CLINICAL TRIALS ARTICLE Clinical Trials 2004; 1: 297–305

Covariate-based constrained randomization of group-randomized trials

Lawrence H Moulton[a]

Group-randomized study designs are useful when individually randomized designs are either not possible, or will not be able to estimate the parameters of interest. Blocked and/or stratified (for example, pair-matched) designs have been used, and their properties statistically evaluated by many researchers. Group-randomized trials often have small numbers of experimental units, and strong, geographically induced between-unit correlation, which increase the chance of obtaining a "bad" randomization outcome. This article describes a procedure – random selection from a list of acceptable allocations – to allocate treatment conditions in a way that ensures balance on relevant covariates. Numerous individual- and group-level covariates can be balanced using exact or caliper criteria. Simulation results indicate that this method has good frequency properties, but some care may be needed not to overly constrain the randomization. There is a trade-off between achieving good balance through a highly constrained design, and jeopardizing the appearance of impartiality of the investigator and potentially departing from the nominal Type I error. *Clinical Trials* 2004; **1**: 297–305. www.SCTjournal.com

Introduction

Group-randomized study designs are sometimes employed in trials of health interventions instead of individually randomized designs. Many factors may lead to the choice of a group-randomized design. These factors fall into the following categories: 1) the relative lack of feasibility of carrying out the intervention at the individual level; and 2) the desire to obtain information on intervention effects at the group level. Many authors have given advice on how to weight these factors in making the decision about the level of randomization and hence of treatment allocation [1–4]. Once the decision has been made to undertake a group-randomized trial, design aspects relating to stratification and randomization need to be addressed. The question of whether to use a highly stratified design (most often a pair-matched design), a completely randomized design, or a blocked design intermediate to these extremes, has been the subject of a good deal of statistical work [5–8]. This article is focused on the utility of further restricting the randomization procedure on the basis of relevant covariates.

Two aspects of group-randomized trials render them especially susceptible to the ill effects of an "unlucky" or "bad" randomization outcome, that is, one that has clear imbalance on one or more important variables. One aspect is that the studies are typically small, with perhaps only 4–20 groups to be randomized. Although the groups may contain thousands of participants, if there is between-group variability of characteristics, a completely randomized (at the group level) design can have a non-negligible probability of resulting in substantial imbalance on one or more characteristics across the trial arms. If these characteristics are also related to the treatment outcome, this can render interpretation of the trial results difficult. Even if there is some adjustment for the characteristics, there will be uneasiness that residual confounding remains due to other factors that are correlated with these characteristics.

The other feature of group-randomized trials, which, coupled with small numbers of units, can

[a]Departments of International Health and Biostatistics, The Johns Hopkins University, Bloomberg School of Public Health, Baltimore, MD, USA
Author for correspondence: Lawrence H Moulton, Departments of International Health and Biostatistics, The Johns Hopkins University, Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD, 20912, USA. E-mail: lmoulton@jhsph.edu

 10.1191/1740774504cn024oa

cause inferential problems, is that the units are often geographically contiguous, or nearly so. This can occur in agricultural experiments, of course, but the correlation patterns can be more difficult to detect in human communities. The placement of clinics or roads with respect to these units can alter reporting levels of disease cases, or even the incidence of infection (for example, in some areas, the spread of HIV by truck drivers). The accompanying spatial correlation of the response variables can affect the size of a statistical test, as seen below in the simulations.

In individually randomized trials in which participants enter sequentially, several mechanisms have been proposed to ensure balance on characteristics of trial participants across the arms of the trial. The most common technique is to use series of permuted blocks within each stratum so as to ensure balance with respect to enrolment sequence. More complicated methods have been proposed to balance on other covariates as well, including minimization techniques [9] and dynamic and adaptive allocation schemes [10,11]. All of these devices may be placed under the rubric of constrained or restricted randomization designs, although as mentioned below, these terms also have more specific connotations. Usually sample sizes in the individually randomized designs to which these devices are applied are sufficiently large so as to ensure, with high probability, that there will be a close balance on the relevant individual characteristics.

