Biostatistics (2004), **5**, 2, *pp.* 263–275
Printed in Great Britain

Optimal multivariate matching before randomization

ROBERT GREEVY*
Department of Statistics, The Wharton School, University of Pennsylvania, 400 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA

BO LU
Center for Statistical Sciences, Department of Community Health, Brown University School of Medicine, Providence, RI 02912, USA,

JEFFREY H. SILBER
Department of Pediatrics, University of Pennsylvania, School of Medicine, 3535 Market Street, STE 1029, Philadelphia, PA 19104-3309, USA

PAUL ROSENBAUM
Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340, USA
rosenbaum@stat.wharton.upenn.edu

SUMMARY

Although blocking or pairing before randomization is a basic principle of experimental design, the principle is almost invariably applied to at most one or two blocking variables. Here, we discuss the use of optimal multivariate matching prior to randomization to improve covariate balance for many variables at the same time, presenting an algorithm and a case-study of its performance. The method is useful when all subjects, or large groups of subjects, are randomized at the same time. Optimal matching divides a single group of $2n$ subjects into n pairs to minimize covariate differences within pairs—the so-called nonbipartite matching problem—then one subject in each pair is picked at random for treatment, the other being assigned to control. Using the baseline covariate data for 132 patients from an actual, unmatched, randomized experiment, we construct 66 pairs matching for 14 covariates. We then create 10 000 unmatched and 10 000 matched randomized experiments by repeatedly randomizing the 132 patients, and compare the covariate balance with and without matching. By every measure, every one of the 14 covariates was substantially better balanced when randomization was performed within matched pairs. Even after covariance adjustment for chance imbalances in the 14 covariates, matched randomizations provided more accurate estimates than unmatched randomizations, the increase in accuracy being equivalent to, on average, a 7% increase in sample size. In randomization tests of no treatment effect, matched randomizations using the signed rank test had substantially higher power than unmatched randomizations using the rank sum test, even when only 2 of 14 covariates were relevant to a simulated response. Unmatched randomizations experienced rare disasters which were consistently avoided by matched randomizations.

Keywords: Experimental design; Matched experiment; Network optimization; Randomized experiment.

*To whom correspondence should be addressed.

1. GAINS FROM BLOCKING IN RANDOMIZED EXPERIMENTS

In a stratified-block randomized experiment, subjects are grouped into blocks of equal size, and a fixed fraction of each block is randomly assigned to each treatment under study. In the simplest case, there are two treatments, blocks are pairs of two subjects, and one subject is picked at random in each pair to receive each treatment.

Blocking on relevant covariates before randomization can increase balance on these covariates, increase the efficiency of estimation and the power of hypothesis tests, and reduce the required sample size for fixed precision or power (Fisher, 1935, Chapter 4; Cochran and Cox, 1957, Section 4.26; Cox, 1958, Chapter 2; Palta, 1985; Matts and Lachin, 1988; Piantadosi, 1997, Section 9.3). Even when covariance adjustment is used to control chance imbalances in covariates in randomized experiments, the adjusted estimate is more precise if the covariates are more nearly balanced (Snedecor and Cochran, 1980, Section 18.2, p. 368, expression 18.2.3). What if the covariates are irrelevant? Blocking or pairing on irrelevant covariates wastes a small amount of computer time but does not harm statistical efficiency or power (Cochran and Cox, 1957, Section 4.26; Chase, 1968).

In experiments, pairing has typically been based on one or two covariates divided into coarse categories. In contrast, in nonrandomized observational studies, optimal multivariate matching on many covariates at once is, nowadays, quite common. See Rosenbaum and Rubin (1985) for a case-study of multivariate matching on 20 covariates in an observational study, Rosenbaum (1989) for discussion of optimal matching, Gu and Rosenbaum (1993) for an evaluation by simulation. Here, we discuss an algorithm for optimal multivariate matching before randomization in experiments and illustrate it by comparing unpaired with paired randomization using baseline covariates from an experiment.

The method we describe is useful when all subjects (or large groups of subjects) are randomized at the same time. For instance, it will often be applicable to group-randomized designs in which whole communities are randomly assigned to treatment or control (e.g. Green *et al.*, 1995; Christian *et al.*, 2000). With a little planning, the method is applicable to many studies that induce brief but interesting effects, such as pain or headache, by randomizing healthy volunteers or patients with chronic symptoms that pose no urgent danger (e.g. Ashina *et al.*, 2000). The method may be used in randomized PET and fMRI studies (e.g. Ernst *et al.*, 2001; Sperling *et al.*, 2002) and randomized studies of the effects of drugs on nonhuman primates (e.g. Lori *et al.*, 2000). Also, the method is applicable to most studies in experimental psychology or experimental economics that randomly assign interesting but harmless treatments to undergraduates (e.g. Thaler *et al.*, 1997).

The method we describe is not useful for clinical trials that gradually accrue patients over a period of years. A sequential strategy useful for such trials was proposed by Pocock and Simon (1975); their method makes assignment decisions one at a time, looking only at the covariate data for previously assigned patients. That method would not typically be used if all subjects were randomized at once, because in this situation, there is no need to accept a suboptimal assignment by limiting the covariate information to previously assigned subjects. Randomization inference for Pocock–Simon randomizations is substantially more complex than for the randomized paired designs we develop.

