Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

# Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates

Donald B. Rubin & Neal Thomas

**To cite this article:** Donald B. Rubin & Neal Thomas (2000) Combining Propensity Score Matching with Additional Adjustments for Prognostic Covariates, Journal of the American Statistical Association, 95:450, 573-585

**To link to this article:** http://dx.doi.org/10.1080/01621459.2000.10474233

Published online: 17 Feb 2012.

# Combining Propensity Score Matching With Additional Adjustments for Prognostic Covariates

Donald B. RUBIN and Neal THOMAS

Propensity score matching refers to a class of multivariate methods used in comparative studies to construct treated and matched control samples that have similar distributions on many covariates. This matching is the observational study analog of randomization in ideal experiments, but is far less complete as it can only balance the distribution of observed covariates, whereas randomization balances the distribution of all covariates, both observed and unobserved. An important feature of propensity score matching is that it can be easily combined with model-based regression adjustments or with matching on a subset of special prognostic covariates or combinations of prognostic covariates that have been identified as being especially predictive of the outcome variables. We extend earlier results by developing approximations for the distributions of covariates in matched samples created with linear propensity score methods for the practically important situation where matching uses both the estimated linear propensity scores and a set of special prognostic covariates. Such matching on a subset of special prognostic covariates is an observational study analog of blocking in a randomized experiment. An example combining propensity score matching with Mahalanobis metric matching and regression adjustment is presented that demonstrates the flexibility of these methods for designing an observational study that effectively reduces both bias due to many observed covariates and bias and variability due to a more limited subset of covariates. Of particular importance, the general approach, which includes propensity score matching, was distinctly superior to methods that focus only on a subset of the prognostically most important covariates, even if those covariates account for most of the variation in the outcome variables. Also of importance, analyses based on matched samples were superior to those based on the full unmatched samples, even when regression adjustment was included.

KEY WORDS: Bias reduction; Causal inference; Covariance adjustment; Discriminant matching; Mahalanobis metric matching; Matched sampling; Nonrandomized studies; Observational studies.

## 1. INTRODUCTION

### 1.1 Combining Propensity Score Matching and Prognostic Covariate Adjustment

Matched sampling is a methodology for reducing bias due to observed covariates in comparative observational studies. In a typical matching setting, covariate data are available for a large sample of potential control subjects but only a relatively small treated sample, and the goal of matching is to select a subset of the control sample that has covariate values similar to those in the treated sample.

When there are several covariates, one approach is to select control subjects based on their Mahalanobis distance from the treated subjects (Carpenter 1977; Cochran and Rubin 1973; Rubin 1976a, 1980), defined for covariate values $\mathbf{X}_1$ and $\mathbf{X}_2$ as $\{(\mathbf{X}_1 - \mathbf{X}_2)'\mathbf{S}_c^{-1}(\mathbf{X}_1 - \mathbf{X}_2)\}^{1/2}$, where $\mathbf{S}_c$ is the control sample covariance matrix. It is defined analogously for subsets of the covariates using corresponding submatrices of $\mathbf{S}_c$. Carpenter (1977) and Rubin (1980) found that Mahalanobis matching with bivariate normal data is effective, especially when combined with regression adjustment on the matched pair differences (Rubin 1979). As the number of covariates increases, however, the ability of this matching to find close matches on any specific covariate decreases, as does its ability to achieve similar treated and matched control means (Gu and Rosenbaum 1993; Rosenbaum and Rubin 1985b).

An alternative method, proposed by Rosenbaum and Rubin (1983), which is closely related to discriminant matching (Rubin 1976a,b), is to match on the propensity score, $e = e(\mathbf{X})$, defined for each subject as the probability of receiving treatment given the covariate values, $\mathbf{X}$, and thus a scalar function of $\mathbf{X}$. Rosenbaum and Rubin showed that if exact matching on $e(\mathbf{X})$ can be achieved, then $\mathbf{X}$ will have the same distribution in the treated and matched control samples, thus eliminating bias in treatment effect estimates due to differences in $\mathbf{X}$. In typical practice, the propensity scores are not known, and even when they are, exact matches are not available. A common practical approach, which we study, is to estimate $e(\mathbf{X})$ using the sample linear discriminant or a linear logistic regression model of the treatment/control indicator variable on $\mathbf{X}$. The estimated propensity score, $\hat{e}$, is then used in place of the propensity score.

We consider both propensity score and Mahalanobis metric matching, and a combination of these methods using caliper matching (Rosenbaum and Rubin 1985b). For the combined method, all control subjects within intervals (calipers) surrounding each treated subject's $\hat{e}$ are identified as potential matches, and then Mahalanobis metric matching is applied to a subset of key covariates, $\mathbf{X}^{(s)}$, to make final selections from these potential matches, where $\mathbf{X}' \equiv (\mathbf{X}^{(s)'}, \mathbf{X}^{(r)'})$. Additional adjustments may also be performed using regression methods applied to the matched datasets.

Analytic and simulation results, which extend earlier results of Rubin and Thomas (1992a, 1992b, 1996), show that combining propensity score matching with matching on $\mathbf{X}^{(s)}$ can eliminate most of the bias with respect to a linear model relating outcome variables to $\mathbf{X}$ and can substantially reduce bias with respect to more general models

Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. Neal Thomas is Principal Research Biostatistician, Bristol-Myers Squibb Company, Wallingford, CT 06492 (E-mail: *neal_thomas@attglobal.net*). The authors thank the associate editor and referees for exceptionally helpful comments.

with outcomes nonlinearly related to $\mathbf{X}^{(s)}$. Metric matching using $\mathbf{X}^{(s)}$, which may be functions of several original covariates, and model-based regression adjustments, should thus be viewed as complementary methods rather than competitive alternatives to propensity score matching and subclassification, as evaluated, for example, by Drake (1993).

The theory, example, and simulations that we present lead to two broader conclusions. First, the general approach to multivariate matching, which includes successful balancing of $\hat{e}$, is superior to methods that focus only on a subset of the prognostically most important covariates, even if those covariates account for most of the variation in the outcome variables. Second, analyses based on multivariate matched samples are superior to those based on the full unmatched samples, even if regression adjustment is used.

### 1.2 Example

We illustrate the effect of matching and regression adjustments on a dataset with 7,847 control children and 96 treated children whose mothers were exposed to a hormone during pregnancy (Reinisch, Sanders, Lykke-Mortensen, and Rubin 1995). The covariates, $\mathbf{X}$, in Table 1 were collected from medical records during a previous study and are derived from the same database utilized by Rosenbaum and Rubin (1985a). The $\mathbf{X}^{(s)}$ covariates chosen by behavioral researchers to receive special treatment in the matching are SEX of the child, age of the mother (MOTAGE), and an ordinal measure of socioeconomic status (SES).

Because collecting follow-up data is expensive and reducing the covariate differences between the treated and control samples is desirable, a subset of 96 control subjects was selected by matching for follow-up. The matching presented here illustrates the methods with real data but differs somewhat from that actually used. We also illustrate 1 (treated)-to-5 (control) matching, which can be practical in situations in which both the covariate and outcome data are available for the full treated and control samples for little or no additional cost. More complex methods that assign a variable number of control subjects to different treated subjects, or more than one treated subject to the same control subject, are alternative methods that can successfully extend matching techniques to difficult matching settings in which subsets of the treated subjects have very few potential matches (Rosenbaum 1989, 1991).

*Table 1. Matching Variables in the Hormone Exposure Data*

| *Label* | *Description* |
|---|---|
| SEX | Male indicated by a 1, female by a 0 |
| MOTAGE | mother's age |
| SES | Socioeconomic status (10 ordered categories) |
| ANTIH | Exposure to antihistamines |
| BARB | Exposure to barbiturates |
| CHEMO | Exposure to chemotherapy agents |
| CIGAR | Mother smoked cigarettes |
| MBIRTH | Multiple births |
| PSYDRUG | Exposure to psychotherapeutic drugs |
| RESPIR | Respiratory illness |
| SIB | First child indicated by a 1; other children, by a 0 |
| CAGE | Calendar time of birth |
| LGEST | Length gestation (10 ordered categories) |
| PBC415 | Pregnancy complication index |
| PBC420 | Pregnancy complication index |
| MOTHT | Mother's height (4 ordered categories) |
| MOTWT | Mother's weight gain (6 ordered categories) |
| PSC | Estimated linear propensity score defined in Section 2.2 |

NOTE: Additional matching is performed on the variables in the top portion of the table. The variables in the middle portion of the table are dichotomous; those in the bottom portion are ordinal or continuous.

We present the results of four matching methods, each of which applies exact matching to the SEX covariate, in strict analogy with blocking in a randomized experiment, before additional matching criteria are considered:

1. Matching on the estimated linear propensity scores, $\hat{e}$.
2. Identifying all potential matches within relatively coarse $\hat{e}$ calipers, followed by Mahalanobis metric matching on two special matching covariates, MOTAGE and SES.
3. Mahalanobis metric matching on the two special covariates, $\mathbf{X}^{(s)}$, only.
4. Mahalanobis metric matching on all of the covariates, $\mathbf{X}$.

The matched data sets produced by these four methods are compared in Section 4.

The propensity scores were estimated using logistic regression with linear terms for each covariate listed in Table 1, and the matching was done on the logistic scale, so that matching on $\hat{e}$ means matching on the logit of $\hat{e}$.

All of the matching was performed using nearest-remaining-neighbor matching, beginning with the treated subject with the highest (and thus most difficult to match propensity score) and proceeding to the subject with the lowest propensity score, selecting $m$ $(= 1 \text{ or } 5)$ matches for each treated subject. Rubin (1973a) found that this simple matching method produces near-optimal balancing of the sample means of a single matching covariate. The theoretical results in Section 3 predict that there is essentially no additional benefit for multivariate balancing from obtaining closer individual matching on $\hat{e}$ provided that close mean matching on the $\hat{e}$ is attained; this is consistent with results of Gu and Rosenbaum (1993) showing that optimal matching, although producing closer $\hat{e}$ than the nearest-neighbor algorithm, did not improve multivariate balance.

