How many Stratification Factors is "Too Many" to Use in a Randomization Plan?

Terry M. Therneau, PhD

Abstract

The issue of stratification and its role in patient assignment has generated much discussion, mostly focused on its importance to a study [1,2] or lack thereof [3,4]. This report focuses on a much narrower problem: assuming that stratified assignment is desired, how many factors can be accomodated? This is investigated for two methods of balanced patient assignment, the first is based on the minimization method of Taves [5] and the second on the commonly used method of stratified assignment [6,7]. Simulation results show that the former method can accomodate a large number of factors (10-20) without difficulty, but that the latter begins to fail if the total number of distinct combinations of factor levels is greater than approximately $n/2$. The two methods are related to a linear discriminant model, which helps to explain the results.

1 Introduction

This work arose out of my involvement with lung cancer studies within the North Central Cancer Treatment Group. These studies are small, typically 100 to 200 patients, and there are several important prognostic factors on which we wish to balance. Some of the factors, such as performance or Karnofsky score, are likely to contribute a larger survival impact than the treatment under study; hazards ratios of 3–4 are not uncommon. In this setting it becomes very difficult to interpret study results if any of the factors are significantly out of balance.

With these small samples, however, there is concern that including too many stratification factors will place us in the situation of the old proverb that "He that attempts to please everyone pleases no one," i.e., that in trying to balance on all of the factors there may be balance on none. To investigate this simulations were performed using two methods of treatment assignment, dynamic allocation and stratified assignment. The imbalance of a single factor was examined in cases where the chosen proceedure is attempting to balance on the relevant factor concurrently with from 0 to 11 others.

It must be pointed out that the issue of stratification is a controversial one. Peto et. al. [3] hold the view that it is unnecessary, especially in large trials where complete randomization is fairly efficient, and also because retrospective stratification by means of an analysis involving covariates takes care of possible imbalances. The opposing view, articulated in Begg and Iglewicz [2], among others, is that balancing provides a more efficient comparison of treatments for trials of the typical size and, more important, that trials in which the

prognostic factors are well balanced are far more convincing to a scientific audience than sophisticated covariate analysis alone. This report addresses a more narrow problem: assuming that balance is desired, how many factors can be managed?

2 Methods

The North Central Cancer Treatment Group and the Mayo Comprehensive Cancer Center use a method of dynamic allocation to assign patients. This is based on a technique found in Pocock and Simon [7] and in Taves [5]. Minor changes have been made that allow easier computation by the randomization personnel. The numer k of categorical randomization variables or *factors* on which balance is desired usually ranges from 1 to 5, a factor commonly has between 2 and 4 levels. The use of the word *allocation* rather than *randomization* is purposeful: with this and other minimization techniques the treatment assignment of nearly every patient, given the prior assignments, is fully determined. A coin or other randomizing device is rarely used.

As an example, with three factors the method works as follows: A new subject arrives with values for the three factors of $F_1 = 1$, $F_2 = 0$, and $F_3 = 0$. A table like the one shown in Table 1 is constructed and filled in with the number of patients who fall into each category. The row labels of the table are based on the new patient's characteristics, the counts within it are based on the patients randomized to this point. The columns of the table correspond to the treatment group. A given patient may count multiple times in a column. The column totals T_1, T_2, etc. are compared, and the subject is assigned

Table 1: Example of a Working Table

to the treatment with the smallest total. If there is a tie between two or more treatment groups, then the treatment with the smallest total number of patients on study (ignoring the factors) is chosen. If there is still a tie, a random choice is made bewteen the tied groups.

Note that the construction and contents of the table did not depend on the number of levels for each factor, but only on the realized values of them for the patient about to be assigned.

If the additional rule to "break ties based on the treatment totals" is dropped this has little effect on the performance of dynamic allocation. It turns out that the column totals in table 1 rarely are tied unless the overall treatment totals are also tied. To see this, assume that treatment group A has more total patients than B. Then treating each row of the table as a sample from the total study, in expectation the total sum for A will be larger than that for B.

Another common way to control balance is by blocked randomization within cells. The sample space is divided into the $L_1 * L_2 * \ldots * L_k$ unique cells formed

by the k stratification factors, where L_i is the number of levels for factor number i. Balance is maintained separately within each cell using randomized blocks or some other simple method. This method will be refered to as stratified assignment.

