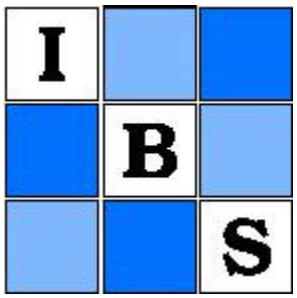

BIOMETRICS *31*, 103–115
March 1975

SEQUENTIAL TREATMENT ASSIGNMENT WITH BALANCING FOR PROGNOSTIC FACTORS IN THE CONTROLLED CLINICAL TRIAL

STUART J. POCOCK

Statistical Laboratory, SUNY at Buffalo, Amherst, New York 14226, U.S.A.

RICHARD SIMON

National Cancer Institute, Bethesda, Maryland 20014

SUMMARY

In controlled clinical trials there are usually several prognostic factors known or thought to influence the patient's ability to respond to treatment. Therefore, the method of sequential treatment assignment needs to be designed so that treatment balance is simultaneously achieved across all such patient factors. Traditional methods of restricted randomization such as "permuted blocks within strata" prove inadequate once the number of strata, or combinations of factor levels, approaches the sample size. A new general procedure for treatment assignment is described which concentrates on minimizing imbalance in the distributions of treatment numbers within the levels of each individual prognostic factor. The improved treatment balance obtained by this approach is explored using simulation for a simple model of a clinical trial. Further discussion centers on the selection, predictability and practicability of such a procedure.

1. INTRODUCTION

In any clinical trial patients will become available for treatment sequentially. If several treatments are to be compared, one must ensure that no bias is incurred by any subjective allocation of treatments to patients. The simplest approach is to use a randomization procedure whereby each patient has an equal probability $1/N$ of receiving any one of the N treatments. For considerations of statistical power it is desirable to have the treatment groups of equal size, and for most clinical trials it is important to balance across the treatments the effects of prognostic factors which are thought to influence the effect of treatment on the patient. For these reasons, various "stratified randomization procedures" have been previously developed.

One method of ensuring equal numbers of patients is to use permuted blocks of treatments. In this design patients are formed in blocks of size $b \times N$ as they enter the trial and a random permutation of treatments, b of each, is determined for each block. This ensures that equal treatment numbers are achieved after every block of patients. Efron [1971] pointed out that such a procedure means that after $bN - 1$ treatments have been assigned in a block the last one is pre-determined. This can be a considerable source of bias if $b \times N$ is small and any one investigator has complete knowledge of treatment assignments. To control this bias Efron suggests an alternative "biased coin" design.

Though the loss in power resulting from treatment groups of unequal size is not likely to compromise the trial, a serious imbalance among the treatment groups with regard to a factor of prognostic importance can have severe consequences. For example, in a trial of chemotherapeutic agents for advanced breast cancer patients the presence or absence of

liver metastases, the patient's age and time from mastectomy to first recurrence of disease all have an effect on the patient's ability to respond to treatment. Patients with liver metastases have an extremely poor prognosis and if one treatment has more patients with liver metastases than another, considerable bias could result in any comparison of treatments. It may be possible to remove such bias by introducing liver metastases as a covariate in analysis, but the number of variables which can be treated in this manner is limited and a severe imbalance will cause a confounding of the effects of treatment and a prognostic factor. Also, in small studies it is not always clear what covariate model is appropriate. Though covariate analysis could be profitably used in many cases, it is a wide-spread practice even in large clinical trials to rely solely on "comparability" of the treatment groups as a basis for inference.

The usual method of achieving balance with respect to prognostic factors is to divide each factor into several levels and to consider treatment assignment separately for patients having each particular combination of such factor levels. Such groups of patients are commonly referred to as randomization groups or *strata*. Treatment assignment is performed entirely separately for each stratum, a permuted block design of the type mentioned above being used. In fact, using purely random treatment assignment for each stratum is equivalent to simple random assignment, so that some equalization of treatment numbers within each stratum is essential. This whole procedure is analogous to performing a factorial experiment, without being able to control the factor levels of the experimental units.

The main difficulty in this approach is that the number of strata increases rapidly as the number of prognostic factors increases. For example, with four prognostic factors each with three levels the number of strata equals $3 \times 3 \times 3 \times 3 = 81$. Not only is it unwieldly to run such a trial with 81 strata, but if the trial requires only 100 or so patients some strata will probably contain no patients and many more will contain only one. In such a situation the whole procedure may fail to achieve its basic aim, since in most strata the first permuted block of treatments will be only partially assigned and considerable imbalance across treatments for any factor could still exist. This point will be referred to again in the example of section 4.

Sedransk [1974] has suggested an improvement on the "permuted blocks within stratum" design whereby the permutations used in all the strata are associated with one another in a specified manner derived from consideration of fractional factorial designs. The procedure is most effective as compared with permuted blocks if the number of strata is somewhat less than the number of patients and the patients are very evenly distributed across strata. However, the latter condition is rarely met in actual clinical trials.