When close balance is desired on many variables, however, and there is a limited number of experimental units, cross-stratification on the variables will not be possible. An alternative is to produce composite scores that can be used to construct strata. If relationships between baseline covariates and the outcome variable have been studied, then a summary propensity score [12] for each unit could be used to pair-match them. Graham *et al.* [13] used a similar strategy, first reducing dimensionality through a principal component analysis, then forming pairs based on the resulting factor scores and assumed confounding effects. Morris and Hill [14], likewise, sequentially applied selection functions of numerous covariates to achieve balance. These procedures can be a useful first step in producing allocations that are balanced on a large number of variables (in both univariate and multivariate senses). Yet there still exists the possibility of a "bad" randomization outcome, with the unit with higher propensity score in each pair being assigned the treatment. Further constraints can preclude such an event.

Why do we randomize? Many trialists have addressed this question. The principal reasons are: 1) to give assurance, both to the investigator and to the general scientific community, of impartiality of treatment assignment; 2) to avoid hidden biases in treatment assignment; 3) to provide a convenient means of implementing treatment assignment. Two other reasons commonly cited are: 4) to provide a basis for statistical inference; and 5) to improve the chance of having a good distribution of relevant characteristics across the treatment arms. This fourth reason is not relevant in the context of model-based inference [15], which is the approach often taken by biostatisticians in the health sciences. In addition, with modern computing, carrying out randomization-based inference with unbalanced designs has become feasible. The fifth reason is relevant when the numbers of randomization units are large, but this usually is not the case with group-randomized trials. It is for this reason that active control of the randomization procedure is advocated so that it results in a plan that is at least balanced on known and measured covariates.

There are many papers in the statistical literature on group-randomized trials regarding whether to match or not, and how to do analysis. There is comparatively little information, however, about strategies other than basic pair-matching or stratification. Indeed, in two recent high-quality books on group-randomized trials, there is little material regarding randomization or the risks incurred in small group-randomized studies [16,17]. This may be because the general principles of blocking and stratification are well known. Attention to those principles, however, may be insufficient when dealing with small numbers of groups of highly variable humans who interact in complicated networks.

Design strategies

Blocking and stratification

The most common technique for allocating treatments is the use of randomly permuted blocks of treatment assignments. This design feature ensures treatment balance at any point in time during the enrolment and randomization phase of a trial, thereby minimizing bias due to secular trends. In clinical trials practice, as opposed to the practice of agricultural field trials, these block effects are rarely estimated, with their existence ignored in the analysis. Kalish and Begg [18], and more recently Chen [19], have investigated the effect of ignoring blocking in the context of permutation analyses, and found the practice to have little effect on size and power. Blocking is also used with stratification to ensure treatment balance within levels of one or more covariates, with separate lists of blocks used for each combination of covariate levels. The deepest stratification occurs in the pair-matched design, with

strata of size 2 determined by criteria such as age, neighborhood of residence, sibship, and so on. In such situations, analysts often account for the matching, especially when the within-pair correlations are considered to be large enough to outweigh the incurred loss in degrees of freedom.

Validity of designs and constrained randomization

When venturing into the territory beyond the conventional stratified designs (randomized blocks, pair-matched, and so on), one runs the risk of producing a design that technically is *biased* or not *valid*. In the design of experiments nomenclature, a design is biased if, across the randomization units, there is any difference in probability of assignment to a given treatment. This problem rarely arises in practice. More problematic is the validity of a design: a completely randomized design is valid if each pair of randomization units has the same probability of being allocated the same treatment [20]. Other criteria exist for other designs, but the main idea is one of whether there is independence of treatment assignment between units. For example, Fisher [21] describes as invalid an extreme situation where plants of one type (one experimental condition) are assigned positions on one side of a greenhouse and the plants of the other type assigned to the other side by a single coin flip. He noted there could be confounding variables such as different lighting or air current conditions on the two sides. What he did not state is that even a "randomization properly carried out" could result in the same unfortunate allocation.

