection bias factor $E(F)$ as a function of sample size (n) for the permuted block design with block length 10, the biased-coin design BCD(2/3), the random allocation rule, and the urn design UD(0, 1).

TWO-SAMPLE PERMUTATION TESTS

As reviewed by Lachin [7], there is a fundamental difference between a population and a randomization model as a basis for constructing a statistical test of the difference between treatment groups. Under a population model, the patients in the trial are assumed to be a random sample from a well-defined homogeneous population. In this case, the distribution of a test statistic under the null hypothesis can be generated on the basis of the assumption that the patients' responses are independent and identically distributed random variables. However, in a clinical trial there is usually no sampling basis for a population model, and such a model can only be postulated. Instead, a randomization model [16, chap. 1] can be utilized to test the null hypothesis H_0 that there is no difference between a and b among the patients entered into the trial. With this approach, the distribution of the significance test is generated by the experimental randomization design actually employed. For any given sequence of patients' responses, one can tabulate all possible patterns of treatment assignments to patients using the restricted randomization rule and calculate the corresponding probabilities of treatment assignments. This will generate the null permutational distribution of a test statistic and a test for H_0.

An example of a permutation test is presented in Table 1 of ref. 7. For complete randomization, the unconditional reference set consists of the 2^n possible sequences of assignments, each of which has equal probability. Likewise, the conditional reference set consists of the ${}_nC_{n_a}$ equiprobable sequences with n_a or n assignments. In Table 2 we present the same example to illustrate the permutation test with a UD(0, 1) randomization. There are four patients and the realized treatment assignments are abba. The patients' treatment responses are $\{Y_j\} = \{2, 1, 5, 6\}$ with corresponding ranks $\{c_j\} = \{2, 1, 3, 4\}$. The test statistic S used here is the ordinary Wilcoxon rank sum statistic computed as the sum of the ranks in group a minus its expected value. Note that the $\{Y_j\}$ and their corresponding ranks $\{c_j\}$ are treated as constants, while the treatment assignments generated by the randomization utilized in the trial are treated as the random component of the trial. The observed value of S is 1. We now wish to see how unlikely it is to observe a value equal to or larger than 1.

If the assignments were made by complete randomization, each of the 2^4 sequences in the unconditional reference set have equal probability ($=1/16$) and the unconditional p value is 4/16 (see Table 1 of ref. 7). Unlike complete randomization, however, with the urn design each of the 2^4 possible patterns of assignments in the reference set does not have the same probability of occurance, and some sequences are eliminated (probability = 0). For example, Pr(*abba*) = (1/2)(1)(1/2)(2/3) = 1/6, whereas Pr(*abba*) = 1/12. Also, the sequence *aaba* could not occur since Pr(*aa*) = 0. Table 2, therefore, shows the resulting permutational probability distribution for S based on the urn design. Unconditionally, there are eight possible sequences for which $S \geq 1$ with probability (1/12) + (1/6) = 3/12. By coincidence, this is the same as the unconditional p value (1/4) obtained assuming complete randomization.

For restricted randomization designs, such as the urn design, Cox [17] suggested the use of a conditional permutation test, whereby the significance

Table 2 Exact Unconditional and Conditional Permutational Tests Based on the UD(0,1) with Realized Treatment Assignments *a*, *b*, *b*, *a* and Observed Responses $Y_j = \{2, 1, 5, 6\}$[a]

Possible Assignments (Unconditional Reference Set)	UD(0,1) Probability	Test Statistic S
aaaa	0	0
aaab	0	−1.5
aaba	0	−0.5
aabb[b]	0	−2
abaa	1/12	1.5[c]
abab[b]	1/6	0
abba[b]	1/6	1[c]
abbb	1/12	−0.5
baaa	1/12	0.5
baab[b]	1/6	−1
baba[b]	1/6	0
babb	1/12	−1.5
bbaa[b]	0	2[c]
bbab	0	0.5
bbba	0	1.5[c]
bbbb	0	0

[a]For each permutation, the test statistic is the Wilcoxon rank sum statistic computed as the sum of the ranks $c_j = \{2, 1, 3, 4\}$ in group *a*, less the mean rank (2.5) times n_a for that permutation. For the actual assignment *abba*, $S = 6 - 2(2.5) = 1$.

