Biometrics 69, 949–959
December 2013
DOI: 10.1111/biom.12077

Repeated Randomization and Matching in Multi-Arm Trials

Zhenzhen Xu,[1,*] John D. Kalbfleisch[2]

[1]Food and Drug Administration, CBER, Rockville, Maryland 20852-1448, U.S.A
[2]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, U.S.A.
[*]*email:* zhenzhen.xu@fda.hhs.gov

Summary. Cluster randomized trials with relatively few clusters have been widely used in recent years for evaluation of health-care strategies. The balance match weighted (BMW) design, introduced in Xu and Kalbfleisch (2010, *Biometrics* **66**, 813–823), applies the optimal full matching with constraints technique to a prospective randomized design with the aim of minimizing the mean squared error (MSE) of the treatment effect estimator. This is accomplished through consideration of M independent randomizations of the experimental units and then selecting the one which provides the most balance evaluated by matching on the estimated propensity scores. Often in practice, clinical trials may involve more than two treatment arms and multiple treatment options need to be evaluated. Therefore, we consider extensions of the BMW propensity score matching method to allow for studies with three or more arms. In this article, we propose three approaches to extend the BMW design to clinical trials with more than two arms and evaluate the property of the extended design in simulation studies.

Key words: BMW design; Clustered randomized trial; Experimental design; Optimal matching; Propensity score matching; Randomization tests; Repeated randomization.

1. Introduction and Motivating Example

The balance match weighted (BMW) design introduced in Xu and Kalbfleisch (2010) applies the optimal full matching with constraints technique to a randomized study, and aims to minimize the mean squared error (MSE) of the treatment effect estimator. This approach involves considering several (M) randomizations of the participating units into the two treatment arms. With each randomization, the optimal full matching with constraint is used in order to identify the best blocking for that randomization. The randomization and corresponding full matching that leads to the smallest total distance is then selected. The distance measure used for the matching was based on estimated propensity scores, as has been proposed in observational studies (Rosenbaum and Rubin, 1984; Rosenbaum, 2002; Hansen, 2004), but other distance measures could have been used. A simulation study demonstrated good MSE properties for the BMW design as compared to other approaches in the literature.

Repeating randomization to achieve better covariate balance between treatment groups has previously been proposed. Cox (2009) suggests "re-randomizing, say, 20 times and choosing the design with most balance." Rubin (2008) recommends re-randomizing treatment allocation when the one obtained has substantial potential for conditional bias given the observed covariate values, and continuing to do so until satisfied. Morgan and Rubin (2012) propose a re-randomization procedure suggesting, for example, a balance criterion to be applied to treatment and control covariate means, and then repeatedly randomizing until the criterion is met. The BMW design, however, performs multiple (say M) randomizations, and selects the one that leads to the best matching. This avoids the setting of a criterion in advance and also takes advantage of matching with potential gains in robustness.

Cluster randomized trials, in which social units are selected as the units of randomization, have been increasingly used in the past three decades to evaluate the effects of interventions at a community level. In such studies, there are typically rather few participating units and often several variables that describe the properties of these units. Further, the units available for randomization and study are generally known in advance of the trial and there is the opportunity to balance the design with respect to these variables. A trial of this type, the INSTINCT trial (Scott et al., 2012) motivated our interest in this problem, and led to Xu and Kalbfleisch (2010) as well as this extension. This trial investigated the effect of a multifaceted educational program directed at hospital emergency departments in enhancing the appropriate use of tissue plasminogen activator (tPA) in the treatment of ischemic stroke. Twenty-four hospitals were randomized to intervention or control using blocking to achieve a degree of balance. Various covariates characterized the participating hospitals and raised the question as to how best to account for this diversity in the design.

In some instances, more complex treatment programs are evaluated in cluster randomized trials. For example, a drug may be applied at different dosage levels or a factorial design might be used to compare simultaneously two treatment strategies. In such cases, the BMW design, which was introduced for two treatment groups, needs to be extended. In a two-arm BMW design, the problem of finding the best blocking can be reduced to solving a standard combinatorial optimization problem for which fast algorithms exist (Rosenbaum, 2002). However, the problem of obtaining an optimal matching with three or more groups has been shown to be a NP (non-deterministic polynomial time) complete problem. The most notable characteristic of such problems is that the time

required to solve them using any currently available algorithm increases very quickly as the size of the problem grows, and in many instances, it is not possible to determine the optimal solution even for relatively small problems. For example, consider a small study with m subjects in each of three arms. The number of possible matchings into block of size m is $(m!)^2$ which is 576 when $m = 4$; 4,518,400 when $m = 6$, and 1.317×10^{13} when $m = 10$. Enumeration of these matchings quickly becomes impossible. In order to circumvent this problem, we develop some ad hoc approaches which may not lead to the optimal tripartite matching, but to solutions that are close to optimum. Although the balance achieved by using our approach is not completely optimal, it is typically close to optimal which we establish by carrying out comparisons with the true optimum in samples of size $m = 6$. Further, with larger sample sizes, we find that the results of this approach compare favorably to approximate solutions based on an integer programming formulation implemented with a commercial software package.

The problem of matching with three or more groups also arises in observational studies. For example, in assessing the effect of an experimental factor, some investigators use a second control group in an effort to detect the hidden biases in the unobserved covariates (Seltser and Sartwell, 1965; Chang et al., 1997; Wells et al., 1997; Bo and Rosenbaum, 2004). Campbell (2009) argues for this approach noting that although matching can adjust the differences in observed covariates, bias may still exist due to some unobserved covariates; comparison with two distinct control groups offers some protection since the control groups may differ from each other substantially on the unobserved covariates. Bo and Rosenbaum (2004) considered this situation and proposed an algorithm to match three groups into balanced incomplete blocks of size two. One drawback of the incomplete block design is that the direct comparison of treatments A and B only occurs in a subset of the blocks. This approach is potentially much less efficient than approaches that match in blocks of size three.

The rest of the article is organized as follows. Notation, the three matched designs (incomplete blocks of size two, asymmetric and symmetric tripartite matching) and the analysis models are presented in Sections 2 and 3. Section 4 gives results of a simulation study comparing the performance of the 3-arm BMW design based on different matching algorithms with one another as well as with the completely randomized design. Section 5 further extends the BMW design to more than three arms and considers, in particular an application to a 2×2 factorial design. A case study is outlined in Section 6, and Section 7 considers aspects of the statistical analysis of the BMW design, including a discussion of randomization tests. The article concludes with some discussion in Section 8.

2. Methods

Suppose that N subjects are available for study and consider a randomization $Z = (Z_1, \ldots, Z_N)^{\mathrm{T}}$ into three treatment assignments A, B, C, where $Z_i = 1,\ 2$ or 3 if subject i is assigned to A, B, or C, respectively. We consider balanced designs in which $N/3$ subjects are assigned to each treatment and suppose that each of the possible $N!/[(N/3)!]^3$ assignments is equally likely to be chosen. It is convenient to define the random sets $\mathcal{A} = \{i : Z_i = 1\}$, $\mathcal{B} = \{i : Z_i = 2\}$, and $\mathcal{C} = \{i : Z_i = 3\}$. Suppose further that associated with subject i, there is an observed vector of covariates, $X_i = (X_{1i}, \ldots, X_{ri})^{\mathrm{T}}$ with $X_{1i} = 1$, for $i = 1, \ldots, N$.

For any two subjects, i and j, we assume that a distance measure, $D(i, j) \geq 0$, has been specified that measures the discordance between the covariates associated with those subjects. Specifically, we suppose that $D(i, j) \geq 0$ specifies the distance. In this article, we utilize a distance measure that is based on the estimated propensity score given the randomization of subjects to the three treatment groups. However, other distance measures could be used such as the Mahalanobis distance suggested by Greevy et al. (2004). In order to reduce imbalance in the between-group comparisons due to chance correlations between the X's and the treatment assignment, we consider matching subjects into blocks in such a way as to minimize the between subject distances within blocks. We consider three approaches for carrying out this matching with three groups, and later consider extensions to more than three groups.