2. OPTIMAL DIVISION OF A SINGLE GROUP INTO PAIRS

Matching before randomization in an experiment differs from matching in an observational study. In an observational study, existing treated and control groups are matched—one searches for the best control for a treated subject—whereas in a randomized experiment, subjects are divided into pairs before they are assigned to treatment or control. That is, for an observational study, one seeks the best pairing of two existing groups—an algorithmic problem called optimal bipartite (i.e. two part) matching—whereas in an experiment, one seeks the best pairing of subjects from a single group—an algorithmic problem called

optimal nonbipartite matching; see Papadimitriou and Steiglitz, (1998, Sections 11.2 and 11.3). For this reason, we use an algorithm and Fortran code due to Derigs (1988) for optimal nonbipartite matching.

Optimal nonbipartite matching of $2n$ subjects begins with an $(2n) \times (2n)$ distance matrix, which gives the distance between every pair of subjects in terms of their baseline covariates. One distance for matching is the Mahalanobis distance (Rubin, 1979), which may be applied to the covariates themselves, or else to their ranks to limit the impact of a few extreme observations. In nonbipartite matching, the covariance matrix in the Mahalanobis distance is computed from all $2n$ subjects.

If desired, certain pairings can be forbidden by defining an infinite distance between two subjects who must not be paired. Infinite distances between individuals of different gender require men to be matched to men, women to women. Infinite distances are more flexible than nonoverlapping strata. One can insist that matched subjects differ by no more than five years in age by setting an infinite distance between subjects who differ in age by more than five years. In this case, a 40 year old could be matched to either a 37 year old or a 44 year old, but there would be an infinite distance between the 37 year old and the 44 year old, so they could not be matched. The Mahalanobis distance may still incorporate age, so in addition to absolute requirement that age not differ by more than five years, the 37 year old would be judged slightly better than the 44 year old as a match for the 40 year old.

An optimal nonbipartite matching divides the $2n$ subjects into n pairs of two subjects to minimize the sum of the n distances within pairs. A naive approach is to pair the two people with the smallest distance, set them aside, pair the two remaining people with the smallest distance, etc., a so-called *greedy algorithm*. Greedy algorithms do not produce optimal nonbipartite matchings. Suppose there were eight people with ages 24, 35, 39, 40, 40, 41, 45, 56, so the distribution is symmetric about 40, and is most dense near its center, as is true, for example, of the Normal distribution. Greedy would first pair (40, 40), then (39, 41), then (35, 45), then (24, 56), so the total absolute difference within pairs is $0 + 2 + 10 + 32 = 44$. In contrast, an optimal matching would pair $(24, 35)$, $(39, 40)$, $(40, 41)$, $(45, 56)$, so the total absolute difference within pairs is $11 + 1 + 1 + 11 = 24$. Multivariate matching is more complex, but the same problem arises: by not looking ahead, greedy ends by producing many poorly matched pairs.

3. METHODS USED IN THE CASE-STUDY

3.1 *Logic of the comparison*

To illustrate the performance of the matching procedure, we compare the probabilities of covariate imbalance with complete randomization and with optimal matching followed by randomization within pairs. We start with an actual, unmatched randomized experiment using nine continuous and five binary baseline covariates for $2n = 132$ subjects from the experiment, and optimally match the subjects into $n = 66$ pairs. We repeatedly randomize the subjects, producing 10 000 unmatched experiments and 10 000 matched experiments, and compare covariate balance. The experiment is just a source for a distribution of baseline covariates.

3.2 *The pretreatment covariates*

The ACE-Inhibitor After Anthracycline (AAA) Study (Silber *et al.*, 2001, 2003) concerned children who had survived at least four years after cancer diagnosis, at least two years after the completion of all cancer treatment, and who had certain defined forms of decline in cardiac systolic performance after treatment with an anthracycline. It was hoped that the treatment, enalapril, would improve cardiac function. A total of 69 children were randomly assigned to enalapril and 66 to placebo. To permit an equitable comparison of matched and unmatched designs, we discarded three enalapril children using random numbers, leaving 132 subjects, of whom 66 will be randomized to treatment and 66 to control.

Table 1. *Mahalanobis distances before and after matching*

	Quartile			Mean
	25%	50%	75%	
Before matching ($n = 8646$)	21.0	27.4	34.7	28.2
After matching ($n = 66$)	7.0	8.9	12.0	9.9