[b]Conditional reference set

[c]$S \geq 1.0$

level is computed conditionally on the difference between the final numbers of patients assigned to *a* and *b*, or some other indicator of the final imbalance. For the above example, the observed treatment assignment pattern is *abba*. If we restrict ourselves to those sequences of treatment assignments where the difference between the final numbers of patients to *a* and *b* is zero, then only six patterns remain in the conditional reference set of permutations (Table 2). The *p* values of this conditional test are 2/6 for complete randomization and 1/4 for the UD(0, 1).

With a small sample size, there can be a substantial difference between the conditional and unconditional *p* values. For example, had the actual sequence of assignments in Table 2 been *abaa*, then $S = 1.5$, and the unconditional *p* value computed from Table 2 is 1/12. The conditional reference set, however, changes from that shown in Table 2 and now includes the four possible sequences with three *a*s and one *b* {*aaab*, *aaba*, *abaa*, *baaa*}. Over this set, the conditional *p* value is 1/2.

The marked difference between these unconditional and conditional *p* values is due to the fact that the sequences with an imbalance (such as abaa) are less likely under restricted randomization and thus contribute differentially to the unconditional *p* value. The conditional *p* value, however, is computed only from sequences with the same imbalance as that observed. Therefore, the conditional *p* value is preferred. This is similar to the traditional argument

for conditioning in a population model wherein n_a and n_b are ancillary statistics that provide no information regarding the true treatment difference.

The calculation of an exact permutational p value becomes rather unwieldy, even for a moderate-sized trial. Recently, Mehta, Patel, and Wei [18] have studied an efficient algorithm for computing exact p values with various restricted randomization rules and test statistics. Their method can easily handle a trial with n up to 50 or to 80, depending on the nature of the scores used in the rank statistic.

For a large trial, Smythe and Wei [19] and Wei, Smythe, and Smith [13] have studied the asymptotic null permutational distribution of linear rank statistics for testing the equality of the two treatment groups based on the randomization model. More specifically, suppose that $\{Y_1, \ldots, Y_n\}$ is the sequence of responses actually observed from the patients. Let the corresponding scores of the Ys be denoted by $\{c_1, \ldots, c_n\}$ with overall mean $\bar{c}$, where c_j may be a function of the rank of Y_j among all ys. Formally, the c_j should be written as c_{jn} to designate that the jth score may change with n. Also, let $\{\tau_1, \ldots, \tau_n\}$ be a sequence of binary indicators for treatment assignment, $\tau_j = 1$ if a, 0 if b. The linear rank statistic used to test the null hypothesis H_0 is

$$S = \sum_{j=1}^{n} (c_j - \bar{c}) \left[\tau_j - \frac{1}{2}\right]. \tag{7}$$

As described in ref. 7, under a population model or a simple randomization model, appropriate choice of scores $\{c_j\}$ yields the algebraic equivalent of the chi-square test for 2 × 2 tables, the Wilcoxon rank sum test, and the log-rank and Peto–Peto–Prentice–Wilcoxon tests for survival data, among many others.

For the UD(α, β), Smythe and Wei [19] and Wei, Smythe and Smith [13] showed that if

$$\frac{\max\limits_{1 \le j \le n} (c_j - \bar{c})^2}{\sum\limits_{j=1}^{n} (c_j - \bar{c})^2} \approx 0, \qquad \text{for large } n, \tag{8}$$

then the distribution of the test statistic S can be approximated by a normal distribution with mean 0 and variance $V = \frac{1}{4} \sum\limits_{j=1}^{n} b_j^2$, where

$$b_j = (c_j - \bar{c}) - \sum_{l=j+1}^{n} \frac{[2\alpha + (j-1)\beta]\,\beta(c_l - \bar{c})}{[2\alpha + (l-1)\beta][2\alpha + (l-2)\beta]}, \qquad 1 \le j \le n,$$
$$b_n = (c_n - \bar{c}). \tag{9}$$