The "permuted block within stratum" design severely restricts the number of prognostic factors which can be balanced across treatments in a clinical trial. However, in almost any trial the medical investigators are aware of many such factors which they wish to allow for, this being especially true as methods of analysis become more sophisticated. The problem becomes even more severe when one considers the increasing tendency towards multicenter (or cooperative) trials. It is desirable that each medical institution in such a trial should have equal numbers of patients on each treatment, both to balance out any "institution effect" and also to keep the institutions more interested in the trial. If "institution" were to be introduced as a further prognostic factor, possibly with 30 or more levels, the total number of strata may then be in the hundreds and one would have achieved little more than purely random treatment assignment.

The procedures described above are often unable to balance simultaneously across treatments the effects of several prognostic factors. The remainder of this paper describes

an alternative approach to this problem. The method to be presented is motivated by the feeling that any real improvement in the effectiveness of a balanced design when the number of prognostic factors is large cannot result from techniques based on pre-specified within-stratum treatment assignments.

2. DEFINITION OF THE GENERAL METHOD

The method will first be described in the general case.

Consider a clinical trial in which patients enter sequentially, each in turn requiring a treatment assignment. Suppose there are N treatments and M prognostic factors for which treatment balance is required, the number of levels of these factors being $n_1, n_2, \cdots, n_M$.

The procedure can best be described by considering an arbitrary point during the trial. At such a point let x_{ijk} be the number of patients with level j of factor i who have been assigned treatment k for $j = 1, 2, \cdots, n_i$; $i = 1, 2, \cdots, M$ and $k = 1, \cdots, N$.

Consider the next patient entering the trial. Let $r_1, \cdots, r_M$ be the levels of factors $1, \cdots, M$ for this patient. The choice of treatment for this new patient is determined in the following manner.

For each treatment k one considers the new $\{x_{ijl}\}$, denoted $\{x_{ijl}{}^k\}$, that would arise if that treatment were assigned to the patient. This is achieved by adding one onto x_{ir_ik} for $i = 1, \cdots, M$. That is, define

$$x_{ijl}{}^k = x_{ijl} \quad \text{for} \quad j \neq r_i \quad \text{or} \quad l \neq k$$

$$x_{ir_ik}{}^k = x_{ir_ik} + 1$$

where the superscript k refers to the particular treatment under consideration. Let $D(\{z_l\}_{l=1}^N)$ be some function which measures the "amount of variation" in any set of non-negative integers $\{z_l\}_{l=1}^N$. Then, if treatment k were assigned to the next patient $d_{ik} = D(\{x_{ir_il}{}^k\}_{l=1}^N)$ would be the resultant *"lack of balance"* among treatment assignments for patients with level r_i of factor i.

Let $G_k = G(d_{1k}, \cdots, d_{Mk})$ where G is some function from $\mathbf{R}^M \to \mathbf{R}$ which combines the d_{ik}. Then G_k represents the *"total amount of imbalance"* in treatment numbers which would exist at all the factor levels of the new patient if treatment k were assigned to that patient.

One can rank treatments according to their values G_k, treatment (1) having minimal G_k, treatment (2) having the next smallest G_k, etc., so that $(s) < (t)$ iff $G_{(s)} \leq G_{(t)}$. In the case of ties a random ordering can be determined.

The assigned treatment T can be determined from the following set of probabilities:

$$\text{prob } (T = (k)) = p_k \quad \text{where} \quad p_1 \geq p_2 \geq \cdots \geq p_N \quad \text{and} \quad \sum p_k = 1.$$

The values of p_k can be fixed constants or functions of $\{G_{(k)}\}$. This ordering of probabilities means that treatments with small values of G_k have a higher probability of being chosen.

The entire procedure is repeated when the next patient enters the trial.

In this section the method has been described in as general a way as possible. In the next section appropriate examples are given for the functions D, G and $\{p_k\}$ defined above. A particular numerical example is then described to illustrate the method.

3. THE SELECTION OF A PARTICULAR PROCEDURE

3.1 The Choice of D

As mentioned in section 2, $D(\{z_l\}_{l=1}^{N})$ measures the "amount of variation" in any set of non-negative integers $\{z_l\}_{l=1}^{N}$. Four possible formulae for D are considered here:

a) *The Standard Deviation* or *Variance* of $\{z_l\}$.

b) *The Range* of $\{z_l\}$ is a simpler measure. If one is essentially interested in comparing pairs of treatments in analysis it may be preferable since D would then be measuring the most imbalance in any pair. Also, with only two treatments the range ($= |z_1 - z_2|$) is equivalent to the standard deviation.