The nine continuous covariates were: MCI = maximal cardiac index, L/min/m^2; WS = left ventricular end systolic wall stress, gm/cm^2; V_{cfc} = rate adjusted velocity of fiber shortening; SF = shortening fraction %; EF = ejection fraction %; QT$_c$, milliseconds; ANTDOSE = total anthracycline dose, mg/m^2; BMI = body mass index; AGE = age at baseline. The five binary covariates were: Male, RADS = heart irradiation, AD = anthracycline dose $\geqslant$ 300 mg/m^2, Black, and Hispanic. Anthracycline dose appears twice, as a continuous and as a binary variable. It is common in randomized experiments to present tables and statistical tests describing subjects at baseline before treatment, as a demonstration that the randomization succeeded in producing initially comparable groups; see, for instance, Tables 1 and 2 in Silber *et al.* (2003). Although the randomization in the actual experiment appears to have been entirely successful by any standard criteria, two of these baseline variables did show some imbalance: race differed significantly ($P = 0.024$ in their Table 1) with 1 black on enalapril and 9 on placebo, and QT$_c$ differed somewhat ($P = 0.056$ in their Table 2). One expects such imbalances from randomization when many covariates are examined—after all, 1 out of 20 covariates should, by accident, differ significantly at the 0.05 level—but one hopes for as much covariate balance as possible. Section 4 compares balance in matched and unmatched experiments.

3.3 *Covariate distances between subjects*

We replaced each of the nine continuous covariates by their ranks, ranking from 1 to 132, and appended the five binary covariates, making a 132×14 matrix. We then computed the Mahalanobis distance between each pair of subjects, yielding a symmetric 132×132 matrix of distances. If subjects are paired, there are 66 distances within the 66 pairs. The optimal matching minimizes the total of the 66 distances over all possible ways to pick 66 pairs.

In terms of the Mahalanobis distance, matched children were much more similar than a pair of children picked at random. Table 1 compares the distances among the $\binom{132}{2} = 8646$ possible pairings of two of the children with the 66 pairings produced by optimal matching. The means and quartiles are about one-third as large in matched pairs compared to two children picked at random. Whether or not this translates into something useful is the topic of Section 4.

3.4 *Creating and evaluating the experiments*

We created 10 000 unmatched randomizations splitting 132 children into two groups of 66, and we created 10 000 matched randomizations in which one child in each of 66 matched pairs was picked at random for each group. For each experiment, and for each of the 14 covariates, $k = 1, \ldots, 14$, there are 66 values, say $y_{k1}, \ldots, y_{k,66}$, of the covariate among the 66 treated subjects, and 66 values, say $x_{k1}, \ldots, x_{k,66}$, of the covariate among the 66 controls. The merged set of 132 values of the kth covariate never changes; the experiments differ only in how randomization partitions the 132 children into two groups of 66. Are the partitions better when formed by randomizing within matched pairs?

We answer this question in various ways. All ways compare the 66 treated children and the 66 control children in terms of the comparability of the two groups with respect to the distributions of the

14 covariates. Some measures focus on the typical imbalance in individual covariates, say the absolute value of the difference in means (Section 4.1) or odds ratios for binary covariates (Section 4.3). Others in Section 4.2 focus on significance levels or P-values as in Section 3.2.

A common suggestion is to control imbalances in covariates using a model such as covariance adjustment; however, theory shows covariance adjustment is more efficient—in the sense of smaller variance of an unbiased estimate—when the covariate imbalance is smaller. We compare the efficiency of covariance adjustment estimates in matched and unmatched randomized experiments; see Section 4.4. Familiar least squares calculations show that relative efficiency is a function of covariate imbalance, and we calculate this function in our simulated experiments.

Because we are interested in a fair comparison of two competing methods of randomization, we always apply the same measure of covariate imbalance to both types of randomized experiment—i.e., if we changed both the measure of imbalance and the method of randomization at the same time, we would not know whether a difference was produced by a change in method or a change in measure. In particular, we do not use measures of covariate imbalance for matched pairs, because they cannot be computed for the unmatched randomizations. This issue is most prominent when computing P-values, because we compute unmatched P-values for both matched and unmatched randomizations. This might be a concern if only P-values were used as measures of covariate imbalance. However, all of our measures end up pointing in the same direction; hence, idiosyncrasies in any one measure cannot explain consistent results over many measures.

For each randomization, there is an opposite randomization which reverses the roles of the y and the x, making treated subjects into controls and controls into treated subjects. As a consequence, it makes sense to ask how balanced the y and x are, but not in a way that attaches importance to which group is labeled y and which is labeled x. All measures we report are unchanged by relabeling.

4. MATCHED AND UNMATCHED RANDOMIZATIONS

4.1 *Absolute difference in means*

The first measure of covariate imbalance for the nine continuous covariates was simply the absolute value of the difference in means, $M_k = \left|\frac{1}{66}\sum_{j=1}^{66} y_{kj} - \frac{1}{66}\sum_{j=1}^{66} x_{kj}\right|$, a measure of the typical difference. As absolute values are taken, M_k is always nonnegative, whereas, without absolute values, M_k would have expectation 0 over repeated randomizations. We average M_k over the 10 000 randomizations, and then compute the ratio of the two averages, matched/unmatched, which is called the 'bias ratio' in Table 2. In Table 2, the absolute difference in means is, on average, 25% to 35% smaller when randomization is performed within matched pairs than when performed without matching.