For the UD(0, β) design, eq. (9) then reduces to

$$b_j = (c_j - \bar{c}) - \sum_{l=j+1}^{n} \frac{(c_l - \bar{c})(j-1)}{(l-1)(l-2)}, \qquad 1 \le j \le n,$$
$$b_n = (c_n - \bar{c}). \tag{10}$$

However, if complete randomization is utilized, then $b_j = (c_j - \bar{c})$ and the variance V is simply $\frac{1}{4} \sum\limits_{j=1}^{n} (c_j - \bar{c})^2$. In each case, the liner rank statistic $W = S/V^{\frac{1}{2}}$ is asymptotically distributed as standard normal.

The condition (8) on the scores c_j is rather mild. For example, if c_j is the rank of Y_j among $\{Y_1, \ldots, Y_n\}$, (8) is trivially satisfied and the test statistic S is the usual Wilcoxon test statistic.

For the urn design UD(0, β), the large-sample approximation to the null permutational distribution of the test statistic S given the final difference $d_n = n_a - n_b$ is given in Wei, Smythe, and Smith [13] and later is justified by Smythe [20]. For large n, the conditional distribution of S given d_n can be approximated asymptotically by a normal distribution with mean

$$\mu = \frac{d_n\left[\Sigma_j b_j \bar{b}_j\right]}{2\left[\Sigma_j \bar{b}_j^2\right]\sqrt{n}} \tag{11}$$

and variance

$$V = \frac{\Sigma_j b_j^2}{4}\left[1 - \frac{\left[\Sigma_j b_j \bar{b}_j\right]^2}{\left[\Sigma_j \bar{b}_j^2\right]}\right] \tag{12}$$

where $\bar{b}_j$ is obtained from eq. (9) with $n^{-\frac{1}{2}}$ in place of $(c_j - \bar{c})$. Under this conditional distribution, the statistic $W = (S - \mu)/V^{\frac{1}{2}}$ is asymptotically distributed as standard normal [20]. This approximation is surprisingly good even for a moderate-sized trial, say $n = 20$ [18].

AN ILLUSTRATION OF PERMUTATION TESTS

As an illustration, we now apply these tests to the data presented in Table 3 from the V.A. Cooperative Urologic Research Group (VACURG) Trial of estrogen ($\tau = 1$) versus placebo ($\tau = 0$) in the treatment of prostatic cancer [21,22]. For each patient, Table 3 presents the death or censoring time and death-censoring indicator variable. The sequence of randomized treatment assignments (a and b) presented in Table 3 was generated using the UD(0, 1) design with no stratification. Note that the actual randomization procedure for this study has not been reported in the literature.

Table 3 Survival or Censoring Time with Indicator (δ) for Survival (δ = 1) or Censoring (δ = 0) from the VACURG Study [21, 22] with Patients Arranged in Sequence According to a UD(0,1) Randomization. Additional Hypothetical Data Include a Trend Variable

Treatment	Time	δ	Trend
0	84	1	0.2460
1	84	1	8.9887
0	32	1	−0.1433
1	20	1	7.8792
0	142	1	2.4582
1	61	1	10.3714
1	45	1	8.1106
1	63	1	4.2394

Table 3 (continued)

Treatment	Time	δ	Trend
0	178	0	2.6265
1	151	0	10.5294
0	5	1	2.7364
0	89	1	1.6887
0	173	0	9.0533
1	75	1	7.6867
1	30	1	6.0447
1	163	0	−1.2359
0	89	1	4.8903
1	117	1	9.5883
1	0	1	3.2329
1	68	1	−0.9706
1	0	1	9.8879
0	166	0	6.6266
0	111	1	5.8767
1	55	1	9.0169
0	133	1	7.5006
0	163	0	12.1874
0	110	1	0.0114
0	192	0	11.0004
0	98	1	10.2382
1	199	0	5.6023
0	95	1	15.1311
1	172	1	16.3684
0	155	0	0.4070
1	140	1	10.9962
0	93	1	19.7615
1	37	1	8.6910
0	155	0	16.6296
0	70	1	3.2987
1	157	0	24.3988
1	19	1	5.9929
0	29	1	13.8394
0	112	1	5.3194
1	144	0	9.4257
1	14	1	10.6037
0	46	1	19.9926
1	77	1	9.0209
0	65	1	18.3835
1	4	1	15.8348
1	143	0	22.2053
0	45	1	10.1928
0	60	1	16.9262
0	142	0	20.4779
1	6	1	19.2374