c) *An Upper Limit of Acceptable Treatment Imbalance* could be defined for each level of each factor. This limit could depend on the factor, but consider the case where it is a constant U. Then,

$$D = \begin{cases} 0 & \text{if range of } \{z_l\} \leq U, \\ 1 & \text{if range of } \{z_l\} > U. \end{cases}$$

This measure only scores imbalance once it is greater than some specified amount. This may be a desirable feature since it means that a simple randomization is followed except when substantial imbalance would be introduced. That is, one introduces a restriction on treatment allocation when "things start to go wrong" for one or more prognostic factors. This is because if $d_{ik} = 0$ for all i and k then the ranking (1) to (N) is determined randomly and each treatment will then have an equal probability of being assigned.

d) *A Sign Rule* can be used in the case of two treatments. This needs to be defined in terms of the actual $d_{i,k}$ rather than defining a general function D. Thus,

$$d_{i1} = \begin{cases} 1 & \text{if } x_{ir_i1} > x_{ir_i2} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad d_{i2} = \begin{cases} 1 & \text{if } x_{ir_i2} > x_{ir_i2} \\ 0 & \text{otherwise} \end{cases}$$

That is, if at level r_i of factor i treatment 1 has more patients than treatment 2 this works against the assignment of treatment 1, and vice versa.

3.2 The Choice of G

One reasonable way of combining $\{d_{ik}\}_{i=1}^{M}$ is to sum them. That is,

$$G_k = G(d_{ik}, \cdots, d_{Mk}) = \sum_{i=1}^{M} d_{ik}.$$

However, the situation may arise where some prognostic factors are considered more important than others. One can then make G_k a weighted sum of $\{d_{ik}\}$ so that

$$G_k = G(d_{ik}, \cdots, d_{Mk}) = \sum_{i=1}^{M} w_i d_{ik}$$

where $\{w_i\}$ are constants.

3.3 The Choice of $\{p_k\}$

The formula for $\{p_k\}$ determines the extent to which one wishes to bias treatment assignment in favor of those treatments with small G_k. The following types of formula are suggested:

a) Let $p_1 = p$ and $p_k = (1 - p)/(N - 1)$ for $k = 2, \cdots, N$ where p is some constant which must be greater than $1/N$ for the bias to be in the right direction. If p were equal to $1/N$ then each treatment would have an equal probability of being selected. If p were

equal to 1 then treatment (1) would be automatically assigned. This would mean that for any particular order of patients, treatment assignment would be deterministic except when treatments had equal G_k. This may make the procedure too predictable by the investigators entering patients but would also achieve the least imbalance for prognostic factors across treatments. The choice of an actual value for p is discussed in section 5.3.

In the case of more than two treatments, this method gives all but the "preferred" treatment an equal chance of being assigned.

b) A formula which would take into account the complete ranking of $\{G_k\}$ is

$$p_k = q - \frac{2(Nq - 1)}{N(N + 1)} k \quad \text{for} \quad k = 1, \cdots, N$$

where q is some constant. It can be easily shown that $\sum p_k = 1$ and q must lie between $1/N$ and $2/(N - 1)$. The larger the value of q, the more bias is used in the treatment assignment. If $N = 2$, this method is identical to a) above. As an example, let $N = 4$ and $q = 1/2$. Then $p_1 = 0.4$, $p_2 = 0.3$, $p_3 = 0.2$ and $p_4 = 0.1$.

c) One can define other formulae for $\{p_k\}$ which depend not only on the ranking (1) to (N) but on the values $\{G_{(k)}\}$. For example, let

$$p_k = \frac{1}{N - t}\left[1 - \frac{tG_{(k)}}{\sum G_{(k)}}\right] \quad \text{for} \quad k = 1, \cdots, N$$

where t is a constant between 0 and 1. However, such a formula would seem to be an unnecessarily complicated way of biasing treatment assignment.

3.4 An Illustrative Example

Consider a clinical trial with three treatments. Suppose there are three prognostic factors for which treatment balance is required, these factors having two, two, and three levels.

Consider the application of a particular procedure of the type defined in sections 2 to 3.3 which is designed to achieve the required balance. Let D be defined by the range (section 3.1b), let G_k be a weighted sum of $\{d_{ik}\}$ weights being 2, 1, and 1, respectively, and let $\{p_k\}$ be defined as in section 3.3a with $p = 2/3$ (i.e. $p_1 = 2/3$, $p_2 = 1/6$, $p_3 = 1/6$).

For the first patient G_k is the same for all k and therefore a treatment is assigned at random. The technique is best illustrated by taking an arbitrary point in the trial. Suppose 50 patients have been entered into the trial and that their distributions by treatment, levels of each prognostic factor, and the combination of treatment and factor level are as shown in Table 1. Suppose the next patient is at levels 1, 2, and 2, respectively, for the three factors. The problem is to determine to which treatment that patient should be assigned.