For simplicity, the simulation experiments were evaluated with two treatment groups, and the k stratification factors had either 2 or 3 levels with corresponding probabilities of 1/2, 1/2 or 1/3, 1/3, 1/3. An experiment consisted of randomizing either 100, 200 or 400 patients to a 'trial' using $k = 1, 2, \ldots,$ or 12 factors. The imbalance I for any experiment is computed as the number of patients with factor $F_1 = 0$ on treatment A minus the number with $F_1 = 0$ on treatment B. By symmetry, I will be positive or negative with equal likelihood and $E(I) = 0$.

If only factor F_1 is considered, the best concievable assignment scheme would achieve an imbalance of 0 whenever possible, i.e., whenever the total number of patients with $F_1 = 0$ was even. When there was an odd number with $F_1 = 0$ the imbalance would be ± 1. The least effective assignment method is represented by a randomization scheme that paid no attention whatever to factor 1. (Of course, a maliciously designed method could do even worse by striving for imbalance, but this is not of interest here).

Thus, the simulation experiments quantify how the balance on one particular factor will be degraded by addition of other "extraneous" factors to the balancing scheme. The relative placement of the expected imbalance between the best case and worst case values provides a measure of randomization "efficiency" with respect to that factor. A strategy whose expected imbalance for F_1 is no better than it would be if that factor had been ignored has by

definition an efficiency of 0 — there is no return benefit for the investigator's time and effort to collect and use the stratification data. A strategy, if any, that matches the ideal imbalance would have an efficiency of 1.

Experience with dynamic allocation has revealed that when there are two treatment groups, the total number on treatment is nearly always balanced after every second patient. To make randomization within cells not appear worse only because of a global treatment imbalance, the method used within a cell was a randomized block of length 2.

3 Results

3.1 Best case imbalance

Let $p = P(F_1 = 0)$ and $q = 1 - p$. By adding the binomial expansions for $(p+q)^n$ and $(p-q)^n$, we see that the probability that the study will contain an even number of patients with $F_1 = 0$ is $.5 + (q-p)^n/2 = .5 + \epsilon$, where n is the number of subjects. Thus the imbalance I takes on values of 0, 1, and -1 with probabilities $.5 + \epsilon$, $.25 - \epsilon/2$ and $.25 - \epsilon/2$, respectively.

For $p = .5$ or $n \to \infty$ we have $\epsilon = 0$ exactly. For $p = .05$, an extreme case that might arise if F_1 had several levels, and $n = 100$ the probability of an even total is .500027. For any realistic n and p, then, the best case will have $E(|I|) = E(I^2) \approx 0.5$.

Table 2: Simple Randomization

3.2 Worst case imbalance

If the chosen factor F_1 were ignored in the randomization scheme then there are two important possibilities for it's imbalance, depending on whether the overall assignment scheme is based on simple randomization or restricted randomization. In the case of simple randomization each subject is assigned independently, e.g., by a coin flip. For the purposes of balance, we can view each assignment as the realization of a multinomial with probabilities as shown in table 2. The imbalance I will be the difference in counts between the two cells in the top row. By symmetry, $E(I) = 0$, and simple calculation using the moments of a multinomial yields $E(I^2) = np$.

If the overall study design uses some form of restricted randomization, then at the end of assignment there will be $n/2$ patients in treatment group A and $n/2$ in B. Since factor F_1 is not used in the balancing process, each of the $n/2$ patients on treatment A will independently have $F_1 = 0$ with probability p, similarly for the patients on treatment B. Thus I is the difference between two independent binomial variables, and has variance $npq/2 + npq/2 = npq$. The exact distribution of I can be enumerated by a simple computer program.

3.3 Stratified assignment

The k stratification variables or factors divide the subjects into $c = L_1 * L_2 * \ldots * L_k$ cells. When assignment is completed, any of these cells which has an even number of patients will be balanced between treatment groups A and B, and any with an odd number will have one extra assignment to either A or B. Application of formula (1) of Hallstrom and Davis [8] yields

$$\begin{aligned} E(I^2) &= \sum_i Pr(\text{cell}_i \text{ has an odd number of subjects}) \\ &= \sum_i \left[1 - (1 - 2p_i)^n\right]/2, \end{aligned} \tag{1}$$

where the sum is over those cells with $F_1 = 0$, and p_i is the probability that a new patient's stratification factors will lie in the ith cell. For stratified assignment with a blocksize greater than 2 the variance would be larger, since cells with an even number of subjects may also be unbalanced.