This proposed design generalizes the two-arm BMW design introduced in Xu and Kalbfleisch (2010). In the two-arm design with specified parameter k and M, study subjects are first randomized to two treatment groups and the matrix of distances is created. The optimal full matching with constraint k is then obtained (Olsen, 1997; Hansen, 2004) and the total distance Δ_k recorded. The process is repeated M times and the randomization yielding the minimum total distance $\Delta_k^* = \min(\Delta_{1k}, \Delta_{2k}, \ldots, \Delta_{Mk})$ is selected.

In the proposed three arm BMW design, we need to match across three treatments. Otherwise it is identical to the two armed design and is described as follows:

Step 1. Randomize one third of the N subjects to each treatment groups $\mathcal{A}$, B, and $\mathcal{C}$.

Step 2. Obtain the matching of the subjects into blocks that leads to the minimum total distance, Δ, between subjects within blocks.

Step 3. Repeat the above two steps M times and choose the randomization (and corresponding matching) with minimum total distance $\Delta^* = \min(\Delta_1, \Delta_2, \ldots, \Delta_M)$.

In Sections 2.1 and 2.2, we consider three methods of matching that can be used in this design. In Section 2.3, we discuss distance measures and propose in particular a distance measure based on propensity score matching, and in Section 2.4, we discuss an integer programming formulation of the tripartite matching problem.

2.1. *Incomplete Block Design with Disjoint Pairs*

If the three treatment comparisons, A with B, A with C, and B with C, are equally important, then Bo and Rosenbaum (2004) suggest an incomplete block design with matched pairs; thus they propose blocks of size two with one third of the $N/2$ blocks assigned at random to each of the three treatment comparisons. Given sets $\mathcal{A}, B, C$, we consider the collection $P_{\mathcal{A},B,C}$ of all possible matchings of size (p_{12}, p_{13}, p_{23}), for which there are p_{12} pairs of the form $\mathcal{A} \times \mathcal{B}$, p_{13} pairs of the form $\mathcal{A} \times \mathcal{C}$ and p_{23} pairs of the form $\mathcal{B} \times \mathcal{C}$, where all the pairs are disjoint and p_{12}, p_{13}, p_{23} are specified constants. In this $\mathcal{A} \times \mathcal{B}$ is the Cartesian product set $\{(i, j) : i \in \mathcal{A}, j \in \mathcal{B}\}$. Let

$\omega \in \mathcal{M}$ be one pair of a particular matching $\mathcal{M} \in P_{\mathcal{A},\mathcal{B},\mathcal{C}}$, we measure the quality of $\mathcal{M}$ as $\Delta_{\mathcal{M}} = \sum_{\omega\in\mathcal{M}} D(\omega)$. An optimal incomplete block matching minimizes $\Delta_{\mathcal{M}}$ over all $\mathcal{M} \in P_{\mathcal{A},\mathcal{B},\mathcal{C}}$. If this minimum is finite, then the optimal matching problem is *feasible*; otherwise, it is *infeasible*. In our application of a randomized experiment, we select $p_{12} = p_{13} = p_{23} = N/6$.

Bo and Rosenbaum (2004) show that the solution to this optimal matching problem can be related to the solution of an equivalent nonbipartite matching problem that is easily solved. Let Φ denote the collection of all possible matchings of the set $\mathcal{A}\bigcup\mathcal{B}\bigcup\mathcal{C} = \{1, 2, \ldots, N\}$ into $N/2$ disjoint pairs, and define the distance measure,

$$D^*(i, j) = \begin{cases} D(i, j), & \text{if } Z_i \neq Z_j; \\ +\infty, & \text{if } Z_i = Z_j, \end{cases}$$

where $i, j = 1, \ldots, N$. Let $\mathcal{M} \in \Phi$ be a matching. The total distance of $\mathcal{M}$ is defined as $\Delta_{\mathcal{M}} = \sum_{\omega\in\mathcal{M}} D^*(\omega)$. The optimal nonbipartite matching problem is to minimize $\Delta_{\mathcal{M}}$ over Φ. A nonbipartite matching $\mathcal{M}$ is called *feasible*, if $\Delta_{\mathcal{M}} < \infty$. Bo and Rosenbaum (2004) (*Claim 1*) proved that P is an optimal nonbipartite matching with $\Delta(P) < +\infty$ if and only if P is also an optimal, feasible tripartite matching into incomplete blocks as described above.

2.2. *Tripartite Matchings with Matched Triples*

We propose matching three groups in triples instead of balanced incomplete blocks, which has the advantage of a direct comparison of all treatments within each block. The optimal tripartite matching is very difficult to obtain and is commented on further in Section 2.4. However, we suggest an alternative approach that builds on the bipartite matchings described in Xu and Kalbfleisch (2010) and yields nearly optimal matched triples.

Suppose first that the comparisons A with B, A with C, and B with C are equally important. Given sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$, we consider the collection $\Phi_{\mathcal{A},\mathcal{B}}$ of all possible matchings of size m_{12} from $\mathcal{A} \times \mathcal{B}$, $\Phi_{\mathcal{A},\mathcal{C}}$ of matchings of size m_{13} from $\mathcal{A} \times \mathcal{C}$, and $\Phi_{\mathcal{B},\mathcal{C}}$ of matchings of size m_{23} from $\mathcal{B} \times \mathcal{C}$. In this, m_{12}, m_{13}, and m_{23} are fixed; in the present application with $N/3$ in each group, we consider $m_{12} = m_{13} = m_{23} = N/3$.

The quality of $\mathcal{M} \in \Phi_{A,B}$ is measured by the total distance, $\Delta(\mathcal{M}) = \sum_{\omega\in\mathcal{M}} D(\omega)$. An optimal pair matching corresponds to the minimum distance $\Delta^*_{\mathcal{A},\mathcal{B}} = \min_{\mathcal{M}\in\Phi_{\mathcal{A},\mathcal{B}}} \Delta(\mathcal{M})$. Let $\mathcal{M}_{\mathcal{A},\mathcal{B}}$ be an optimal matching so that $\Delta^*_{\mathcal{A},\mathcal{B}} = \Delta(\mathcal{M}_{\mathcal{A},\mathcal{B}})$. Similarly, by considering the set of all matchings in $\Phi_{\mathcal{A},\mathcal{C}}$ and $\Phi_{\mathcal{B},\mathcal{C}}$, we can determine their respective optimal matchings and corresponding minimum distances; in an obvious notation, this gives $\Delta^*_{\mathcal{A},\mathcal{C}} = \Delta(\mathcal{M}_{\mathcal{A},\mathcal{C}})$ and $\Delta^*_{\mathcal{B},\mathcal{C}} = \Delta(\mathcal{M}_{\mathcal{B},\mathcal{C}})$, respectively. Given the optimal pair matchings, $\mathcal{M}_{\mathcal{A},\mathcal{B}}$ and $\mathcal{M}_{\mathcal{B},\mathcal{C}}$, the subjects in A are also paired with those in C through their individual matchings with B. Let $\mathcal{M}^+_{\mathcal{A},\mathcal{C}}$ represent this induced matching. In this case, we refer to B as the reference group; the corresponding tripartite matching is denoted by $\mathcal{M}_{\mathcal{B}}$ and the minimum total distance with B as reference is $\Delta^*_{\mathcal{M}_{\mathcal{B}}} = \Delta^*_{\mathcal{A},\mathcal{B}} + \Delta^*_{\mathcal{B},\mathcal{C}} + \sum_{\omega\in\mathcal{M}^+_{\mathcal{A},\mathcal{C}}} D(\omega)$. Similarly, with A and C as reference groups, the optimal tripartite matchings are denoted by $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{C}}$ with respective minimum total distances, $\Delta^*_{\mathcal{M}_{\mathcal{A}}} = \Delta^*_{\mathcal{A},\mathcal{C}} + \Delta^*_{\mathcal{A},\mathcal{B}} + \sum_{\omega\in\mathcal{M}^+_{\mathcal{B},\mathcal{C}}} D(\omega)$ and $\Delta^*_{\mathcal{M}_{\mathcal{C}}} = \Delta^*_{\mathcal{M}_{\mathcal{B},\mathcal{C}}} + \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} + \sum_{\omega\in\mathcal{M}^+_{\mathcal{A},\mathcal{B}}} D(\omega)$. The reference group associated with the smallest total distance, $\min\{\Delta^*_{\mathcal{M}_{\mathcal{A}}}, \Delta^*_{\mathcal{M}_{\mathcal{B}}}, \Delta^*_{\mathcal{M}_{\mathcal{C}}}\}$ is the *optimal reference group*, and the corresponding matching is referred to as the *optimal symmetric tripartite matching*.