First, consider the results of assigning treatment 1:
For factor 1 level 1, treatment numbers would then be 10, 10, 9 ie range $(=d_{11}) = 1$
For factor 2 level 2, treatment numbers would then be 10, 11, 9 ie range $(=d_{21}) = 2$
For factor 3 level 2, treatment numbers would then be 5, 5, 3 ie range $(=d_{31}) = 2$
Therefore, $G_1 = 2d_{11} + d_{21} + d_{31} = 2 + 2 + 2 = 6$. This measures treatment imbalance at the appropriate factor levels if treatment 1 were assigned. Assigning treatment 2 would result in $d_{12} = 2$, $d_{22} = 3$, $d_{32} = 3$ so that $G_2 = 10$ and assigning treatment 3 would result in $d_{13} = 1$, $d_{23} = 2$, and $d_{33} = 1$ so that $G_3 = 5$.

G_3 is the smallest of $\{G_k\}$. Therefore, treatment 3 is assigned to the 51st patients with probability 2/3, treatments 1 and 2 being assigned with probability 1/6 each. If p were equal to 1 instead of 2/3, then treatment 3 would automatically be assigned.

TABLE 1
50 PATIENTS DISTRIBUTED BY TREATMENT AND FACTOR LEVELS

	Factor	1		2		3			TOTAL
	Level	1	2	1	2	1	2	3	
Treatment									
1		9	8	8	9	8	4	5	17
2		10	7	6	11	8	5	4	17
3		9	7	7	9	8	3	5	16
Total		28	22	21	29	24	12	14	50

This example is much simpler than many actual clinical trials, as there would only be 12 strata if the usual "permuted blocks with stratum" technique were adopted. However, the technique described could be used for any number of patients, treatments, and prognostic factors.

4. COMPARATIVE EVALUATION OF NEW AND TRADITIONAL METHODS

The statistical properties of the procedures described in sections 2 and 3 need to be explored in order to establish their effectiveness and appropriateness in actual clinical trials. For the general situation any theoretical evaluation using probability theory would appear to be so complex as to be virtually impossible. In fact, even with a simplified model of a trial, an analytical approach becomes extremely unwieldly except in the most trivial situation. Therefore, the use of simulation based on a particular model for patient entry into a trial would seem the most feasible way of comparing the properties of new and traditional methods of treatment assignment.

4.1 A Model for a Clinical Trial

The model for a trial needs to be general enough to give some insight into the methods of treatment assignment under a variety of circumstances, but specific enough so that one is not swamped by an excessive number of model parameters. The model chosen is as follows:

a) The number of patients entered into the trial, N, is fixed. $N = 50$ is arbitrarily chosen for the simulations. This should be large enough to illustrate the general properties of the methods but small enough to allow an adequate number of computer simulations.

b) There are two treatments. This is the most common situation in controlled clinical trials. However, it seems reasonable to suppose that conclusions drawn from this model about the relative merits of the various methods of treatment assignment can be generalized to trials with more than two treatments.

c) There are M prognostic factors, each at two levels, for which balanced treatment numbers are desired. The probability of a patient being in level 1 of a factor equals 1/2 and there is no association between factors. In an actual trial, prognostic factors are not usually split equally between their two levels, and factors are usually associated with one another to some degree. However, to introduce variable level probabilities and measures of factor association would complicate the model unnecessarily, as would consideration of factors with more than two levels.

d) Patients enter the trial sequentially in purely random fashion, the factor levels of any one patient being independent of those for any other. Random patient entry may not be realistic in some trials, but there are so many ways in which patient entry could deviate from randomness that such alternatives cannot be considered here.

This model has the advantage that all M prognostic factors are treated identically so that their results can be combined in simulations. This symmetry means that M, the number of prognostic factors, is the only variable quantity in the model for patient entry into the trial.

4.2 Methods of Treatment Assignment

The purpose of this simulation example is to compare the traditional methods of randomization mentioned in section 1 with one example of the new method of treatment assignment defined in sections 2 and 3. These methods can be briefly summarized as follows:

(a) *Purely Random Assignment*

Treatment 1 is assigned to a patient with probability 1/2, assignments being made independently for each patient. Therefore, the distribution of patient numbers by treatment and level of an arbitrary factor can be illustrated diagramatically as follows:

		Factor Level	
		1	2
Treatment	1	X_{11}	X_{12}
	2	X_{21}	X_{22}

N

X_{11} , X_{12} , X_{21} , X_{22} have a joint Multinomial distribution with index N and equal probabilities.

(b) *Random Permuted Blocks*

Permuted blocks of treatment assignments of size 2b are often used to ensure equal treatment numbers after every 2b additional patients. The order of treatments in each block of 2b is randomly determined. If N is divisible by 2b then the completed trial will have equal treatment groups. In this situation X_{11} and X_{21} will be independent random variables each with a binomial distribution index $N/2$ probability 1/2.