With the second method, calipers were formed around $\hat{e}$ for each treated subject by identifying all remaining unmatched control subjects within $\pm.2$ of the treated $\hat{e}$, with $\hat{e}$ standardized to have unit variance in the control sample; the choice of .2 is based on Cochran (1968) and Cochran and Rubin (1973).

Figure 1 contains a histogram of the $\hat{e}$ in the full treated sample with a partial histogram of the $\hat{e}$ in the full control sample overlaid. The control sample histogram is truncated on the left because control subjects with $\hat{e}$ much lower than the lowest treated $\hat{e}$ are not relevant to the matching; the histogram is truncated at the top because there are so many more control subjects than treated subjects; the number of control subjects in each bin is indicated above it. The histogram bin width is .4, corresponding to the caliper matching width for matching method 2, although the bins used for caliper matching are centered around each treated value

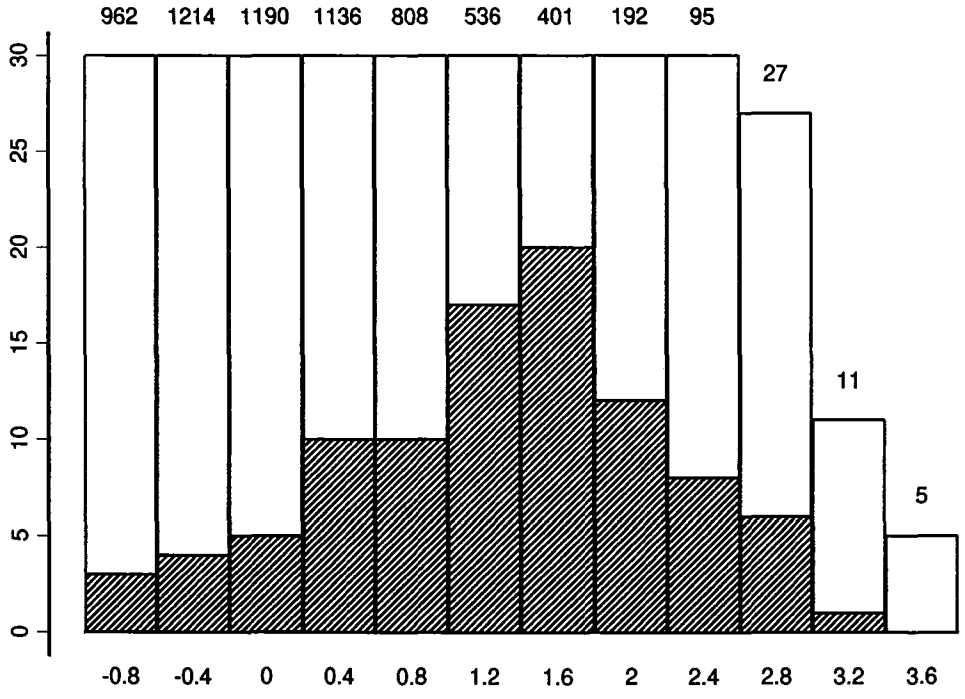


*Figure 1. The Logit of the Estimated Propensity Scores in the Treated Sample Are Displayed in the Shaded Histogram. The logit of the estimated propensity scores in the control sample are displayed in the unshaded histogram. The control histogram is truncated on the left at −1.0, and above at 30. The number of control subjects in each bin is listed above the bin.*

and are not fixed intervals. The large number of control subjects with $\hat{e}$ near most of the treated subjects demonstrates the considerable potential for matching on covariates other than $\hat{e}$.

Rosenbaum and Rubin (1985b) and Gu and Rosenbaum (1993) found good performance for coarse propensity score caliper matching followed by Mahalanobis matching on $\mathbf{X}$, but the latter found that the method did not perform as well as propensity score matching alone with high dimensional $\mathbf{X}$ and large initial biases, implying few potential matches for many of the treated subjects. Matching methods that treat the $\hat{e}$ as just one of many matching covariates may fail to eliminate bias in some covariates because close mean matching may not be attained on $\hat{e}$ (Rosenbaum and Rubin 1985b).

### 1.3 The Analogy Between Propensity Score Matching and Randomization, and Between Special Matching and Experimental Blocking

When close matching on $\hat{e}$ is attained, the treated and matched control samples are "balanced" in the sense that the differences between the treated and matched control means of the covariates included in $\hat{e}$ are even smaller than would be expected if the data had been generated by randomization (Rubin and Thomas, 1992b, 1996). Propensity score matching is also comparable to randomization in that it does not distinguish between covariates strongly related to the outcome variables and covariates poorly correlated with the outcome variables.

The results of Rubin and Thomas (1992b, 1996) and Gu and Rosenbaum (1993), combined with the results of Cochran (1968) and Cochran and Rubin (1973), show that much of the balance resulting from matching on $\hat{e}$ alone can be achieved with relatively coarse matching. Consequently, when the control subjects outnumber the treated subjects, additional matching on a subset $\mathbf{X}^{(s)}$ will often be possible. Mahalanobis metric matching on $\mathbf{X}^{(s)}$ within $\hat{e}$ calipers can achieve the balancing properties of matching on $\hat{e}$, while improving the comparability of the distributions of $\mathbf{X}^{(s)}$. This method is roughly analogous to blocking in a randomized study to achieve improved balance on key covariates. The choice of covariates is a "design" issue determined by the knowledge of subject area researchers. A method directly analogous to blocking in a randomized study is to perform propensity score matching on $\mathbf{X}^{(r)}$ separately within blocks defined by $\mathbf{X}^{(s)}$; we did this with the covariate SEX. Within each block, standard results for propensity score matching apply, but the number of blocks can be excessive when some $\mathbf{X}^{(s)}$ have many distinct values.

Another analogy to randomized experiments is that once outcome variables are observed, model-based (regression) adjustments can be performed on the matched samples. Because of the reduced extrapolation involved with matched samples, such model-based adjustments also perform better than when applied to the unmatched samples, and the combination is effective at reducing bias due to $\mathbf{X}$, as demonstrated by our multivariate results in Section 5 and by the univariate and bivariate results of Rubin (1973b, 1979).

The analogy between an observational study with matching on both $\hat{e}$ and a subset of prognostic covariates, and a randomized block experiment is limited. In particular, matching on $\hat{e}$ in an observational study adjusts only for linear trends due to *observed* $\mathbf{X}$, is not invariant to nonlinear transformations of the $\mathbf{X}$, and is successful only when there exists an adequately large control pool of subjects

from which to select the matches. In contrast, randomization stochastically balances the distribution of *all* covariates.

### 1.4 Overview

Herein we extend the theoretical approximations in Rubin and Thomas (1992a,b) to produce distributional results describing the samples produced when matching on both $\hat{e}$ and a subset of covariates, $\mathbf{X}^{(s)}$. Section 2 presents terminology and notation. Section 3 contains analytic approximations describing the sampling properties of matched datasets formed by combining matching on $\hat{e}$ with close matching on $\mathbf{X}^{(s)}$, and concludes with a discussion of the practical implications. The datasets resulting from the different matching methods are compared in Section 4 and evaluated assuming various nonlinear models in Section 5. Settings in which 1–1 matched sampling is a financial necessity are considered, as well as settings in which costs are low enough that five control subjects can be selected for each treated subject.

## 2. NOTATION AND TERMINOLOGY

### 2.1 Covariate Distributions

Suppose that $N_\mathrm{t}$ treated and $N_\mathrm{c}$ control subjects are available and that there are $p$ fully recorded covariates, $\mathbf{X}' = (X_1, \ldots, X_p)$, for each subject, where $R = N_\mathrm{c}/N_\mathrm{t}$. The covariates are arranged with the $s$ special covariates subject to additional matching listed first and the $r$ remaining covariates listed second, $\mathbf{X}' = (\mathbf{X}^{(s)'}, \mathbf{X}^{(r)'})$, where $s + r = p$. The $\mathbf{X}$ values among the treated and control subjects are assumed to be independent with finite moments and mean vectors and variance–covariance matrices denoted by

$$\boldsymbol{\mu}_\mathrm{t} = \begin{pmatrix} \boldsymbol{\mu}_\mathrm{t}^{(s)} \\ \boldsymbol{\mu}_\mathrm{t}^{(r)} \end{pmatrix}, \qquad \boldsymbol{\Psi}_\mathrm{t} = \begin{bmatrix} \boldsymbol{\Psi}_{\mathrm{t}(s,s)} & \boldsymbol{\Psi}_{\mathrm{t}(s,r)} \\ \boldsymbol{\Psi}_{\mathrm{t}(r,s)} & \boldsymbol{\Psi}_{\mathrm{t}(r,r)} \end{bmatrix}, \quad (1)$$

and

$$\boldsymbol{\mu}_\mathrm{c} = \begin{pmatrix} \boldsymbol{\mu}_\mathrm{c}^{(s)} \\ \boldsymbol{\mu}_\mathrm{c}^{(r)} \end{pmatrix}, \qquad \boldsymbol{\Psi}_\mathrm{c} = \begin{bmatrix} \boldsymbol{\Psi}_{\mathrm{c}(s,s)} & \boldsymbol{\Psi}_{\mathrm{c}(s,r)} \\ \boldsymbol{\Psi}_{\mathrm{c}(r,s)} & \boldsymbol{\Psi}_{\mathrm{c}(r,r)} \end{bmatrix}. \quad (2)$$