Sometimes investigators may have two or more treatment options to compare against a common control. In this case, the matching can focus on balancing the covariate differences between each treatment group and the control. Suppose that group C is the control or reference group. As defined above, we can determine the optimal pair matchings $M_{\mathcal{A},\mathcal{C}}$ and $\mathcal{M}_{\mathcal{B},\mathcal{C}}$; these can be combined to obtain the *optimal asymmetric tripartite matching given the reference group* C, which minimizes the distance measure

$$\Delta^*_{\mathcal{M}_{\mathcal{C}}} = \Delta^*_{\mathcal{M}_{\mathcal{A},\mathcal{C}}} + \Delta^*_{\mathcal{M}_{\mathcal{B},\mathcal{C}}}. \tag{1}$$

Note that the distance between the corresponding members of A and B, $\sum_{\omega\in\mathcal{M}^+_{\mathcal{A},\mathcal{B}}} D(\omega)$, is not included in (1) since the adjustment of the covariate imbalance between the two treatment groups A and B is not of primary interest.

2.3. *Choice of Distance Measure*

The matchings described above can be carried out using any measure of distance between the subjects available for the experiment. In our presentations here, we make use of a propensity score distance that is a generalization of that used in Xu and Kalbfleisch (2010). Other measures of distance (e.g., Mahalanobis distance) could also be used.

The method of propensity score matching has been widely used in observational studies to control for bias (Rosenbaum and Rubin, 1984; Gu and Rosenbaum, 1993; Ming and Rosenbaum, 2000; Rosenbaum, 2002; Hansen, 2004). Rosenbaum and Rubin (1984) proved that treatment assignment and the observed covariates are conditionally independent given the propensity score, which implies that adjustment for the scalar propensity score is sufficient to remove bias due to all observed covariates. Although the true propensity score is typically known from the randomization scheme in randomized experiments, matching on the estimated score may still have some substantial advantages. Indeed, it has been shown that matching based on an estimated propensity score has advantage over the use of the true propensity score (Robins, Mark, and Newey, 1992).

Again, let Z_i, $i = 1, \ldots, n$ be the random assignment of the N subjects to the three treatment groups. Although we know that $Pr(Z_i = t) = 1/3, t = 1, 2, 3$, is independent of the $\mathbf{X}_i$, we can nonetheless consider the extent to which $\mathbf{X}_i$ is predictive of the observed randomization $z = (z_1, \ldots, z_N)$. For this purpose, we consider a baseline category model

$$\begin{aligned} \delta_{it} &= Pr(Z_i = t \mid \mathbf{X}_i) \\ &= \frac{\exp\{\boldsymbol{\alpha}_t^{\mathrm{T}}\mathbf{X}_i\}}{1 + \exp\{\boldsymbol{\alpha}_1^{\mathrm{T}}\mathbf{X}_i\} + \exp\{\boldsymbol{\alpha}_2^{\mathrm{T}}\mathbf{X}_i\}}, \end{aligned} \tag{2}$$

where $t = 1, 2, 3$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1r})^{\mathrm{T}}$, and $\boldsymbol{\alpha}_2 = (\alpha_{21}, \ldots, \alpha_{2r})^{\mathrm{T}}$ are regression coefficients, and $\boldsymbol{\alpha}_3 = \mathbf{0}$, so that the third group is regarded as the reference. This model applied to the observed randomization yields estimates

$\hat{\boldsymbol{\alpha}}_1$, $\hat{\boldsymbol{\alpha}}_2$ and a corresponding estimate, $\hat{\delta}_{it}$, of the propensity score that individual i is assigned to group t, $t = 1, 2, 3$. The Euclidean distance between the estimated propensity scores for subject i and subject j is

$$D(i, j) = \sqrt{(\hat{\delta}_{i1} - \hat{\delta}_{j1})^2 + (\hat{\delta}_{i2} - \hat{\delta}_{j2})^2 + (\hat{\delta}_{i3} - \hat{\delta}_{j3})^2}, \quad (3)$$

and we refer to this as the generalized propensity score distance.

2.4. *Integer Programming Formulation of the Tripartite Matching*

The asymmetric tripartite matching design given a reference group is an optimal design since both component parts are optimal and this design can be computed efficiently. The symmetric tripartite matching, however, is typically not an optimal solution, but rather a solution that is approximately optimum. In fact, the problem that we are attempting to solve can be simply formulated as an integer programming problem, and such problems have been extensively studied. Let $D_{ijk} = D(i, j) + D(j, k) + D(i, k)$ and $W_{ijk} = 1$ if ijk is a block in the design and $W_{ijk} = 0$ otherwise. Under the constraints that each individual in every set is matched once and only once, the optimal tripartite matching of the sets $\mathcal{A}, \mathcal{B}, \mathcal{C}$ into blocks of three corresponds to the solution of the following problem:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}}\sum_{k\in\mathcal{C}} W_{ijk} D_{ijk}, \\
\text{Subject to} \quad & \sum_{i\in\mathcal{A}}\sum_{j\in\mathcal{B}} W_{ijk} = 1, \ \ k \in \mathcal{C}; \\
& \sum_{i\in\mathcal{A}}\sum_{k\in\mathcal{C}} W_{ijk} = 1, \ j \in \mathcal{B}; \\
& \sum_{j\in\mathcal{B}}\sum_{k\in\mathcal{C}} W_{ijk} = 1 \ i \in \mathcal{A};
\end{aligned}$$

where $W_{ijk} \in \{0, 1\}$ for all i, j, k. This very simple mathematical formulation is deceptive, however, in that its solution is computationally very difficult unless N is very small. A discussion of this and related problems can be found, for example, in Bandelt, Crama, and Spieksma (1994) and Burkard, Dell'Amico, and Martello (2009).

There are a number of commercial software packages that address integer programming problems such as this, and these also typically lead to approximate solutions. We compare this alternative approach with the symmetric tripartite matching in Section 4.

3. Assessment of the BMW Designs

For assessment purposes, we consider the following linear model: let Y represent the response and suppose that conditional on a given treatment assignment Z and an r-vector of covariates $\mathbf{X}$,

$$Y = \alpha + \beta_1 I(Z = 1) + \beta_2 I(Z = 2) + \gamma^{\mathrm{T}}\mathbf{X} + \varepsilon, \quad (4)$$

where $I(\cdot)$ is the indicator function, β_1 and β_2 are the treatment effects with $Z = 3$ as reference, $\gamma = (\gamma_1, \ldots, \gamma_r)^{\mathrm{T}}$ represents the confounding effects, $\varepsilon \sim N(0, \sigma^2)$ is the error, and $\mathrm{cov}(\mathbf{X}) = \Sigma$. It is further assumed that X and ε are mutually independent.