4.2 *P-values from the rank sum test*

Table 2 also describes the two-sided P-values for Wilcoxon's two sample rank sum test comparing the x and the y; see Section 3.4 for discussion. Because the rank sum test is a randomization test for a study with complete (unmatched) randomization, we expect 5% of these P-values to be less than 0.05 with unmatched randomization, and Table 2 confirms this, with about 500 P-values less than 0.05 in 10 000 randomizations. In the final column of Table 2, the rank sum P-value is used solely as a comparable measure of covariate imbalance in the matched samples. Here, for all nine covariates, the P-value is less than 0.05 in substantially less than 0.5% of randomizations, and 0.2% is more nearly typical. The balance is better with matched randomizations.

In each of the 20 000 randomized experiments, we computed the minimum of the nine two-sided P-values from Wilcoxon's rank sum test applied to the nine continuous covariates. With nine *independent*

Table 2. *Comparing covariate balance with matched and unmatched randomization for nine continuous covariates in 10 000 experiments*

Covariate	Bias ratio	Unmatched: # of $P \leqslant 0.05$ in 10 000 randomizations	Matched: # of $P \leqslant 0.05$ in 10 000 randomizations
MCI	65%	544	21
WS	65%	458	21
V_{cfc}	79%	512	21
SF	68%	479	12
EF	64%	511	6
QTC	72%	507	17
ANTDOSE	62%	438	2
BMI	75%	503	37
AGE	63%	506	5

Table 3. *Minimum P-value over nine continuous covariates for 10 000 unmatched and 10 000 matched randomizations*

	Median	Low quartile	# < 0.10	# < 0.05	# < 0.01
Unmatched	0.081	0.034	5829	3416	807
Matched	0.272	0.183	732	139	1

covariates, one would expect $1 - (1 - 0.05)^9 = 37\%$ of randomizations to yield a minimum P-value less than 0.05. Table 3 describes the distribution of the minimum P-value. In $34\% = 3416/10\,000$ of the unmatched randomizations, the minimum P-value was less than 0.05, and in $8\% = 807/10,000$ it was less than 0.01, whereas with matched randomizations, the corresponding percentages were 1.4% instead of 34% and 0.01% instead of 8%. The worst of nine imbalances is much better with matched randomization compared to unmatched randomization. In fact, the worst imbalance for nine covariates with matching is smaller than expected for a single covariate without matching, in that only 1.4% of the minimum P-values with matching was less than 0.05, whereas 5% are expected by chance for a single covariate without matching.

It is important to keep in mind that Tables 2 and 3 hold the two methods of randomization to the same standard, finding better covariate balance with matched randomization. The picture would, of course, be different if the two methods were held to different standards; see Section 3.4. If the rank sum test were applied to unmatched randomizations and the signed rank test were applied to matched randomizations, then, because both are randomization tests for the corresponding randomizations, theory would lead one to expect 5% rejections at the 0.05 level. However, the signed rank test for matched pairs examines only differences *not controlled* by the matching, and so it is not the appropriate way to measure what differences *were controlled* by the matching. The signed rank and rank sum tests are compared in Section 4.5.

4.3 *Binary covariates*

For each of the 20 000 randomizations, for each of the five binary covariates, there is a 2×2 contingency table, $treatment \times covariate$. The odds ratio can be computed from a 2×2 in two ways, so that one odds ratio is the reciprocal of the other; we select the odds ratio greater than one, which is analogous to absolute values of mean differences in Section 4.1. For the kth of the five variables, we calculated the

upper quartile ω_{mk} of these odds ratios over the 10 000 matched randomizations, and the upper quartile ω_{uk} over the 10 000 unmatched randomizations, where by definition, $\omega_{mk} \geqslant 1$ and $\omega_{uk} \geqslant 1$. Balance is an odds ratio of 1. The five unmatched ω_{uk} ranged from 1.48 to 2.62, while the five matched ω_{mk} ranged from 1.00 to 1.35, so the matched randomizations were better balanced for all five binary covariates, $1 \leqslant \max_k \omega_{mk} \leqslant \min_k \omega_{uk}$.

The two-sided P-value from Fisher's exact test is twice the smaller of the two one-sided P-values. For unmatched randomizations, Fisher's exact test behaved as statistical theory says it should behave: because of discreteness, it gave P-values $\leqslant 0.05$ in slightly fewer than 5% of the 10 000 randomizations for each variable. In sharp contrast, none of the $5 \times 10\,000$ P-values for matched randomization was significant at either 0.10 or 0.05. The binary variables were better balanced by matched randomization.

4.4 *An efficiency measure*

Whether or not matching has been used, one might adjust for the 14 covariates using covariance adjustment; see Rubin (1979) who compares several approaches, including conventional unmatched covariance adjustment applied to both matched and unmatched samples. Covariance adjustment is more efficient when the covariates are more nearly balanced, and the gain can be measured in a simple way.