Table 3 (continued)

Treatment	Time	δ	Trend
1	65	1	12.8644
0	130	1	18.6262
1	128	1	17.5272
1	177	0	16.0931
0	156	0	14.7935
0	146	0	7.9850
1	171	1	24.3326
0	26	1	14.7333
1	93	1	22.2685
1	33	1	12.3432
1	5	1	30.7030
0	76	1	22.2811
0	38	1	20.5690
1	26	1	9.9499
0	61	1	25.4126
0	131	0	15.0021
0	28	1	14.9356
0	38	1	19.9086
1	140	0	10.1829
1	136	0	21.1393
0	113	1	18.1153
0	125	0	17.7858
1	66	1	20.2225
0	120	0	21.8095
1	13	1	23.0034
1	117	1	26.4385
0	56	1	28.6514
1	107	1	25.9123
1	108	0	12.2057
0	148	0	20.9661
1	0	1	21.3069
1	12	1	19.7782
0	114	1	27.3316
0	117	0	17.9777
0	119	0	14.0875
0	103	1	20.4445

Table 4 presents the results of various analyses of these data. For each scoring function, the linear rank test was applied using the UD(0, 1) unconditional permutational variance obtained from eq. (10), and the UD(0, 1) conditional test based on the conditional mean and variance in eqs. (11) and (12). These are compared to the test using the conditional complete randomization variance, that is, the UD(1, 0) variance (eq. (3) in ref. 1). As shown in refs. 1 and 7, the unconditional complete randomization test is asymptot-

Table 4 Linear Rank Statistics (S) Applied to the Data of Table 3, Their Null Expectation (E), Variance (V), Standardized Deviate (Z), and p Value, Using the Complete Randomization UD(1, 0) Permutational Distribution and Using Both the Unconditional and Conditional Urn Randomization UD(0, 1) Permutational Distributions

	Rank statistic	Unconditional UD(0, 1) complete randomization			Conditional UD(1, 0) urn randomization[a]			Conditional UD(0, 1) urn randomization			
	S	V	Z	p	V	Z	p	E	V	Z	p
Dead vs. alive—binary scores	2.56	4.601	1.194	0.232	4.656	1.187	0.235	0.045	4.649	1.167	0.243
Survival times											
Log-rank scores	−6.247	15.370	−1.593	0.111	15.646	−1.579	0.114	−0.034	15.642	−1.571	0.116
Wilcoxon scores	−5.490	7.042	−2.069	0.039	7.211	−2.044	0.041	−0.019	7.210	−2.037	0.042
Time trend measure—rank scores											
Null distribution	23.0	14685.0	0.190	0.849	11063.2	0.219	0.827	−0.0017	10101.6	0.398	0.690
Alternative distribution[+]	−302.0	14685.0	−2.492	0.013	11008.7	−2.878	0.004	−0.0017	10085.1	−2.841	0.004

[a]Expectation (E) is zero.

[b]5 subtracted from all values for TREND in Table 3 for observations with Treatment group = 1.

ically equivalent to both the conditional test and the population model–based test. The scoring functions employed are the simple binary score (1, 0) corresponding to simple mortality (i.e., dead vs. alive at the time last seen), which is equivalent to the chi-square test for a 2×2 table; and the log-rank scores and the modified-Wilcoxon scores for time of death (or censoring) [7]. In each case, the numerators of the rank statistic are equivalent for complete randomization and the UD(0, 1), each using $E(\tau_j) = 1/2$. Therefore, the differences between the statistics for the two randomization designs reflect differences in the variances.

