*Drug Information Journal*, Vol. 32, pp. 423–432, 1998
Printed in the USA. 
0092-8615/98


# A METHOD OF BIASED COIN RANDOMIZATION, ITS IMPLEMENTATION, AND ITS VALIDATION

JAMES W. FRANE

Associate Director, Biostatistics, Genentech, Inc., South San Francisco, California

*Many randomized studies in small patient populations and studies in early research (such as Phase I and Phase II trials) have small to moderate numbers of patients. In such studies the use of simple randomization or blocking on only one or two factors can easily result in imbalance between treatment groups with respect to one or more potentially prognostic variables. Baseline adaptive randomization methods (such as biased coin methods) can be used to virtually guarantee balance between treatment groups with respect to several covariates. One such method, which has been implemented in Splus, is discussed in detail. The impact of the baseline adaptive randomization method on the nominal distribution of the analysis of covariance test statistic is also discussed. Rather than relying solely on the assumption that the distribution of the analysis of covariance test statistic has its nominal distribution when adaptive randomization is used, a mechanism in Splus has been developed to perform a randomization test taking into account all of the constraints imposed by the chosen adaptive randomization procedure.*

*Key Words:* Biased coin randomization; Randomization test; Splus

## MOTIVATION

A FREQUENTLY OCCURRING problem in small studies is that simple randomization can easily result in imbalance between treatment and control groups with respect to one or more significant prognostic variables. Sometimes analysis of covariance or multiway analysis of variance are used in an attempt to adjust for between-group differences with respect to the covariates. The value of analysis of covariance and multiway analysis of variance is well-recognized as a valid means of reducing the error sum of squares and thereby increasing power. Use of analysis of covariance and multiway analysis of variance to adjust for between-group differences, however, does not lead to unambiguous evaluation of study results when there is imbalance between groups with respect to covariates (eg, 1).

When there is imbalance between groups with respect to covariates, the bias in analysis may also be exacerbated when the relationship of the covariate to the dependent variable is nonlinear (and so raises the likelihood of unequal slopes among treatment groups). Such a nonlinearity is more likely to reduce power but not introduce bias in the case of balance between treatment groups with respect to the statistical distributions of the covariates.

In very large studies, it is unlikely that there will be any important differences between treatment groups with respect to prog-

Reprint address: James W. Frane, Genentech, One DNA Way, South San Francisco, CA 94080.

nostic variables. Hence, concern here is primarily for studies of small and moderate size where it is likely that there could be notable differences between treatment groups with respect to one or more prognostic variables. Even in large studies, however, important imbalance may occur when planned and unplanned subset analyses are performed.

On those rare occasions when all patients to be randomized are known before any patient is randomized to treatment, stratified randomization might be used if the number of covariates and blocking factors is small. The concern here is for studies where patients are enrolled one at a time and for studies with several covariates and blocking factors. For such studies, several papers have been written describing what could be called baseline adaptive randomization procedures. These are sequential randomizations where the randomization of each new patient is adjusted for the covariate values of both the new patient and for the covariate values of all previously enrolled patients, for example, Harville (2), Pocock and Simon (3), Taves (4), Pocock (5), Simon (6), Begg and Iglewitz (7), Atkinson (8), Smythe and Wei (9), Smith (10), Halpern and Brown (11), and Hannigan and Brown (12). Related work has been done by Hollander and Pena (13) and Begg (14). It is important to note that these randomizations are not adjusted for the results of the endpoint of the trial in question, hence the expression "baseline adaptive randomization."

While many papers provide a certain degree of general guidance and in some instances specific results, the results are not so general as to permit the unqualified use of baseline adaptive randomization procedures. In addition, constraints are made in the procedures so that approximately equal numbers of treatment and control patients are assigned at each study center and so that the total number of treatment and control patients are approximately equal at all times during study enrollment.

The author's personal experience has been with studies with continuous outcomes. (The general strategy and software, however, can easily be made to accommodate categorical response and censoring.) Covariates can be categorical (eg, sex, study center, etiology of disease) and/or continuous (eg, age, weight, body mass index).

The procedure discussed here is a generalization of the biased coin method originally proposed and studied by Efron (15,16). Patients in the procedure are assigned one at a time, that is, when each new patient is presented for randomization, the values of all covariates for this patient and previously randomized patients are known as well as the treatment assignments for the previous patients.