For both allocation methods (a) and (b) the distribution of treatment numbers within any level of any factor does not depend on the number of factors M, since these are ignored when making treatment assignments. This is no longer true for (c) and (d).

(c) *Permuted Blocks within Strata*

Random permuted blocks of treatment assignments of size 2b are used for each of the 2^M strata. The probability distributions for X_{11} , etc., for any factor can be defined analytically but become very cumbersome even for quite small values of M, so that it becomes more feasible to use simulation to estimate these distributions. b has been set equal to 1 in the results shown in section 4.4, but it can easily be shown that treatment balance deteriorates to that achieved for (a) as b increases.

(d) *The New Procedure*

The general procedure defined in section 2 is to be illustrated using an example similar to that in section 3.4 applied to the model of section 4.1. Thus, D is defined by the range

and this time G_k is the *unweighted* sum of $\{d_{ik}\}$. $\{p_k\}$ is defined as in section 3.3a and values 1 and 3/4 are considered for p. Again, simulation is necessary to estimate the distribution of X_{ij} .

4.3 Measures of Treatment Imbalance

The purpose of methods 4.2c and d is to achieve a pattern of treatment assignments in a trial which is balanced for the various prognostic factors under consideration. The word "imbalance" has been used rather loosely so far and is now defined more precisely.

Choose an arbitrary prognostic factor and define X_{kj} as the number of patients on treatment, k, at level, j, of this factor, there being two treatment and two factor levels (see section 4.1). Let $q_1 = X_{11}/(X_{11} + X_{21})$ and $q_2 = X_{12}/(X_{12} + X_{22})$, these being the proportions in each factor level on treatment 1. Then, the *amount of treatment imbalance for that factor* is defined as $|q_1 - q_2|$.

This definition is quite reasonable on intuitive grounds, and the following additive model provides an interpretation of $|q_1 - q_2|$. Let y be some measure of response to treatment for a patient, and suppose y depends on treatment, k, and factor level, j, as follows:

$$y = c + t_k + l_j + \epsilon$$

where c is a constant, t_k and l_j are treatment and factor level effects and ϵ is a random variable with zero mean. Then, if $\bar{y}_1$ and $\bar{y}_2$ are the observed means of y for patients on treatments 1 and 2 in a trial, it can easily be shown that Expectation $(\bar{y}_1 - \bar{y}_2) = (t_1 - t_2) + (q_1 - q_2)(l_1 - l_2)$. If the trial were perfectly balanced $\bar{y}_1 - \bar{y}_2$ should be an unbiased estimate of $t_1 - t_2$. The bias in this estimate is therefore proportional to $|q_1 - q_2|$.

One could consider more complex forms of imbalance induced by interactions between factors. Method 4.2c will allow for such interactions provided one has a substantial number of patients per stratum and method 4.2d can be adapted to do so (see section 5.2).

Another form of imbalance in a trial is to have unequal treatment numbers overall. No bias can be incurred by this, but as such inequality increases, the precision of estimated treatment effects will deteriorate. If N_1 and N_2 are the total numbers on the two treatments, then $|N_1 - N_2|$ is a measure of total treatment imbalance.

4.4 The Results of Simulation

The purpose of this section is to present results which provide a comparison of the methods of treatment assignment defined in section 4.2. This is based on the model of section 4.1 using the measures of treatment imbalance of section 4.3. The total size of trial to be used is fixed at 50 patients.

Consider $|q_1 - q_2|$, the amount of treatment imbalance for a factor. For purely random assignment (4.2a) and random permuted blocks (4.2b) the probability distribution of $|q_1 - q_2|$ can be calculated exactly using properties of the binomial distribution, and the right-handed cumulative distribution functions grouped in intervals of 0.02 for $|q_1 - q_2|$ are shown in Figure 1(a). The two distributions are very similar and since both methods ignore prognostic factors there is a high probability of incurring substantial imbalance.

For the "permuted blocks within stratum" method (4.2c) the probability distribution of $|q_1 - q_2|$ has been estimated by 100 computer simulations of a trial for each number of prognostic factors from $M = 1$ to 8 (Figure 1b). Because of the symmetry of the problem, when $M \geq 2$ the results for all M prognostic factors have been used in estimation. It is clearly seen that as the number of prognostic factors is increased the ability of the method to achieve treatment balance for a factor deteriorates. In fact for $M = 8$ the distribution