Since the treatment assignments were generated by the UD(0, 1), the proper analysis is that using the UD(0, 1) permutational variance. However, the results are virtually identical using the variances based on the complete randomization distribution and based on the UD(0, 1), either unconditionally or conditionally.

This raises the general question of whether the treatment assignment rule used in a trial can be ignored in the analysis under the randomization model. That is, if the urn design was actually utilized in allocating patients to treatment groups, can a valid analysis be performed by acting as though complete randomization were used in the trial? In general, the answer is no. Mehta, Patel, and Wei [18] have generated several data sets with various time trends and demonstrated that the design should not be ignored in analyzing such data.

To illustrate this phenomenon, Table 3 also presents a variable (TREND) that displays a moderate time trend as represented by a Spearman rank order correlation of 0.761 with order of entry into the trial. The values shown in Table 3 were generated under H_0 (no treatment effect). To introduce a treatment effect under the alternative hypothesis, the value $5\tau_j$ was subtracted from each observation. The analysis of these time trend measures using the Wilcoxon rank sum statistic (rank scores) is also presented in Table 4. Under both the null and alternative, there are substantial differences between the Z values for the unconditional and conditional UD(0, 1) analyses versus the UD(1, 0) complete randomization analysis. Also, these differences are more pronounced under the alternative hypothesis. In each case, the UD(0, 1) p value is substantially smaller than the complete randomization p value. A similar example of the effect of a time trend was also given by Halpern and Brown [23]. Such calculations with various degrees of time trend indicate that the ratio of the UD(0, 1) Z value to that for complete randomization increases as the magnitude of the time trend increases, as measured under H_0 by the Spearman rank order correlation with order of entry.

This is important because the null distribution of a two-sample test statistic based on complete randomization is the same as that generated under the population model. In analyses where a UD(0, 1) permutational analysis is approximately equivalent to a complete randomization UD(1, 0) permutational analysis, it may be reasonable to assume that the observations arose from a homogeneous population. However, if there is an obvious discrepancy in the significance levels generated by the design actually used versus complete randomization for a given variable, then this indicates that the homogeneous population model assumptions may be violated. Often, this discrepancy will arise due to a time trend among the observations, in which case it would be

more plausible to invoke a time heterogeneous population model. This would suggest that additional analyses, for example, regression models, should incorporate temporal sequence of entry into the model.

We caution, however, that a simple monotonic temporal trend in the scores, as illustrated in Table 3, is only one way in which a difference between the UD(0, 1) and the UD(1, 0) complete randomization (or population model) analyses could arise. Therefore, it is highly recommended that the principal analyses of outcome measures be based on the permutational distribution under the randomization design employed, in this case, the UD(α, β).

Finally, we note that the urn design can easily be generalized to the case of multiple-group comparisons. The mechanism of assignments is exactly as described previously, where β balls of each treatment other than the chosen one are added to the urn after each assignment. The balancing and randomization properties of this generalization are described by Wei [11,14]. Also, the permutation test of the hypothesis that there is no difference among two or more treatment groups based on the UD(α, β) was investigated by Wei, Smythe, and Smith [13].

PROSPECTIVELY STRATIFIED RANDOMIZATION

In clinical trials often there are one or more prognostic factors that are known or thought to affect the patients' responses to treatment, and each factor has several levels. A stratum is defined as a group of patients who have one particular combination of factor levels in common. If the number of stratification factors is small, each stratum is generally treated as a separate independent subtrial in which the treatment assignments may be based on a separate UD(α, β) randomization. For example, in multicenter trials it is customary that the randomization be stratified by clinical center. The randomization and balancing properties of this prospectively stratified urn scheme have been studied by Wei [11].

Let n_{ak}, n_{bk}, and $n_k = n_{ak} + n_{bk}$ refer to the sample sizes in the kth strata, $k = 1, \ldots, K$. For a prospectively stratified UD(α, β), eqs. (1)–(3) can be used to calculate the probability of imbalance within any particular strata, or in aggregate over all strata combined. As for complete randomization [1], using eq. (3) it is easy to show that the probability of an total imbalance, say $d = |\Sigma_k[n_{ak} - n_{bk}]|$, using an UD($\alpha$, β) separately within each stratum, is asymptotically the same as a single-stratum design with $n = \Sigma n_k$.