There may be concern about the cost and effort of implementing a baseline adaptive randomization procedure. One should recall, however, that protocols typically specify lists of inclusion and exclusion criteria so that obtaining information at a study center about the prognostic variables is frequently an easy additional task. Moreover, the additional effort required to execute a baseline adaptive randomization procedure is often small in comparison to the total of other efforts in the course of a study.

The author's baseline adaptive randomization procedure was implemented and validated using the Splus version of S (17).

## BLOCKING

Input to the baseline adaptive randomization procedure for each patient includes patient identification, study center, treatment blocking, and the values of one or more covariates. (As each new patient's screening data are added to the [blinded] randomization database, the treatment assignment is left pending until the randomization is performed and then the randomization database is updated with the randomized treatment code. There is necessarily some time lapse between the time of patient screening and the time of randomization.)

In addition to the data for each patient, the procedure is invoked with options including the following:

- Study drug is sent to centers in blocks. All of the drug in each block must be assigned before assignment of any drug in the following block at that center is permitted. For example, if there are two active and two placebo patients in each block, then there must be two active and two placebo assignments at a study center before either the third active or third placebo assignment can be made at that study center, and
- Conditional on completing the block as discussed above, there is a required balance in the overall number of patients for each treatment at any time during patient enrollment. For example, one can specify that the maximum difference in the total number of placebo and active patients over all study centers cannot exceed four.

The biased coin randomization is not, however, constrained to equal or nearly equal numbers of patients assigned to each treatment group. For example, one may (for a variety of reasons) choose to have approximately two treated patients for every one placebo patient. In this case, the balance in the overall number of patients for each treatment at any time during patient enrollment is controlled in terms of the actual and target numbers of patients randomized to each group. Specifically, if there is a two-to-one randomization and if the total number of patients at a given point during enrollment is 30, then the target numbers of patients in the two groups are 20 and 10. The constraint applies to actual enrollment numbers that differ from 20 and 10. If the constraint limit is four, then the actual numbers must be between 16 and 24 and six and 14, respectively.

## BIASED COIN RANDOMIZATION

The following is just one of many possible ways in which a biased coin randomization procedure can be implemented. The specific procedure described here can also easily be modified to provide other variations of biased coin randomization.

In a study, before the biased coin randomization can get fully underway, several patients must be randomly assigned without constraints. If there are $m$ treatment groups then the first $m$ patients are assigned with simple randomization, one to each of the groups without constraints. Then the next $m$ patients are also randomly assigned without constraints.

Subsequent randomizations are constrained with respect to:

- The continuous and categorical covariates,
- The number of patients in each treatment group at each study center, and
- The total number of patients over all study centers in each treatment group.

Conditional on satisfying both of the constraints described in the last section regarding the number of patients in each treatment group, the randomization of each new patient is performed using a variation of the biased coin procedure first proposed and defined by Efron (15). The essence of any biased coin procedure is that the probability of assignment to the various treatment groups varies according to the imbalance between groups with respect to one or more covariates. Other biased coin procedures have fixed probabilities defining the bias (eg, assigning the current patient with probability 2/3 to the group that results in the best balance). The author's procedure is distinguished from others insofar as the degree of bias varies with the degree of imbalance that exists at the time each patient is randomized.

There are several ways that one might choose to define imbalance. The procedure discussed here uses p-values for t-tests and analysis of variance (ANOVA) for continuous prognostic variables and uses frequency table p-values for categorical prognostic variables.

The procedure is defined as follows. For the sake of simplicity, temporarily ignore blocking constraints (such as equal or nearly equal numbers of treated and placebo patients at each study center). Also consider the case of a single continuous prognostic variable with a normal distribution and suppose that there are two groups. Let p denote

the p-value for the t-test between the two groups with respect to the prognostic variable. Let $p_1$ and $p_2$ denote the t-test p-values alternatively assuming assignment of the current patient to groups one and two, respectively. These t-tests are computed using the values of the prognostic variable for all patients previously randomized and for the patient currently being randomized. The current patient is then assigned to group one with probability $p_1/(p_1 + p_2)$ and to group two with probability $p_2/(p_1 + p_2)$. Thus, the randomization biases the assignment in the direction of providing balance between groups with respect to the prognostic variable. (Note that it is not necessary that the use of the t-test is rigorously justified. One only needs to know that the p-values in some sense measure the imbalance between the two groups with respect to the prognostic variable.)

If there are three groups, one uses ANOVA to obtain three p-values. Assignments are made to the three groups with probabilities $p_1/(p_1 + p_2 + p_3)$, and so forth.