$E(I^2)$ is equal to the expected number of cells with $F_1 = 0$ which recieve an odd number of assignments. If factor F_1 is independent of the others, then $E(I^2) = pE(n^*)$, where n^* is the total number of odd cells and p is the probability that $F_1 = 0$. If $c >> n$, then with high probability every cell will have either 0 or 1 subjects, and $E(n^*) \to n$. That is, for a large number of factors stratified assignment will behave like simple randomization.

An important special case is when all of the cells have equal probability. (This does not imply that $p = .5$, since factor F_1 may have > 2 levels.) Then $E(n^*)$ is c times the probability that a binomial$(n, 1/c)$ is odd giving

$$E(I^2) = p\,\frac{c}{2}\left[1 - \left(1 - \frac{2}{c}\right)^n\right].$$

Figure 1: Imbalance versus number of stratification factors for n = 100. Lower dotted line, ideal imbalance of $\sqrt{.5}$; upper dotted lines, imbalance for worst case.

Some algebra shows that for $c = n/2$, $E(n^*) \approx n/4$. If $p = .5$ this suggests that stratified assignment would have about half the variance of restricted randomization.

3.4 Simulation results

The imbalance distribution for stratified assignment and for dynamic allocation was approximated by simulation, using 500 replications. The main results are shown in figure 1 for sample size of $n = 100$ (50 patients per treatment group). The root mean square imbalance $\sqrt{E(I^2)}$ for factor $F_1 = 0$ is plotted versus the total number of factors in the study. Three comparison lines are also present. The lower represents the ideal balance of .5, and the upper the

expected absolute imbalance for the worst case of F_1 ignored and simple randomization. The middle line is intermediate, with F_1 ignored but assuming restricted randomization for the study as a whole. From this figure it is apparent that the stratified method has completely degraded by the time there are 11 factors (2048 cells), and is equivalent to unstratified simple randomization. The method has lost about half of its efficiency, as compared to restricted randomization's error of $\sqrt{100/4} = 5$, with 5 factors (32 cells). The value for 6 factors is is close agreement with the argument in the prior section, that the squared error would be about half of npq for $n/2 = 50$ cells.

Dynamic allocation, on the other hand, is only mildly effected by the other factors. With 11 competing factors, it has lost only about 20 percent of its efficiency for balance on F_1, with respect to the ideal method for that factor. Figure 2 shows the actual distribution of the imbalance for dynamic allocation with 2, 5, and 10 factors. As the number of factors grows the distribution grows wider and the absolute imbalace increases, but the maximum imbalance is still very well controlled. Figure 3 shows the same information for stratified allocation. The expected imbalance distribution with F_1 ignored and restricted randomization is included on both figures.

Because dynamic allocation is focused on the univariate margins, it might be expected to do well in the situation above. Figure 4 shows results from the same set of simulations as figure 1, but it displays the imbalance for the smaller subgroup where both factor F_1 and factor F_2 are equal to 0. The first point on each plot is rather high — both methods do poorly when one of the salient factors has been ignored. For 2 factors, stratified allocation attains nearly the minimum possible imbalance. As more factors are added

Figure 2: Distribution of the imbalance for factor 1 by number of factors for dynamic allocation.

Figure 3: Distribution of the imbalance for factor 1 by number of factors for stratified allocation.

