Biometrics 66, 813–823
September 2010

DOI: 10.1111/j.1541-0420.2009.01364.x

Propensity Score Matching in Randomized Clinical Trials

Zhenzhen Xu* **and John D. Kalbfleisch****

Department of Biostatistics, University of Michigan, Ann Arbor, Michighan 48109, U.S.A.
**email:* zzxu@umich.edu
***email:* jdkalbfl@umich.edu

Summary. Cluster randomization trials with relatively few clusters have been widely used in recent years for evaluation of health-care strategies. On average, randomized treatment assignment achieves balance in both known and unknown confounding factors between treatment groups, however, in practice investigators can only introduce a small amount of stratification and cannot balance on all the important variables simultaneously. The limitation arises especially when there are many confounding variables in small studies. Such is the case in the *INSTINCT* trial designed to investigate the effectiveness of an education program in enhancing the tPA use in stroke patients. In this article, we introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal matching with constraints technique to a prospective randomized design and aims to minimize the mean squared error (MSE) of the treatment effect estimator. A simulation study shows that, under various confounding scenarios, the BMW design can yield substantial reductions in the MSE for the treatment effect estimator compared to a completely randomized or matched-pair design. The BMW design is also compared with a model-based approach adjusting for the estimated propensity score and Robins-Mark-Newey E-estimation procedure in terms of efficiency and robustness of the treatment effect estimator. These investigations suggest that the BMW design is more robust and usually, although not always, more efficient than either of the approaches. The design is also seen to be robust against heterogeneous error. We illustrate these methods in proposing a design for the *INSTINCT* trial.

Key words: Clustered randomized trial; Experimental design; Optimal full matching; Propensity score matching; Randomization study.

1. Introduction and Motivating Example

Cluster randomized trials have been widely used in the past three decades for the evaluation of health care and educational strategies, in which intact social units are selected as the units of randomization. On average, randomized treatment assignment avoids bias, achieves balance of both known and unknown confounding factors between intervention groups, and provides valid comparisons of competing intervention strategies. There is much literature that discusses design methods for cluster randomizations such as the completely randomized design (Abdeljaber et al., 1991), matched-pair design (COMMIT, 1995), stratified design (Graham et al., 1984), and minimization design (Pocock and Simon, 1975). However, investigators can only introduce a small amount of stratification in practice, which does not ensure balance on all important variables, and post hoc adjustment for many confounders is also problematic. These limitations are particularly important when there are many confounding variables in a small study.

Tissue plasminogen activator (tPA) is a clot-busting drug, which has been found to be an effective treatment for the prevention of post-stroke disability if administered within a three-hour time window of the onset of an ischemic stroke (NINDS, 1995). However, the use of tPA has remained relatively low. A randomized clinical trial, *INSTINCT*, was designed in order to investigate the effectiveness of an education program administered to hospital emergency departments in enhancing tPA therapy for stroke patients. Historical data were collected from 24 participating hospitals in Michigan regarding previous stroke volume and demographic variables. Hospitals were the units of randomization and those assigned to the treatment group received educational interventions designed to promote appropriate tPA use, whereas the other hospitals served as controls. The primary outcome is the frequency of appropriate tPA use in each hospital. Stroke volume at baseline (low versus high), population density (urban versus rural), age, and gender mix are cluster-level factors thought to be strongly associated with outcome. Among these, stroke volume measured as number of stroke discharges and population density were classified as binary. Percentage of female (male) stroke patients who are older than 65 years is used as a continuous measure. It is possible to create balance on stroke volume and population density through stratified randomization, however, it is not feasible to balance on all covariates at the same time. As a result, direct estimation of the treatment effect may be subject to bias due to possible imbalance on confounding factors. To resolve this problem, this article describes and evaluates a new randomization design based on propensity score matching.

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). The propensity score is defined as the conditional probability of a subject being assigned to the treatment group given the observed covariates. Rosenbaum and Rubin (1984) showed that exact matching of treated and control subjects on the propensity score will balance all the observed covariates. In nonrandomized experiments, the propensity score function is always unknown but the sample estimates of the propensity score can be used. On the other hand, in a randomized clinical trial, the true propensity score is often a known function from the randomization scheme. For example, in the simplest randomized trial, subjects are assigned to treatment or control by the flip of a fair coin and the propensity score is equal to one half for all the subjects and the two treatment groups are perfectly matched on the true propensity score (Joffe, 1999). However, especially in small studies, substantial chance imbalances may still exist and yield some (conditional) bias in the direct treatment effect estimator. Although methods based on the estimated propensity score have not been widely used in the randomized studies, it could have some substantial advantages over the methods by using the true scores under certain scenarios. Robins, Mark, and Newey (1992) has shown that there are even theoretical advantages to using estimated propensity scores.

We introduce a new randomization design, the balance match weighted (BMW) design, which applies the optimal full matching with constraints technique (Olsen, 1997) to the given randomization with the general aim of reducing the mean squared error (MSE) of the treatment effect estimator. In this design, treated and control subjects are matched into subsets based on their estimated propensity score and an overall estimate is constructed using a weighted sum of the subset-specific estimates. In contrast to the existing stratified design, which first stratifies and then randomizes within strata, the BMW design first randomly assign the units to treatments and then stratifies on the randomized sample. In an implementation of the design, this randomization-stratification process is repeated M times in order to choose a randomization that gives a good overall balance. In general, the BMW design has two advantages. First, it reduces the chance imbalance between the treatment groups in observed covariates through optimal matching, and hence decreases the (conditional) bias in the resultant estimator. Second, it controls for the increase in variance due to matching by using the full matching with constraints technique (Olsen, 1997), in which the choice of the constraint, k, adjusts for the trade-off between the potential gain in bias reduction and possible loss in precision. We examine various strategies for selecting M and k, seeking a good choice that yields good results with respect to MSE. It is obvious that MSE performance also depends on the inherent degree of confounding, so we compared the BMW design with the completely randomized design and matched-pair design under different confounding scenarios. If there is no confounding, the three design methods perform equally well. However, if there is considerable confounding, the BMW design can result in a substantial reduction in the MSE of the treatment effect estimator.

The design we propose is appropriate for the situation where all units are available for randomization at the onset, and cannot be applied to clinical trials with staggered entry. Pocock and Simon (1975) proposed a sequential strategy, minimization design, which makes the assignment decision one unit at a time, based solely on the covariate information of previously assigned subjects. On the other hand, the minimization design is not well suited for trials where all observational units are available for randomization at the onset.