Let $\mathbf{W}$ be the 132×15 matrix whose rows correspond to the 132 subjects, and whose columns are the 14 covariates plus a column of 1's for a constant term. Let $\mathbf{V}$ be the 132×1 vector which is 1 for someone assigned to treatment and -1 for someone assigned to control, so $\sum_{i=1}^{132} V_i = 0$ as there are always 66 treated subjects. Covariance adjustment is a linear least squares regression of some outcome on the predictors $[\mathbf{V}, \mathbf{W}]$ with the first of the 16 estimated regression coefficients, that is the coefficient of $\mathbf{V}$, reported as half the estimated treatment effect. Under Gauss–Markov assumptions—a linear model with additive errors that are uncorrelated with constant variance σ^2—the 16 estimated coefficients have covariance matrix,

$$\sigma^2 \begin{bmatrix} \mathbf{V}^T\mathbf{V} & \mathbf{V}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{V} & \mathbf{W}^T\mathbf{W} \end{bmatrix}^{-1}$$

whose (1, 1) element is

$$\frac{\sigma^2}{\mathbf{V}^T\left[\mathbf{I} - \mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{W}^T\right]\mathbf{V}} = \frac{\sigma^2}{\mathbf{V}^T\mathbf{Q}\mathbf{V}} \text{ with } \mathbf{Q} = \mathbf{I} - \mathbf{W}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\mathbf{W}^T, \tag{1}$$

by properties of the inverse of a partitioned matrix (e.g. Rao, 1973, p. 33). The variance $\sigma^2/\mathbf{V}^T\mathbf{Q}\mathbf{V}$ of the estimated treatment effect in (1) is smallest when the covariates are most nearly balanced: that is, when $\mathbf{V}$ is orthogonal to the columns of $\mathbf{W}$ so that $\mathbf{W}^T\mathbf{V} = \mathbf{0}$. It follows that $\mathbf{V}^T\mathbf{Q}\mathbf{V}$ is a measure of covariate balance directly relevant to the efficiency of covariance adjustment: as $\mathbf{V}^T\mathbf{Q}\mathbf{V}$ becomes larger, the sample becomes more balanced, and the standard error of the covariance adjusted estimated treatment effect becomes smaller. A 5% increase in $\mathbf{V}^T\mathbf{Q}\mathbf{V}$ is effectively the same as a 5% increase in sample size. Reversing the group labels makes $\mathbf{V}$ into $-\mathbf{V}$, so $\mathbf{V}^T\mathbf{Q}\mathbf{V}$ is unchanged.

For each randomization, we compute $\mathbf{V}^T\mathbf{Q}\mathbf{V}$, where larger values are preferred. In Table 4, $\mathbf{V}^T\mathbf{Q}\mathbf{V}$ is higher for the matched rather than unmatched randomizations, equivalent to an increase in effective sample size of about 7% on average. The unmatched randomizations are not only less efficient, but also less stable in their performance—the interquartile range is more than twice as large—and the worst performance is much worse. The worst of 10 000 matched randomizations is comparable to the 25% percentage point of unmatched randomizations.

Table 4. *Effect of matching on the efficiency of covariance adjustment in 10 000 randomizations*

	min	Quartile 25%	Quartile 50%	Quartile 75%	Max	Mean
Unmatched randomizations	92.83	114.85	118.46	121.48	129.80	117.87
Matched randomizations	114.00	124.87	126.50	127.84	131.32	126.20
Gain in efficiency due to matching	22.8%	8.7%	6.8%	5.2%	1.2%	7.1%

In short, the use of matched rather than unmatched randomization prior to covariance adjustment in this example is, on average, equivalent to about a 7% increase in sample size with no additional cost. The average tells only part of the story, however. Matched randomization is much less likely than unmatched randomization to produce a bad randomization that results in substantial losses in efficiency.

4.5 *Power of randomization tests*

Does matching affect the power of randomization tests? As a small illustration, we created an artificial response variable R in the following way. The response is intended to simulate the post-treatment maximum cardiac index MCI, which was the primary endpoint in the experiment. We cannot use the actual post-treatment MCI because this is known only for the randomization actually performed, whereas we need values for every possible randomization. The artificial MCI was equal to the baseline MCI, minus 5 if the subject received heart irradiation prior to the experiment (RADS $= 1$), plus a treatment effect of either $\tau = 0$ or $\tau = 2.5$ if the subject was randomized to enalapril, plus independent Normal errors with mean zero and standard deviation 5. Notice that we matched on 14 covariates, but only two were relevant, and our algorithm did not know which were relevant. Each randomization changes the pattern of treatment effects, and we drew new Normal errors for each randomization. In each randomization-simulation step, we tested the null hypothesis of no treatment effect by an appropriate randomization test, the rank sum test for unmatched randomizations, the signed rank test for matched randomizations, recording whether or not the null hypothesis was rejected. All tests were two-sided, 0.05 level tests. When the treatment effect was $\tau = 0$, the proportion of rejections estimates the levels of tests that aim to have level 0.05. When the treatment effect was $\tau = 2.5$, the proportion of rejections estimates the power of the test. We increased precision using antithetic variables.

Both tests had the correct level: with $\tau = 0$, the fraction of rejections was 5.1% for both the signed rank test applied to matched randomizations and the rank sum test applied to unmatched randomizations. The power was higher with matched randomizations: with $\tau = 2.5$, the signed rank test rejected the null hypothesis for 71.9% of matched randomizations, while the rank sum test rejected the null hypothesis for 58.3% of unmatched randomizations. Standard errors were $\pm 0.2\%$.