For the analytic approximations in Section 3, the distributions of $\mathbf{X}$ are modelled by a conditional normal distribution, similar to the general location model (Olkin and Tate 1961). Specifically, the treated and control distributions of $\mathbf{X}^{(r)}$ conditional on $\mathbf{X}^{(s)}$ are assumed to be multivariate normal with parallel linear regressions on $\mathbf{X}^{(s)}$, with common $s$ by $r$ matrix of coefficients denoted by $\boldsymbol{\beta}$. The residuals from these regressions, $\mathbf{X}^{(r|s)} = \mathbf{X}^{(r)} - \boldsymbol{\beta}'\mathbf{X}^{(s)}$, have means given by $\boldsymbol{\mu}_\mathrm{t}^{(r|s)} = \boldsymbol{\mu}_\mathrm{t}^{(r)} - \boldsymbol{\beta}'\boldsymbol{\mu}_\mathrm{t}^{(s)}$ and $\boldsymbol{\mu}_\mathrm{c}^{(r|s)} = \boldsymbol{\mu}_\mathrm{c}^{(r)} - \boldsymbol{\beta}'\boldsymbol{\mu}_\mathrm{c}^{(s)}$ conditional on $\mathbf{X}^{(s)}$ in the treated and control groups. The treated and control variance–covariance matrices of $\mathbf{X}^{(r|s)}$ (or $\mathbf{X}^{(r)}$) conditional on $\mathbf{X}^{(s)}$ are denoted by $\boldsymbol{\Psi}_{\mathrm{t}(r|s)}$ and $\boldsymbol{\Psi}_{\mathrm{c}(r|s)}$ and are assumed to be constant (i.e., not dependent on $\mathbf{X}^{(s)}$) and proportional, $\boldsymbol{\Psi}_{\mathrm{t}(r|s)} = \sigma^2\boldsymbol{\Psi}_{\mathrm{c}(r|s)}$, for some positive scalar $\sigma^2$. The Mahalanobis distance in the control metric between the means of the $\mathbf{X}^{(r|s)}$ in the treated and control populations is given by

$$B^2_{r|s} = (\boldsymbol{\mu}_\mathrm{t}^{(r|s)} - \boldsymbol{\mu}_\mathrm{c}^{(r|s)})'\boldsymbol{\Psi}^{-1}_{\mathrm{c}(r|s)}(\boldsymbol{\mu}_\mathrm{t}^{(r|s)} - \boldsymbol{\mu}_\mathrm{c}^{(r|s)}).$$

This model has been used extensively in the discrimination literature (Daudin 1986; Krzanowski 1975, 1980). Rubin and Thomas (1996) showed that the analytic expressions obtained with stringent assumptions similar to these can still provide useful approximations for matching even when the normality and proportionality assumptions are clearly violated.

Although we estimate the propensity score using linear logistic regression, the distributional assumptions in this section need not give rise to a linear logit model for the population propensity score (e.g., if the variance matrices differ, then the logit model contains linear and quadratic terms). Models that include quadratic and interaction terms should also be explored. As all efforts to select a correct model may be only partially successful, it is important to understand performance when propensity scores are imperfectly modeled. Here we examine the common practice of matching on estimated linear propensity scores under nonideal conditions.

### 2.2 Matching Methods Using Linear Propensity Score Estimators

Estimators of the linear coefficients motivated by normal-theory discriminant analysis are $\mathbf{S}_\mathrm{p}^{-1}(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})$ and $\mathbf{S}_\mathrm{c}^{-1}(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})$, where $\bar{\mathbf{X}}_\mathrm{t}, \mathbf{S}_\mathrm{t}$ and $\bar{\mathbf{X}}_\mathrm{c}, \mathbf{S}_\mathrm{c}$ are the means and variance–covariance matrices of $\mathbf{X}$ in the treated and control samples and $\mathbf{S}_\mathrm{p}$ is the pooled sample variance–covariance matrix. The logistic regression estimator is also often used. Theoretical approximations are based on the "semi-estimated" discriminant coefficient, $\boldsymbol{\Psi}_\mathrm{c}^{-1}(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})$, which replaces population means with sample means. Simulation experience of Rubin and Thomas (1996), as well as asymptotic calculations, demonstrate that approximations for matching methods based on the semi-estimated coefficient are quite accurate for matching methods that use the sample linear discriminant or logistic regression estimates.

Denote the standardized population discriminant with respect to the control variance–covariance matrix by

$$Z = \{(\boldsymbol{\mu}_\mathrm{t} - \boldsymbol{\mu}_\mathrm{c})'\boldsymbol{\Psi}_\mathrm{c}^{-1}(\boldsymbol{\mu}_\mathrm{t} - \boldsymbol{\mu}_\mathrm{c})\}^{-1/2}\{\boldsymbol{\Psi}_\mathrm{c}^{-1}(\boldsymbol{\mu}_\mathrm{t} - \boldsymbol{\mu}_\mathrm{c})\}'\mathbf{X}, \quad (3)$$

with $Z \equiv 0$ when $\boldsymbol{\mu}_\mathrm{t} = \boldsymbol{\mu}_\mathrm{c}$, and denote the corresponding standardized semi-estimated discriminant by

$$Z^* = \{(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})'\boldsymbol{\Psi}_\mathrm{c}^{-1}(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})\}^{-1/2}\{\boldsymbol{\Psi}_\mathrm{c}^{-1}(\bar{\mathbf{X}}_\mathrm{t} - \bar{\mathbf{X}}_\mathrm{c})\}'\mathbf{X}. \quad (4)$$

Our analytic results apply to methods that match on $\mathbf{X}^{(s)}$ and $Z^*$ ($\approx \hat{e}$), such as methods 1 and 2 of Section 1.2. Such matching methods are invariant to affine transformations of $\mathbf{X}$ that fix $\mathbf{X}^{(s)}$, which permits considerable analytic simplification when combined with conditionally ellipsoidal matching covariates (Rubin and Thomas 1992a).

### 2.3 Close Mean and Variance Matching on Special Covariates and the Estimated Propensity Score

For the analytic approximations that follow, we assume that the matching has produced approximate equality of the treated and matched control sample means and variances of $(\mathbf{X}^{(s)}, Z^*)$; although not all of our analytic approximations require close matching of these sample moments, we assume it throughout to simplify the presentation. After matching is completed, the sample moments of $(\mathbf{X}^{(s)}, Z^*)$ can be examined to determine if close matching has been achieved.

A useful guide for determining the feasibility of close matching of the sample moments of $(\mathbf{X}^{(s)}, Z^*)$, developed by Rubin (1976b) and Rubin and Thomas (1992b, 1996), can be obtained from the moments of the unmatched data assuming proportional multivariate normality of $\mathbf{X}$. Specifically, consider the maximum possible fractional bias reduction,

$$\theta^*_{\max} = \Omega(R/m)(pV + B^2)^{-1/2}, \tag{5}$$

where $V = (\sigma^2 + R^{-1})N_t^{-1}, B^2 = (\boldsymbol{\mu}_t - \boldsymbol{\mu}_c)'\boldsymbol{\Psi}_c^{-1}(\boldsymbol{\mu}_t - \boldsymbol{\mu}_c), \Omega(R/m)$ is the expectation of the $R/m$ upper tail of a standard normal variate, approximated by

$$\Omega(R/m) \approx 2^{\pi/2}(2\pi)^{-1/2}(R/m)^{(1-\pi/4)}(1 - m/R)^{\pi/4},$$

and $\sigma^2$ in $V$ is the ratio of the treated to control variance of the best linear discriminant (with respect to $\boldsymbol{\Psi}_c$) and the constant in Section 2.1 when $\boldsymbol{\Psi}_t$ and $\boldsymbol{\Psi}_c$ are proportional (Rubin 1976b).

The condition $\theta^*_{\max} > 1$ ensures that matching methods that select exclusively from the upper tail of the distribution of $Z^*$ in the control sample can eliminate the expected difference in the treated and matched control sample means of $Z^*$, which in turn implies that the expected differences in the sample means of $\mathbf{X}$ are also eliminated. Unless $\theta^*_{\max} > 1$, close matching of the sample moments of $\mathbf{X}^{(s)}$ and $Z^*$ is not possible. Consequently, if $\theta^*_{\max} < 1$, it may not be worthwhile to expend resources collecting outcome data or continuing investigation in this setting, at least without limiting the goals of the study. Even when $\theta^*_{\max} > 1$, however, the additional matching on $\mathbf{X}^{(s)}$ will not necessarily produce closely matched first or second sample moments of $\mathbf{X}^{(s)}$ or second moments of $Z^*$, and thus a value of $\theta^*_{\max}$ larger than 1 is typically required to yield good matching with special covariates.

Figure 1 exemplifies the typical situation with most treated subjects having numerous potential $\hat{e}$ matches and the treated subjects with the largest $\hat{e}$ having little opportunity for additional matching. In the hormone matching example described in Section 4, where matching on $\hat{e}$ and additional covariates was successful, estimates of $\theta^*_{\max}$ are 2.13 with 1–1 matching and 1.45 for 1–5 matching. Based on figure 1 of Rubin and Thomas (1996), which displays bias reductions achieved for nearest-remaining-neighbor matching on $\hat{e}$ for a variety of different treated and control distributions, we anticipate that $\theta^*_{\max}$ of 1.5 or larger indicates potential for successful matching on $\hat{e}$ along with additional matching on $\mathbf{X}^{(s)}$.

Note that the moments required to evaluate $\theta^*_{\max}$ can sometimes be estimated using the unmatched sample moments even before any individual subject's outcome data are collected. As illustrated in Section 4, this is also true for the approximations developed in Section 3, which can provide useful information when designing an observational study, especially when deciding whether to go forward and collect outcome data.