The rest of the article is organized as follows. Notation and models are presented in Section 2. The BMW design is outlined in Section 3 and Section 4 gives results of a simulation study comparing the performance of the BMW design with the completely randomized design, a matched-pair design, the model-based approach by adjusting for the estimated propensity score, and the Robins-Mark-Newey E-estimation procedure. Its performance under heterogeneous error is also investigated. Section 5 outlines a case study and the article concludes with discussion in Section 6.

2. Methods

In this section, we present the notation and problem formulation as well as introduce some optimal matching techniques employed in the proposed design.

2.1 *Optimal Matching*

Consider a study with the aim of assessing the effect of treatment. Let N denote the number of subjects available for the study. We assume that N is even and $N/2$ subjects are randomized to each of the treatment and control groups, but the method we propose could allow imbalance in the randomized assignment. Thus, we suppose that a randomization process divides the N subjects into a set T of $N/2$ subjects to be treated and a set C of $N/2$ subjects to receive the control. We also assume that a vector of r covariates, $X = (X_1, X_2, \ldots, X_r)^{\mathrm{T}}$, is observed for each individual.

Similarity of covariates is measured through an estimated propensity score. Writing $Z = 1$ for the treated subjects, and $Z = 0$ for the control subjects, the (estimated) propensity score distance between the treated unit i and control unit j is given by

$$d_{i,j} = |\widehat{\delta}_i - \widehat{\delta}_j|, \tag{1}$$

where $\widehat{\delta}_i$ is the estimate of the true propensity score, $\delta_i = Pr(Z = 1 \mid X_i)$, and is obtained from a model such as the logistic regression model

$$\begin{aligned}\delta_i &= Pr(Z = 1 \mid X_i; \alpha)\\ &= \exp\left(\alpha_1 + \sum_{j=2}^{r} \alpha_j X_{ij}\right)\\ &\qquad \bigg/ \left\{1 + \exp\left(\alpha_1 + \sum_{j=2}^{r} \alpha_j X_{ij}\right)\right\}.\end{aligned} \tag{2}$$

In a randomized clinical trial, the true propensity score δ_i is typically determined by the randomization scheme and known. We consider the estimated propensity score $\widehat{\delta}_i$ in defining the distances with the aim of producing a design that reduces the actual observed imbalance between treated and control subjects. Matching assembles treated and control units that are as similar as possible into the same stratum using the overall estimated propensity score distance measure. Given T and C, we consider the collection $P_{C,T}$ of all possible matchings, where a matching corresponds to a collection of S strata comprised of matched subsets $\{(C_1, T_1), (C_2, T_2), \ldots, (C_S, T_S)\}$, in which, $C_1, C_2, \ldots, C_S$ is a partition of $C, T_1, T_2, \ldots, T_S$ is a partition of T, and $1 \leqslant S \leqslant N/2$. As is often done (e.g., Rosenbaum, 2002), we measure the quality of a particular matching as

$$\Delta = \sum_{s=1}^{S} w(|T_s|, |C_s|) \bullet \overline{T_s \times C_s}, \tag{3}$$

where

$$\overline{T_s \times C_s} = \sum_{(i,j)\in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j| / |T_s \times C_s|$$

is the average distance between the $|T_s \times C_s|$ possible pairs in the sth strata, and $w(.,.)$ is a weight function. Thus, Δ is a weighted sum of average distances and an optimal matching minimizes Δ over $P_{C,T}$.

A full matching is one in which each stratum is comprised of one treated (or control) subject matched to one or more control (or treated) subjects so that $\min(|T_s|, |C_s|) = 1$, for $s = 1, 2, \ldots, S$. Rosenbaum (2002, Lemma 2) showed that if the weight function in (2) is *neutral* or *favors small subclasses*, then there is always a full matching that is optimal. Among the class of full matchings with the weight function $w(|T_s|, |C_s|) = |T_s| + |C_s| - 1$, equation (3) reduces to

$$\Delta = \sum_{s=1}^{S} (|T_s| + |C_s| - 1) \bullet \overline{T_s \times C_s} = \sum_{s=1}^{S} \sum_{(i,j)\in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j|. \tag{4}$$

In this article, we use this total distance measure to evaluate the quality of a matching. One potential drawback of the optimal full matching is that some of its matched subsets can be very unbalanced with many controls to one treatment or vice versa. The imbalance among full matching subsets decreases the precision of the estimated treatment effect. One remedy for this is to constrain the full matching so that the ratio of the number of treated versus the number of controls in each stratum is between a lower and upper bound. To accomplish this, we choose an integer $k \in \{1, 2, \ldots, N/2 - 1\}$ and consider the optimization problem

$$\text{Minimize } \Delta = \sum_{s=1}^{S} \sum_{(i,j)\in T_s \times C_s} |\widehat{\delta}_i - \widehat{\delta}_j|, \tag{5}$$

over the class of full matchings subject to $k^{-1} \leqslant |T_s|/|C_s| \leqslant k$. We refer to the solution to this optimization problem as the *optimal full matching with constraint* k. When $k = 1$, we obtain the best matched-pair design with one treated unit and one control unit in each stratum. This assignment leads to a treatment effect estimator with minimum variance in the linear model discussed in the next section, but can result in relatively large bias. When $k = N/2 - 1$, there is no constraint on the balance in the relative numbers of treated and control units in any matched subset, the covariates are optimally balanced so the bias of treatment effect estimator tends, in this case, to be small, but the variance is larger. The BMW design we propose searches for the optimal full matching with constraint k. The choice of k represents a trade-off between bias and variance. In the next section, we examine the MSE as a measure of this trade-off with a class of linear models. For a specific model in this class, we can choose k to generate a BMW design that achieves minimum MSE. It is observed that the choice of k does not depend much on the specific model.

2.2 *Model*

To appreciate the effect of treatment on response in a pooled sample and matched sample, respectively, consider the following model: Let $Y_i, i = 1,2, \ldots, N$, represent responses of the unit i, conditional on a given treatment assignment T, C, and X,

$$Y_i = \alpha + \beta I(i \in T) + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i, \tag{6}$$

where $I(\cdot)$ is the indicator function, β denotes the true treatment effect, $\gamma_1, \gamma_2, \ldots, \gamma_r$ are the confounding effects, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N)$ is the vector of the measurement errors with $E[\varepsilon | T, C, X] = 0, \mathrm{var}[\varepsilon | T, C, X] = \sigma^2 I, \sigma^2 < +\infty$ and I is the $N \times N$ identity matrix.