5. THEORETICAL COMPARISONS

In the case study in Section 4.4, covariance adjustment of matched experiments was, on average, about 7% more efficient than covariance adjustment of unmatched experiments. Here, we further explore this comparison in a few tractable theoretical cases. As it turns out, the gain in efficiency from matching before randomization can be substantially larger or substantially smaller than 7%. In the situations we examine, the gain in efficiency increases as the number of covariates increases, and it decreases as the sample size increases.

In the notation of Section 4.4, $E\left(\mathbf{V}^T\mathbf{QV}\right) = \text{tr}\left\{E\left(\mathbf{VV}^T\right)\mathbf{Q}\right\}$, where the expectation is over the randomization distribution, so we need to compute $E\left(\mathbf{VV}^T\right)$ for unmatched and matched randomizations. Because $V_i^2 = 1$, the $2n \times 2n$ matrix $E\left(\mathbf{VV}^T\right)$ has 1's along the diagonal. Now consider the off-diagonal, $i \neq j$, so that $V_i \times V_j$ is ± 1. For unmatched randomization, an easy counting argument shows $E\left(V_i \times V_j\right) = -1/(2n-1)$. For matched randomization, number the $2n$ subjects so adjacent subjects are matched, $2k-1$ matched to $2k$, $k = 1, \ldots, n$; then $V_{2k-1} = -V_{2k}$, so $E\left(\mathbf{VV}^T\right)$ is block diagonal, with n blocks, each a 2×2 matrix with 1's on the diagonal and -1's off the diagonal.

Table 5 displays the percentage increase in $E\left(\mathbf{V}^T\mathbf{QV}\right) = \text{tr}\left\{E\left(\mathbf{VV}^T\right)\mathbf{Q}\right\}$ due to optimal matching before randomization for several structured covariate matrices, with $k = 3, 7, 11$ covariates and $2n = 16, 32, 64, 128$ observations. This percent increase in $E\left(\mathbf{V}^T\mathbf{QV}\right)$ is the percentage increase in effective sample size; see Section 4.4. In all cases, the covariates are ± 1 and are given by complete or fractional factorial designs from Box *et al.* (1978, p. 410, Table 12.15). A design $r \times 2^{k-p}$ is a two-level (fractional) factorial with k factors, in 2^{k-p} distinct treatment combinations, each combination replicated r times, for a total of $r \times 2^{k-p}$ observations. For instance, 8×2^3 is a complete factorial in three factors, eight treatment combinations, replicated eight times, for 64 observations. Similarly, $1 \times 2^{7-3}$ is a $\frac{1}{8}$ fraction of a 2^7 factorial, with seven factors, in $2^{7-3} = 16$ observations, with each treatment combination appearing only once. When $r \geqslant 2$ with r even, there are at least two observations with exactly the same covariate values for all covariates, so an exact matching on the covariates exists, and $E\left(\mathbf{V}^T\mathbf{QV}\right) = 2n$ for the matched experiment. When $r = 1$, no two observations have the same covariate values, and no exact matching exists. Indeed, when $r = 1$, the matching is very poor, because in a good unreplicated fractional factorial, the treatment combinations are extremely dispersed among corners of the 2^k cube with vertices ± 1; see the figure in Box *et al.* (1978, p. 387, Table 12.4). For example in the $1 \times 2^{7-2}$ design with seven factors in 32 observations, every one of the $\binom{32}{2} = 496$ pairs of two observations differ in at least two covariates. If the ± 1 had been generated by coin flips, rather than by the fractional factorial, then covariates would have been closer. For instance, suppose the first subject had covariate values specified by a particular pattern of seven of the ± 1. How close would the remaining 31 subjects be to this first subject if the covariates had been generated by coin flips? Take the pattern for the first subject as given. From the binomial distribution, the chance that another subject will differ from the first subject by 0/7 or 1/7 covariates is $0.0625 = \left(\frac{1}{2}\right)^7 + 7\left(\frac{1}{2}\right)^7$. The chance that at least one of the 31 other subjects differs from the first subject on at most one covariate is then $1 - (1 - 0.0625)^{31} = 0.865$. In other words, 86.5% of the time, coin flips would produce a closer match for the first subject than would the $1 \times 2^{7-2}$ design. In short, when $r \geqslant 2$, matching is unnaturally easy, but when $r = 1$ matching is unnaturally difficult. Together, the cases $r \geqslant 2$ and $r = 1$ give an indication of the range of what is possible.

Table 5 exhibits interesting patterns. The case-study in Section 4.4 had $k = 14$ covariates and $2n = 132$ observations, so it most closely resembles the last two situations in Table 5, with $k = 11$ covariates and $2n = 128$ observations. The 7.1% average gain in Section 4.4 is just about half-way between the 9.5% gain for the exact match in the $2 \times 2^{11-5}$ design in Table 5 and the 4.4% gain for the poor match $1 \times 2^{11-4}$ design. In Table 5, the gain from matching before randomization is greater when there are more covariates and when there are fewer observations. With three covariates and 128 observations, the 2.4% gain is probably not worth the effort, but substantial increases, often above 10%, in effective samples size are common in Table 5, even in the unnaturally 'poor match' situation.