## 3. ANALYTIC APPROXIMATIONS

### 3.1 Decomposition for a Generic Outcome Variable

Consider a linear combination of $\mathbf{X}, Y = \boldsymbol{\gamma}'\mathbf{X}$, standardized to have unit variance in the control distribution; $Y$ can be thought of as the conditional expectation of a generic outcome variable given $X$. Let $\mathcal{Z}$ be the standardized regression of $Y$ on $(\mathbf{X}^{(s)}, Z)$ in the control distribution and let $\mathcal{W}$ be the standardized residual of this regression, where $\mathcal{W}$ is a linear combination of $X$ orthogonal to $(\mathbf{X}^{(s)}, Z)$. The variable $Y$ can be represented as

$$Y = \rho\mathcal{Z} + (1 - \rho^2)^{1/2}\mathcal{W}, \tag{6}$$

where $\rho$ is the multiple correlation coefficient of $Y$ on $(\mathbf{X}^{(s)}, Z)$; $\mathcal{Z} \equiv 0$ when $\rho = 0$, and $\mathcal{W} \equiv 0$ when $\rho = 1$. The means of $\mathcal{Z}$ and $\mathcal{W}$ in the treated and control samples are denoted by $\bar{\mathcal{Z}}_t, \bar{\mathcal{W}}_t$, and $\bar{\mathcal{Z}}_c, \bar{\mathcal{W}}_c$, and the corresponding means for the $mN_t$ matched controls are denoted by $\bar{\mathcal{Z}}_{mc}, \bar{\mathcal{W}}_{mc}$. The sample variances of a variable are denoted by $v_t(\bullet), v_c(\bullet)$, and $v_{mc}(\bullet)$, in the treated, control, and matched control samples.

### 3.2 The Moments of X in Matched Samples With Close Matching

Assuming the distributions and matching methods described in Sections 2.1 and 2.2, and close mean and variance matching on $\mathbf{X}^{(s)}$ and $\hat{e}$ as described in Section 2.3, we use the decomposition in (6) to provide analytic approximations for the first and second sample moments of $Y$ in terms of the corresponding moments of $\mathcal{Z}$ and $\mathcal{W}$. Without close matching, the relationship between the moments of $\mathcal{Z}$ in the treated and control samples are not easily characterized because they depend on the distributions of $\mathbf{X}^{(s)}$, which can differ arbitrarily in the treated and control samples under the assumptions in Section 2.1. The approximations for the effect of the matching on the first two sample moments of each linear combination of $\mathbf{X}$ yield a complete characterization of all of its matched sample means, variances, and covariances; derivations are in the Appendix.

*Results for the Expectation of the Sample Means.*

$$\frac{\mathrm{E}(\bar{Y}_t - \bar{Y}_{mc})}{\mathrm{E}(\bar{Y}_t - \bar{Y}_c)} = \frac{\mathrm{E}(\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_{mc})}{\mathrm{E}(\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_c)} \approx 0; \tag{7}$$

also, $\mathrm{E}(\bar{Y}_t - \bar{Y}_{mc}) \approx 0$ when $\mathrm{E}(\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_c) = 0$; that is, when there is no initial bias, the matching does not create bias.

*Results for the Variance of the Difference in the Sample Means.*

$$\operatorname{var}(\bar{Y}_{\mathrm{t}} - \bar{Y}_{\mathrm{mc}}) = \rho^2 \operatorname{var}(\bar{\mathcal{Z}}_{\mathrm{t}} - \bar{\mathcal{Z}}_{\mathrm{mc}}) + (1-\rho^2)\operatorname{var}(\bar{\mathcal{W}}_{\mathrm{t}} - \bar{\mathcal{W}}_{\mathrm{mc}}); \quad (8)$$

also, when $\rho \neq 0$,

$$\operatorname{var}(\bar{\mathcal{Z}}_{\mathrm{t}} - \bar{\mathcal{Z}}_{\mathrm{mc}}) \approx \frac{r-1}{r + B^2_{r|s}/V}\,(m^{-1} - R^{-1})N_{\mathrm{t}}^{-1}; \quad (9)$$

and when $\rho \neq 1$,

$$\operatorname{var}(\bar{\mathcal{W}}_{\mathrm{t}} - \bar{\mathcal{W}}_{\mathrm{mc}}) \approx \left(1 - \frac{1}{r + B^2_{r|s}/V}\right)(m^{-1} - R^{-1})N_{\mathrm{t}}^{-1}. \quad (10)$$

The quantity $\operatorname{var}(\bar{\mathcal{W}}_{\mathrm{t}} - \bar{\mathcal{W}}_{\mathrm{mc}})$ is the same for each choice of the linear combination $Y$; that is, for all $\gamma$.

*Results for the Expectation of Sample Variances.*

$$\mathrm{E}\{v_{\mathrm{mc}}(Y)\} = \rho^2 \mathrm{E}\{v_{\mathrm{mc}}(\mathcal{Z})\} + (1-\rho^2)\mathrm{E}\{v_{\mathrm{mc}}(\mathcal{W})\}; \quad (11)$$

also, when $\rho \neq 0$,

$$\left|\frac{\mathrm{E}\{v_{\mathrm{mc}}(\mathcal{Z})\}}{\mathrm{E}\{v_{\mathrm{t}}(\mathcal{Z})\}} - 1\right| \approx \left|\frac{(1-\sigma^2)(r-1)}{r + B^2_{r|s}/V}\right| \Big/ \mathrm{E}\{v_{\mathrm{t}}(\mathcal{Z})\}; \quad (12)$$

and when $\rho \neq 1$,

$$\mathrm{E}\{v_{\mathrm{mc}}(\mathcal{W})\} \approx 1 + \frac{\sigma^2 - 1}{r + B^2_{r|s}/V}. \quad (13)$$

The quantity $\mathrm{E}\{v_{\mathrm{mc}}(\mathcal{W})\}$ is the same for each choice of the linear combination $Y$.

### 3.3 Comments on the Results

*Remark A.* Result (7) shows that (a) the bias in the matched sample means of any linear combination of $\mathbf{X}$ is determined by the bias in the matched sample means of $\mathcal{Z}$, a linear combination of $\mathbf{X}^{(s)}$ and $Z$, and (b) this bias is eliminated when close matching is achieved. Note that (7) implies that the expected difference in the sample means of any linear combination of $\mathbf{X}$ uncorrelated with $(\mathbf{X}^{(s)}, Z)$ is 0.

*Remark B.* Provided that close matching on $\hat{e}$ is possible, (7)–(10) show that methods based on propensity score matching are very effective in reducing the variability of differences in the sample means of many variables. Typical applications have numerous covariates, so $\rho$ tends to be modest, implying that the variance of the difference in the sample means of most $Y$ variables is heavily weighted by the contribution of the $\mathcal{W}$ component in (8). Moreover, the variance in (9) is less than or equal to the variance in (10). With numerous covariates and moderate or large sample sizes, $B^2_{r|s}/V$ tends to be large, so (9) is near 0 and the first factor in (10) approaches 1. Thus $\operatorname{var}(\bar{Y}_{\mathrm{t}} - \bar{Y}_{\mathrm{mc}})$ is dominated by (10) for most linear combinations of $\mathbf{X}$. The term $(m^{-1} - R^{-1})$ in (9) and (10) is a finite sampling correction arising from the selection of the matched controls from the full control sample. When $\sigma^2 \approx 1$ and the control sample is relatively large so that $R^{-1}$ is small, (10) reduces to $1/(mN_{\mathrm{t}})$, which is no more than one-half of the variance that results from a corresponding randomized design (see also Hill, Rubin, and Thomas 1999).

*Remark C.* As with the variance of the difference in the sample means, the expected value of the matched control sample variance of most linear combinations of $\mathbf{X}$ is determined largely by $\mathrm{E}\{v_{\mathrm{mc}}(\mathcal{W})\}$. The approximation in (12) shows that the matched control sample variance of the component of each linear combination of $\mathbf{X}$ determined by $\mathbf{X}^{(s)}$ and $Z$ is similar to the corresponding treated sample variance, because the numerator in the righthand side of (12) is typically close to 0 and $\mathrm{E}\{v_{\mathrm{t}}(\mathcal{Z})\}$ is close to 1. This is not surprising in view of the close matching assumed for $\mathbf{X}^{(s)}$ and $Z^*$ (or $\hat{e}$). But the approximation in (13) demonstrates that the matching does little to reduce differences in sample variances for the $\mathcal{W}$ component; the expected value of $v_{\mathrm{c}}(\mathcal{W})$ is exactly 1 by construction. Any differences between the matched control and the treated sample variances are thus reduced by only a small amount, so the matching is not expected to reduce the bias (conditional or unconditional) due to nonlinear relationships in $\mathbf{X}^{(r)}$ for most covariates, other than due to $\mathbf{X}^{(s)}$.

*Remark D.* The multivariate normal assumption for the $\mathbf{X}^{(r)}$ distributions ensures the approximately constant expectation of each covariate uncorrelated with the $Z^*$ and $\mathbf{X}^{(s)}$, conditional on the sample means. Multivariate normality also provides simplification by ensuring constant conditional variances for these covariates. The sensitivity of conclusions based on these approximations to violations of normality increases as the matching on $\hat{e}$ becomes increasingly different from random sampling of the full control sample, that is, as the treated and control distributions become more different. The robustness of the approximations based on the proportional normal model was examined by Rubin and Thomas (1996) in the case without special covariates, where the general conclusion is that the approximations work quite well.

*Remark E.* The approximation for the moments of $\mathcal{Z}$ in (9) and (12) are upper bounds provided that close matching is attained. The bounds are increasingly conservative as $Y$ becomes a function of $\mathbf{X}^{(s)}$ alone; the conservative nature of the approximations is explained in the Appendix, where the results are derived.