In addition, the extension of the permutational linear rank test of H_0 with a prospectively stratified randomization is presented in eq. (7) in ref. 7. For the kth stratum, let S_k be the corresponding stratum-specific rank test with null expectation μ_k and variance V_k. For an unconditional UD(α, β) test, $\mu_k = 0$ and V_k is presented in eq. (9). For a conditional test, μ_k and V_k are as presented in eqs. (11) and (12). The stratified test is then provided by

$$W = \frac{\Sigma_k \omega_k(S_k - \mu_k)}{[\Sigma_k \omega_k^2 V_k]^{\frac{1}{2}}}, \tag{13}$$

which is distributed as standard normal under H_0 for any set of weights $\{\omega_k\}$. The choice of weights is discussed in ref. 7.

POSTSTRATIFIED (SUBGROUP) ANALYSES AND MISSING DATA

Subgroup Analyses

It is a common practice in clinical trials that various poststratified or subgroup analyses are performed using factors that were not used in the beginning of the trial to stratify the randomization of treatments to patients. These poststratified analyses may be conducted to explore whether there is a treatment effect within a particular stratum or whether there is an interaction between treatment and the various strata of a particular factor. They may also provide a covariate-adjusted assessment of treatment effects. In the latter case, the investigator may want an aggregate test for the equality of two treatment groups or an estimate of treatment difference combined over the strata to yield an overall evaluation of treatment effects.

Under complete randomization, the permutational distribution of a test statistic computed within any subset of patients is the same as if that subset had been obtained by prestratified randomization, and therefore the stratum-specific statistics are independent. Likewise, under a population model, the statistics for different strata are independent. Therefore, in these cases, a poststratified or subgroup analysis is straightforward. However, for the randomization model with a restricted randomization rule the analysis is not obvious.

For the urn design, the large-sample theory of two-sample poststratified permutation tests for testing the null hypothesis H_0 and other hypotheses has been investigated by Davis [24]. More specifically, suppose that in a single-stratum trial, the poststratification factor in which the investigator is interested has L levels. Also, let H_l be the hypothesis that there is no difference between a and b for the lth stratum, $l = 1, \ldots, L$. For the jth patient, let $v_j = (v_{1j}, \ldots, v_{Lj})'$, where $v_{lj} = 1$, if the level of the factor for this patient is l, and 0 otherwise. Furthermore, let the scores of the Ys be denoted by c_j, $j = 1, \ldots, N$. The scores are a function of the responses $\{Y\}$ and the indicator vector $\{v\}$. For example, c_j may be the rank of Y_j among Ys for patients who fall into the same level of the factor as the jth patient. The mean of the scores within the lth stratum can be denoted as $\bar{c}_l$.

For simplicity we describe the unconditional UD(0, 1) test. Consider the test statistic S_l for testing H_l within the lth stratum, where

$$S_l = \sum_{j=1}^{n} v_{lj}(c_j - \bar{c}_l)[\tau_j - E(\tau_j - {}^1/_2)] \tag{14}$$

and

$$\bar{c}_l = \frac{\sum_{j=1}^{n} v_{lj}\, c_j}{\sum_j v_{lj}} \tag{15}$$

for $l = 1, \ldots, L$. Let $B_j = (b_{j1}, \ldots, b_{jl})$, where

$$b_{jl} = v_{lj}(c_j - \bar{c}_l) - \sum_{i=j+1}^{n} \left[\frac{v_{lj}(c_j - \bar{c}_l)(j-1)}{(i-1)(i-2)}\right], \qquad j = 1, \ldots, n. \tag{16}$$

Then under the mild conditions on the cs equivalent to eq. (8), for large n the distribution of the vector of rank statistics $S = (S_1, \ldots, S_L)'$, can be

approximated by a multivariate normal with mean 0 and $L \times L$ covariance matrix Λ, where

$$\lambda = \frac{1}{4}\sum_{j=1}^{N} B_j B_j'. \tag{17}$$

Note that eq. (16) is an obvious extension of eq. (10). Likewise, the b_j from eq. (9) can be so modified for the UD(α, β). Unfortunately, the poststratified permutation test has not been developed to condition on the numbers of *a*s and *b*s in each stratum.