If a design is not valid, it runs the risk of changing the Type I error from its nominal value. In addition, it may bring into question whether the investigator has selected a design that will give an advantage to a treatment comparison of interest. The term constrained, or restricted, randomization has come to refer to those designs that go beyond the basic design constraints to specify classes of randomization outcomes that satisfy certain balancing criteria, while retaining validity of the design. Application of these approaches, which involve group-theoretic considerations of permutation groups, is not often a trivial task, and may be impossible, even when dealing with one simple covariate of spatial pattern along one geographic axis [22].

Example 1: hypothetical constrained design

Suppose we have an intervention study with an HIV incidence outcome that will take place in four villages that have baseline HIV prevalence survey results (or, for example, a composite socio-economic score): 2, 4, 10, 13%. In practice, one would use more villages, but the key problems are easier to illustrate with this limited number. One possible allocation strategy is to use no constraints whatsoever, selecting at random one of the six possibilities shown in Table 1. There is, however, a 1 in 3 chance of either the intervention villages being those with the lowest baseline prevalences (A: 2 and 4%), or with the highest prevalences (F: 10 and 13%). Each of these two extreme situations will have a mean absolute difference of 8.5 percentage points in prevalence in the two treatment groups, even greater than the overall mean prevalence of 7.25%. Adjustment for these initial differences could be made at the time of analysis, but is not a very satisfactory solution for reasons that are mentioned above.

A convenient way to handle this difficulty is to specify, in advance of randomization, that only those treatment allocations that result in exact balance on mean prevalence will be permissible. Imposing this constraint in this situation would preclude any allocation being selected. If the constraint is relaxed to require only a mean difference of less than one percentage point, this would mean selecting either allocation C or D in Table 1 with equal probability. Note, however, that those villages with 2 and 13% (and 4 and 10%) are always linked together. This might be of concern if, for example, the villages with 2 and 13% were adjacent to each other, and in the far north of the country, while those with 4 and 10% were near each other in the far south. One could argue that there are effectively only two experimental units in such a situation, not four, and that the analysis should proceed accordingly. One might further relax the constraint to "mean difference of less than 3 percentage points" and add the criterion of geographic

Table 1 All possible allocations in a hypothetical trial of two conditions (intervention, control) and four experimental units (villages)

	Villages				
Allocation	Intervention		Control		Mean difference
A	2	4	10	13	−8.5
B	2	10	4	13	−2.5
C	2	13	4	10	0.5
D	4	10	2	13	−0.5
E	4	13	2	10	2.5
F	10	13	2	4	8.5

The baseline HIV prevalence (%) labels each village; overall balance is measured by the mean difference in prevalence between intervention and control villages. Example: In allocation A, the villages with HIV prevalences of 2 and 4% are assigned to the intervention arm, and the villages with 10 and 13% are control villages; the mean difference is $[(2+4)-(10+13)]/2 = -8.5$.

balance defined as "one intervention unit in the north, and one in the south". This would lead to randomly selecting from allocation B or E in Table 1, but this still is subject to the criticism that there are really only two effective units of randomization. However, if geography were not a factor, then just using the criterion of less than three percentage points' difference would permit selection from B, C, D or E. This would achieve good balance, retain four randomization units, and reduce the basis for outside accusations of having rigged the randomization. We say reduce, not eliminate, for a critic might note that the units with 2 and 4% prevalences (or 10 and 13%) never have the chance to be in the same treatment arm. This is less serious an accusation; indeed, we often design studies expressly to achieve this result. For example, in this simple case, there is another obvious way to achieve this design result: pair-match the units with prevalences (2%, 4%) and (10%, 13%) and randomize within each pair. Then, if the pairing were ignored at the time of analysis [5], the two situations would be indistinguishable. As the second example (below) indicates, however, such an alternative may not always be evident.

Covariate-based constrained randomization

General approach

We focus on those designs that seek to achieve global and simultaneous balance, or near-balance, on one or more covariates, or functions of covariates, that could be related to the outcome(s) of the trial. There is a trade-off involved in this approach: we do not want to constrain the design to such a high degree that the investigator is open to accusations of manipulation in favor of his or her hypothesis. The next section addresses this trade-off.