*Remark F.* The $\mathbf{X}^{(s)}$ should be included when computing $\hat{e}$, as assumed by the theoretical results in Section 3. If they are not included, then there can be differences in the expected values of the treated and matched control sample means of some covariates even if close matching is attained for $\mathbf{X}^{(s)}$ and $\hat{e}$.

## 4. EXAMPLE

The mothers of children exposed to the hormone are substantially different than those of the unexposed children with respect to the covariates in Table 1. The Mahalanobis distance between the sample means measured with respect

to $\mathbf{S}_c$ is 1.44 standard deviations, which is primarily due to three covariates, as summarized in the rightmost column of Table 2. The correlations between covariates are small to moderate ($< 0.4$). There are also differences in the variances of some of the covariates, as displayed in the rightmost column of Table 3. Because of the very large matching ratio, $R = 82$, the approximate value of $\theta^*_{\max}$ is 2.13 when $m = 1$, and 1.45 when $m = 5$, so the control pool is adequate to produce a matched control sample with means similar to the treated sample despite the large initial differences.

Table 2 summarizes the differences in the treated, control, and matched control sample means. The differences in the means of the continuous covariates are standardized by their treated sample standard deviations; the differences in the binary covariates are proportions. The differences in the $\hat{e}$ are reported at the bottom of the table. All of the matching methods except method 3, which uses only $\mathbf{X}^{(s)}$, produce large reductions in the differences in the sample means, which determines the bias due to linear trends between outcome variables and the covariates. Method 3 produces almost exact matches for the $\mathbf{X}^{(s)}$ it uses for Mahalanobis metric matching (the lowest treated–matched control correlation for a special variable was .98), but results in large differences in several other covariates potentially related to the outcome variables.

From Table 2, matching methods 1 and 2, which use $\hat{e}$, successfully reduce the differences in $\hat{e}$, although the remaining standardized bias of .10 for the 1–1 matched control sample produced by method 2, which includes matching on $\mathbf{X}^{(s)}$, may be larger than desired. The result in (7) predicts that in expectation, the difference in the sample means of each covariate is essentially eliminated by these two propensity score methods. The differences in the means are generally consistent with our predictions for the performance of the propensity score methods based on the approximations in (8)–(10). If a much larger difference occurs than is predicted by its standard error, then that covariate should be examined, particularly for the possibility of adding interaction and quadratic terms in the linear propensity score model. The values of the independent-sample $Z$ statistics vary considerably less than a standard normal deviate (not displayed), consistent with the theoretical prediction that their variances in the matched samples are less than one half that in a corresponding randomized design. (Hill et al. 1999 has an application in this setting.)

From Table 2, the Mahalanobis matching on all of the covariates, method 4, did a reasonable job eliminating differences in individual sample means. But the remaining difference in the means of $\hat{e}$ is large enough to be of concern and is reflected in the unacceptably large remaining mean differences in $\mathbf{X}^{(s)}$, especially for the 1–5 matching. The other notable feature of method 4 is its tendency to place very heavy weight on binary covariates representing rare events; this failing of Mahalanobis metric matching has also been noted by Rosenbaum and Rubin (1985b) and Gu and Rosenbaum (1993).

Table 3 summarizes the differences in the sample variances of the continuous and ordinal covariates, which are important because, along with the means, they determine the bias due to quadratic trends between the outcome variables and the covariates. The MOTAGE covariate has a much larger variance in the control sample, as does the pregnancy complication covariate, PBC420. The results in (11)–(13) predict that matching on $\hat{e}$ will produce only modest improvements in the sample variances. The results in Table 3 are typically consistent with this prediction; in particular, matching only on $\hat{e}$, method 1, did not reduce the differences in the variance of MOTAGE, although both methods 1 and 2 reduced those differences in PBC420, MOTHT, and MOTWT somewhat more than predicted. Method 2, matching on $\hat{e}$ and $\mathbf{X}^{(s)}$, successfully eliminated the differences in the variances of $\mathbf{X}^{(s)}$ as well as the differences in their covariances (not displayed).

Mahalanobis matching on all of the covariates, method 4, reduced the largest differences in the variances, but the differences for $\mathbf{X}^{(s)}$ are still unacceptably large. As with the sample means, by attempting to closely match all of the covariates, Mahalanobis matching produces some improvement in most covariates, but substantial differences remain, sometimes in the most prognostically important covariates.

## 5. BIAS REDUCTION DUE TO THE MATCHED SAMPLING–SIMULATED DATA

We evaluate the bias and the mean squared error (MSE) of some common estimators based on the different matched samples in Section 4 using data simulated from several linear and nonlinear models relating outcomes to the covariates $\mathbf{X}$. Sections 5.1 and 5.2 present results from evaluations that sample a large number of potential simulated models and use MSE and bias to summarize the bias reduction for the large class of models. Section 5.3 focuses on bias reduction in the most important prognostic covariates.

### 5.1 Simulated Outcomes Nonlinear Function of the Covariates

Following Rubin (1973b, 1979), we use the exponential function to create models with varying amounts of curvature and nonadditivity. For a generic vector $\mathbf{X}$ of dimension $J$, we form a multiplicative model, $c_1(\boldsymbol{\lambda})e^{\boldsymbol{\lambda}'\mathbf{X}}$, and an additive model, $c_2(\kappa)\sum_{j=1}^{J}\lambda_l e^{\pm\kappa X_j}$, where $c_1(\boldsymbol{\lambda}), c_2(\kappa)$, and $\|\boldsymbol{\lambda}\|$ are chosen to produce additive and multiplicative functions of the covariates with variances close to 1 in the treated sample. These functions are applied to both $\mathbf{X}^{(s)}$ and $\mathbf{X}^{(r)}$ and are denoted by $f_s(\mathbf{X}^{(s)})$ and $f_r(\mathbf{X}^{(r)})$. The functions are combined additively to form expected values given $\mathbf{X}$ of an outcome variable, $Y = w_s f_s(\mathbf{X}^{(s)}) + (1 - w_s) f_r(\mathbf{X}^{(r)})$, where $w_s$ is a weight that determines the relative contribution of $\mathbf{X}^{(s)}$ to the outcome. An additive treatment effect is used for each model, and there is no bias associated with unmeasured covariates.

We report results with both $f_s$ and $f_r$ having the additive form and with both having the multiplicative form. The amount of nonlinearity in the multiplicative function

*Table 2. Standardized Bias in the Treated and Matched Control Sample Means*

| | Method 1 | | Method 2 | | Method 3 | | Method 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *1–1* | *1–5* | *1–1* | *1–5* | *1–1* | *1–5* | *1–1* | *1–5* | *Unmatched* |
| SEX | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | −.02 |
| MOTAGE | .07 | −.11 | −.02 | −.08 | 0 | 0 | .13 | .20 | .92 |
| SES | .02 | −.03 | −.01 | −.02 | 0 | 0 | .04 | .19 | .74 |
| CAGE | .01 | −.08 | −.19 | −.06 | −.10 | −.05 | .06 | 0 | −.05 |
| LGEST | .02 | .01 | .11 | .06 | −.02 | −.03 | −.10 | −.04 | −.03 |
| PBC415 | .02 | −.03 | .13 | .10 | .70 | .54 | .12 | .10 | .23 |
| PBC420 | −.09 | .10 | .06 | −.01 | 2.71 | 1.62 | .19 | .25 | .94 |
| MOTHT | −.15 | −.02 | −.10 | 0 | .49 | .46 | −.03 | .09 | .34 |
| MOTWT | −.01 | .02 | 0 | −.04 | −.07 | .04 | −.15 | −.08 | .01 |
| ANTIH | .03 | 0 | .03 | .01 | .28 | .24 | .02 | .03 | .18 |
| BARB | .02 | .02 | .02 | 0 | .17 | .12 | .01 | .01 | .11 |
| CHEMO | 0 | .02 | −.02 | 0 | .11 | .09 | 0 | .01 | .04 |
| CIGAR | −.05 | .02 | −.06 | −.02 | −.28 | −.19 | .03 | −.01 | −.17 |
| MBIRTH | .03 | −.01 | −.01 | −.01 | .01 | .01 | 0 | 0 | 0 |
| PSYDRUG | .04 | .02 | .05 | 0 | .19 | .15 | 0 | 0 | .12 |
| RESPIR | −.01 | .01 | −.03 | 0 | .04 | .02 | 0 | .01 | .02 |
| SIB | .04 | .01 | .01 | .04 | .42 | .25 | .01 | .03 | .01 |
| PSC | 0 | .01 | .10 | .06 | 2.31 | 1.49 | .21 | .39 | 1.44 |

is determined by the magnitude of $\boldsymbol{\lambda}$. We select a magnitude $\|\boldsymbol{\lambda}\|$ that results in a model with $Y$ nearly linear in $\mathbf{X}$, and a $\|\boldsymbol{\lambda}\|$ that produces a model with linear $r^2 \approx .85$ when $Y$ is linearly regressed on $\mathbf{X}$; this amount of nonlinearity is difficult to detect in realistic multivariate settings. The nonlinearity in the additive function is determined by $\kappa$. We select a positive $\kappa$ to produce a nearly linear function and a function with linear $r^2$ of .85. The sign of the exponent for each covariate is randomly selected with probability one-half of a negative multiplier. The two levels of nonlinearity are applied to $f_s$ and $f_r$ separately, producing $2^2$ combinations, and these functions are combined using two choices of weights, $w_s$, selected so that the $r^2$ between $Y$ and $f_s(\mathbf{X}^{(s)})$ is high ($> .90$) or moderate ($.4 < r^2 < .6$). Thus eight types of additive and eight types of multiplicative models are evaluated.