6. How many covariates?

How many covariates and which covariates should be used in matching? This decision is made with the covariates in hand, before randomization, before the outcomes are available. Obviously, if the

Table 5. *Gain in efficiency of covariance adjustment from matched randomization: percentage increase in effective sample size*

Covariates k	Sample size, 2*n*	Design	Comment	Gain in efficiency (%)
3	16	2×2^3	Exact match	25.0
3	32	4×2^3	Exact match	10.7
3	64	8×2^3	Exact match	5.0
3	128	16×2^3	Exact match	2.4
7	16	$2 \times 2^{7-4}$	Exact match	87.5
7	16	$1 \times 2^{7-3}$	Poor match	17.2
7	32	$2 \times 2^{7-3}$	Exact match	29.2
7	32	$1 \times 2^{7-2}$	Poor match	13.0
11	32	$2 \times 2^{11-7}$	Exact match	55.0
11	32	$1 \times 2^{11-6}$	Poor match	16.2
11	64	$2 \times 2^{11-6}$	Exact match	21.2
11	64	$1 \times 2^{11-5}$	Poor match	9.8
11	128	$2 \times 2^{11-5}$	Exact match	9.5
11	128	$1 \times 2^{11-4}$	Poor match	4.4

experimenter knew with certainty which covariates will matter for the outcomes measured later, and if there was general scientific consensus about which covariates matter, then matching would focus on these covariates; however, an experimenter may work in a field in which certainty and consensus are, at times, unavailable.

In Sections 4.1–4.3, multivariate matching before randomization substantially improved the comparability of treated and control groups at baseline for 14 covariates simultaneously. In Sections 4.4 and 5, multivariate matching before randomization meaningfully increased the efficiency of covariance adjustments, increasing the effective sample size, at no additional cost. In the artificial illustration in Section 4.5, only two of the 14 covariates mattered for the simulated outcome, and the matching algorithm had no information about which two were important and which 12 were irrelevant, yet matching on all 14 covariates substantially increased the power of a conventional randomization test while preserving the level of the test. In this case study, matching on 14 covariates was much better than complete randomization.

How many covariates should be used in matching? Obviously, there is no one answer for all contexts, but the case study leads to two relevant suggestions. First, the common practice in randomized experiments is either complete randomization, without matching, or matching on one or two covariates. The case study suggests that it is practical to match on substantially more than two covariates, with substantial benefits.

Second, as our case study exemplifies, the baseline covariates in an experiment may be studied empirically, before randomization, to decide which of several competing randomized designs has the best operating characteristics over repeated randomizations. Suppose that an experimenter has the baseline covariate information for the subjects in an experiment and is about to randomize those subjects to treatment or control. In principle, this experimenter could use the covariates to carry through the calculations we performed in our case study to decide whether to use complete randomization or matched randomization, or alternatively, to decide whether to match on 5 or 25 covariates. (Although the initial programming of the case study was time-consuming, once programmed, the computer's calculating effort

was comparable to many Markov chain Monte Carlo analyses that are, today, performed routinely.) A less ambitious experimenter might simply construct the optimal match with 5 covariates and with 25 covariates, and, before randomizing, examine the covariates in the resulting two sets of matched pairs. In short, a general rule is not needed; rather, the experimenter can take a look at the matched pairs before randomization.

7. Analysis of a matched randomization

In a matched randomized experiment, a variety of methods of analysis are available. For instance, Wilcoxon's signed rank test is a randomization test for the null hypothesis of no treatment effect (Lehmann, 1998, Section 3.2, p. 123). If the treatment effect is constant, not varying from person to person, then an exact, randomization based confidence interval for the constant effect may be formed by inverting the test. Alternatively, if the treatment effect is not constant, then the signed rank test may be inverted in a different way to obtain exact, randomization based confidence intervals for a measure of the magnitude of effect attributable to treatment (Rosenbaum, 2003, Section 3).

Multivariate matching reduces covariate imbalances, but the pairs are imperfectly matched. One might wish to reduce these imperfections with some form of covariance adjustment, and still use randomization as the 'reasoned basis for inference,' in Fisher's (1935) phrase. That is, one might wish to obtain an exact, randomization based confidence interval for a constant effect, with covariance adjustment for imperfections in the matching of covariates, without assuming the covariance adjustment model is true, and without distributional assumptions. This is possible, and the straightforward steps are described in Rosenbaum (2002, Section 4).

Subjects may be randomized in pairs, but a few subjects may fail to provide the needed outcome data, leaving a treated subject with no matching control or a control with no matching treated subject. If the missing data were missing completely at random, then it is possible to retain the broken pairs using methods proposed by Wei (1982).