Constraining criteria

The acceptability criteria that limit the possible allocations can be of varying levels of specificity. For continuous covariates, a simple caliper-type criterion should be sufficient: one could specify that the group means for each trial arm be within a quarter standard deviation of each other, or within $\pm 10\%$, say. Covariates that are dichotomous at the individual level (for example, gender), could be balanced within ± 10 percentage points. Group-level covariates, such as geographic location, or presence of a clinic in the community, could be balanced either exactly or with some specified maximum range for the difference. For example, if the design were pair-matched, it might be required that to make a pair, both groups should have a clinic, or should not have a clinic. Alternatively, the constraint could be global, only requiring that each trial arm have the same number of clinics, or $\pm$ one clinic difference between them. Techniques that weight sets of covariates, such as factor analysis or the use of propensity scores as mentioned above, also could be employed.

Overly constrained designs

A design that has constrained the randomization on the basis of one or more covariate values may yield a set of potential allocations that has a pair of units always in the same treatment arm, as in the most highly restricted designs in Example 1. Alternatively, it may be that a given pair of units is *never* in the same arm, which might not seem fair to a critic if the basic stratification structure of the design does not preclude it. A quick check on these possible extreme situations can be done by counting the number of times any given pair of randomization units (or units in different blocks) receives the same treatment allocation. The results may be stored, for example, in the lower triangle of a $2m \times 2m$ matrix, where m is the number of units in each treatment arm. When the problem is too large to enumerate, and the "loop" method mentioned above is used, this matrix could be constructed when a suitably large number of acceptable designs is reached. Examination of the matrix for under- or over-represented pairs would then reveal any potential causes for concern in the design.

Proposed randomization algorithm

1. Form a list of all the possible allocations. For a pair-matched design, this will have 2^m entries, where m is the number of pairs; for a completely randomized (at the group level) design, there will be $\binom{2m}{m}$ entries, where $2m$ is the total number of groups.
2. Making a pass through all of these entries, select those allocations that meet the specified criteria. These criteria could mean achieving some level of balance on a given set of covariates.
3. Make a matrix whose elements are the number of times, from among those allocations identified in step 2, each pair is together.
4. Accept the constrained list of possibilities and go to step 5; or relax or tighten criteria and go to step 2; or change the stratification and go to step 1.
5. Randomly select one allocation from among the ones that have been selected as being acceptable in step 2.

An alternative approach is to put a loop in the computer program executing the algorithm, randomly generating plans until an acceptable one is found. This may be necessary when the total number of possible allocations is too great to enumerate in a feasible amount of time. Step 3 may be omitted if the proportion of potential allocations that is acceptable is high, that is, the constraints are not very stringent. To avoid accusations of having rigged the outcome, it may be best to ensure, at step 3, that no two units are always together in the same arm, or always separated (except for pair-matched strata).

Example 2: Baltimore drug intervention study

A recent example of covariate-based constrained randomization utilizing the proposed randomization algorithm is afforded by a study undertaken as part of the Aid First Initiative in Baltimore, MD. In order to assess the impact of family and community network mobilization efforts, this trial intervenes at the census tract level. The primary outcome measure is the incidence rate of admission to treatment facilities for drug dependence. From among all Baltimore City census tracts, 20 had been identified as being of particular concern. Census data provided the following relevant covariates for each tract: total population, family income, % vacant houses, % males employed, % high school education, % receiving Public Assistance, % 15–64 years of age, % African-American. One more criterion was geographic balance: the investigators would not have been satisfied if, for example, the randomization resulted in all intervention units being in Baltimore's west side, with the control units in the east side.

The initial constraints considered for the randomization scheme were:

1. Perfect balance on geographic areas that had received prior city-level special attention or not; five of each type must be in each of the control and intervention groups.
2. Balance to within 10% on the eight census data covariates. Specifically, for each covariate, if for a given allocation pattern the mean of the intervention group divided by the mean of the control group was greater than 1.1 or less than 1/1.1, that pattern would not be considered for potential selection.
3. Approximate geographic balance. The city was divided into quadrants each containing five of the 20 study tracts; the criterion was that each quadrant should have at least two intervention units.