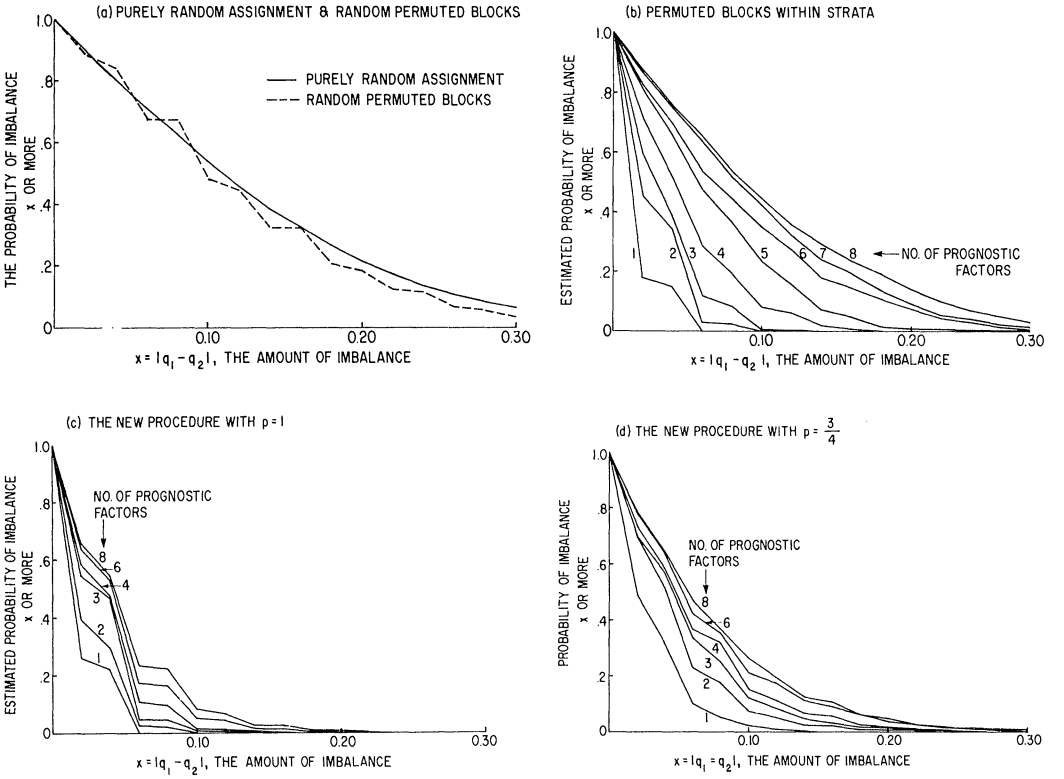

FIGURE 1

PROBABILITY DISTRIBUTIONS OF THE AMOUNT OF TREATMENT IMBALANCE FOR A FACTOR

of $|q_1 - q_2|$ is similar to that achieved for purely random assignment. The major reason for this deterioration is that once M is greater than 4 a large number of the 2^M strata contain 0 or 1 patients so that a large proportion of treatments are assigned randomly.

For the new procedure (4.2d) the distribution of $|q_1 - q_2|$ is again estimated by simulation for $M = 1$ to 8 and the results are shown in Figures 1(c) and 1(d) for $p = 1$ and 3/4, respectively. The graphs for $M = 5$ and 7 are deleted for clarity of presentation. For $p = 1$ treatment balance for a factor is achieved quite effectively, even with a substantial number of prognostic factors.

By setting $p = 3/4$ one allows a purely random treatment assignment to be made with probability 1/2. This lack of restriction means that with probability 1/4 the choice of treatment is not optimal as determined by $\min_k \{G_k\}$, and as a consequence there is an increased risk of treatment imbalance as compared with $p = 1$. In practice, one has to select that p which strikes a balance between minimizing $|q_1 - q_2|$ and avoiding predictability of treatment assignment (see section 5.3).

Now, consider the probability distribution of $|N_1 - N_2|$, the overall imbalance in treatment numbers for methods 4.2(a) to (d). Exactly equal treatment numbers are achieved for "random permuted blocks". Figure 2(a) shows the right handed c.d.f. of $|N_1 - N_2|$ for "purely random assignment". Figures 2(b), 2(c) and 2(d) show the estimated distributions for "permuted blocks within strata" and the new procedure with $p = 1$ and 3/4, all for $M = 1$ to 8. Again, some curves are omitted for clarity of presentation.