If there is more than one prognostic variable, then p-values are computed first for each of the prognostic variables individually. Let $q_{ij}$ denote the p-value for the jth prognostic variable in the ith group. Then let $p_i = \min(q_{i1}, q_{i1}, \ldots)$. Probabilities for treatment group assignments are then made as described previously. Note that this strategy focuses on achieving balance with respect to the prognostic variable for which between-group imbalance would be greatest, that is, the strategy maximizes the minimum between-group p-values among all prognostic variables.

Table 1 provides a specific example with hypothetical p-values associated with randomizing a patient in a study with three treatment groups and four prognostic variables. Assignment of the new patient to Group One would yield an analysis of variance p-value of 0.43 for the first prognostic variable. The minimum p-value obtained by assignment of the new patient to Group One is for the third prognostic variable, that is, 0.15. Similarly, p-values and their minimums are obtained for groups two and three. Thus, the three groups have minimum p-values of 0.15, 0.19, and 0.11, respectively. The probability of randomization to the three groups is therefore:

**TABLE 1**
**Hypothetical p-values for Randomizing a Patient**

| | Group 1 | Group 2 | Group 3 |
|---|---|---|---|
| Covariate 1 | 0.43 | 0.91 | 0.11 |
| Covariate 2 | 0.37 | 0.22 | 0.17 |
| Covariate 3 | 0.15 | 0.23 | 0.24 |
| Covariate 4 | 0.82 | 0.19 | 0.14 |
| Minimum p | 0.15 | 0.19 | 0.11 |
| Randomization p | 0.33 | 0.42 | 0.24 |

$$0.15/(0.15 + 0.19 + 0.11) = 0.15/0.45 = 0.33,$$
$$0.19/0.45 = 0.42, \text{ and } 0.11/0.45 = 0.24$$

The section on blocking described constraints such as blocking. When determining the treatment assignment of each patient, the biased coin randomization is performed within the limits of the constraints. Thus, for example, if all treatments in a block have been assigned except one, then the remaining unassigned treatment is assigned to the next patient without further explicit randomization.

## ANALYSIS OF RESULTS

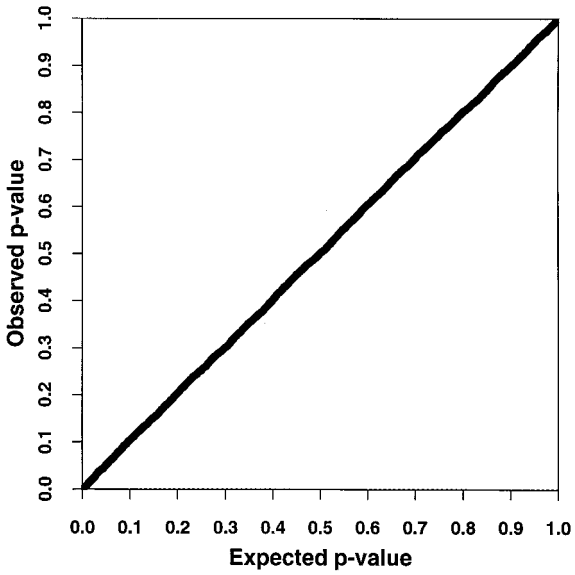
There are three issues regarding the analysis. First, for the most rigorous analysis of study results, one should use study center and other blocking factors in the analysis as one would do in any study. Second, the prognostic variables used in the adaptive randomization should be used as covariates. Ignoring the fact that biased coin randomization has been used for a continuous response variable one would use an analysis of covariance with a number of continuous covariates, blocking factors, and categorical covariates.

A serious question is how the biased coin part of the randomization should affect the analysis. Halpern and Brown (11) discussed this issue in detail and concluded that it is likely in many instances that the analysis need not be adjusted. Nevertheless, in the light of

modern computing power there seems to be an obligation to develop analytic methods adjusted for all of the features in the randomization process. Because of the complexity of the entire set of constraints, exact methods would seem to be difficult. Hence, the author has resorted to a randomization test.

The randomization test entails rerandomizing patients to treatment groups. The rerandomization is done with exactly the same constraints as were done in the original study including constraining the order of enrollment of the patients to be the same as in the original study. Everything is the same except the outcomes of tossing the biased coin. The rerandomization is repeated a few thousand times and the empirical distribution of the ANCOVA p-value testing treatment effect is compared with the nominal p-value obtained using analysis of covariance with all of the blocking factors and all of the prognostic variables used as covariates. The randomization test has also ben implemented in Splus.