(1) *Pooled Samples*. Under model (4), the common treatment effect estimators obtained from the completely randomized design and ignoring X are $\widehat{\beta}_{1,\mathrm{pool}} = \overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{C}}$, $\widehat{\beta}_{2,\mathrm{pool}} = \overline{y}_{\mathcal{B}} - \overline{y}_{\mathcal{C}}$ and $\widehat{\beta}_{3,\mathrm{pool}} = \overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{B}}$, respectively, where $\overline{y}_{\mathcal{G}} = 3 \times \sum_{i\in\mathcal{G}} y_i/N$ for $\mathcal{G} = \mathcal{A}, \mathcal{B}, \mathcal{C}$. The conditional expectation for $\widehat{\beta}_{1,\mathrm{pool}}$ is

$$E[\widehat{\beta}_{1,\mathrm{pool}}|\mathcal{H}] = \beta_1 + \gamma^{\mathrm{T}}(\overline{\mathbf{X}}_{\mathcal{A}} - \overline{\mathbf{X}}_{\mathcal{C}}), \quad (5)$$

where $\mathcal{H} = \{X_i, Z_i, i = 1, 2, \ldots, N\}$ and $\overline{X}_{\mathcal{G}} = 3 \times \sum_{i\in\mathcal{G}} X_i/N$ for $\mathcal{G} = \mathcal{A}, \mathcal{B}, \mathcal{C}$. The mean squared error for $\widehat{\beta}_{1,\mathrm{pool}}$ is

$$\mathrm{MSE}(\widehat{\beta}_{1,\mathrm{pool}}|\mathcal{H}) = \gamma^{\mathrm{T}}(\overline{\mathbf{X}}_{\mathcal{A}} - \overline{\mathbf{X}}_{\mathcal{C}})(\overline{\mathbf{X}}_{\mathcal{A}} - \overline{\mathbf{X}}_{\mathcal{C}})^{\mathrm{T}}\gamma + \frac{6\sigma^2}{N}, \quad (6)$$

$$\begin{aligned}
\mathrm{MSE}(\widehat{\beta}_{1,\mathrm{pool}}) &= \gamma^{\mathrm{T}}\mathrm{Cov}(\overline{\mathbf{X}}_{\mathcal{A}} - \overline{\mathbf{X}}_{\mathcal{C}})\gamma + \frac{6\sigma^2}{N} \\
&= \frac{6}{N}\gamma^{\mathrm{T}}\Sigma\gamma + \frac{6}{N}\sigma^2. \quad (7)
\end{aligned}$$

From (7), the MSE is comprised of two parts: the first is due to the conditional bias in (6) arising from the imbalance in the observed covariates $\mathbf{X}$; and the second is the conditional variance of $\widehat{\beta}_{1,\mathrm{pool}}$. The properties of the estimator $\widehat{\beta}_{2,\mathrm{pool}}$ and $\widehat{\beta}_{3,\mathrm{pool}}$ are similar.

(2) *Matched Samples*. Under model (4), estimating the treatment effect for the matched sample involves computation of the average of the within-pair differences.

(i) The incomplete block (ICB) design results in a set $\mathcal{M}_{13}$ of $N/6$ pairs from $\mathcal{A} \times \mathcal{C}$, $\mathcal{M}_{12}$ of $N/6$ pairs from $\mathcal{A} \times \mathcal{B}$ and $\mathcal{M}_{23}$ of $N/6$ pairs from $\mathcal{B} \times \mathcal{C}$. Let $\overline{y}_{\mathcal{A}13} = 6 \times \sum_{(i,j)\in\mathcal{M}_{13}} y_i/N$ and $\overline{y}_{\mathcal{C}13} = 6 \times \sum_{(i,j)\in\mathcal{M}_{13}} y_j/N$ be respectively the mean response of $\mathcal{A}$ treated and $\mathcal{C}$ treated subjects in the matching $\mathcal{M}_{13}$. Then, in an obvious notation, the treatment effect estimator of $\mathcal{A}$ versus $\mathcal{C}$ is $\hat{\beta}_1^{\mathrm{ICB}} = \frac{2}{3}(\overline{y}_{A13} - \overline{y}_{C13}) + \frac{1}{3}[(\overline{y}_{A12} - \overline{y}_{B12}) + (\overline{y}_{B23} - \overline{y}_{C23})]$, which has conditional expectation

$$\begin{aligned}
E[\widehat{\beta}_1^{\mathrm{ICB}}|\mathcal{H}] = \beta_1 + \frac{1}{3}\gamma^{\mathrm{T}}[&2(\overline{\mathbf{X}}_{A13} - \overline{\mathbf{X}}_{C13}) \\
&+ (\overline{\mathbf{X}}_{A12} - \overline{\mathbf{X}}_{B12}) + (\overline{\mathbf{X}}_{B23} - \overline{\mathbf{X}}_{C23})],
\end{aligned}$$

where $\overline{\mathbf{X}}_{A13} = 6\sum_{(i,j)\in\mathcal{M}_{13}} \mathbf{X}_i/N$, is the average of the covariates of subjects in treatment group $\mathcal{A}$ in the matching $\mathcal{M}_{13}$, etc. The mean squared error is

$$\begin{aligned}
\mathrm{MSE}(\widehat{\beta}_1^{\mathrm{ICB}}) = \frac{1}{9}\gamma^{\mathrm{T}}\mathrm{Cov}^*[&2(\overline{\mathbf{X}}_{A13} - \overline{\mathbf{X}}_{C13}) + (\overline{\mathbf{X}}_{A12} - \overline{\mathbf{X}}_{B12}) \\
&+ (\overline{\mathbf{X}}_{B23} - \overline{\mathbf{X}}_{C23})]\gamma + \frac{8\sigma^2}{N}. \quad (8)
\end{aligned}$$

The estimators of $\hat{\beta}_2^{\mathrm{ICB}}$ and $\hat{\beta}_3^{\mathrm{ICB}}$ have similar properties. Note that "Cov*" refers to the covariance matrix under the BMW sampling scheme for the ICB design.

(ii) The asymmetric and symmetric tripartite matching design (ATM and STM) with matched triples lead to a set $\mathcal{M}_{123}$ of $N/3$ matched triples from $\mathcal{A} \times \mathcal{B} \times \mathcal{C}$. As before, let $\overline{y}_{\mathcal{G}} = 3 \times \sum_{i \in \mathcal{G}} y_i/N$ for $\mathcal{G} = \mathcal{A}, \mathcal{B}, \mathcal{C}$. The treatment effect estimator comparing A with C from the asymmetric and symmetric designs are $\widehat{\beta}_1^{\mathrm{ATM}}$ and $\widehat{\beta}_1^{\mathrm{STM}}$, respectively, where $\widehat{\beta}_1^{\mathrm{ATM}} = \widehat{\beta}_1^{\mathrm{STM}} = \overline{y}_{\mathcal{A}} - \overline{y}_{\mathcal{C}}$ and

$$\mathrm{MSE}(\widehat{\beta}_1^{\mathrm{STM}}) = \gamma^{\mathrm{T}} \mathrm{Cov}^*(\overline{\mathbf{X}}_{\mathcal{A}} - \overline{\mathbf{X}}_{\mathcal{C}})\gamma + \frac{6\sigma^2}{N}, \quad (9)$$

with a similar equation for $\mathrm{MSE}(\widehat{\beta}_1^{\mathrm{ATM}})$.

The formulas (7)–(9) are quite similar. The calculations in (7), however, are straightforward, whereas for (8) and (9), because of Cov*, the expressions must be evaluated by simulation as in the next section. The BMW design essentially chooses a randomization and post randomization stratification in order to achieve good balance between the treatment groups. In doing this, it reduces the bias in the conditional treatment effect estimators and, as a consequence, reduces the bias term in the mean squared error. Note that with the ATM and STM designs, the variance term in the MSE is still $6\sigma^2/N$ as in (9). The ICB design, on the other hand, reduces the bias of the estimator but, as (8) shows, the incomplete blocking has the effect of increasing the variance term to $8\sigma^2/N$. Therefore, it is to be expected that the ICB design can be less efficient than the pooled design when the confounding effects are small. The marginal MSEs for these estimators are evaluated by simulation in the next section.

4. Simulation Results

In order to assess the performance of the BMW design based on each of these three matching algorithms, we compared them to one another and to the completely randomized design in a simulation study. In doing so, we considered a wide variety of settings and, for each setting, estimated the mean squared error of treatment effect estimators based on 1000 replications.