Figure 4: Imbalance of the F1 = F2 = 1 cell by number of factors, n = 100. Lower dotted line, ideal imbalance of $\sqrt{.5}$; upper dotted lines, imbalance for worst case.

its behavior degrades, and any benefit of the stratification process has been lost when there are 10 factors. Dynamic allocation controls the two unvariate margins $F_1 = 0$ and $F_2 = 0$ very well, giving a hypergeometric behavior to the combination $F_1 = F_2 = 0$, which attains about 50% efficiency. The dashed line on figure 4 shows the behavior of dynamic allocation when the balance is performed on all 2 factor interactions rather than on the univariate margins. That is, for 6 factors, the allocation was performed as if each of the $\binom{6}{2} = 15$ pairs were a separate 4 level factor, rather than using the six 2 level factors directly. The results are encouraging, especially since in actual practice only the few factors whose interactions are of particular interest would need to be treated in this way. Table 3 (see below) shows that the univariate margins

remain well balanced with this proceedure.

Multiple other simulations were done, including sample sizes of 200 and 400, multi-level factors, correlated factors, additional randomization, and a relationship to linear models. Most of the results are summarized in Table 3. The explanations below are indexed to the table.

1. The case shown in figures 1–3, $n = 100$, root mean square imbalance on the univariate margin $F_1 = 0$.

2. The same data as case 1, but using $E(|I|)$ rather that $\sqrt{E(I^2)}$ as the measure of imperfection. The values are somewhat smaller, but the overall trend is the same. Because the absolute value is harder to work with, this example is its only use in the paper.

3. The case shown in figure 4, $n = 100$, RMS imbalance on the bivariate margin $F_1 = 0$ $F_2 = 0$. For stratified assignment and $k \geq 2$ we can apply the exact results by treating F_1 and F_2 as a single 4 level factor with $p = .25$.

4. The first line shows the RMS imbalance on the bivariate margin $F_1 = 0$ $F_2 = 0$, using dynamic allocation on all pairwise factor combinations. The second line shows the imbalance for the univariate margin $F_1 = 0$.

5. RMS imbalance for $n = 400$, and 2 to 12 binary factors. As n is increased, the plots for stratified assignment have the same initial behavior, but reach a higher asymptote. The behavior of dynamic allocation, on the other hand, is essentially unchanged by n. The reason for this

stability can be seen by looking at Table 1: the outcome of a calculation is invariant to whether the average cell has 4–6 or 104–106 as its entry.

6. RMS imbalance for $n = 100$ and 2 to 12 three level factors. The behavior of stratified assignment is more closely related to the total number of cells $L_1 * L_2 * \ldots * L_k$ than to the total number of factors k. Dynamic allocation, however, is largely independent of the number of levels per factor, for the same reason that it is independent of n.

7. The table shows results for $n = 100$ and 1 to 12 binary factors, where adjacent factors have a correlation of $\sqrt{2}/2 = .71$. Correlation among the stratification factors is somewhat beneficial to stratified assignment; it appears to give a smaller effective number of cells. (Or equivalantly, cells with very low probability which add little to the sum in equation 1). For dynamic allocation correlation makes very little difference, for correlation, unless it is extreme, has only a small effect on the marginal distribution of any one factor.

8. In practice, stratification totals will often be kept by a computer program, and the computer system may suffer periods of unavailability during which an ordinary coin flip or similar method used for assignment. Dynamic allocation appears to be quite robust against such outage, as shown by this simulation in which 20% of the assignments were made randomly.

   About half of the random assignments will agree with what the computer would have recommeded had it been available. If we ignore the

fact that computer down time will normally occur in blocks, this simulation is equivalent to augmenting the dynamic allocation proceedure with a biased coin rule with $p = 0.9$, i.e., each assignment suggested by the dynamic allocation proceedure is accepted with probability p. Proponents of the biased coin method often suggest $p = 2/3$. Results for this value are also in the table, and show that the proceedure's ability to maintain balance is badly compromised.

9. Stratified allocation is equivalent to the following procedure: Fit a linear discriminant model in the stratification factors F_1, F_2, etc. to patients 1 to n, including all $2, 3, \ldots, k$-fold interaction terms, with treatment group as the dependent variable. Use the stratification factors of patient $n+1$ to compute class probabilities based on the model, and then assign the patient to the group with the lowest probability. Use treatment totals to break any ties. (This is equivalent to a linear model with one term for each of the $L_1 * \ldots * L_k$ cells, so prediction depends only on the contents of the cell). This equivalence also holds, trivially, for dynamic allocation with a single factor. For multiple factors, the discriminant analysis approach based only on main effect terms does not give patient by patient assignments identical to dynamic allocation. However, it has identical overall behavior in terms of imbalance: the table shows the results of such a "linear model" allocation under the conditions shown in case 1. Thus, another way to look at figure 1 is in terms of a main effects versus a full interaction model, and case 4 above would be similar to a model with all pairwise interactions.