There are 184,756 possible combinations of 20 units chosen 10 at a time, half of which are unique. All combinations were enumerated [23] in the GAUSS System matrix manipulation language [24], and each was checked for whether the above criteria were satisfied. Only 46 allocations (23 unique) met the specified criteria. Using these, a 20×20 matrix was formed whose elements contained the number of times each pair of randomization units was designated for inclusion in the same arm of the trial. There were four pairs of units that never were together in the same arm of a trial under any of the 23 unique allocation patterns: (1,17); (1,12); (12,14); (14,17), where the numbers in parentheses refer to the randomization unit numbers. Inspection of the census tract data indicated that unit 12 had the highest percentage of vacant houses and unit 1 had the second highest; thus, it was difficult for both to be allocated to the same arm. Similarly, unit 17 had the highest percentage receiving public assistance, and unit 14 had the second highest. Given these data configurations, we decided to relax the balance criteria for these variables, allowing $\pm 25\%$ imbalance between the treatment arm means for each of these, instead of the 10% criterion employed for the other variables. Running the allocation counting program again, we ended up with 148 combinations that met the new criteria, which afforded 22 opportunities for the pair (1,17) to be in the same arm, 22 for (1,12), six for (12,14), and 10 for (14,17). Clearly, this is not optimal, but we considered these opportunities to be of sufficient number considering the desire to avoid an unreasonable balance on the variables of interest. At the minimum, we could not be accused of setting up a randomization scheme that gave no opportunity for these pairs to arise, which would have meant a complete dependence between the units: knowing that unit 1 was in one arm would have given certain knowledge that units 12 and 17 were in the other, for example. As it turned out, the allocation that was randomly selected from among the 148 potential ones was very well balanced, every variable coming within 10% balance between the two arms. Still, as per the results of the following simulations, it might be the case that the nominal size of the final trial test is altered, if there is substantial correlation between the outcomes of the units involved.

Simulations using the Baltimore study structure

Rationale

It is clear that the imposition of a set of global, covariate-based constraints can render a design for which a randomization-based justification for the standard ANOVA no longer holds. Less clear is the

effect this may have on the frequentist properties of the design. In general, one may expect the greatest problems when two situations coincide: the design produces great unevenness in the opportunities units have to be in the same or different arms of a trial, and outcomes among some set(s) of units are highly correlated. The simulations serve to give a rough idea of what may be the consequences of highly restricting a design, followed by an analysis that is conditional on the actual selected allocation. To investigate these, we carried out a series of simulations based on the final set of restrictions used in the Baltimore study of Example 2 that resulted in only 148 possible allocations from which to choose. The results are not meant to furnish a precise guide for practitioners, but to give a qualitative idea of the consequences of such designs. In those situations where investigators feel the need for a highly constrained design, it may be useful for them to construct their own simulations that more closely match their particular circumstances.

Analysts, however, have the option to carry out randomization-based inference that uses the restricted design structure. In the Baltimore study, this would mean basing it on the 148 potential allocations. The advantage of such an approach is that we would expect the corresponding randomization test to have approximately the nominal test size regardless of the correlation structure. Drawbacks include the small number of allocations in this instance, making the randomization distribution perhaps too discrete, and the high computational complexity for the calculation of confidence intervals, as described by Tukey [25]. Whether the analysis uses the randomization distribution or treats the actual allocation as an ancillary statistic, the trial will have benefited from the balance on important baseline covariates.

Setup

For simplicity, for each simulation we generate 20 Gaussian outcomes for the census tracts (units). Many actual trials have binomial or Poisson responses at the individual level, but their mean at the group level will be approximately Gaussian if the group is sufficiently large and/or the risks are not too near zero or one, or rates are not too low. We construct the standard *t*-test statistic comparing the two groups of 10 tracts based on a pooled variance estimate, and take as our simulation outcome the estimated size of the test under various conditions. Three factors are modified: the type of units that may be correlated, the number of units that may be correlated, and the degree of correlation. In group-randomized trials, spatial correlation of the responses often is introduced via geographic proximity or other social or geographic features of the involved communities. These correlations, however, will in general not be known, or not feasible to model. Group-randomized trials with hundreds of groups might permit some spatial modeling of the correlation structure, but even then it will be difficult due to the complexity of human networks of disease transmission. The simulation setup corresponds to this situation of effective ignorance, assigning a fraction of the groups to have correlated variates generated, but not accounting for this correlation at the analysis stage. This is the most common type of analysis, assuming a random effects structure with an effect level for each group, but independence across groups.