2.2.1 *Pooled sample.* Under model (6), the common treatment effect estimator based on the unstratified pooled sample is $\widehat{\beta}_{pool} = \overline{y}_T - \overline{y}_C$, which has conditional expectation

$$E[\widehat{\beta}_{pool} \mid T, C, X] = \beta + \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT} - \overline{X}_{jC}), \tag{7}$$

where the subscripts C and T mean that the averages are computed over the control and treatment groups, respectively. The MSE (conditional on T, C, and X) is

$$MSE(\widehat{\beta}_{pool} \mid T, C, X) = \left\{ \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT} - \overline{X}_{jC}) \right\}^2 + 2\sigma^2/N. \tag{8}$$

2.2.2 *Matched sample.* Under model (6), estimating the treatment effect for the matched sample involves the computation of a weighted sum. In the sth matched subset (T_s, C_s), the treatment effect estimator is $\widehat{\beta}_{strata,s} = \overline{y}_{T_s} - \overline{y}_{C_s}$, which has conditional expectation

$$E[\widehat{\beta}_{strata,s} \mid T, C, X] = \beta + \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT_s} - \overline{X}_{jC_s}). \tag{9}$$

The overall estimate can be constructed using a weighted sum,

$$\widehat{\beta}_{strata} = \sum_{s=1}^{S} w_s \widehat{\beta}_{strata,s}, \tag{10}$$

where $\sum_s w_s = 1, w_s \geqslant 0$. It should be noted that this stratified estimator can be modified to accommodate different weighting methods. Two common choices are weighting in proportion to the number of subjects that each subset contains, $(|T_s| + |C_s|)/N$ (Cochran, 1968), or the inverse variance weighting, $(1/|T_s| + 1/|C_s|)^{-1} / \sum_{t=1}^{S} (1/|T_t| + 1/|C_t|)^{-1}$. For the purpose of this discussion, the former weighting method is considered, but it can be easily modified to handle the latter. It follows that the MSE of the stratified estimator (conditional on T, C, and X) can be written as

$$\begin{aligned} & MSE(\widehat{\beta}_{strata} \mid T, C, X) \\ &= \left\{ \sum_{s=1}^{S} \frac{(|T_s| + |C_s|)}{N} \sum_{j=1}^{r} \gamma_j (\overline{X}_{jT_s} - \overline{X}_{jC_s}) \right\}^2 \\ &\quad + \sum_{s=1}^{S} \frac{(|T_s| + |C_s|)^2}{N^2} \left(\frac{1}{|T_s|} + \frac{1}{|C_s|} \right) \sigma^2. \end{aligned} \tag{11}$$

With no confounding effects or chance imbalance in covariates, the pooled estimator is the unbiased estimate of the treatment effect with minimum variance. In the presence of confounding, stratification reduces the bias but increases the variance. We use the MSE to measure the trade-off between bias and variance.

3. The BMW Design

In a randomized trial with fixed small sample size N and many confounding covariates, it may be impossible to produce balance on all of the variables simultaneously. In order to reduce the actual observed imbalance as well as increase precision of the estimator, we propose the BMW design. The design with specified parameters k and M is defined algorithmically as follows:

Step 1. Randomize half of the subjects to the treatment group, and half to control to obtain sets T and C;

Step 2. Compute the estimated propensity scores and create the $|T| \times |C|$ matrix of estimated propensity score distances;

Step 3. Obtain the optimal full matching with constraint k and record the total distance Δ_k;

Step 4. Repeat *Steps* 1 to 3 M times; choose the randomized sample with minimum total distance $\Delta_k^* = \min(\Delta_{1k}, \Delta_{2k}, \ldots, \Delta_{Mk})$. The choice of M is discussed below.

It is clear that the choice of k represents a trade-off between bias and variance. We use MSE as a measure of the trade-off. The choice of k ($k \in (1, 2, \ldots, N/2 - 1)$), which minimizes the MSE of the treatment effect estimator, depends on the confounding effect $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_r)$. If γ were known and M is fixed, it would be possible to compute the MSE for each k based on the BMW design in *Step 1* to *4* above. It would be possible then to select the k that minimizes the MSE. In practice, of course, the true value of γ is unknown; therefore in the next section we use a simulation study to evaluate the effects of k on reducing the MSE under a variety of assumptions about the size of the confounding effects. We find that $k = 2$ is a suitable choice under most of the confounding scenarios considered.

Clearly, the larger M is, the better the matching the BMW design attains. In the next section, we examine how the MSE depends on M and find that most of the gain is attained by relatively small M of 10 or 20 in the cases considered, and we recommend a value of M in this range. It should also be noted that, as M increases, the BMW design becomes more deterministic.

The implementation of *Step* 3, which searches the optimal full matching with constraint k (Olsen, 1997) is conducted using the `Optmodel` Procedure in `SAS` (Version 9.1.3.2). A similar program `Optmatch` in `R` has also been developed (Hansen, 2004).

There are alternative ways to adjust for the covariate imbalance resulting from randomization. Since small sample sizes do not allow for control of all variables by model-based method, one possible approach, suggested by an associate editor, is to adjust the estimated propensity score in a regression model such as:

$$Y_i = \alpha + \beta I(i \in T) + \gamma \widehat{\delta_i} + \varepsilon_i. \qquad (12)$$

Let $\widehat{\beta}_{MB}$ denote the ordinary least squares estimate of β from (12). Our simulations and investigations suggest that the model-based approach seems to work well if the model for the propensity score is *appropriately* specified, where, by appropriately specified, we mean that the regression model for the propensity score includes the same regression parameters and is of the same form as the true model for the outcome variable Y. For example, if the true model is $Y_i = \alpha + \beta I(i \in T) + \gamma_1 X_i + \gamma_2 X_i^2 + \varepsilon_i$ and we specify $\text{logit}(\delta_i) = \text{logit}(Pr(Z = 1 \mid X_i; \alpha)) = \alpha_1 + \alpha_2 X_i + \alpha_3 X_i^2$, then regression adjustment using $\widehat{\delta_i}$ will tend to work well. In fact, if the confounding effects are large, $\widehat{\beta}_{MB}$ tends to be somewhat more efficient than the estimator obtained from the BMW approach. On the other hand, the BMW approach is more robust if the propensity score model is *inappropriately* specified as, for example, if the same true model of Y holds and we specify $\text{logit}(\delta_i) = \text{logit}(Pr(Z = 1 \mid X_i; \alpha)) = \alpha_1 + \alpha_2 X_i$. This is examined further in the simulations of Section 4.

Robins et al. (1992) proposed another procedure based on the propensity score in observational studies. Their approach is designed to provide a consistent estimator, $\widetilde{\beta_E}$, when the model for propensity score $\widehat{\delta_i}$ is *correctly* specified. This estimator is

$$\widetilde{\beta_E} = \sum_{i=1}^{n} Y_i (Z_i - \widehat{\delta_i}) \Big/ \sum_{i=1}^{n} Z_i (Z_i - \widehat{\delta_i}). \qquad (13)$$

At the suggestion of a reviewer, we also evaluate this approach in the simulations of the next section.