Case		Number of Factors					
		2	4	6	8	10	12
1	stratified†	1	2	3.9	5.9	6.7	7.0
	dynamic	0.8	1.0	1.2	1.4	1.5	1.6
2	stratified	0.7	1.6	3.2	4.7	5.5	5.8
	dynamic	0.5	0.8	0.9	1.1	1.1	1.2
3	stratified†	0.7	1.4	2.8	4.2	4.8	4.9
	dynamic	2.4	2.4	2.7	2.7	2.7	2.7
4	bivariate†	0.7	1.0	1.5	1.9	2.1	
	univariate	0.9	1.1	1.3	1.4	1.4	
5	stratified†	1	2	4	7.8	11.8	13.5
	dynamic	0.8	1.0	1.1	1.4	1.5	1.6
6	stratified†	1.2	3.5	5.4	5.7	5.8	5.8
	dynamic	0.8	1.1	1.3	1.5	1.6	1.8
7	stratified	1.0	1.9	2.7	3.6	4.5	6.0
	dynamic	0.8	1.0	1.2	1.3	1.5	1.7
8	bias=.9	1.0	1.3	1.4	1.7	1.9	2.0
	bias=2/3	2.5	3.0	3.2	3.5	3.7	3.9
9	linear	0.7	1.0	1.2	1.4	1.6	1.7

Table 3: Imbalance for Various Cases. A †marks exact results, others are by simulation.

4 Discussion

Across several configurations using various sample sizes and 2 or 3 level factors, there was a fairly consistent rule of thumb for the stratified strategy: when the number of cells is approximately $n/2$, then the efficiency for balancing on any single factor is about 1/2. For any particular configuration, a more precise comparison can be obtained by comparing equation 1 to the limit npq. With a sufficient number of factors, performance is actually worse than using an appropriate unstratified assignment method for the study. Studies that use stratified assignment should not attempt to balance on more than a few important predictors.

Dynamic allocation is able to deal with a large number of factors: an extension of figure 1 (not shown) showed that .5 efficiency for this case occurs with approximately 30 factors. If a particular factor*factor substratum is deemed important, then that pair of stratification variables can be entered as a single factor with a larger number of levels. Similar behavior would be expected from other procedures that work with only the first order or linear terms in their balance attempts, such as that of Begg and Igelwicz [2].

The scope of this report has been purposefully restricted to a single question: whether the methods actually achieve their goal of creating balance between the treatments. Before choosing a method of assignment other issues need to be explored as well, these include the actual need and utility of study balance, and appropriate analysis methods when a deterministic rather than random assignment of patients has occured.

References

[1] Simon R: Restricted Randomization Designs in Clinical Trials. Biometrics 35:503-12, 1979

[2] Begg CB, Iglewicz, B: A Treatment Allocation Procedure for Sequential Clinical Trials. Biometrics 36:81-90, 1980

[3] Peto R, Pike MD, Armitage P, Breslow NE, Cox DR, Howard SV, Mantel N, McPherson K, Peto J, Smith PG: Design and analysis of randomized clinical trials requiring prolonged observation on each patient. 1: Introduction and design. British Journal of Cancer 34:585-612, 1976

[4] Lachin JM, Matts JP, Wei LJ: Randomization in Clinical Trials: Conclusions and Recommencations. Controlled Clinical Trials 9:365-374, 1988

[5] Taves DR: Minimization: A new method of Assigning Patients to Treatment and Control Groups. Clin Pharmacol Ther 15:443-53 1974

[6] Zelen M: The randomization and stratification of patients to clinical trials. J Chron Dis 27365-375, 1974

[7] Pocock SJ, Simon R: Sequential Treatment Assignment with Balancing for Prognostic Factors in the Controlled Clinical Trial. Biometrics 31:103-15, 1975

[8] Hallstrom A, Davis K: Imbalance in Treatment Assignments in Stratified Blocked Randomization. Controlled Clinical Trials 9:375-382, 1988