This large-sample permutation test based on S can be used to test the hypothesis H_l separately within the lth stratum using the normal deviate $W_l = S_l/\lambda_{ll}^{\frac{1}{2}}$, where λ_{ll} is the lth diagonal element of Λ, or to test the L hypotheses $(H_1, H_2, \ldots, H_L)$ simultaneously using $S'\Lambda^{-1}S$, which is distributed as chi-square on L df under H_0. Furthermore, if there is no obvious interaction between treatment and strata, then the $\{S_l\}$ can be combined in a linear fashion using a vector of weights $\omega = (\omega_1, \ldots, \omega_L)'$ in

$$W = \frac{\omega' S}{(\omega' \Lambda \omega)^{\frac{1}{2}}} \tag{18}$$

in order to make an overall inference about the treatment difference [7, eq. (7)].

These methods can be generalized in an obvious manner to a poststratified analysis on one factor (say F) in a trial that employed a randomization stratified by another factor (say G). In this case, the poststratified analysis on F is conducted separately within each level of G, and then the results are pooled over levels of G using eq. (13), where S_k is actually the numerator of eq. (18) for the kth level of G and the V_k is likewise the denominator. Since this is an unconditional analysis, $\mu_k = 0$.

Missing Data

As described by Lachin [7], when some patients' responses are missing, a permutation test of treatment effect can be justified as a post hoc stratified subgroup analysis under the missing-at-random assumption. This assumption states that missingship is statistically independent of treatment. If this assumption is plausible, a test can then be performed as previously described using the single stratum of patients with observed responses.

THE USE OF THE URN DESIGN WHEN THE NUMBER OF STRATA IS LARGE

In a stratified randomization, when the number of strata is large, each stratum may contain very few patients. In this case, it is difficult to use the prospectively stratified design, that is, to each stratum as a separate independent trial. However, we still can use the urn scheme to construct an overall

treatment assignment rule that forces treatment balance simultaneously across all factor levels. A detailed illustration is presented in Wei [14].

Consider that there are Q factors at f_i levels for each, $i = 1, \ldots, Q$. Stratum-specific randomization would require a separate UD(α, β) or "urn" for each of the $\pi_{i=1}^{Q} f_i$ strata. Here, however, only $\Sigma_i f_i$ "marginal" urns are used, one for each stratifying factor level. For an income patient, the urn chosen is the one that has the greatest proportionate imbalance for whichever of the patient's factor levels. A ball from that urn is chosen and replaced, and then β balls of the opposite color are added to all the urns for that patient's factor levels.

For example, suppose three factors are used: clinic with ten levels, pretreatment (none, radiotherapy, or chemotherapy) and sex (male or female). Thus, 15 "urns" are used, each with white and red balls to represent treatments a and b, respectively. Suppose the next patient is a male from clinic 5 who had prior chemotherapy. Therefore, we would only consider using the clinic 5 urn, the male urn or the pretreatment-chemotherapy urn. Suppose the respective imbalances of white to red balls are 2 : 0, 5 : 6, and 8 : 11 in each urn. Here the greatest proportionate imbalance (2 : 0) is in the clinic 5 stratum for which the actual numbers of white and red balls are 1 and 3, respectively. Thus, for this patient, using the clinic 5 urn, treatment a will be assigned with probability 0.25, b with probability 0.75. After selection, β balls of the other color are added to that patients stratum factor level urns: the clinic 5, male, and pretreatment-chemotherapy urns.

CONCLUSIONS

The properties of the urn procedure can be summarized as follows. The urn procedure is relatively easy to implement, especially via computer. It forces a small-sized trial to be balanced but approaches complete randomization as the sample size increases. It has less vulnerability to selection bias than does the permuted-block design, biased-coin design, or random allocation rule. As n increases, the potential for selection bias approaches that of complete randomization for which the expected selection bias is zero. Likewise, as n increases, the potential for accidental bias approaches that of complete randomization. The urn design can also be extended to the prospectively stratified trial when the number of strata is either small or large.