4. Simulation Study

In order to assess the performance of the BMW design, we first carried out a simulation study to compare it with a completely randomized design and a matched-pair design. In doing so, we considered a wide variety of settings and, for each setting, estimated the MSE based on 1000 replications.

4.1 *Structure of the Simulation*

For each of N subjects, we generated a set of r covariates $X_1, X_2, \ldots, X_r$, where the covariates were drawn independently from various distributions as described below. Given a randomization of subjects to the two treatment groups, the responses were generated conditional on the treatment assignment ($Z_i = 0$ or 1) and the covariates (X_{ij}), where $Pr(Z_i = 1 \mid X_{ij}) = 0.5$. Specifically, the response was obtained from

$$Y_i = \beta Z_i + \sum_{j=1}^{r} \gamma_j X_{ij} + \varepsilon_i, \qquad (14)$$

where $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ and $i = 1, 2, \ldots, N$. In the simulations, we considered the following:

- The true treatment effect was taken to be $\beta = 0.7$
- The true confounding effects were $\gamma_j = \gamma, j = 1, \ldots, r$ where $\gamma = 0.5, 1.0, 1.5$. Note that the results we obtain do not depend on the choice of β. When the covariates follows symmetric distributions, the results do not depend on the signs of the components of γ either.
- For the first three settings, we considered $r = 4$ covariates selected from the following distributions: $(i) X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $(ii) X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$; $(iii) X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5); $X_3, X_4 \overset{i.i.d}{\sim}$ Bernoulli(0.66).

- For the fourth case, we considered $r = 8$ covariates: $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim}$ Bernoulli(0.5)
- We consider sample size of $N = 30, 60$.

The completely randomized design assigns half of the units at random to each of the two treatment groups. For this design, the treatment effect estimator is $\widehat{\beta}_{pooled} = \overline{Y}_T - \overline{Y}_C$ and the corresponding MSE (conditional on T, C, and X) is given in (8). We also consider a matched-pair design in which each unit is matched (so much as possible) to another unit based on the first covariate X_1. One unit in each pair is then randomly assigned to treatment and one to control. The BMW design, as described in the preceding section, creates an optimally matched sample for each constraint k, where $k = 1, 2, \ldots, N/2 - 1$, and for each choice of M, this leads to the weighted treatment effect estimator $\widehat{\beta}_{strata}$ in (10) along with its MSE (11). We further consider $\widehat{\beta}_{MB}$, from the model-based approach by adjusting for the estimated propensity score (12) and the Robins-Mark-Newey E estimator $\widetilde{\beta_E}$ (13). Finally, we examine the possible effects of homoscedastic error on the BMW design by allowing the error variance to depend on the first covariate X_1.

4.2 *Results*

The average MSEs based on 1000 replications are summarized in Table 1. From Cochran (1968), the true unconditional MSE of $\widehat{\beta}_{pool}$ is $2\sigma_y^2/N$, where σ_y^2 refers to the overall variability in outcome Y. In this, one part, $\sum_j \gamma_j^2 \mathrm{var}(\overline{X}_{jT} - \overline{X}_{jC})$, is due to variability in the observed covariates $X_1, X_2, \ldots, X_r$ and the other to the conditional variations of Y given $X_1, X_2, \ldots, X_r$. Formally, the unconditional MSE is (from (8))

$$\begin{aligned} MSE(\widehat{\beta}_{pool}) &= E\left[\left\{\sum_{k=1}^{K} \gamma_k(\overline{X}_{kT} - \overline{X}_{kC})\right\}^2 + \frac{2}{N}\sigma^2\right] \\ &= \sum_{k=1}^{K} \gamma_k^2 \mathrm{var}(\overline{X}_{kT} - \overline{X}_{kC}) + \frac{2}{N}\sigma^2 \\ &= \frac{2\sigma_y^2}{N}. \end{aligned} \tag{15}$$

With prerandomization matching or postrandomization stratification on covariates, the average MSE values are also obtained in the simulation. A similar formula to (15) can be obtained for the matched-pairs design, but formulas for the BMW design are complicated. For the BMW design, the average MSE for each constraint $k = (1, 2, \ldots, N/2 - 1)$ were examined in the simulations, but only those for $k = 1, 2, 3$ are displayed since the MSE changes little when k increases over three. The percent reduction in MSE is $100 \times (MSE - MSE^*_{BMW})/MSE$, where MSE^*_{BMW} corresponds to the minimal value of MSE for each k in the BMW design, and MSE refers to the MSE value for the design to which BMW is being compared (e.g., the completely randomized design or the matched-pair design).

It is interesting to examine how the MSE of the treatment effect estimator is affected by various parameter settings. Overall, the BMW design shows significant reductions in MSE as compared to both the completely randomized and matched-pair designs.

4.2.1 *Confounding effects* γ_j. Table 1 reveals that as the confounding effects, measured by $\sum_j \gamma_j$, increase, the average MSEs generally increase. However, the MSE in the BMW design increases much more slowly than the MSE in the completely randomized or matched-pair design. This suggests that the BMW design becomes much more effective in reducing the MSE when confounding effects increase. Specifically, as we raise $\sum_{j=1}^r \gamma_j$ from 2.0 to 6.0 for Bernoulli-distributed covariates (Table 1), the MSE reduction of the BMW design with $k = 2$ compared to the matched-pair design varies dramatically from 5.96% to 53.77% for $M = 5$, from 7.50% to 54.59% for $M = 10$, and from 9.36% to 56.10% for $M = 20$. An even larger reduction in MSE arises when comparing the BMW design with the completely randomized design.

4.2.2 *The choice of the constraint k.* We now examine the MSE as a function of k. When the model contains four covariates of various forms (Table 1) and there is relatively little confounding such as $\sum_{j=1}^r \gamma_j = 2.0$, then the MSEs corresponding to $k = 1$ are slightly smaller than those corresponding to $k = 2$. As $\sum_{j=1}^r \gamma_j$ increases, however, a greater reduction in MSE due to constraint $k = 2$ becomes apparent. Intuitively, for a small sample with strong confounding effects, bias reduction is more important than variance reduction, so the larger value of k ($k = 2$) is more efficient. However, when the number of covariates is $r = 8$, the constraint $k = 2$ minimizes the MSE for all confounding effects considered.