4.1. *Structure of the Simulation and Results*

For the ith subject, we generated a set of r covariates $\mathbf{X}_i = (X_{i1}, \ldots, X_{ir})^{\mathrm{T}}$, $i = 1, 2, \ldots, N$, where the covariates were drawn independently from various distributions as described below. Given a randomization of $N/3$ subjects to each of the three treatment groups, the responses were generated from the model (4) with $\alpha = 0$ and $\varepsilon_i \sim \mathcal{N}(0, 1)$. In the simulations, we considered the following:

- $\beta_1 = \beta_2 = 0.5$. Note that the results do not depend on the choice of β.
- $\gamma_j = \gamma$, $j = 1, \ldots, r$, where $\gamma = 0.5, 1.0, 1.5$. If the covariates follow symmetric distributions, the results do not depend on the signs of the components of γ.
- $r = 4$ where: (i) $X_1, X_2, X_3, X_4 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5) ; or (ii) $X_1, X_2 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5) ; $X_3, X_4 \overset{\text{i.i.d}}{\sim} N(0, 0.25)$.
- $N = 24$ or 36, and $M = 100$.

Table 1 presents the MSEs for the ATM, STM, and ICB designs. Since the STM and ICB designs are symmetric in all treatment effect estimators, we only show the percent reduction in MSE in β_1 for those methods. With the ATM design, however, there is symmetry in the estimates of β_1 and β_2, but the estimate of the contrast, $\beta_{AB} = \beta_1 - \beta_2$, is different. Therefore, we report results for $\hat{\beta}_1$ and $\hat{\beta}_{AB}$.

The three-arm BMW designs typically provide important gains in efficiency by reducing the MSE of the treatment effect estimators as compared to the completely randomized design. This is especially so for the STM and ATM designs.

The BMW designs based on ICB or STM can be applied to situations when the three pairwise comparisons are deemed equally important. The STM design, however, is substantially more effective than the ICB design in reducing MSE. This is especially the case when the confounding effects are moderate (e.g., $\gamma = 0.5$ or 1.0). For example, if $N = 36$ and the common confounding effect is $\gamma = 0.5$, the estimated treatment effect, $\hat{\beta}_1$, from the ICB design is less efficient (reduction in MSE $= -12.3\%$) than the completely randomized design. The extra blocks created by the ICB design lead to a loss of efficiency, and this limitation arises especially when the confounding effects are not too strong. In this case, the increase in variance outweighs the bias reduction. On the other hand, our proposed symmetric tripartite matching algorithm utilizes all the subjects assigned to each treatment group in the direct comparisons and, at the same time, reduces the bias by efficiently matching the subjects into triples.

Simulation results for the ATM design are similar to those for the STM design. As expected, we observe some gain in efficiency in estimating the AC and BC comparisons as compared to the STM though, as Table 1 illustrates, the gains are relatively small. Similarly there are relatively small reductions in the efficiency of estimating the AB comparison.

The effects of study size N on the performance of the BMW design is examined in Table 1, which presents the results $N = 36$ and $N = 24$. These reveal that the performance of the ICB design is poorer with smaller sample size, whereas the STM and ATM designs have similar efficiency gains for the sample sizes considered.

In the implementation of the BMW design, we have used $M = 100$. To gain some insight as to whether this choice is adequately large, we carried out a series of simulations in which we first generated Y_i, X_i for $N = 24$ subjects based on the model (4) with $r = 1$, $X \sim N(0, 0.25)$ and $\varepsilon \sim N(0, 1)$. We then applied the BMW design with $M = 100$ repeatedly for 100 times on these same 24 subjects. Each time, we obtained a new "optimal" symmetric tripartite matching and for each such matching, we evaluated the covariate imbalance $\bar{X}_A - \bar{X}_C$, $\bar{X}_B - \bar{X}_C$, and $\bar{X}_A - \bar{X}_B$, at the optimum. These variables have mean 0, and the standard deviations were 7.33E−05, 6.95E−05, and 5.33E−05, respectively. As indicated in section 3, the BMW design reduces the MSE of the treatment effect estimator through a reduction in the conditional bias of the estimator. We repeated these calculations many times and

Table 1

Percent reductions in the MSE of treatment effect estimator for the BMW designs based on incomplete blocks(ICB), symmetric tripartite matching (STM), and asymmetric tripartite matching (ATM), and compared to a completely randomized design (CR)

			MSE percent reduction (%)			
			ICB vs CR	STM vs CR	ATM vs CR	
γ	M	MSE (CR)	$\hat{\beta}_1 = \hat{\beta}_{AC}$	$\hat{\beta}_1 = \hat{\beta}_{AC}$	$\hat{\beta}_1$ or $\hat{\beta}_2$	$\hat{\beta}_{AB} = \hat{\beta}_1 - \hat{\beta}_2$
$N = 24$						
			$X_1, X_2, X_3, X_4 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5)			
(0.5, 0.5, 0.5, 0.5)	100	0.312	−11.95	15.52	15.23	15.42
(1.0, 1.0, 1.0, 1.0)	100	0.487	18.05	37.02	38.18	34.58
(1.5, 1.5, 1.5, 1.5)	100	0.806	40.20	53.61	55.56	47.96
			$X_1, X_2 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{\text{i.i.d}}{\sim} N(0, 0.25)$			
(0.5, 0.5, 0.5, 0.5)	100	0.288	−19.11	10.12	10.36	9.14
(1.0, 1.0, 1.0, 1.0)	100	0.403	7.11	28.74	29.38	27.28
(1.5, 1.5, 1.5, 1.5)	100	0.600	29.24	44.37	45.44	42.23
$N = 36$						
			$X_1, X_2, X_3, X_4 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5)			
(0.5, 0.5, 0.5, 0.5)	100	0.207	−12.34	15.35	15.89	13.43
(1.0, 1.0, 1.0, 1.0)	100	0.326	18.95	37.61	38.45	36.18
(1.5, 1.5, 1.5, 1.5)	100	0.535	40.93	54.63	55.52	51.39
			$X_1, X_2 \overset{\text{i.i.d}}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{\text{i.i.d}}{\sim} N(0, 0.25)$			
(0.5, 0.5, 0.5, 0.5)	100	0.193	−18.17	10.64	10.99	9.59
(1.0, 1.0, 1.0, 1.0)	100	0.268	8.38	28.53	29.37	26.71
(1.5, 1.5, 1.5, 1.5)	100	0.417	33.95	47.11	48.53	43.57

Sample size $N = 24$ and 36 subjects. Number of replications = 1000.

found that this summary is typical of the results seen. This suggests that $M = 100$ is large enough for implementing the BMW design in the cases considered here.

4.2. *Evaluating STM*

In order to evaluate how close STM is to the true optimal tripartite matching, we performed a simulation study to compare them for a small sample size of $18 = 3 \times 6$, for which an exact solution can be obtained. We generated the response from

$$Y_i = \beta_1 I(Z_i = 1) + \beta_2 I(Z_i = 2) + \gamma X_i + \varepsilon_i, \quad i = 1, 2, \ldots, 18, \tag{10}$$

where $X_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 0.25)$ and $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. We used a dynamic programming algorithm in Matlab to search for the optimal tripartite matching. Even with this relatively small sample size, the calculations to obtain the true optimum matching is extensive, taking approximately 10 minutes for a single CPU to search for the optimal solution on each simulation. We used a cluster of CPUs to conduct 1000 simulations and for each simulated sample, we apply the STM with $M = 100$ and true optimal tripartite matching method, respectively.

The minimum distances obtained by the STM and optimal tripartite matching are very close. The distribution of the difference in minimum distances obtained from the two algorithms based on the 1000 simulation runs has a maximum of 2.73×10^{-02}, mean of 8.56×10^{-04}, median of 2.51×10^{-06}, minimum of 0. The average MSE of the treatment effect estimator is 0.33 for both approaches. This suggests that the proposed STM design leads to results that are nearly optimal, at least for the sample size $N = 18$. More extensive simulations with larger sample sizes would be interesting, but would require more efficient methods than those currently available to obtain the true optimum matching.

As an alternative to STM, we also investigated the use of the commercial optimization software, Gurobi 5.0 (Bixby, Gu, and Rothberg, 2010), to solve the integer programming formulation as described in Section 2.4. It should be noted that Gurobi and other such software only gives an approximate solution to such problems, again because of the computational burden of obtaining exact results. We compared STM with the solution that Gurobi gave in 100 examples, where each example has $N = 24$. With the default searching algorithm and default stopping criterion specified in Gurobi, the software leads to very similar results to those obtained from STM. In 58 of the 100 examples, STM with $M = 100$ yields smaller total distances than Gurobi. The distribution of the difference in minimum distances obtained from the two algorithms based on the 100 examples has a maximum of 0.46, a mean and median of −0.013, and minimum of −0.47. This investigation suggests that the STM leads to competitive solutions and has the clear advantage that it does not require specialized software.