For the sake of simplicity, analyses are sometimes presented without including all of the blocking factors and covariates, particularly when there is no evidence of a statistically significant effect for one or more covariates or blocking factors. Naturally, there should be reluctance to remove such variables from the ANCOVA model. Forsythe and Stitt (18) and Halpern and Brown (11) have shown that the alpha level of a test can be reduced (with consequent loss of statistical power) when randomization has been per-

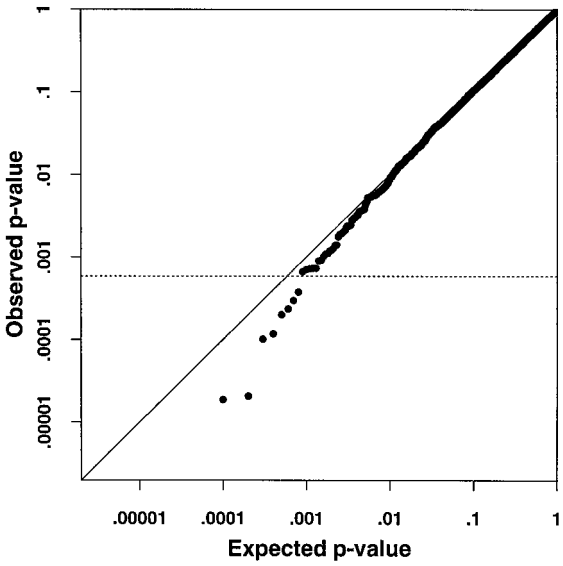



**Number of Samples = 10000.**
**Size of Each Sample = 71.**

**Proportion of p-values < 0.05 = 0.0487. Binominal test p = 0.5663.**
**Proportion of p-values < 0.01 = 0.0107. Binominal test p = 0.4814.**
**Proportion of p-values < 0.005 = 0.0053. Binominal test p = 0.6701.**
**Proportion of p-values < 0.001 = 0.0015. Binominal test p = 0.1503.**

**Nominal ANCOVA p = 0.0006 for primary efficacy test.**

**FIGURE 1. Observed versus expected p-values (sorted).**




**Number of Samples = 10000.**
**Size of Each Sample = 71.**

**Proportion of p-values < 0.05 = 0.0487. Binominal test p = 0.5663.**
**Proportion of p-values < 0.01 = 0.0107. Binominal test p = 0.4814.**
**Proportion of p-values < 0.005 = 0.0053. Binominal test p = 0.6701.**
**Proportion of p-values < 0.001 = 0.0015. Binominal test p = 0.1503.**

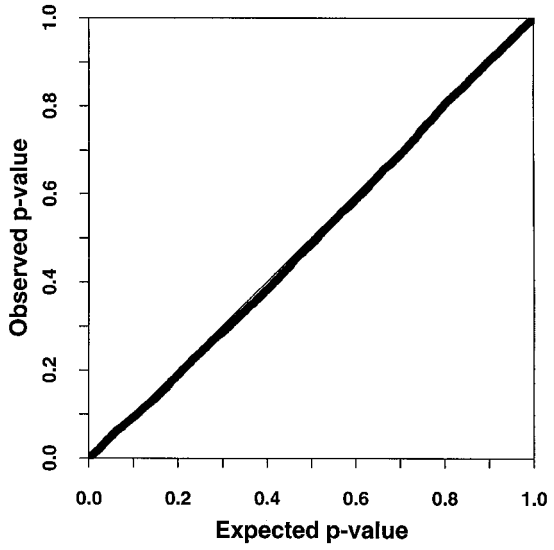
**Nominal ANCOVA p = 0.0006 for primary efficacy test.**

**FIGURE 2. Observed versus expected p-values (sorted).**

formed to balance on a baseline characteristic that is correlated with response but has not been included in the analysis of covariance or analysis of variance model. There would seem, however, to be no objection to elimination of nonstatistically significant effects from the ANCOVA model as long as the analysis is also performed using the complete set of covariates and constraints imposed by the adaptive randomization. For the reduced set of covariates and blocking factors, however, both the naive analysis (the usual ANCOVA) and the analysis with the randomization test should be performed.

It has been suggested by some that a randomization test could be performed assuming that the patients come in a random order, that is, doing the randomization test where the order of enrollment of the patients is rerandomized. While one feels obligated to "never say never," it would seem that this practice would require justification in each specific instance since patient characteristics at the time of enrollment may (and often do) vary over calendar time. In such circumstances, randomly permuting patient enrollment order would not constitute a plausible reexecution of the original study.