In the highly constrained Baltimore study design, the pairs of units that had the highest probabilities of being in the same treatment arm are: (12,17): 144 out of 148 potential allocations; (1,10): 126; (3,12): 120; (1,14): 120; (10,14): 118; (3,17): 58. Thus, (1, 10, 14) are often grouped together, as are (3,12,17). In the first set of simulations, trivariate Gaussian responses for one or both of these sets were generated, with uniform correlation among all members of a triplet. The next set of simulations focused on pairs that had the lowest probabilities of occurring together: (12,14): 6; (14,17): 10; (1,12): 22; (1,17): 22; (8,15): 26; (9,15): 26; (9,20): 26. These were put in the sets: (1,12,14,17) and (8,9,15,20). Note, however, that not all pairs within these sets have a low chance of occurring together – for example, 8 and 9 are in the same arm in 112 of the 148 allocations. Finally, a set of simulations was run involving correlations among units that can occur together with 50% chance, that is, in 74 of the 148 allocations. Among these were the pairs: (1,20); (2,18); (9,16); (10,11). These were put into the sets: (1,2,18,20) and (9,10,11,16).

Correlation was introduced by generating a multivariate Gaussian response (proc rndmn, Gary King, GAUSS language algorithm, 1999) for the given sets of 3 or 4 units, with correlation coefficient equal to 0, 0.25, 0.5, or 0.75; or, where possible, 0, -0.25, or -0.5 (actually, -0.499999, due to the need to have a positive definite covariance matrix). A nominal test size of 0.05 was used for testing the null hypothesis of equality of means for the Gaussian outcomes in each of the two arms. Each configuration was simulated one million times.

Results

In the absence of correlation, we expect the rejection level to be 0.05. As can be seen in Table 2, the more extreme the correlation, the greater the

Table 2 Proportions of simulated group-randomized trials with rejection of the null hypothesis of no difference between two treatment arms by varying levels of correlation

Units to be correlated	Proportion rejected			
Units with high probability of being in the same arm				
	ρ:	0.00	0.25	0.50	0.75
(1,10,14)		0.051	0.057	0.063	0.069
(1,10,14) & (3,12,17)		0.052	0.063	0.075	0.088
	ρ:	0.00	−0.25	−0.50	
(1,10,14)		0.051	0.045	0.039	
(1,10,14) & (3,12,17)		0.052	0.040	0.028	
Units with low probability of being in the same arm					
	ρ:	0.00	0.25	0.50	0.75
(1,12,14,17)		0.051	0.048	0.044	0.041
(1,12,14,17) & (8,9,15,20)		0.051	0.046	0.041	0.035
	ρ:	0.00	−0.25		
(1,12,14,17)		0.051	0.055		
(1,12,14,17) & (8,9,15,20)		0.051	0.057		
Units with average probability of being in the same arm					
	ρ:	0.00	0.25	0.50	0.75
(1,2,18,20)		0.051	0.050	0.049	0.049
(1,2,18,20) & (9,10,11,16)		0.051	0.050	0.049	0.048
	ρ:	0.00	−0.25		
(1,2,18,20)		0.051	0.052		
(1,2,18,20) & (9,10,11,16)		0.051	0.053		

Note: constraints on the covariance matrix determinant eliminated $\rho = -0.50$ and/or -0.75 from some simulations. Nominal test size is 0.05, all data are generated under the null hypothesis. Within each specified set of unit identifiers, multivariate Gaussian responses are simulated with the given correlation amongst all set members.

departure from this nominal level, although none of the examples are particularly worrisome in this regard. The greatest difference from 0.05 occurred when the two sets of the triplets that had a high chance of being in the same arm were highly correlated ($\rho = 0.75$ within each set); the rejection level was 0.088. Conversely, when units with a low chance of ending up in the same arm were positively correlated, the rejection level fell below the nominal level, that is, the test became more conservative. These results were reversed for the case of negative correlation. Correlation among units that had a 50% chance of being in the same arm (were in the same arm for 74 of the 148 possible allocations), had negligible effect.