4.2.3 *Number of replication M.* The MSE is obviously a decreasing function of M for given γ and k. However, when it comes to percent reduction in MSE attained from using the BMW design as compared to the completely randomized design or matched-pair design, simulations suggest an interesting interplay between the number of replications, M, and the confounding effect $\sum_j \gamma_j$. The results suggest that if there is little confounding ($\sum_{j=1}^r \gamma_j = 2.0$) and the covariates are independently Bernoulli distributed (Table 1), the percent reduction in MSE of the BMW design versus the matched-pair design increases from 7.96% to 10.29% to 13.46% for M from 5 to 10 to 20, with $k = 1$. If there is relatively more confounding ($\sum_{j=1}^r \gamma_j = 6.0$), the percent reduction in MSE increases more modestly from 53.77% to 54.59% to 56.10% with M, while using matching with constraints $k = 2$. Similar trends are seen in comparing the BMW design with the completely randomized design or using different covariate distributions. We conclude that, when confounding effects are relatively strong, the BMW design even with relatively small M is very effective in reducing MSE. A good compromise value of M is $M = 10$ for the cases considered.

4.2.4 *Covariate settings.* There are four covariate settings examined in the simulation studies. The results suggest that, in situations where existing designs often fail in producing balance across covariates, the BMW design provides a useful approach. Gains in efficiency are substantial when the covariates are Bernoulli variables with important, but somewhat more modest gains, when the covariates include continuous variables. For given γ, the gains due to the BMW design are similar for symmetric and asymmetric Bernoulli distributions for the covariates. Finally, when the number of Bernoulli covariates increases from four to eight, the BMW design achieves a larger reduction in MSE.

Table 1

Percent reductions in the MSE of treatment effect estimator for the BMW design compared to a completely randomized design (CR) and matched-pair design (MP). Sample size $N = 30$ subjects. Number of replications = 1000.

			MSE	*MSE* percent reduction(%) (*BMW* vs. *CR* design)			*MSE*	*MSE* percent reduction(%) (*BMW* vs. *MP* design)		
γ	$\sum_{j=1}^{4}\gamma_j$	M	(*CR*)	$k=1$	$k=2$	$k=3$	(*MP*)	$k=1$	$k=2$	$k=3$
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$										
(0.5, 0.5, 0.5, 0.5)	2	5	0.166	12.21	10.30	6.87	0.158	7.96	5.96	2.37
		10		14.43	11.77	7.14		10.29	7.50	2.64
		20		17.45	13.54	8.81		13.46	9.36	4.40
(1.0, 1.0, 1.0, 1.0)	4	5	0.280	35.61	43.58	39.67	0.239	24.57	33.90	29.33
		10		40.37	44.45	41.74		30.15	34.92	31.75
		20		50.39	48.66	46.21		41.87	39.86	36.99
(1.5, 1.5, 1.5, 1.5)	6	5	0.450	45.39	61.58	57.94	0.374	34.29	53.77	49.39
		10		52.19	62.26	59.02		42.47	54.59	50.69
		20		58.43	63.52	60.64		49.97	56.10	52.64
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$										
(0.5, 0.5, 0.5, 0.5)	2	5	0.155	8.77	5.67	1.09	0.148	4.45	1.21	−3.59
		10		9.46	5.85	1.66		5.17	1.40	−2.99
		20		12.17	7.74	3.52		8.01	3.38	−1.05
(1.0, 1.0, 1.0, 1.0)	4	5	0.218	24.37	30.79	27.29	0.190	13.20	20.58	16.56
		10		28.89	32.40	29.18		18.39	22.42	18.73
		20		32.85	33.09	30.13		22.94	23.22	19.82
(1.5, 1.5, 1.5, 1.5)	6	5	0.316	35.91	50.61	47.45	0.252	19.56	38.01	34.04
		10		42.98	52.08	48.45		28.43	39.85	35.29
		20		48.35	51.58	48.30		35.17	39.22	35.10
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.66)$										
(0.5, 0.5, 0.5, 0.5)	2	5	0.165	12.11	12.08	7.31	0.158	8.03	8.00	3.01
		10		14.93	12.99	8.78		10.98	8.96	4.55
		20		16.13	12.69	8.72		12.24	8.64	4.48
(1.0, 1.0, 1.0, 1.0)	4	5	0.267	32.21	40.76	36.77	0.229	20.97	30.94	26.29
		10		37.92	43.13	39.39		27.63	33.71	29.34
		20		41.88	44.14	41.22		32.25	34.88	31.48
(1.5, 1.5, 1.5, 1.5)	6	5	0.430	50.98	61.68	59.36	0.352	40.15	53.20	50.37
		10		50.63	59.12	55.57		42.75	52.60	48.48
		20		55.05	59.33	56.08		47.87	52.84	49.07
$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim} Bernoulli(0.5)$										
(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	4	5	0.204	17.35	23.93	18.68	0.187	10.06	17.21	11.49
		10		18.63	24.30	19.63		11.44	17.62	12.53
		20		22.65	25.22	19.42		15.82	18.62	12.30
(1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)	8	5	0.390	28.74	52.41	52.21	0.363	23.39	48.83	48.62
		10		35.80	56.12	53.11		30.97	52.82	49.58
		20		43.23	57.60	54.22		38.96	54.41	50.78
(1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)	12	5	0.725	35.07	66.86	68.47	0.664	29.12	63.83	65.58
		10		46.71	71.55	69.76		41.83	68.94	66.99
		20		52.71	73.14	70.29		48.38	70.68	67.57

4.2.5 *Sample size N.* Sample size has an impact on the performance of the BMW design, and as sample size becomes very large, we would expect the relative gains to decrease as randomization itself guarantees substantial balance among the covariate values. Our simulation results reveal, however, that when the sample size increases from 30 to 60, the percent reduction in MSE from the BMW design decreases only very little. This suggests a possible value for this approach even in larger studies. Computational aspects are easily accommodated for the larger sample sizes; for example, the processing time for the simulations with $N = 60$ increases by about 40% over those for $N = 30$.

It is also of interest to compare the BMW design with the model-based approach adjusting for the estimated propensity score and Robins-Mark-Newey E-estimation procedure in terms of efficiency and robustness of the treatment effect

Table 2

Percent reductions in the MSE of treatment effect estimator for the BMW design compared to the model-based adjustment approach adjusting for the estimated propensity score (*MB*) *and E estimation procedure* (*E-est*), *where the propensity score model is appropriately and inappropriately specified, respectively. Number of replications* = 1000.