Table 2

Percent reductions in the MSE of the treatment effect estimator for the symmetric quadripartite matching (SQM) and asymmetric quadripartite matching (AQM), compared to a completely randomized design (CR) for a 2×2 factorial experiment

			% MSE reduction (MSE)		
		MSE (CR)	SQM versus CR	AQM versus CR	
γ	M	$\hat{\beta}_2$	$\hat{\beta}_2$	$\hat{\beta}_1$	$\hat{\beta}_2$
(0.5, 0.5, 0.5, 0.5)	100	0.116	9.2% (0.105)	8.8% (0.106)	8.9% (0.105)
(1.0, 1.0, 1.0, 1.0)	100	0.166	26.7% (0.122)	28.6% (0.118)	25.4% (0.120)
(1.5, 1.5, 1.5, 1.5)	100	0.245	39.6% (0.148)	40.8% (0.145)	39.7% (0.147)

Sample size $N = 40$ subjects. Number of replications = 1000. We consider four confounding covariates, X_1, X_2, X_3, and X_4, where $X_1, X_2 \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(0.5); X_3, X_4 \overset{\text{i.i.d}}{\sim} N(0, 0.25)$.

5. The BMW Design in Trials with Four or More Arms

The BMW design can be further extended to trials with four or more arms. With four arms, we can apply an asymmetric quadripartite matching (AQM) design that generalizes the ATM design when there is a predefined reference group; alternatively, the STM design can be extended to a symmetric quadripartite matching (SQM) design when these four groups are to be compared in a symmetric way. On the other hand, Bo and Rosenbaum (2004) noted that the nonbipartite matching algorithm that gives rise to the incomplete block design for three arms does not extend to four or more treatments.

Traditional designs in cluster randomized trials have generally studied the effect of one variable at a time. In many instances, however, there may exist two or more factors, and then it is inefficient to study each variable individually. Table 2 summarizes a simulation study for a 2×2 factorial design (four treatment groups) with $N = 40$ subjects to compare treatment E with its control $\bar{E}$ and treatment F with its control $\bar{F}$. The model generating the data is additive with no interaction between the treatments and β_1 and β_2 represent the respective treatment main effects; the additive error term is $N(0, 1)$. Each point in the simulation corresponds to 1000 replications. Both the SQM design and the AQM design are considered; in the AQM, the group $\bar{E}\bar{F}$ is taken as the reference. There were four covariates with distributions defined in Table 2.

6. Planning a 2×2 Factorial Design for tPA Usage in Stroke

In this section, we consider the use of the generalized BMW design in planning a hypothetical cluster randomized trial that would be a follow up to the INSTINCT trial discussed in Section 1. This trial would be, designed to investigate the effectiveness of an educational program and a promotional campaign administered to hospital emergency departments with the aim of enhancing the appropriate use of tPA therapy for ischemic stroke patients. With treatments arranged in a factorial structure, this would provide a confirmatory study for the original trial with respect to the education program, but also allow a separate investigation of the second factor. Twenty four hospitals are recruited and are to be randomized to four experimental groups EC, $E\bar{C}$, $\bar{E}C$, $\bar{E}\bar{C}$, where E refers to the education program and C refers to the promotional campaign. The primary outcome is the frequency of appropriate tPA use in each hospital. The cluster-level factors thought to be strongly associated with the outcome consist of stroke volume (low vs. high), population density (urban vs. rural) as well as age and gender mix. We choose $M = 100$ in proposing a design for the tPA study.

We randomly assigned the 24 hospitals to the four experimental groups EC, $E\bar{C}$, $\bar{E}C$,$\bar{E}\bar{C}$, and estimate the sample-based probability of being assigned to each group for every hospital. The hospitals were then optimally matched by using the asymmetric quadripartite algorithm with group $\bar{E}\bar{C}$ as the predefined reference group, which gave a minimum total Euclidean distance of 4.65. We then randomized the hospitals an additional 99 times and recorded the minimum distance measures for each time. The 86th randomization produced the smallest distance as $\Delta^*_{\mathcal{M}_{\bar{E}\bar{C}}} = \Delta^*_{\mathcal{M}_{EC,\bar{E}\bar{C}}} + \Delta^*_{\mathcal{M}_{E\bar{C},\bar{E}\bar{C}}} + \Delta^*_{\mathcal{M}_{\bar{E}C,\bar{E}\bar{C}}} = 1.80$. The corresponding BMW design is presented in Table 3, showing the matches of each group with the control group $\bar{E}\bar{C}$. This leads to six quadruples assigning subjects to treatments, the first group being 5, 6, 2, 17, the second being 9, 20, 3, 18, etc.

When the comparisons among the four groups are equally important, we might adopt the symmetric quadripartite matching algorithms to assign the 24 hospitals to the experimental groups and search for the optimal solution of the quadripartite matching with respect to the optimal reference group for each randomization. This process was repeated 100 times and among those, the 55th randomization produced the smallest total Euclidean distance $\Delta^*_{\mathcal{M}} = \Delta^*_{\mathcal{M}_{EC,\bar{E}\bar{C}}} + \Delta^*_{\mathcal{M}_{E\bar{C},\bar{E}\bar{C}}} + \Delta^*_{\mathcal{M}_{\bar{E}\bar{C},\bar{E}C}} + \sum_{\omega \in \mathcal{M}^+_{EC,E\bar{C}}} D(\omega) + \sum_{\omega \in \mathcal{M}^+_{\bar{E}\bar{C},E\bar{C}}} D(\omega) + \sum_{\omega \in \mathcal{M}^+_{EC,\bar{E}\bar{C}}} D(\omega) = 4.93$; for that randomization, group $\bar{E}C$ was selected as the optimal reference group. Table 3 also shows the results produced by the BMW design based on quadripartite matching algorithm.

7. Some Comments on Analysis Issues

Although this article is primarily about design, it is useful to include some comments on the statistical analysis. There is much literature that proposes the use of the randomization distribution as opposed to parametric (e.g., Gaussian)

Table 3
Optimal matched blocks for the asymmetric and symmetric quadripartite matching for the 2×2 *factorial design in the case study*

Strata	ID	X_1	X_2	X_3	X_4	ID	X_1	X_2	X_3	X_4
Asymmetric quadripartite matching algorithm (AQM)										
			EC					$\bar{E}\bar{C}$		
1	5	0.19	0.13	0	1	17	0.24	0.12	1	0
2	9	0.14	0.06	1	1	18	0.09	0.05	1	1
3	10	0.26	0.18	1	0	8	0.22	0.14	0	0
4	13	0.13	0.09	0	0	15	0.10	0.06	0	0
5	21	0.18	0.14	0	1	16	0.10	0.07	1	1
6	22	0.30	0.17	1	0	7	0.24	0.19	0	1
			$\bar{E}C$					$\bar{E}\bar{C}$		
1	6	0.19	0.07	0	0	17	0.24	0.12	1	0
2	20	0.08	0.06	1	1	18	0.09	0.05	1	1
3	1	0.15	0.13	0	0	8	0.22	0.14	0	0
4	14	0.13	0.07	0	0	15	0.10	0.06	0	0
5	23	0.11	0.07	1	1	16	0.10	0.07	1	1
6	19	0.25	0.15	1	1	7	0.24	0.19	0	1
			$E\bar{C}$					$\bar{E}\bar{C}$		
1	2	0.17	0.11	1	0	17	0.24	0.12	1	0
2	3	0.13	0.06	1	1	18	0.09	0.05	1	1
3	11	0.22	0.14	1	0	8	0.22	0.14	0	0
4	4	0.12	0.06	0	1	15	0.10	0.06	0	0
5	12	0.07	0.06	1	1	16	0.10	0.07	1	1
6	24	0.23	0.16	0	1	7	0.24	0.19	0	1
Symmetric quadripartite matching algorithm (SQM)										
			EC					$E\bar{C}$		
1	3	0.13	0.06	1	1	2	0.17	0.11	1	0
2	5	0.19	0.13	0	1	24	0.23	0.16	0	1
3	8	0.22	0.14	0	0	17	0.24	0.12	1	0
4	14	0.13	0.07	0	0	9	0.14	0.06	1	1
5	16	0.10	0.07	1	1	20	0.08	0.06	1	1
6	21	0.18	0.14	0	1	1	0.15	0.13	0	0
			$\bar{E}C$					$\bar{E}\bar{C}$		
1	11	0.22	0.14	1	0	10	0.26	0.18	1	0
2	4	0.12	0.06	0	1	6	0.19	0.07	0	0
3	22	0.30	0.17	1	0	19	0.25	0.15	1	1
4	15	0.10	0.06	0	0	18	0.09	0.05	1	1
5	23	0.11	0.07	1	1	12	0.07	0.06	1	1
6	7	0.24	0.19	0	1	13	0.13	0.09	0	0