Discussion

Some statisticians will prefer not to constrain a design beyond basic blocking and stratification maneuvers. But when faced with a multimillion dollar study and only a handful of communities to be randomized, there is much pressure to achieve as much statistical power as possible. There is a limit to the variance reduction that can be achieved by blocking and stratification in these situations, commonly seen in group-randomized trials. Freedman *et al.* [7] considered the possibility of further restricting a pair-matched design by requiring a continuous covariate to be higher in the intervention group for about half the pairs, and lower in the other pairs. In their simulations, substantial gains in efficiency of such an approach were possible when the covariate was strongly related to the outcome. They could have added another requirement, for example, that of marginal balance on the covariate across the trial arms. In the end, however, they did not use any restrictions other than the original pairing, citing the desire for a method that was "easily understood by non-statisticians". We note that there are other ways to make a randomization "understood", or politically acceptable. After highly, constraining a sexual health intervention trial in Tanzania based on the method described here, Hayes and colleagues publicly randomized the trial using a lottery-like table tennis ball selection mechanism, with very satisfactory results (R. Hayes *et al.*, personal communication and unpublished manuscript).

The example of the randomization performed for the Baltimore Aid First Initiative raises the question: at what point of restriction might the objectivity of the investigator be called into question? After all, using the original constraint criteria, only 0.025% of the total number of potential randomization results were deemed balanced enough to be selected. However, there were still 46 possible allocations, which is as many as found in small, less-restricted group-randomized trials, for example, one with 10 units in a pair-matched design. Yet when the

number of possibilities becomes very small, say 2 or 4, there may be more suspicion that the investigator has manipulated the design to his or her advantage. The main drawback of the original constraint scenario was not that there were too few randomization outcomes, but that some pairs of units could not have received the same treatment.

An algorithm has been presented for implementing and checking covariate-based constrained randomization. It is not difficult to implement and assure the desired degree of balance on relevant covariates. When the number of experimental units is sufficiently large and/or the number of covariates is small, the standard techniques of blocking and stratification may be sufficient. However, even in such cases it may be wise first to enumerate or simulate the proposed scheme, and identify the probability of obtaining a "bad" randomization. This was done in a group-randomized trial of pneumococcal vaccine with 38 geographically defined units, with eligible population size as the primary covariate of interest [26]. Many group-randomized trials, however, will have smaller numbers of units, and more covariates for which an imbalance at the start of the trial could prove an embarrassment and complicate inference at the trial's conclusion.

The simulations presented here indicate that when units with highly correlated outcomes have a high probability of all being included in the same trial arm, the actual Type I error can be inflated. Investigators can take steps to diminish this possibility. If they are aware that certain units are likely to have correlated responses, they can specify in the design that there should be balance among them as to treatment assignment. This was done in the Baltimore drug intervention study example by specifying near-balance by geographic quadrant. They can also relax the criteria for balance, thereby reducing assignment linkage between units. Another factor in favor of the investigator is that units would tend to have a higher probability of joint inclusion only if they were balanced on covariates related to the outcome – but this balance would mitigate against highly positively correlated outcomes. Perhaps the more likely scenario is that units with positive correlation of their outcomes would be separated into different treatment arms more often than would be the case for other units. This would tend to lower the Type I error below the nominal level, and decrease power slightly as well, although we might still expect substantial gains in power due to the assurance of balance on important covariates [7].

Although the focus of this manuscript is group-randomized trials, a highly constrained design could be of use in other situations with small numbers of experimental units whose covariates are all known before randomization. Consider a study that divides a classroom of 20 students into two experimental conditions, with data on the students' relevant covariates (age, weight, etc., depending on the intervention and response variables). After forming whatever strata will be accounted for at the time of analyses for variance reduction, such a study might benefit by adding further constraints on the remaining baseline covariates.