			MSE percent reduction(%) (*BMW* vs. *MB*)				*MSE* percent reduction(%) (*BMW* vs. $E-est$)		
γ	*M*	*MSE* (*MB*)	$k=1$	$k=2$	$k=3$	*MSE* $(E-est)$	$k=1$	$k=2$	$k=3$
Where propensity score *inappropriately* specified (17)–(18)									
$X \overset{i.i.d}{\sim} Normal(0, 0.25)$									
(0.5, 0.5)	10	0.185	0.65	14.75	12.25	0.334	45.06	52.85	51.47
(1.0, 1.0)	10	0.365	−0.15	30.03	32.31	0.964	62.10	73.52	74.39
(1.5, 1.5)	10	0.665	5.80	41.88	46.12	2.013	68.90	80.81	82.21
Where propensity score *appropriately* specified (15)–(16)									
$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$									
(0.5, 0.5, 0.5, 0.5)	10	0.165	15.01	15.74	6.79	0.211	33.41	33.98	26.97
(1.0, 1.0, 1.0, 1.0)	10	0.166	−0.87	6.02	1.44	0.528	68.38	70.54	69.10
(1.5, 1.5, 1.5, 1.5)	10	0.166	−29.84	−2.49	−11.31	0.971	77.85	82.52	81.01
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.66)$									
(0.5, 0.5, 0.5, 0.5)	10	0.152	7.19	5.08	0.48	0.247	42.97	41.68	38.85
(1.0, 1.0, 1.0, 1.0)	10	0.152	−8.99	0.16	−6.41	0.492	66.32	69.15	67.12
(1.5, 1.5, 1.5, 1.5)	10	0.153	−32.00	−9.29	−18.78	0.916	77.99	81.78	80.19
$X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$									
(0.5, 0.5, 0.5, 0.5)	10	0.148	5.41	1.64	−2.74	0.203	30.71	27.95	24.74
(1.0, 1.0, 1.0, 1.0)	10	0.148	−4.52	0.64	−4.09	0.387	59.89	61.88	60.06
(1.5, 1.5, 1.5, 1.5)	10	0.148	−21.56	−2.15	−9.91	0.689	73.82	78.00	76.33

estimator. Therefore, we evaluate the MSE property of the three approaches under two scenarios, one where the propensity score model is *appropriately* specified and one where it is not.

4.2.6 *Propensity score appropriately specified.* Under this scenario, we specify the true model and propensity score model as follows:

$$Y_i = \alpha + \beta I(i \in T) + \sum_{j=1}^{4} \gamma_j X_{j,i} + \varepsilon_i. \tag{16}$$

$$\text{logit}(\delta_i) = \text{logit}\{Pr(Z = 1 \mid X_i; \alpha)\} = \alpha_0 + \sum_{j=1}^{4} \alpha_j X_{j,i}. \tag{17}$$

From the results summarized in Table 2, we see that the MSE obtained by the model-based approach remains relatively constant as the confounding effects increase, provided the terms in the propensity score model mimic that in the true model for *Y*. If there is relatively little confounding ($\sum_{j=1}^{r} \gamma_j < 6.0$), the MSEs in the BMW design are slightly smaller than those from the model-based approach. As $\sum_{j=1}^{r} \gamma_j$ increases, however, a somewhat greater reduction in MSE is obtained through the model-based approach. Both the BMW design and the model-based estimate perform much better than the E-estimation procedure in the context of these small randomized experiments.

4.2.7 *Propensity score inappropriately specified.* In practice, the true model for outcome *Y* is unknown, and due to the small sample size, it is difficult to determine what model is best; consequently, adjustment for many potential confounders may not work well. The simulation studies in Table 2 suggest that when the propensity score model does not mimic the correct regression terms in the true model, the BMW design provides a more robust approach than the model-based approach. For illustration purposes, we looked at a true model and propensity score model as follows:

$$Y_i = \alpha + \beta I(i \in T) + \gamma_1 X_i + \gamma_2 X_i^2 + \varepsilon_i. \tag{18}$$

$$\text{logit}(\delta_i) = \text{logit}\{Pr(Z = 1 \mid X_i; \alpha)\} = \alpha_1 + \alpha_2 X_i, \tag{19}$$

where $X_i \overset{i.i.d}{\sim} Normal(0, 1)$. As the confounding effects γ_j increases from 0.5 to 1.5, the percent reduction in MSE of the BMW design compared to the model-based approach increases from 14.75% to 41.88%, for $M = 10$. Again, the E-estimation procedure does not perform well in this context. This suggests that the BMW design is more robust than the model-based approach when the propensity score model is *inappropriately* specified, as would often be the situation in practice.

4.2.8 *Heteroscedastic errors.* In the clustered randomized trials with few but relatively large clusters, the hemoscedasticity error assumption is unlikely to hold. To investigate the effects of this, we allowed the error distribution of the outcome to vary by the first covariate X_1 in our simulation studies. In particular, in the model (6), we specified $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ if $X_1 = 1$ and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 0.25)$ if $X_1 = 0$, where $X_1, X_2 \overset{i.i.d}{\sim}$ Bernoulli(0.5) and $X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$. The results in Table 3 suggest that the relaxation of the

Table 3

Percent reductions in the MSE of treatment effect estimator for the BMW design compared to a completely randomized design (CR) and matched-pair design (MP) under the heteroscedastic and hemoscedastic error assumption, respectively. Number of replications = 1000. $X_1, X_2 \overset{i.i.d}{\sim} Bernoulli(0.5); X_3, X_4 \overset{i.i.d}{\sim} N(0, 0.25)$.

				MSE percent reduction(%) (*BMW* vs. *CR* design)				*MSE* percent reduction(%) (*BMW* vs. *MP* design)		
γ	$\sum_{j=1}^{8} \gamma_j$	M	*MSE* (*CR*)	$k=1$	$k=2$	$k=3$	*MSE* (*MP*)	$k=1$	$k=2$	$k=3$
				$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$ if $X_1 = 1$ and $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, 0.25)$ if $X_1 = 0$						
(0.5, 0.5, 0.5, 0.5)	2	10	0.089	14.19	13.32	8.83	0.075	−0.90	−1.93	−7.20
(1.0, 1.0, 1.0, 1.0)	4	10	0.152	41.40	49.28	44.78	0.121	26.63	36.50	30.87
(1.5, 1.5, 1.5, 1.5)	6	10	0.258	52.57	65.36	65.05	0.166	26.48	46.31	45.83
				$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,1)$						
(0.5, 0.5, 0.5, 0.5)	2	10	0.155	9.46	5.85	1.66	0.148	5.17	1.40	−2.99
(1.0, 1.0, 1.0, 1.0)	4	10	0.218	28.89	32.40	29.18	0.190	18.39	22.42	18.73
(1.5, 1.5, 1.5, 1.5)	6	10	0.316	42.98	52.08	48.45	0.252	28.43	39.85	35.29