X_1, percent of females greater than 65 years old among all females in the census tract (%); X_2, percent of males greater than 65 years old among all males in the census tract (%); X_3, stroke volume (low vs. high); X_4: population density (urban vs. rural). $M = 100$.

distribution-based analyses. These ideas go back to Fisher (1926). Randomization based procedures are particularly relevant when the experimental units are highly correlated as in agricultural trials or when, as in ESP trials, the modeling of individual responses is difficult or impossible. Repeated randomization, as we propose here, improves the covariate balance across treatment groups and leads to more precise estimates of treatment effects. Traditional Gaussian distribution-based parametric approaches that do not take the repeated randomization into account result in overly conservative statistical inferences. Randomization-based inference, however, is still valid. Morgan and Rubin (2012) also discuss this point.

In what follows, we examine a randomization approach to inference in a simple BMW design with two arms and matching into pairs, so that one member of each pair is in each treatment group. This is not because this simple structure is needed to carry out the tests; randomization tests are generally straightforward to execute. But the nested simulations required to assess the properties of the randomization approach make this simpler framework useful. We also expect that results found in this case would extend at least qualitatively to more complicated situations.

We suppose that there are N units and that the random vector $(W_1, \ldots, W_N)$ denotes the selected randomization with $W_i = 1$ or 0 indicating the assignment of unit i and

let $S_1, \ldots, S_{N/2}$ denote the $N/2$ (random) matched pairs. Let $W_{\text{obs},i}, i = 1, \ldots, N$ and $S_{\text{obs},i}, i = 1, \ldots, N/2$ denote the corresponding observed values. By analogy with the linear model when it looks to be approximately appropriate, we might select the paired t statistic given by

$$T = \frac{\bar{Y}_1 - \bar{Y}_0}{S_d/\sqrt{N/2}},$$

where S_d is the sample variance of the $N/2$ pair differences. Although not strictly necessary, it is convenient to suppose that unit i has two potential outcomes, $y_i(1)$ and $y_i(0)$ depending on whether it is assigned to group 1 or group 0. The observed outcome values are then

$$Y_{\text{obs},i} = y_i(1)W_{\text{obs},i} + y_i(0)(1 - W_{\text{obs},i}) \tag{11}$$

and from this, we can obtain the observed value T_{obs} of the t-statistic.

A randomization test of the sharp null hypothesis of no treatment effect at the individual level (i.e., $y_i(1) = y_i(0)$ for each subject i) can then be obtained. Under this hypothesis, the vector of observed outcomes Y_{obs} remains fixed for every treatment assignment. We create the appropriate reference distribution by repeating the BMW design a large number B times. Each time, the subjects are randomized to two groups $M = 100$ times and the optimal randomization $W_1, \ldots, W_N$ and corresponding pairs are determined; this gives rise to a value of the T statistic. The proportion of the B trials that result in a T statistic that is at least as large as T_{obs} provides an estimate of the exact one tailed p-value corresponding to the randomization test. A two sided p-value is obtained in the usual way by taking twice the minimum tail area.

Size and Power of the randomization test based on the BMW design can be evaluated with simulations. We report the results of such simulations here for treatment effects, $\beta = 0, 0.5, 0.7, 1.0$ in the linear model

$$Y = \beta Z + \gamma_1 X_1 + \gamma_2 X_2 + \varepsilon \tag{12}$$

with confounding variables $X_1 \overset{\text{i.i.d}}{\sim}$ Bernoulli (0.5), $X_2 \overset{\text{i.i.d}}{\sim} N(0, 0.25)$ and random error $\varepsilon \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1)$. We repeat the following steps 1000 times:

(i) Simulate data X_1, X_2, generate 100 randomizations, apply the BMW design to obtain the observed randomization W_{obs} and associated pairs.
(ii) Generate the observed values of Y_β according to the model (12) for each value of β and compute the test statistic $T_{\text{obs}}(\beta)$.
(iii) For each Y_β, carry out the randomization test as described in the previous paragraph and compute the associated p-values, P_β.

For a tests of size 0.05, the estimated power for each value of β is the proportion of values P_β that are less than or equal to 0.05. For comparison, we also perform the randomization test for the completely randomized design and for a variation of the design proposed by Morgan and Rubin (2012), which we denote M&R. In the M&R design, we choose M randomizations and then select the one that gives the smallest Mahalanobis distance between the means of the X variables in the two samples. In addition, the size and power of the standard normal analysis strategies for these three designs are evaluated. The results are summarized in Table 4.

Table 4
Powers of designs and analysis strategies: randomization test for BMW design; paired t-test for BMW; randomization test for CR design; two sample t-test for CR; randomization test for M&R design; two sample t test for M&R design

Type of test	Power of a one-tailed 5% test			
	$\beta = 0.0$	$\beta = 0.5$	$\beta = 0.7$	$\beta = 1.0$
BMW randomization	0.047	0.335	0.533	0.792
BMW paired t test	0.029	0.269	0.458	0.755
CR randomization	0.041	0.258	0.380	0.637
CR 2 sample t	0.047	0.276	0.436	0.670
M&R randomization	0.049	0.377	0.563	0.839
M&R 2 sample t	0.019	0.224	0.403	0.711

By design, the randomization tests based on the BMW design, the completely randomized design and Morgan and Rubin's design all lead to correct one-sided type I error (except for the discreteness) of 5%. When the treatment effect is relatively large (e.g., $\beta = 0.7$ or 1.0), however, the BMW and M&R designs yield substantially larger powers than the completely randomized design. The randomization test based on the M&R design appears slightly more powerful than the BMW design.

In contrast to the randomization test, standard analyses based on the student t are conservative when applied to the BMW and M&R designs that involve repeated randomizations. For example, the paired t-test applied to the BMW design yields a true type one error rate of 2.9%, and the two sample t test for M&R design results in an even more conservative value of 1.9%. On the other hand, the two sample t test produces valid inferences when applied to the completely randomized design. In our experience, the standard parametric procedures are always conservative, but more so with the M&R design than with the BMW design. This is discussed further below.

We can also consider tests of the hypothesis $\beta = \beta_0$ for a given β_0, which we interpret as specifying that $y_i(1) = y_i(0) + \beta_0$ for all i. This can be reduced to the previous case of testing $\beta = 0$ by subtracting β_0 from the observed values of Y in the treatment group $Z = 1$. Having done this, the test proceeds as before and gives rise to a (two tailed) p-value which we denote by $SL(\beta_0)$ with SL standing for "significance level." A 95% confidence interval based on the randomization test is given by $\{\beta_0 : SL(\beta_0) \geq 5\%\}$.

The randomization test based on the BMW design for more than two arms can be conducted in a similar way. Typically, the underlying reference set of the randomization test is very large. For example, if $N = 30$, there are $\binom{30}{15,15} = 1.55 \times 10^8$ or $\binom{30}{10,10,10} = 5.55 \times 10^{12}$ possible results of a randomization

to two or three groups, respectively. In either case, the reference set is very large compared to $M = 100$. It is interesting to note that an 'exact' solution of the optimization problem (with $M = \infty$) would yield a nearly deterministic design if the covariates are continuous and there would not be a sufficiently large reference set to carry out a randomization test.