Acknowledgements

This research was supported in part by NICHD R01-HD38209, the Bill and Melinda Gates Foundation as part of the CREATE project, and by the Aid First Initiative (supported by Baltimore Open Society Institute).

References

1. **Cook TD, Campbell DT**. *Quasi-experimentation: design and analysis issues for field settings*. Chicago, IL: Rand-McNally, 1979, 354–356.
2. **Donner A, Klar N**. Statistical considerations in the design and analysis of community intervention trials. *J Clin Epidemiol* 1996; **49**: 435–39.
3. **Fortmann SP, Flora JA, Winkleby MA, Schooler C, Taylor CB, Farquhar JW**. Community intervention trials: reflections on the Stanford Five-City Project Experience. *Am J Epidemiol* 1995; **142**: 576–86.
4. **Gail MH, Mark SD, Carroll RJ, Green SB, Pee D**. On design considerations and randomization-based inference for community intervention trials. *Stat Med* 1996; **15**: 1069–92.
5. **Diehr P, Martin DC, Koepsell T, Cheadle A**. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. *Stat Med* 1995; **14**: 1491–1504.
6. **Donner A**. Sample size requirements for stratified cluster randomization designs. *Stat Med* 1992; **11**: 743–50.
7. **Freedman LS, Green SB, Byar DP**. Assessing the gain in efficiency due to matching in a community intervention study. *Stat Med* 1990; **9**: 943–52.
8. **Martin DC, Diehr P, Perrin EB, Koepsell TD**. The effect of matching on the power of randomized community intervention studies. *Stat Med* 1993; **12**: 329–38.
9. **Pocock SJ, Simon R**. Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 1975; **31**: 103–15.
10. **Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ, Callaghan T**. Dynamic balanced randomization for clinical trials. *Stat Med* 1993; **12**: 2343–50.
11. **Berry DA, Eick SG**. Adaptive assignment versus balanced randomization in clinical trials: a decision analysis. *Stat Med* 1995; **14**: 231–46.
12. **Rosenbaum PR**. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
13. **Graham JW, Flay BR, Johnson CA, Hansen WB, Collins LM**. Group comparability: A multiattribute utility measurement approach to the use of random assignment with small numbers of aggregated units. *Evaluation Review* 1984; **8**: 247–60.

14. **Morris C, Hill J**. The Health Insurance Experiment: design using the finite selection model. In Morton SC, Rolph JE eds. *Public policy and statistics: case studies from RAND*. New York: Springer-Verlag, 2000, 29–53.
15. **Royall RM**. Current advances in sampling theory: implications for human observational studies. *Am J Epidemiol* 1976; **104**: 463–74.
16. **Murray DM**. *Design and analysis of group-randomized trials*. New York: Oxford University Press, 1998.
17. **Donner A, Klar NS**. *Design and analysis of cluster randomisation trials in health research*. London: Arnold, 2000.
18. **Kalish LA, Begg CB**. The impact of treatment allocation procedures on nominal significance levels and bias. *Control Clin Trials* 1987; **8**: 121–35.
19. **Chen H**. *Effect of ignoring randomization constraints in the analysis of clinical trials*. PhD dissertation. Baltimore: The Johns Hopkins University, 1997.
20. **Bailey R**. Restricted randomization. *Biometrika* 1983; **70**: 183–98.
21. **Fisher RA**. *The design of experiments*, fourth edition (first 1935). New York: Hafner-Publishing Co. Inc., 1947.
22. **Bailey RA**. Restricted randomization: a practical example. *J Am Statist Assoc* 1987; **82**: 712–19.
23. **Gentleman JF**. Algorithm AS 88: generation of all ${}_{N}C_{R}$ combinations by simulating nested *DO* loops. *Appl Statist* 1975; **24**: 374–76.
24. **Aptech Systems Inc**. *The GAUSS system*. Maple Valley, WA: Aptech Systems Inc., 1999.
25. **Tukey J**. Tightening the clinical trial. *Control Clin Trials* 1993; **14**: 266–85.
26. **Moulton LH, O'Brien KL, Kohberger R** *et al*. Design of a group-randomized *Streptococcus pneumoniae* vaccine trial. *Control Clin Trials* 2001; **22**: 438–52.