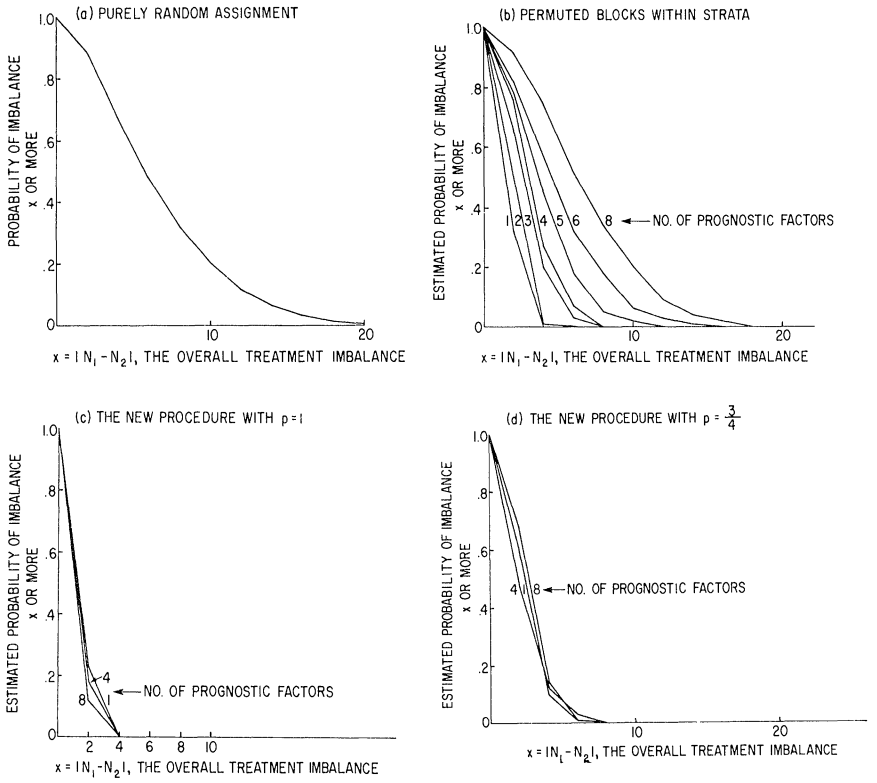

FIGURE 2

PROBABILITY DISTRIBUTIONS OF OVERALL TREATMENT IMBALANCE

For purely random assignment there is some risk that a substantial difference will arise in the overall treatment numbers. For "permuted blocks within strata" the situation deteriorates from the exact equality of "random permuted blocks" as M, the number of prognostic factors, increases. Indeed, for $M = 8$ one achieves almost the same results as for purely random assignment. For the new procedure, both for $p = 1$ and $p = 3/4$, only very small overall treatment differences are likely to arise even as M becomes quite large.

This simulation example has served to illustrate that new procedures for treatment assignment of the type defined in section 2 do result in a more balanced trial than traditional methods, according to the criteria of section 4.3. Though some of the restrictions imposed on the model of section 4.1 cannot be assumed in many actual trials it seems very plausible that these favorable results should apply to all trials with several prognostic factors.

5. FURTHER DISCUSSION OF THE METHOD

In this section the properties of the new method of treatment assignment defined in section 2 are further explored.

5.1 An Actual Trial

The following example is a clinical trial which is currently being carried out by the Eastern Cooperative Oncology Group. Patients with advanced histiocytic lymphomas

were to be entered onto the trial and the three treatments were: 1. Cytoxan + Prednisone, 2. Cytoxan + Prednisone + Vincristine, and 3. Cytoxan + Prednisone + Vincristine + BCNU. There were five prognostic factors for which treatment balance was desired:

Cell Type	Cell Pattern	Prior Radiotherapy	Prior Chemotherapy	Disease Stage
1. Histocytic	1. Diffuse	1. Yes	1. Yes	1. Stage III
2. Mixed	2. Nodular	2. No	2. No	2. Stage IV
3. Unknown	3. Unknown			

There are other factors which are known to influence the patient's ability to respond to treatment but if a "permuted blocks within stratum" design is used the number of factors could not be further increased. Even so, there were $3^2 \times 2^3 = 72$ strata in the experimental design. After a year of patient accrual, 73 patients had been assigned a treatment. These fell into 23 strata leaving 49 strata completely empty. Only 2 strata contained an exact number of permuted blocks, so that 21 strata had unequal treatment numbers. Patients were entered from 19 medical institutions, but institution could not be incorporated in the permuted block design since there would then be $72 \times 19 = 1368$ strata.

It is clear that a "permuted blocks within stratum" design is somewhat inadequate in such a trial, and a more satisfactory arrangement would be to use a method of the type mentioned in section 2. Institution could be included as a further prognostic factor and, as illustrated by the simulation results of section 4, more factors could be added without any rapid deterioration in treatment balance.

5.2 Allowance for Interactions

It may happen that the interaction between two or more prognostic factors may affect response to treatment. For example, in the trial of section 5.1 it could be that the effects of prior radiotherapy and chemotherapy are not additive and that patients with neither therapy respond much better than all others. It is evident that in most clinical trials one cannot reliably achieve treatment balance for all factor interactions of any order, though "permuted blocks within strata" would achieve this with a large sample size per stratum. In the new method of section 2 one can allow for a first order interaction between two factors by replacing them by one factor whose levels consist of all combinations of levels of the original factors. For example, prior radiotherapy and/or chemotherapy would have four levels: 1. yes, yes 2. yes, no 3. no, yes and 4. no, no. One should base one's choice of interactions to be included on experience gained in previous trials and on consideration of the model to be used in statistical analysis.