Estimating statistical power at the time a study is designed is a related issue. One really does not know the precise baseline characteristics of the subjects who will be recruited into the trial in question. One does not know the order in which the patients will be enrolled and randomized. Nevertheless, the same basic software can be used to compute




**Number of Samples = 6100.**
**Size of Each Sample = 98.**

**Proportion of p-values < 0.05 = 0.0503. Binominal test p = 0.9064.**
**Proportion of p-values < 0.01 = 0.0113. Binominal test p = 0.3026.**
**Proportion of p-values < 0.005 = 0.0056. Binominal test p = 0.5240.**
**Proportion of p-values < 0.001 = 0.0010. Binominal test p = 1.0000.**

**Nominal ANCOVA p = 0.0002 for primary efficacy test.**

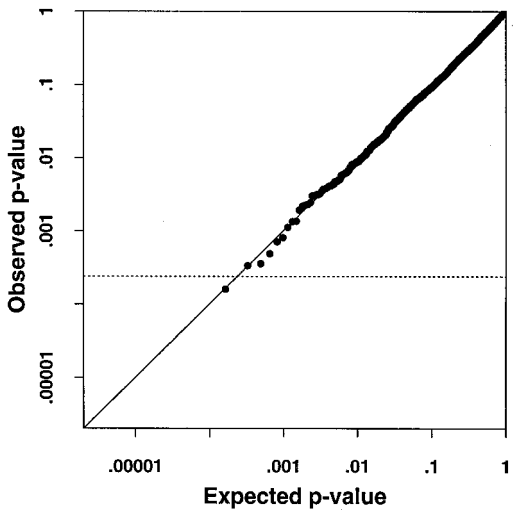
**FIGURE 3. Observed versus expected p-values (sorted).**

an estimate of power on the basis of the muli-variate distribution of the selected characteristics as seen in previous studies of similar patients.

## EXAMPLES

Two examples are presented here. The first example is taken from a study with 71 patients from 26 study centers. By design, two-thirds of the patients were randomized to treatment and one-third to control. The block size within study center was three. (The block size was small since the number of patients per study center was small.) The baseline characteristics for which balance was required included sex, etiology of disease (with three categories), age, and three other continuous covariates. The nominal analysis of covariance p-value using the blocking factor and all of the six continuous and categorical covariates was 0.0006. Figure 1 shows the q-q plot for the p-values obtained from the randomization distribution (n = 10,000) of the p-values from analysis of covariance, that is, from reassigning the patients to treatment via adaptive randomization and recomputing the analysis of covariance p-value for each rerandomization. This figure demonstrates the credibility of the nominal p-value from the analysis of covariance. Figure 2 is similar to Figure 1 except the logarithms of the p-values are used in order to emphasize the small values.

The second example is taken from a study with 98 patients from 20 study centers. Again, the number of patients per study center is




**Number of Samples = 6100.**
**Size of Each Sample = 98.**

**Proportion of p-values < 0.05 = 0.0503. Binominal test p = 0.9064.**
**Proportion of p-values < 0.01 = 0.0113. Binominal test p = 0.3026.**
**Proportion of p-values < 0.005 = 0.0056. Binominal test p = 0.5240.**
**Proportion of p-values < 0.001 = 0.0010. Binominal test p = 1.0000.**

**Nominal ANCOVA p = 0.0002 for primary efficacy test.**

**FIGURE 4. Observed versus expected p-values (sorted).**

small. Patients were randomized in approximately equal numbers to one of two treatments. The block size within each study center was four. There were seven baseline characteristics for which balance was established through adaptive randomization. Figures 3 and 4, like Figures 1 and 2, demonstrate the credibility of the nominal p-value from the analysis of covariance.

There was no statistically significant difference between treatment groups with respect to any baseline characteristic in either example. The number of baseline characteristics for which balance was established by adaptive randomization was not large in either example. Therneau (19) discusses simulation results from adaptive randomization accommodating up to 20 baseline characteristics for two treatment groups with 50 subjects in each group.

## SOME PRACTICAL CONSIDERATIONS

Performing a randomization test for the results from a study using a baseline adaptive randomization is likely to require considerable computing resources. Computations at Genentech are facilitated by the Medical Affairs Solaris system with eight processors. Thus, the rerandomizations were executed in parallel and the results were merged.

The rerandomizations discussed in the previous section were also executed using the UNIX **at** utility to begin processes at times of otherwise low computer use. In addition, some of these rerandomizations were executed during normal working hours using the UNIX **nice** utility to be sure that they did not interfere with other work.