The $\boldsymbol{\lambda}$ are selected uniformly over parameter spaces of appropriate dimension ($s$ or $r$) by drawing $\boldsymbol{\lambda}$ from the standard multivariate normal distribution and then norming $\boldsymbol{\lambda}$ to have the specified $\|\boldsymbol{\lambda}\|$ to achieve the appropriate nonlinearity and variance. The $\boldsymbol{\lambda}$ for $f_s$ are generated independently of the $\boldsymbol{\lambda}$ for $f_r$. Five hundred $\boldsymbol{\lambda}$ were generated for each type of model to obtain accurate estimates of the expected performance of estimators within the uniform parameter space.

Each function is used to produce $Y$ values for each subject in the treated and control samples, which are the mean responses of the subjects conditional on $\boldsymbol{\lambda}$ and the covariate values. Residual variation is not actually generated and added to the $Y$ values; rather, the effect of residual variance is included in the evaluation of the estimators by specifying the proportion of the variance in the outcome variable due to the variation in the regression functions (model $r^2$) and then computing the residual variance required to produce the specified model $r^2$. The residual variances change for different $\boldsymbol{\lambda}$ because the variance in the generated mean responses vary with $\boldsymbol{\lambda}$. After the residual variance is determined, the outcome variable and bias computed from it are rescaled so that the unconditional variance of the response variable equals 1.0 among the treated subjects; that is, the residual variance plus the variance of $Y$ among the treated subjects is rescaled to 1.0. For the 1–1 matching, we consider integrated squared bias only (the model $r^2$ is one), following Rubin (1979), because the variance of the estimators that we study are approximately the same for each

*Table 3. Difference in Treated and Matched Control Sample Variances: $(V_t - V_{mc})/V_t$*

| | Method 1 | | Method 2 | | Method 3 | | Method 4 | | |
|---|---|---|---|---|---|---|---|---|---|
| *Variable* | *1–1* | *1–5* | *1–1* | *1–5* | *1–1* | *1–5* | *1–1* | *1–5* | *Unmatched* |
| MOTAGE | −.80 | −.77 | −.02 | −.09 | 0 | .01 | −.25 | −.35 | −.83 |
| SES | −.06 | .07 | .03 | .06 | 0 | .01 | −.01 | .13 | .14 |
| CAGE | −.01 | .12 | .09 | .15 | .23 | .17 | .28 | .29 | .17 |
| LGEST | .13 | .12 | −.04 | −.05 | −.20 | −.05 | .49 | .17 | −.20 |
| PBC415 | .23 | .29 | .28 | .21 | .15 | −.01 | .23 | .20 | .19 |
| PBC420 | .15 | −.15 | −.42 | −.09 | −.38 | −.88 | −.08 | −.07 | −.71 |
| MOTHT | −.01 | −.20 | −.26 | −.24 | −.62 | −.71 | .10 | .03 | −.54 |
| MOTWT | −.10 | .14 | .12 | .18 | .56 | .34 | .31 | .39 | .23 |
| PSC | −.01 | .03 | −.03 | .01 | −.09 | −.52 | .02 | .06 | −.23 |

matched dataset. For the 1–5 matching, we report results for a model $r^2$ of 1.0 and for a lower model $r^2$ of .25.

### 5.2 Results of Simulation

The two estimators that we evaluate are the difference between sample means and the difference between the sample means adjusted by linear least squares regression estimated by pooling the assumed parallel linear regressions in the treated and control, or matched control, samples (i.e., the common "covariance-adjusted" estimate). The bias of the difference in the sample means conditional on the covariate values is $\bar{Y}_t - \bar{Y}_c$ for the control sample and $\bar{Y}_t - \bar{Y}_{mc}$ in each matched control sample, where $\bar{Y}_t, \bar{Y}_c$, and $\bar{Y}_{mc}$ are the average of the $Y$ values generated for the treated, control, and matched control samples. Letting $\boldsymbol{\Gamma}$ denote the coefficients of the least squares regression of $Y$ on the covariates in the treated and control samples, $\bar{Y}_t - \bar{Y}_c - \boldsymbol{\Gamma}'(\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_c)$ is the bias of the linear regression estimator applied to the unmatched samples, and similarly for each matched sample using the $Y$ values of the treated and matched control subjects, $\bar{Y}_t - \bar{Y}_{mc} - \boldsymbol{\Gamma}'_{mc}(\bar{\mathbf{X}}_t - \bar{\mathbf{X}}_{mc})$.

We summarize the performance of the estimators in Table 4, which gives the square root of the average squared bias with respect to the uniform distribution of the $\boldsymbol{\lambda}$ parameters for each type of model estimated using the response function computed with each of the 500 $\boldsymbol{\lambda}$ values. Because the results have very little variation across the levels of nonlinearity in $\mathbf{X}^{(r)}$, the reported results are averages across these settings of the simulation study. The results from the approximately linear model generated using the additive and multiplicative forms of the response function are similar, so these too are averaged with one column of results for these settings. The average squared bias for the linear regression–based estimators are very small with the data generated from the approximately linear models, as expected, but the squared biases are greater than 0 primarily because the results are averaged across nonlinear functions of $\mathbf{X}^{(r)}$.

The results in Table 4 demonstrate that the estimators based on the matched samples, with the exception of method 3, are more accurate than corresponding estimates based on the unmatched samples, and they are more robust to monotone nonlinearity in the regression functions. Linear regression adjustment consistently improved the performance of simple differences in sample means, although the performance of the simpler estimates is very good, particularly when based on the sample formed by the second matching method. The superior performance of regression adjustments applied to matched samples is consistent with previous theoretical and simulation studies (Rubin 1973b, 1979). We expect the improvements resulting from linear regression adjustments to be much smaller for nonmonotone regression functions. As anticipated, matching and regression adjustment provide less improvement when the covariates are weakly predictive of the outcome variable. Results with $r^2 = .25$ based on the MSE (not displayed), which includes the variance of the estimators and the squared conditional bias, show that the differences in means based on the matched samples are still much more accurate than those based on the unmatched samples, and that the better matching methods are never substantially worse

*Table 4. Average Conditional Bias*

| | *High-weight special variables* | | | *Low-weight special variables* | | |
|---|---|---|---|---|---|---|
| | *Approximate linear* | *Nonlinear additive* | *Nonlinear multiplicative* | *Approximate linear* | *Nonlinear additive* | *Nonlinear multiplicative* |
| **1–1 matching** | | | | | | |
| Method 1 | .04 | .11 | .09 | .07 | .09 | .08 |
| Method 2 | .04 | .05 | .05 | .08 | .09 | .09 |
| Method 3 | .23 | .27 | .29 | .55 | .53 | .63 |
| Method 4 | .07 | .05 | .08 | .08 | .07 | .08 |
| Unmatched | .60 | .55 | .57 | .47 | .44 | .44 |
| Method 1 + reg | .02 | .12 | .09 | .03 | .08 | .06 |
| Method 2 + reg | .01 | .03 | .03 | .03 | .04 | .04 |
| Method 3 + reg | .03 | .18 | .10 | .06 | .13 | .09 |
| Method 4 + reg | .01 | .04 | .04 | .03 | .04 | .03 |
| Unmatched + reg | .02 | .13 | .09 | .03 | .09 | .06 |
| **1–5 matching** | | | | | | |
| Method 1 | .05 | .15 | .10 | .05 | .11 | .08 |
| Method 2 | .04 | .05 | .05 | .05 | .05 | .05 |
| Method 3 | .15 | .19 | .18 | .35 | .35 | .38 |
| Method 4 | .14 | .15 | .13 | .11 | .13 | .10 |
| Unmatched | .58 | .52 | .55 | .45 | .43 | .39 |
| Method 1 + reg | .01 | .12 | .08 | .02 | .09 | .05 |
| Method 2 + reg | .01 | .04 | .03 | .02 | .04 | .03 |
| Method 3 + reg | .02 | .10 | .06 | .04 | .08 | .05 |
| Method 4 + reg | .01 | .09 | .05 | .03 | .07 | .04 |
| Unmatched + reg | .02 | .13 | .09 | .03 | .09 | .06 |

NOTE: Square root of the average conditional squared bias. The treated population standard deviation is 1, because the matching leaves the treated group fixed. Simulation error is beyond the reported decimal value.

than linear regression based on the unmatched control sample.

There are some important and consistent differences across the various matched control samples, although the differences are small when contrasted with the unmatched control sample. Of particular importance, the matching using only $\mathbf{X}^{(s)}$ is inferior in this simulation, even when $\mathbf{X}^{(s)}$ accounts for 90–95% of the variation in the outcome. These results should discourage the approach of ignoring potentially relevant covariates. Matching on $\hat{e}$, method 1, is very effective when the response function is linear in $\mathbf{X}^{(s)}$, but it does not perform as well when the response function is nonlinear, consistent with the theoretical predictions of Section 3. Matching on $\mathbf{X}^{(s)}$ within $\hat{e}$ calipers, method 2, performed well in all settings, especially those with dominant $\mathbf{X}^{(s)}$ in a nonlinear regression. Mahalanobis matching, method 4, also performed well, although the 1–1 matched sample formed with it is noticeably better than the corresponding 1–5 matched sample. Mahalanobis matching can result in substantial bias when it fails to eliminate differences in the means of $\hat{e}$, as happened with 1–5 matching.

### 5.3 The Effect of Matching on Different Covariate Components

The simulation models described in Section 5.1 were also used to assess the reduction in bias due to specific components of the covariate space. The covariates evaluated are the ones most likely to be prognostic of outcome. Results are presented for matching method 2, Mahalanobis metric matching on special covariates within propensity score calipers.