For the family of linear rank tests, which includes the popular log-rank and modified-Wilcoxon tests for survival data, the urn design permits explicit large-sample permutation tests. The urn design also permits a poststratified or subgroup analyses. These tests are not available in standard statistical computing packages, but can be programmed easily.

The UD(α, β) permutation test values may differ substantially from those based on a population model analysis that ignores the design actually employed. In fact, a difference should be expected if there is an obvious time trend in the scores employed in the rank statistic. In general, therefore, the data should be analyzed the way the study was randomized, with the appropriate UD(α, β) permutational analysis, including stratification if appropriate.

For Wei this work was partially supported by grants RO1-CA-45122 from the National Cancer Institute and RO1-AM-35952 from the National Institute of Diabetes, Digestive and Kidney Diseases (NIDDK). For Lachin this work was partially supported by the Diabetes Control and Complications Trial under NO1-DK-2-2206 from the NIDDK.

REFERENCES

1. Lachin JM: Properties of simple randomization in clinical trials. Controlled Clin Trials 9:312–326, 1988
2. Pocock SJ: Allocation of patients to treatment in clinical trials. Biometrics 35:183–197, 1979
3. Matts JP, Lachin JM: Properties of the permuted-block randomization in clinical trials. Controlled Clin Trials 9:327–344, 1988
4. Efron B: Forcing a sequential experiment to be balanced. Biometrika 58:403–417, 1971
5. Wei LJ: A class of designs for sequential clinical trials. J Am Stat Assoc 72:382–386, 1977
6. Wei LJ: The adaptive biased-coin design for sequential experiments. Ann Stat 6:92–100, 1978
7. Lachin JM: Statistical properties of randomization in clinical trials. Controlled Clin Trials 9:289–311, 1988
8. Friedman B: A simple urn model. Commun Pure Appl Math 2:59–70, 1949
9. Stigler SM: The use of random allocation for the control of selection bias. Biometrika 56:553–560, 1969
10. Wei LJ: On the random allocation design for the control of selection bias in sequential experiments. Biometrika 65:79–84, 1978
11. Wei LJ: A class of treatment assignment rules for sequential experiments. Commun Stat Theory Methods 7:285–295, 1978
12. Smith RL: Sequential treatment allocation using biased coin designs. J R Stat Soc B 46:519–543, 1984
13. Wei LJ, Smythe RT, Smith RL: K-treatment comparisons with restricted randomization rules in clinical trials. Ann Stat 14:265–274, 1986
14. Wei LJ: The application of an urn model to the design of sequential controlled trials. J Am Stat Assoc 73:559–563, 1978
15. Blackwell D, Hodges JL Jr: Design for the control of selection bias. Ann Math Stat 28:449–460, 1957
16. Lehmann EL: Nonparametrics: Statistical Methods Based on Ranks. San Francisco: Holden-Day, 1975
17. Cox, DR: A remark on randomization in clinical trials. Utilitas Math 21A:245–252, 1982
18. Mehta CR, Patel NR, Wei LJ: Constructing exact significance tests with restricted randomization rules. Biometrika, 75:295–302, 1988
19. Smith RT, Wei LJ: Significance tests with restricted randomization design. Biometrika 70:496–500, 1983
20. Smythe RT: Conditional inference for restricted randomization designs. Ann Stat, in press
21. Byar DP: Treatment of prostatic cancer: Studies by the Veterans Administration Cooperative Urological Research Group. Bull NY Acad Med 48:751–766, 1972
22. Slud E, Wei LJ: Two-sample repeated significance tests based on the modified Wilcoxon statistic. J Am Stat Assoc 77:862–868, 1982

23. Halpern J, Brown BW: Sequential treatment allocation procedures in clinical trials with particular attention to the analysis of results for the biased coin design. Stat Med 5:219–229, 1986

24. Davis CS: Two-sample post-stratified or subgroup analysis with restricted randomization rules. Commun Stat A, in press