Limitations were encountered regarding

the amount of computation that could be performed in one invocation of Splus. This limitation was circumvented by UNIX shell scripts and the UNIX **make** utility.

While the software for adaptive randomization is capable of handling any number of prognostic variables, there must be some practical limitation on the number of prognostic variables. In practice, the number of categorical prognostic variables seems to be more important than the number of continuous prognostic variables. The limitation becomes more serious as:

- The number of treatment groups gets larger,
- The number of categories in a prognostic factor increases, and
- The number of subjects in a given category gets smaller.

Thus, for example, it is difficult to achieve balance between treatment groups with respect to several etiologies of a disease when one or more etiologies is rare and there are four or five treatment groups.

Another point to keep in mind when performing a randomization test is that each possible random assignment is not equally likely. Thus, the number of rerandomizations necessary to execute the randomization test should be larger than would be the case when all possible randomizations are equally likely.

## CONCLUSIONS

Baseline adaptive randomization can be a valuable feature in small and moderate sized studies. It can also be valuable when subset analyses are required by protocol or even in larger studies when subset analyses are required later by a regulatory agency. Since care must be taken to be sure that patients satisfy inclusion/exclusion criteria for studies, the added effort to perform baseline adaptive randomization is frequently small in comparison with other efforts. It is likely that nominal analysis of covariance will provide a satisfactory analysis (without taking into consideration the constraints imposed by adaptive randomization). Given modern computers, however, randomization tests can easily be used to verify the nominal p-value from analysis of covariance.

*Acknowledgments*—The proposal for Genentech to use biased coin randomization was first made by Maria Koretz in 1982. The software for adaptive randomization was implemented in Splus primarily by Linfeng You. Peter Compton has been a primary mentor at Genentech for maturity in use of the methodology and for the details of the design and the implementation of the Splus software. The author is also indebted to Bradley Efron who devised the fundamental principles of biased coin randomization and to the others who have thoroughly thought through the implications.

## REFERENCES

1. Lord FM. A paradox in the interpretation of group comparisons. *Psychological Bull.* 1967;68:304–305.
2. Harville DA. Nearly optimal allocation of experimental units using observed covariate values. *Technometrics.* 1974;16:589–599.
3. Pocock SJ, Simon R. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics.* 1975;31:103–115.
4. Taves DR. Minimization: A new method of assigning patients to treatment and control groups. *Clin Pharmacol Therapeutics.* 1974;15:443–453.
5. Pocock SJ. Allocation of patients to treatments in clinical trials. *Biometrics.* 1979;35:183–197.
6. Simon R. The Consultant's Forum: Restricted randomization designs in clinical trials. *Biometrics.* 1979;35:503–512.
7. Begg CB, Iglewicz B. A treatment allocation procedure for sequential clinical trials. *Biometrics.* 1980; 36:81–90.
8. Atkinson AC. Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika.* 1982;69:61–67.
9. Smythe RT, Wei LJ. Significance tests with restricted randomization design. *Biometrika.* 1983;70:496–500.
10. Smith RL. Sequential treatment allocation using biased coin designs. *J R Stat Soc B.* 1984;46(3):519–543.
11. 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.* 1986;5:211–229.
12. Hanningan JF, Brown BW. Adaptive randomization biased coin-design: experience in a cooperative group clinical trial. Technical Report No. 74, Division of Biostatistics, Stanford University, Stanford, CA.
13. Hollander M, Pena E. Nonparametric test under restricted treatment-assignment rules. *J Am Stat Assoc.* 1988;83:1144–1151.

14. Begg CB. On inferences from Wei's biased coin design for clinical trails. *Biometrika.* 1990;77(3): 467–484.
15. Efron B. Forcing a sequential experiment to be balanced. *Biometrika.* 1971;58:403–417.
16. Efron B. Randomizing and balancing a complicated sequential experiment. In: *Biostatistics Casebook.* Miller Jr RG, Efron B, Brown Jr BW, Moses LE, eds. New York: Wiley; 1980:19–30.
17. Becker RA, Chambers JM, Wilks AR. *The New S Language: A Programming Environment for Data Analysis and Graphics*. Pacific Grove, California: Wadsworth & Brooks/Cole; 1988.
18. Forsythe AB, Stitt FJ. Randomization or minimization in the treatment assignment of patient trials: validity and power of tests. Technical Report 28, Health Sciences Computing Facility, University of California, Los Angeles; 1977.
19. Therneau T. How many stratification factors are "too many" to use in a randomization plan? *Control Clin Trials.* 1993;14:98–108.