Because the propensity score is the combination of covariates that the subject-matter experts (e.g., clinicians, educators, program administrators) and subjects judged important when selecting appropriate treatments, it will often be prognostically important. The first row of Table 5 contains the bias due to approximately linear and exponential trends applied to the propensity scores. (The additive and multiplicative model coincide with a univariate input variable.) The matching, especially when combined with linear regression, effectively eliminates the very large initial bias due to linear and nonlinear models. Similar results were obtained for the continuous and ordinal special matching covariates, mother's age and SES. The linear covariance adjustment substantially reduces the bias due to monotone nonlinear functions, but these linear adjustments perform much better when applied to the matched dataset. The unadjusted matched differences compare favorably to the covariance adjusted unmatched results.

Results for PBC420 were similar to those for $\hat{e}$; PBC420 had a large initial bias and was thus correlated (.65) with $\hat{e}$. In contrast, PBC415 had a moderate initial bias and is not highly correlated (.18) with the propensity score or $\mathbf{X}^{(s)}$. The reduction in its initial difference between the treated and unmatched means is consistent with theory, but as predicted, the bias due to nonlinear trend was not eliminated by the matching and linear covariance adjustment. Although our results are specific to the models and the dataset in this section, we anticipate that similar results will attain more broadly.

*Table 5. Conditional Bias for Specific Variables: Method 2 with 1–1 Matching*

| | *Approximately linear* | | | | *Non-linear* | | | |
|---|---|---|---|---|---|---|---|---|
| | *Matched* | | *Unmatched* | | *Matched* | | *Unmatched* | |
| | *Unadj* | *Adj* | *Unadj* | *Adj* | *Unadj* | *Adj* | *Unadj* | *Adj* |
| PSC | .10 | 0 | 1.38 | .04 | .10 | .01 | .89 | .25 |
| MOTAGE | .02 | 0 | .88 | .01 | .03 | .01 | .47 | .16 |
| SES | .01 | 0 | .73 | .01 | 0 | .01 | .61 | .09 |
| PBC415 | .15 | .02 | .25 | .02 | .24 | .14 | .29 | .15 |
| PBC420 | .04 | .01 | .89 | .02 | 0 | .05 | .53 | .12 |

NOTE: Square root of the conditional squared bias. "Adj" results are for the estimator which includes covariate regression adjustment. "Unadj" results are without the adjustment.

## 6. SUMMARY

Our example demonstrates the importance of multivariate matching in observational studies, even when many of the covariates have relatively limited prognostic value. Estimators based on multivariate matched samples consistently perform better than estimators based on the unmatched samples. Multivariate matching methods should be designed to eliminate the difference in the sample means of the estimated propensity score, because there will be remaining potential bias due to some covariates if a method fails to eliminate such differences. Matching within propensity score calipers followed by additional matching on key prognostic covariates is an effective method for ensuring that differences in the propensity scores are eliminated while allowing researchers to use information about the relative prognostic value of different covariates. Combined with Mahalanobis matching on special covariates and regression adjustments, propensity score methods can effectively reduce bias due to a large number of measured covariates in an observational study. With prognostically important discrete covariates with a small number of distinct values, the methodology can be applied separately within blocks defined by these covariates.

## APPENDIX: OVERVIEW OF THE DERIVATIONS

### A.1 Reduction to a Canonical Form

Following Rubin and Thomas (1992a), we consider a matching method that is conditionally affinely invariant given $\mathbf{X}^{(s)}$ and $Z^*$ in the sense that it is applied such that after any affine transformation that leaves $\mathbf{X}^{(s)}$ fixed, it produces the same matches as when matching on the untransformed $\mathbf{X}$. Matching methods that use only $\mathbf{X}^{(s)}$ and $Z^*$ are conditionally affinely invariant. By regressing $\mathbf{X}^{(r)}$ on $\mathbf{X}^{(s)}$ and applying an appropriately rotated form of $\mathbf{\Psi}_{c(r|s)}^{-1/2}$ to the residuals, we can assume without loss of generality that the distribution of $\mathbf{X}$ has moments of the form $\boldsymbol{\mu}_c^{(r)} = 0, \boldsymbol{\mu}_t^{(r)} = B_{r|s} r^{-1/2} \mathbf{1}_r, B_{r|s} \geq 0, \mathbf{\Psi}_{c(r,r)} = \mathbf{I}_r, \mathbf{\Psi}_{c(r,s)} = 0, \mathbf{\Psi}_{t(r,s)} = 0$ and $\mathbf{\Psi}_{t(r,r)} = \sigma^2 \mathbf{I}_r$, where $\mathbf{1}_r$ and $\mathbf{I}_r$ are the $r$-dimensional equal angular vector and identity matrix. The $\mathbf{X}^{(s)}$ are unchanged with moments $\boldsymbol{\mu}_t^{(s)}, \boldsymbol{\mu}_c^{(s)}, \mathbf{\Psi}_{t(s,s)}$, and $\mathbf{\Psi}_{c(s,s)}$, which have no special conditions placed on them. In the canonical form, $\mathbf{X}^{(r)}$ are multivariate normal and independent of $\mathbf{X}^{(s)}$ in the

treated and control distributions, and the standardized discriminant $Z$ with respect to the control distribution is proportional to $(\boldsymbol{\mu}_t^{(s)} - \boldsymbol{\mu}_c^{(s)})'\boldsymbol{\Psi}_{c(s,s)}^{-1}\mathbf{X}^{(s)} + B_{r|s}r^{-1/2}\mathbf{1}_r'\mathbf{X}^{(r)}$. Similar simplifying transformations have been used by Efron (1975), Rubin (1976a,b), Ashikaga and Chang (1981), and others.

We consider a linear combination, $U = \boldsymbol{\alpha}'\mathbf{X} = \boldsymbol{\alpha}^{(s)'}\mathbf{X}^{(s)} + \boldsymbol{\alpha}^{(r)'}\mathbf{X}^{(r)}$, which is scaled to have variance equal to 1 in the control distribution. We specialize $U$ to have one of two forms. The first corresponds to a standardized regression of some arbitrary linear combination of $\mathbf{X}$ onto $\mathbf{X}^{(s)}$ and $Z$, which we denoted by $\mathcal{Z}$ in Section 3.1, so $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^{(s)'}, \alpha\mathbf{1}_r')'$ such that $\boldsymbol{\alpha}^{(s)'}\boldsymbol{\Psi}_{c(s,s)}\boldsymbol{\alpha}^{(s)} + r\alpha^2 = 1$, with $\mathbf{X}$ in canonical form. The second form of $U$ corresponds to an $\mathcal{W}$, also defined in Section 3.1, which is a standardized variable uncorrelated with $(\mathbf{X}^{(s)}, Z)$; thus $\boldsymbol{\alpha} = (\mathbf{0}', \boldsymbol{\alpha}^{(r)'})'$ such that $\boldsymbol{\alpha}^{(r)'}\boldsymbol{\alpha}^{(r)} = 1, \boldsymbol{\alpha}^{(r)'}\mathbf{1}_r = 0$.

## A.2 Decomposition Based on the Semi-Estimated Discriminant

A decomposition analogous to (6) is now formed for $U$ based on $\mathbf{X}^{(s)}$ and $Z^*$. With the covariates in canonical form,

$$Z^* \propto (\bar{\mathbf{X}}_t^{(s)} - \bar{\mathbf{X}}_c^{(s)})'\boldsymbol{\Psi}_{c(s,s)}^{-1}\mathbf{X}^{(s)} + (\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)})'\mathbf{X}^{(r)}. \quad \text{(A.1)}$$

The decompositions are computed with respect to the control distribution conditional on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$; the distribution of the $\mathbf{X}^{(r)}$ in the control sample conditional on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$ is multivariate normal with mean $\bar{\mathbf{X}}_c^{(r)}$ and variance–covariance matrix $(1 - N_c^{-1})\mathbf{I}_r$. All of the correlations, moments, and regressions are computed with respect to the control distribution conditional on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$. We use the accurate approximation $\text{var}(U) = \boldsymbol{\alpha}^{(s)'}\boldsymbol{\Psi}_{c(s,s)}\boldsymbol{\alpha}^{(s)} + \boldsymbol{\alpha}^{(r)'}\boldsymbol{\alpha}^{(r)}(1 - N_c^{-1}) = 1$ throughout.

A conditional orthogonal decomposition for $U$ is

$$U = \rho_*\mathcal{Z}^* + (1 - \rho_*^2)^{1/2}\mathcal{W}^*. \quad \text{(A.2)}$$

The component $\mathcal{Z}^*$ is the standardized linear regression of $U$ on $\mathbf{X}^{(s)}$ and $Z^*$ in the control distribution,

$$\mathcal{Z}^* = c(\boldsymbol{\alpha}^{(s)}, \boldsymbol{\alpha}^{(r)}) \times\left\{\boldsymbol{\alpha}^{(s)'}\mathbf{X}^{(s)} + \frac{\boldsymbol{\alpha}^{(r)'}(\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)})}{\|\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)}\|^2}(\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)})'\mathbf{X}^{(r)}\right\}, \quad \text{(A.3)}$$

where

$$c(\boldsymbol{\alpha}^{(s)}, \boldsymbol{\alpha}^{(r)}) = \left[\boldsymbol{\alpha}^{(s)'}\Psi_{c(s,s)}\boldsymbol{\alpha}^{(s)} + (1 - N_c^{-1})\left\{\frac{\boldsymbol{\alpha}^{(r)'}(\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)})}{\|\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)}\|}\right\}^2\right]^{-1/2}.$$

The component $\mathcal{W}^*$ is the standardized residual of the regression; direct calculations show that the correlation of $U$ and $\mathcal{Z}^*$ is $\rho_* = c^{-1}(\boldsymbol{\alpha}^{(s)}, \boldsymbol{\alpha}^{(r)})$.