Some insights into the inaccuracy of the parametric analyses can be obtained through an artificial but instructive example. Suppose that a large number of units are available, all with the model (12) applying, and an experiment is to be conducted that involves selecting N units and assigning them to two treatment groups. In this case, it is possible to achieve near perfect balance in both the M&R and the BMW designs.

The M&R approach. In this case, we can obtain a design in which the treatment and control sample means of the confounding variables are exactly equal. That is $(\bar{X}_1^t, \bar{X}_2^t) = (\bar{X}_1^c, \bar{X}_2^c)$, where superscripts t and c are being used to identify treatment and control groups. In this case, it is easy to see that $\bar{Y}_t - \bar{Y}_c = \bar{\varepsilon}_t - \bar{\varepsilon}_c$ has variance $4\sigma_\varepsilon^2/N$. On the other hand, the usual pooled estimate of the variance used in the two sample t statistic would estimate $4(\gamma_1^2 \text{var}(X_1) + \gamma_2^2 \text{var}(X_2) + \sigma_\varepsilon^2)/N$. The resulting two sample t statistic would be too small by a factor $\sigma_\varepsilon / \left[\gamma_1^2 \text{var}(X_1) + \gamma_2^2 \text{var}(X_2) + \sigma_\varepsilon^2\right]^{0.5}$, resulting in a substantially conservative inference.

The BMW approach. This would lead to a design in which there is a perfect matching of the confounding variables within each pair. In this extreme case, the paired t statistic would be exactly correct.

Of course there will not be an exact matching in either case and the parametric analysis in the BMW design would also be somewhat conservative. It is worth noting, however, that the BMW approach would be expected to be less conservative and would be more robust to mis-specification of the model. In particular, it would lead to a correct parametric inference in the extreme case of this example even if the dependence of the response on the covariates were not linear. The M&R design, on the other hand, could have very poor properties if the relationship is substantially nonlinear.

8. Conclusions and Discussions

It was an observational study with three groups that motivated the incomplete block design of Bo and Rosenbaum (2004), and with some adjustments, the proposed symmetric and asymmetric tripartite matchings might also be used in this context. One aspect of observational studies is that it may not be simple to control the size of the exposure groups and some imbalance may be present. One approach in this case would be to take the smallest group as defining the number of blocks, and use repeated randomizations to select items for matching from the other groups. One might, for example, use full matching techniques that match each subject of the reference group to one or more (say up to two or three) in each of the other groups. Options in this area are currently under investigation.

If one has resources to do a trial on only a small number N of experimental units, but there is a much larger number of units from which to choose, it would be possible to extend these ideas to incorporate first a random selection of N units to be used followed by the BMW design on the chosen units. This process could then be replicated to find a good subset of units to use as a good randomization and matching of them.

Alternative methods have been proposed for adjusting for covariate imbalances. One approach proposed by Greevy et al. (2004) involves matching in order to minimize the total Mahalanobis distance between subjects in the two treatment groups. Xu and Kalbfleisch (2010) compared the performance of the BMW design with this and other approaches in the two-arm trial and found that the BMW design generally outperformed the alternative methods. As noted earlier, the Mahalanobis distance could be used in place of the propensity score distance in the BMW design. It may also be possible to combine discrepancy measures (e.g., propensity based and Mahalanobis) in some way to achieve a better match than either achieves separately.

More work is needed on analysis issues. The randomization distribution certainly presents one option, although it is highly computational. The matched pairs parametric analysis might also be considered if the design has resulted in a fairly close matching of X values. It is conservative, but not severely so, at least in the situations we have investigated. A detailed simulation study would be very useful to examine analysis strategies. It would also be interesting to investigate whether there are simple methods of correcting the parametric analyses to account for the reduced variation induced by the repeated randomizations in the BMW and M&R designs.

One alternative to the repeated randomization approach in the BMW design is to adjust post hoc in a regression model for the effects of important covariates. This works reasonably well provided the model is correct and the number of covariates is not too large. A large advantage of the BMW design is the simplicity of the treatment effect estimator, which is expressed as a difference in two means. It should also be noted that the BMW design retains the advantage of balancing on average over unmeasured confounders since each subject is equally likely to be randomized to each treatment.

The proposed BMW and the M&R designs both involve repeated randomization. The BMW design also makes use of a post randomization matching, which has potential gains in robustness against nonlinear dependence on the covariates as the example at the end of Section 7 indicates. In the context of cluster randomized trials, the blocking of the experimental units is important since the trial can be coordinated within blocks and so, for example, temporal effects can be controlled by entering study units in pairs at the same time.

Acknowledgements

The authors would like to thank Professors Ben Hansen, Douglas Schaubel and Thomas Braun for helpful discussions and Professor Phillip Scott for introducing them to the stroke studies that motivated this work. This work was supported in part by grant (RO1 NS050372) from the National Institute of Neurological Disorders and Stroke and by the University of Michigan Kidney Epidemiology and Cost Center.

The opinions and information in this article are those of the authors, and do not represent the views and/or policies of the U.S. Food and Drug Administration.

References

Bandelt, H., Crama, Y., and Spieksma, F. (1994). Approximation algorithms for multi-dimensional assignment problems with decomposable costs. *Discrete Applied Mathematics* **49**, 25–50.

Bixby, R., Gu, Z., and Rothberg, E. (2010). Gurobi optimization. Available at http://gurobi.com.

Bo, L. and Rosenbaum, P. (2004). Optimal pair matching with two control groups. *Journal of Computational and Graphical Statistics* **13**, 422–434.

Burkard, R., Dell'Amico, M., and Martello, S. (2009). *Assignment Problems*. Philadelphia: Society for Industrial Mathematics.

Campbell, D. (2009). Prospective: Artifact and Control. In *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, Oxford: Oxford University Press 264–286.

Chang, W., Hwang, B., Wang, D., and Wang, J. (1997). Cytogenetic effect of chronic low-dose, low-dose-rate [gamma]-radiation in residents of irradiated buildings. *The Lancet* **350**, 330–333.

Cox, D. R. (2009). Randomization in the design of experiments. *International Statistical Review* **77**, 415–429.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33**, 503–513.

Greevy, R., Lu, B., Silber, J., and Rosenbaum, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5**, 263–275.

Gu, X. S. and Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics* **2**, 405–420.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association* **99**, 609–618.

Ming, K. and Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56**, 118–124.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics* **40**, 1262–1282.

Olsen, S. P. (1997). *Multivariate matching with non-normal covariates in observational studies*. Ph.D. Thesis, University of Pennsylvania, Philadelphia.

Robins, J., Mark, S., and Newey, W. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* **48**, 479–495.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer.

Rosenbaum, P. R. and Rubin, D. B. (1984). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin, D. B. (2008). Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association* **103**, 1350–1353.

Scott, P. A., Meurer, W. J., Frederiksen, S. M., Kalbfleisch, J. D., Xu, Z., Haan, M. N., Silbergleit, R., and Morgenstern, L. B. (2012). A multilevel intervention to increase community hospital use of alteplase for acute stroke (instinct): A cluster-randomised controlled trial. *The Lancet Neurology*. Lancet Neurol. 2013 Feb;12(2):139-48. doi: 10.1016/S1474-4422(12)70311-3.

Seltser, R. and Sartwell, P. (1965). The influence of occupational exposure to radiation on the mortality of American radiologists and other medical specialists. *American Journal of Epidemiology* **81**, 2–22.

Wells, K., Roberts, C., Daniels, S., Hann, D., Clement, V., Reintgen, D., and Cox, C. (1997). Comparison of psychological symptoms of women requesting removal of breast implants with those of breast cancer patients and healthy controls. *Plastic and Reconstructive Surgery* **99**, 680–685.

Xu, Z. and Kalbfleisch, J. (2010). Propensity score matching in randomized clinical trials. *Biometrics* **66**, 813–823.

Received December 2012. *Revised June* 2013.
Accepted June 2013.