Table 4

Optimal matched sample produced by the BMW design with $k = 2$ *and* $M = 10$ *for the case study.* X_1*: percent of females greater than* 65 *years of age among all females in the census tract* (%); X_2*: percent of males greater than* 65 *years of age among all males in the census tract* (%); X_3*: stroke volume* (*low vs. high*); X_4*: population density (urban vs. rural). The estimated propensity score* ($\widehat{\delta}$) *was shown for each subject and the total propensity score distance* $\Delta = 0.202$ *for the stratum.*

	Treatment group					Control group				
Strata	$ID(\widehat{\delta})$	X_1	X_2	X_3	X_4	$ID(\widehat{\delta})$	X_1	X_2	X_3	X_4
1	1 (0.33)	0.15	0.13	0	0	6 (0.35)	0.19	0.07	0	0
2	2 (0.38)	0.17	0.11	1	0	8 (0.35)	0.22	0.14	0	0
	11 (0.40)	0.22	0.14	1	0					
3	3 (0.63)	0.13	0.06	1	1	9 (0.63)	0.14	0.06	1	1
						19 (0.67)	0.25	0.15	1	1
4	4 (0.58)	0.12	0.06	0	1	12 (0.60)	0.07	0.06	1	1
5	14 (0.32)	0.13	0.07	0	0	13 (0.32)	0.13	0.09	0	0
	15 (0.31)	0.10	0.06	0	0					
6	17 (0.41)	0.24	0.12	1	0	10 (0.41)	0.26	0.18	1	0
	22 (0.43)	0.30	0.17	1	0					
7	20 (0.60)	0.08	0.06	1	1	16 (0.61)	0.10	0.07	1	1
						18 (0.61)	0.09	0.05	1	1
8	21 (0.60)	0.18	0.14	0	1	5 (0.61)	0.19	0.13	0	1
9	24 (0.62)	0.23	0.16	0	1	7 (0.62)	0.24	0.19	0	1
						23 (0.62)	0.11	0.07	1	1

hemoscedasticity error assumption has little impact on the performance of the BMW design. The case study in the next section is a case where such heteroscedasticity may be present.

5. Planning an Educational Study for tPA Usage in Stroke

In this section, we consider the use of the BMW design in planning an educational study to increase tPA therapy use for stroke patients as described in the Introduction. As noted there, four covariates were measured on participating institutions, and it was impossible to simultaneously obtain a balance in a matched-pair design. The simulation study in Section 4 suggests that design parameter $k = 2$ and the number of replication $M = 10$ give results that are close to optimum over a broad class of covariate distributions and confounding effects. We therefore choose these parameters in proposing a design for the tPA study.

We randomly assigned the 24 hospitals to two treatment groups, and estimated the sample-based propensity score for each hospital. The hospitals were then optimally matched into subsets with $k = 2$, which gave a minimum total distance of 2.5887. We then randomized the hospitals an additional nine times obtaining distance measures: 2.05, 2.50, 0.20, 1.42, 0.49, 3.00, 1.14, 0.72, and 1.48. The fourth randomization produced the smallest distance. The corresponding BMW design is presented in Table 4, where there were nine matched subsets with treated hospital 1 matched to control 6, treated hospitals 2

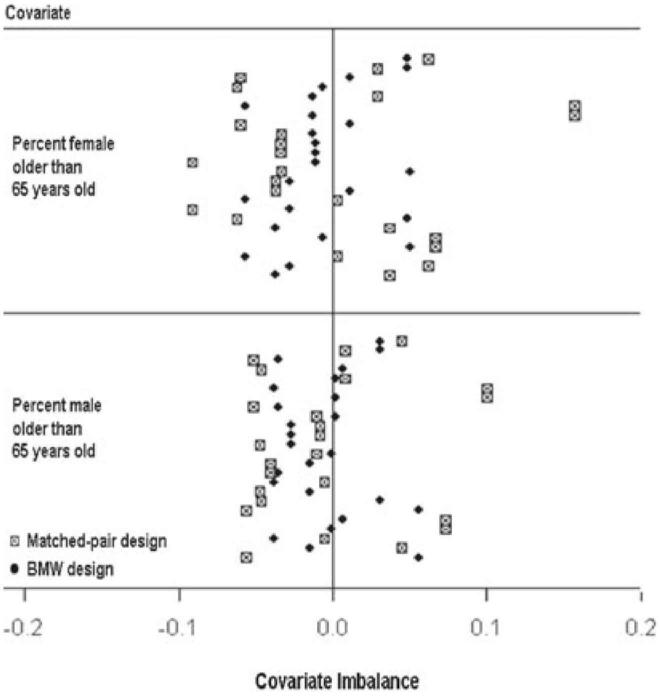

Figure 1. Covariate imbalances from the matched-pair design (matching on the categorical covariates: Population density and stroke volume) and the BMW design. The imbalance value in covariate X for unit i was computed as $Imbalance(X_i) = \sum_{j\in T_s} X_j/|T_s| - \sum_{k\in C_s} X_k/|C_s| = \overline{X}_{T_s} - \overline{X}_{C_s}$, where s is the stratum that unit i belongs to.

and 3 jointly matched to control 8, and so on. For comparison, the data were also randomized by using a matched-pair design, where the 24 hospitals were matched into 12 pairs based on the two binary covariates, rural versus urban population density and low versus high stroke volume. One hospital in each pair was then randomized to treatment and one to control. Figure 1 illustrates treatment to control group imbalance in the two continuous covariates, under the BMW and the matched-pair design.

When γ is known, we can determine the constraint k that minimizes the MSE when using the BMW design. Preliminary data provided estimates of the regression parameters in a logit model for the proportion of stroke cases receiving tPA as -0.63 (stroke volume), 0.02 (population density), 4.33 (percent female older than 65), and -1.23 (percent male older than 65). Since there are 24 hospitals, k can take values from 1 to 11. For $k = 1, M = 10$ randomizations gave a minimum distance of 0.2936. We then repeated the above process with the same randomized samples but with constraints $k = 2, 3, \ldots, 11$ and for each k, searched for the optimal sample with minimal distance. Third, based on the approximate value of γ above, we computed the MSE from (11) as 0.1076, 0.1045, and 0.1114 for the optimal sample with constraint $k = 1, 2, 3$. This suggests that pair matching and matching with constraint $k = 2$ achieve approximately the same level of optimality in terms of minimizing MSE. Compared with the matched-pair design described above, the BMW design reduced the MSE of the treatment effect estimator by 42%.