8. Summary

In our case study, by every measure, for every one of the 14 covariates, randomization within optimally matched pairs produced better covariate balance than did complete randomization. Whether judged by means, odds ratios or P-values, the improvements in covariate balance were substantial. The power of conventional randomization tests of no treatment effect was greater with matched randomizations; see Section 4.5. Although it is possible, under Gauss–Markov assumptions, to remove chance imbalances by covariance adjustment, those adjustments yield more precise estimates when performed in a matched experiment. In our case study, the typical reduction in standard error of this adjusted estimate was equivalent to an increase in sample size of about 7%; see Section 4.4. Moreover, unmatched randomization produced occasional disasters which were consistently avoided by matched randomizations; e.g., rare binary variables totally confounded with treatment in Section 4.3 or a 23% loss of efficiency in Section 4.4. The method entails a small amount of additional computation and is practical when all subjects, or large groups of subjects, are randomized at the same time.

Acknowledgements

This work was supported by a grant from the US National Science Foundation and by grant R01 HL-50424 from the US National Heart, Lung and Blood Institute.

REFERENCES

ASHINA, M., BENDTSEN, L., JENSEN, R. AND OLESEN, J. (2000). Nitric oxide-induced headache in patients with chronic tension-type headache. *Brain* **123**, 1830–1837.

BOX, G. E. P., HUNTER, W. G. AND HUNTER, J. S. (1978). *Statistics for Experimenters*. New York: Wiley.

CHASE, G. R. (1968). On the efficiency of matched pairs in Bernoulli trials. *Biometrika* **55**, 365–369.

CHRISTIAN, P., WEST, K. P., SUBARNA, K. K. *et al.* (2000). Night blindness during pregnancy and subsequent mortality among women in Nepal. *American Journal of Epidemiology* **152**, 542–547.

COCHRAN, W. AND COX, G. (1957). *Experimental Designs*. New York: Wiley.

COX, D. R. (1958). *Planning of Experiments*. New York: Wiley.

DERIGS, U. (1988). Solving nonbipartite matching problems by shortest path techniques. *Annals of Operations Research* **13**, 225–261.

ERNST, M., MATOCHIK, J. A., HEISHMAN, S. J. *et al.* (2001). Effect of nicotine on brain activation during performance of a memory task. *Proceedings of the National Academy of Sciences* **98**, 4728–4733.

FISHER, R. (1935). *Design of Experiments*. Edinburgh: Oliver and Boyd.

GREEN, S. B., CORLE, D. K., GAIL, M. H. *et al.* (1995). Interplay between design and analysis for behavioral intervention trials with community as the unit of randomization. *American Journal of Epidemiology* **142**, 587–593.

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.

LEHMANN, E. L. (1998). *Nonparametrics*. Englewood Cliffs, NJ: Prentice-Hall.

LORI, F., LEWIS, M. G., XU, J. *et al.* (2000). Control of SIV rebound by structured treatment interruptions during early infection. *Science* **290**, 1591–1593.

MATTS, J. AND LACHIN, J. (1988). Properties of permuted-block randomization in clinical trials. *Controlled Clinical Trials* **9**, 327–344.

PALTA, M. (1985). Investigating maximum power losses in survival studies with nonstratified randomization. *Biometrics* **41**, 497–504.

PAPADIMITRIOU, C. H. AND STEIGLITZ, K. (1998). *Combinatorial Optimization: Algorithms and Complexity*. New York: Dover.

PIANTADOSI, S. (1997). *Clinical Trials*. New York: Wiley.

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

RAO, C. R. (1973). *Linear Statistical Inference and its Application*. New York: Wiley.

ROSENBAUM, P. R. (1989). Optimal matching in observational studies. *Journal of the American Statistical Association* **84**, 1024–1032.

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17**, 286–327.

ROSENBAUM, P. R. (2003). Exact confidence intervals for nonconstant effects by inverting the signed rank test. *American Statistician* **57**, 132–138.

ROSENBAUM, P. AND RUBIN, D. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician* **39**, 33–38.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74**, 318–328.

SILBER, J. H., CNAAN, A., CLARK, B. J. *et al.* (2001). Design and baseline characteristics for the ACE-inhibitor after anthracycline (AAA) study of cardiac dysfunction in pediatric oncology long-term survivors. *American Heart Journal* **142**, 577–585.

SILBER, J. H., CNAAN, A., CLARK, B. J. *et al.* (2003). Enalapril to prevent cardiac function decline in long-term survivors of pediatric cancer exposed to anthracyclines. *Journal of Clinical Oncology*.

SNEDECOR, G. W. AND COCHRAN, W. G. (1980). *Statistical Methods*, 7th Edition. Ames, IA: Iowa State University Press.

SPERLING, R., GREVE, D., DALE, A. *et al.* (2002). Functional MRI detection of pharmacologically induced memory impairment. *Proceedings of the National Academy of Sciences* **99**, 455–460.

THALER, R. H., TVERSKY, A., KAHNEMAN, D. AND SCHWARTZ, A. (1997). The effect of myopia and loss aversion on risk taking: an experimental test. *Quarterly Journal of Economics* 647–661.

WEI, L. J. (1982). Interval estimation of location difference with incomplete data. *Biometrika* **69**, 249–251.

[*Received March 3, 2003; revised October 16, 2003; accepted for publication October 30, 2003*]