5.3 Predictability of the Method

As mentioned in section 1, one major consideration in all treatment assignment procedures is to ensure that the investigator entering a patient onto the trial cannot predict in advance what the treatment will be, since this might influence his decision to enter the patient. It is for this reason that in section 2 the optimal treatment as determined by $\min_k \{G_k\}$ is not necessarily the automatic choice. Instead, 'biased-coin' procedures of the type introduced by Efron [1971] are defined by $\{p_k\}$ in section 3.3. In fact, the new method described in this paper can be considered a generalization of Efron's approach to more than two treatments and several prognostic factors.

For any individual trial, there is no obvious decision rule for optimizing one's choice of $\{p_k\}$. If $p_1 = 1$ the chance of treatment imbalance is minimized but predictability is at a maximum, whereas as p_1 tends to 1/(No. of Treatments) the reverse is true. Efron's choice of $p_1 = 2/3$ for two treatments may be too small if the trial is small and has several prognostic factors, but in a large trial one could reduce p_1 still further.

For certain types of trial it may be quite feasible to set $p_1 = 1$. For instance, consider the example of section 5.1. In any such multi-institution trial no single investigator is aware of treatment assignments at other institutions and consequently it is impossible for him to determine for his next patient that treatment k which minimizes G_k . If his institution has fewer patients on one particular treatment he may guess that one will be chosen, but with several other prognostic factors contributing to $\{G_k\}$ he would be wrong in a high proportion of cases. Even for small single institution clinical trials most investigators would not be aware of the detailed rules determining treatment assignment so that the possibility of treatment prediction would be remote, and $p_1 = 1$ might be used.

One further problem to be resolved is the possibility of any "accidental bias" arising if patients enter the trial in a nonrandom order. If the nonrandom entry is only in terms of prognostic factors the method will produce a balanced experiment. However, if the order of patients is associated with some unknown nuisance factor which affects patient response the problem is slightly different. No general principles have been determined in this respect but it is assumed that the results for Efron's method can be generalized, so that the association between treatment and the nuisance factor has a similar expectation to that for purely random treatment assignment.

5.4 Practical Implementation of the Method

In using the methods described in this paper, one problem is the amount of calculation and centralized data storage required to determine each treatment assignment but this is counterbalanced against the many randomization lists required for "permuted blocks within strata". However, it should be feasible to program a small computer to perform all necessary calculations so that patients could be supplied with treatment assignments within seconds.

The method would appear to be particularly applicable to multi-center clinical trials (see section 5.1). In such circumstances the "closed envelope" technique for assigning treatment could not be used and telephone assignments through a central registration office would be most appropriate.

5.5 A General Assessment

This paper has described a new procedure for assigning treatments in clinical trials. Its major advantage is that it enables treatments to be balanced across several prognostic factors more effectively than traditional methods such as "permuted blocks within strata". The procedure is especially useful in small trials (e.g., less than 100 patients) as has been shown in the simulation example of section 4. The method is currently being evaluated for realistic models of factor dependence, and for large clinical trials with many factors. A time-sharing computer program has been developed which permits the procedure to be conveniently used in clinical trials at the National Cancer Institute. The full implications of the method cannot be appreciated until it has been further used in actual clinical trials, but it may eventually be able to replace more traditional procedures.

ACKNOWLEDGMENTS

This research was supported in part by a grant from the National Cancer Institute, CA-10810. We wish to thank the referees for useful comments.

TRAITEMENT EN SÉQUENTIEL ÉQUILIBRANT LES FACTEURS PRONOSTIQUES DANS L'ESSAI CLINIQUE CONTRÔLÉ

RESUME

Dans les essais cliniques contrôlés il y a habituellement plusieurs facteurs de pronostic, connus ou dont on pense qu'ils peuvent influencer la capacité de réponse du malade au traitement. C'est pourquoi il est nécessaire de planifier la méthode d'attribution du traitement séquentiel de telle sorte que l'équilibre du traitement soit réalisé en tenant compte simultanément de tous les facteurs concernant le malade. Les méthodes traditionnelles de randomisation restreinte, telle que les "blocs permutés à l'intérieur des strates" ne conviennent pas quand le nombre de strates ou de combinaisons de niveaux de facteurs atteint l'effectif de l'échantillon. On décrit un nouveau procédé général pour l'attribution du traitement, qui cherche à minimiser le déséquilibre entre les distributions de nombres de traitement à l'intérieur des niveaux de chaque facteur de pronostic. L'équilibre amélioré des traitements, obtenu par cette approche est exploré en utilisant la simulation pour un modèle simple d'essai clinique. Il est discuté également du choix de la prédictibilité et de la praticabilité d'un tel procédé.

REFERENCES

Efron, B. [1971]. Forcing a sequential experiment to be balanced. *Biometrika 58*, 403–17.
Sedransk, N. [1974]. Optimal stratification in clinical trials. In preparation.

Received November 1973

Key Words: Clinical Trials; Treatment assignment; Randomization; Stratification.