6. Discussion

The BMW design is, in essence, applying the optimal full matching with constraints technique to randomization in order to achieve overall balance between treatment groups and control the variance of the treatment comparison and so yield good MSE properties. One of the virtues of this design is that it will not only reduce the chance imbalance in observed covariates but also preserve the advantage of traditional randomized designs in balancing the unobserved covariates on average. Although only partial balance on the observed covariates is achieved by the BMW design, it is substantially better than the balance obtained by random assignment of treatments. When there is considerable confounding in small studies, this improvement in balance can result in a substantial decrease of MSE in the treatment effect estimator.

The BMW design can be revised to allow the user to select other criteria besides MSE to compromise between bias and variance. If variance of the estimator is not a concern, one can modify this design to achieve optimal balance and so reduce conditional bias (i.e., set $k = N/2 - 1$). On the other hand, if the objective is to minimize variance, optimal pair matching with constraint $k = 1$ is the best full matching choice.

We recommend use of a super-population model for analysis, and this is the basis of the simulation comparisons that we have made. It is worth noting, however, that the BMW design with a choice of M that is not too large, can also form the basis of a randomization test. Suppose, for example, that a sample has been collected using the BMW design with given k and M and the value of the test statistic (e.g., t statistic) has been computed. We now repeat the BMW design with the same k and M a large number B of times and each time compute the test statistic based on the fixed outcomes observed. This would lead to a randomization test and confidence intervals following standard methods. This would typically yield a reasonably large reference set as the basis of the test. On the other hand, if M is too large, most of the probability will be concentrated on relatively few designs and the randomization distribution becomes less useful. For example, with continuous covariates, if $M = \infty$ and $k = 1$, then the BMW design will always lead to the same set of matched pairs with the same treatment assignments. With smaller M or discrete covariates, the reference set is larger.

The model-based approach of adjusting for the estimated propensity score and the Robins-Mark-Newey E-estimation procedure could be considered as alternatives to the BMW design. Our simulation studies suggest that, when the propensity score model is *appropriately* specified, the BMW design is more efficient than the model-based approach when the confounding effects are relatively small; the model-based approach, however, becomes more efficient than the BMW design when the confounding effects increase. On the other hand, when the propensity score model is *inappropriately* specified, the BMW design achieves substantial gain over the model-based approach. In the context considered in this article, the E-estimation procedure is the least efficient and robust.

Greevy et al. (2004) proposed another multivariate matching design based on Mahalanobis distance. This approach searches for the optimal multivariate nonbipartite matching followed by randomization within pairs. We also investigated this in a simulation study presented in Table 5. As the

Table 5
Percent reductions in the MSE of treatment effect estimator for the BMW design compared to multivariate nonbipartite matching design (NB). Number of replications = 1000.

				MSE percent reduction(%) (*BMW* vs. *NB* design)		
γ	$\sum_{j=1}^{8}\gamma_j$	M	*MSE* (*NB* design)	$k = 1$	$k = 2$	$k = 3$
		$X_1, X_2, X_3, X_4 \overset{i.i.d}{\sim} Bernoulli(0.5)$				
(0.5, 0.5, 0.5, 0.5)	2	5	0.146	−0.07	−2.27	−6.11
		10		2.47	−0.55	−5.90
		20		5.91	1.44	−3.91
(1.0, 1.0, 1.0, 1.0)	4	5	0.185	2.42	14.49	8.53
		10		9.62	15.79	11.68
		20		24.78	22.18	18.44
(1.5, 1.5, 1.5, 1.5)	6	5	0.250	1.77	30.92	24.36
		10		14.01	32.12	26.28
		20		25.24	34.40	29.20
		$X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8 \overset{i.i.d}{\sim} Bernoulli(0.5)$				
(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)	4	5	0.156	−8.15	0.51	−6.35
		10		−6.41	0.96	−5.13
		20		−1.15	2.18	−5.39
(1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0, 1.0)	8	5	0.222	−25.19	16.39	16.07
		10		−12.76	22.92	17.65
		20		0.26	25.53	19.59
(1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5, 1.5)	12	5	0.338	−39.10	29.01	32.47
		10		−14.16	39.06	35.22
		20		−1.31	42.46	36.37

confounding effects increase, or the number of covariates increase, the BMW design becomes much more effective in reducing MSE compared to Greevy's design. This may be because the Mahalanobis distance is inferior to propensity scores when there are many covariates.

In general terms, the BMW design appears to provide a viable approach in the context of small studies where adjustment for randomization imbalance may be important. Furthermore, the simplicity of this matching-based design allows researchers to perform simple stratified analyses that adjust for imbalance in the randomization, which is appealing.

Finally, simulation shows that the BMW design can substantially reduce the MSE of the treatment effect estimate, as compared to the existing randomized designs in linear models. These investigations could be extended to other regression models, such as the class of general linear models. It should also be noted that the BMW design can be generalized to clinical trials with more than two treatment arms. Baseline category logit model can be used to estimate the probability of a subject being assigned to each treatment arm, and Euclidean distance can be used to measure the quality of a matching.

7. Supplementary Materials

The data for the case study and `SAS` macro implementing the BMW design are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors would like to thank Professors Mary Haan and Phillip Scott for helpful discussions. This work was supported in part by grant (RO1 NS050372) from the National Institute of Neurological Disorders and Stroke.

References

Abdeljaber, M. H., Monto, A. S., Tilden, R. L., Schork, M. A., and Tarwotjo, I. (1991). The impact of vitamin A supplementation on morbidity: A randomized community intervention trial. *American Journal of Public Health* **81,** 1654–1656.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassifications in removing bias in observational studies. *Biometrics* **24,** 295–313.

COMMIT, Research Group. (1995). Community intervention trial for smoking cessation (COMMIT): I. Cohort results from a four-year community intervention. *American Journal of Public Health* **85,** 183–192.

Graham, J. W., Flay, B. R., Johnson, C. A., Hansen, W. B., and Collins, L. M. (1984). A multiattribute utility measurement approach to the use of random assignment with small numbers of aggregated units. *Evaluation Review* **8,** 247–260.

Greevy, R., Lu, B., Silber, J. H., 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.

Joffe, M. M. (1999). Propensity scores. *American Journal of Epidemiology* **150,** 327–333.

Ming, K. and Rosenbaum, P. R. (2004). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics* **56,** 118–124.

NINDS, The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. (1995). Tissue plasminogen activator for acute ischemic stroke. *The New England Journal of Medicine* **333,** 1581–1588.

Olsen, S. P. (1997). *Multivariate Matching with Non-normal Covariates in Observational Studies*. PhD thesis, University of Pennsylvania.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* **31,** 103–115.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modeling the expectation of exposure conditional on confounders. *Biometrics* **48**(2), 479–495.

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.

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

Received December 2008. *Revised September* 2009.
Accepted September 2009.