When the general formula in (A.2) is applied to some choice of $\mathcal{Z}$, the correlation, $\rho_*$, is denoted by $a_*$, and when (A.2) is applied to some choice of $\mathcal{W}$, $\rho_*$ is denoted by $r_*$. It is easy to check that $a_*^2 \to 1$ and $r_*^2 \to 0$ as $N_t, N_c \to \infty$ when $\boldsymbol{\mu}_t^{(r)} \neq \boldsymbol{\mu}_c^{(r)}$.

## A.3 Distribution of $\mathcal{W}^*$

The following theorem characterizes the distribution of the matched control sample moments of any variable $\mathcal{W}^*$ uncorrelated with $Z^*$ and $\mathbf{X}^{(s)}$ in the control sample conditional on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$. The sample means and variances of $\mathcal{Z}^*$ and $\mathcal{W}^*$ are denoted analogous to those of $\mathcal{Z}$ and $\mathcal{W}$ in Section 3. The theorem is used extensively in the derivation of the results in Section 2. The proof of the theorem closely parallels the proof of theorem 3.1 of Rubin and Thomas (1992b), using the additional fact that $\mathbf{X}^{(s)}$ and $\mathbf{X}^{(r)}$ are independent in the canonical representation, and is not presented here.

*Theorem A.1.* For matching methods that depend only on $\mathbf{X}^{(s)}$ and $Z^*$, and covariate distributions satisfying the conditions of Section 2.1,

$$\text{E}(\bar{\mathcal{W}}_t^* - \bar{\mathcal{W}}_{mc}^* | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}) = 0, \quad \text{(A.4)}$$

$$\text{var}(\bar{\mathcal{W}}_t^* - \bar{\mathcal{W}}_{mc}^* | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}) = n_{mc}^{-1}\left(1 - \frac{n_{mc}}{N_c}\right), \quad \text{(A.5)}$$

$$\text{E}\{v_{mc}(\mathcal{W}^*) | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}\} = 1, \quad \text{(A.6)}$$

and

$$\text{E}\{v_t(\mathcal{W}^*) | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}\} = \sigma^2. \quad \text{(A.7)}$$

In addition, for any linear combination, $\mathcal{Z}^*$, of $\mathbf{X}^{(s)}$ and $Z^*$, and corresponding uncorrelated variable $\mathcal{W}^*$, conditional on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$, $(\bar{\mathcal{W}}_{mc}^*, v_{mc}(\mathcal{W}^*), \bar{\mathcal{W}}_t^*, v_t(\mathcal{W}^*))$ are independent of $(\bar{\mathcal{Z}}_{mc}^*, v_{mc}(\mathcal{Z}^*), \bar{\mathcal{Z}}_t^*, v_t(\mathcal{Z}^*))$.

The following corollary of Theorem A.1 is analogous to corollary 3.1 of Rubin and Thomas (1992b), except that the corollary applies to any standardized linear combination of $\mathbf{X}$.

*Corollary A.1.* Consider a standardized linear combination, $U$, and its representation in (A.2). For matching methods that depend only on $\mathbf{X}^{(s)}$ and $Z^*$, and covariate distributions satisfying the conditions of Section 2.1,

$$\text{var}(\bar{U}_t - \bar{U}_{mc}) = \text{var}\{\rho_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*)\} + \{1 - \text{E}(\rho_*^2)\}n_{mc}^{-1}\left(1 - \frac{n_{mc}}{N_c}\right), \quad \text{(A.8)}$$

$$\text{E}\{v_{mc}(U)\} = \text{E}\{\rho_*^2 v_{mc}(\mathcal{Z}^*)\} + \{1 - \text{E}(\rho_*^2)\}, \quad \text{(A.9)}$$

and

$$\text{E}\{v_t(U)\} = \text{E}\{\rho_*^2 v_t(\mathcal{Z}^*)\} + \sigma^2\{1 - \text{E}(\rho_*^2)\}. \quad \text{(A.10)}$$

The equality in (A.8) is obtained by applying (A.2) to the sample means,

$$\bar{U}_t - \bar{U}_{mc} = \rho_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*) + (1 - \rho_*^2)^{1/2}(\bar{\mathcal{W}}_t^* - \bar{\mathcal{W}}_{mc}^*),$$

and computing $\text{var}(\bar{U}_t - \bar{U}_{mc})$ by first conditioning on $(\bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)})$, then applying (A.4), (A.5), and the conditional independence of the moments of $\mathcal{Z}^*$ and $\mathcal{W}^*$,

$$\text{E}(\bar{U}_t - \bar{U}_{mc} | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}) = \text{E}\{\rho_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*) | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}\}$$

and

$$\text{var}(\bar{U}_t - \bar{U}_{mc} | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}) = \text{var}\{\rho_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*) | \bar{\mathbf{X}}_t^{(r)}, \bar{\mathbf{X}}_c^{(r)}\} + (1 - \rho_*^2)n_{mc}^{-1}\left(1 - \frac{n_{mc}}{N_c}\right),$$

establishing (A.8). Similarly, (A.9) and (A.10) follow from (A.6), (A.7), and the independence conditions in Theorem A.1 applied to the representations

$$v_{mc}(U) = \rho_*^2 v_{mc}(\mathcal{Z}^*) + (1 - \rho_*^2)v_{mc}(\mathcal{W}^*) + 2\rho_*(1 - \rho_*)^{1/2}\text{cov}_{mc}(\mathcal{Z}^*, \mathcal{W}^*)$$

and

$$v_t(U) = \rho_*^2 v_t(\mathcal{Z}^*) + (1-\rho_*^2)v_t(\mathcal{W}^*) + 2\rho_*(1-\rho_*)^{1/2}\mathrm{cov}_t(\mathcal{Z}^*, \mathcal{W}^*),$$

where $\mathrm{cov}_t$ and $\mathrm{cov}_{mc}$ are the treated and matched control sample covariances.

### A.4 Derivation of the Results in Section 3

Using the representation in (A.1) with $U = \mathcal{Z}$ and $\rho_* = a_*$,

$$\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_{mc} = a_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*) + (1-a_*^2)^{1/2}(\bar{\mathcal{W}}_t^* - \bar{\mathcal{W}}_{mc}^*). \quad \text{(A.11)}$$

So from (A.4),

$$\mathrm{E}(\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_{mc}) = \mathrm{E}\{a_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*)\},$$

and the latter term is approximately 0 when close matching of the sample moments of $\mathbf{X}^{(s)}$ and $Z^*$ is achieved, yielding (7).

### Approximations for $\mathrm{E}(a_*^2)$ and $E(r_*^2)$

The results in Section 3.2 involving variances require approximations for the expectations of the conditional correlations defined at the end of Section A.2. Specializing the formula for $\rho_*$ to $a_*$ and noting that $\alpha^2 = r^{-1}(1 - \boldsymbol{\alpha}^{(s)'}\Psi_{c(s,s)}\boldsymbol{\alpha}^{(s)})$, it follows that

$$\mathrm{E}(a_*^2) \geq \mathrm{E}\left[r^{-1}\left\{\frac{1_r'(\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_t^{(r)})}{\|\bar{\mathbf{X}}_t^{(r)} - \bar{\mathbf{X}}_c^{(r)}\|}\right\}^2\right] \approx \frac{1 + B_{r|s}^2/V}{r + B_{r|s}^2/V}. \quad \text{(A.12)}$$

The latter approximation is obtained from (A.2) of Rubin and Thomas (1996), with $r$ and $B_{r|s}^2$ substituted for $p$ and $B^2$. The approximation for $\mathrm{E}(a_*^2)$ is very conservative if $\mathcal{Z}$ is a function of the the special matching variables, $\mathbf{X}^{(s)}$. An approximation for $\mathrm{E}(r_*^2)$ is obtained directly from section 4.1 of Rubin and Thomas (1992b) with the same substitutions,

$$\mathrm{E}(r_*^2) \approx \frac{1}{r + B_{r|s}^2/V}. \quad \text{(A.13)}$$

### Results Involving Variances of Matched Sample Means

Applying (A.8) of Corollary A.1 to $\mathcal{Z}$, approximation (A.12), and the approximate close matching of the means, $\bar{\mathcal{Z}}_t^* \approx \bar{\mathcal{Z}}_{mc}^*$,

$$\begin{aligned}\mathrm{var}(\bar{\mathcal{Z}}_t - \bar{\mathcal{Z}}_{mc}) &= \mathrm{var}\{a_*(\bar{\mathcal{Z}}_t^* - \bar{\mathcal{Z}}_{mc}^*)\} + \{1 - \mathrm{E}(a_*^2)\}n_{mc}^{-1}\left(1 - \frac{n_{mc}}{N_c}\right) \\ &\approx \{1 - \mathrm{E}(a_*^2)\}n_{mc}^{-1}\left(1 - \frac{n_{mc}}{N_c}\right) \\ &\approx \left(\frac{r-1}{r + B_{r|s}^2/V}\right)N_t^{-1}(m^{-1} - R^{-1}).\end{aligned}$$

A similar derivation is used for $\mathrm{var}(\bar{\mathcal{W}}_t - \bar{\mathcal{W}}_{mc})$, along with (A.13).

### Results Involving Expectations of Matched Sample Standard Deviations

Applying (A.9) with $U = \mathcal{Z}$, and using the close matching assumption, $v_t(\mathcal{Z}^*) \approx v_{mc}(\mathcal{Z}^*)$,

$$\begin{aligned}\mathrm{E}\{v_{mc}(\mathcal{Z})\} &= \mathrm{E}\{a_*^2 v_{mc}(\mathcal{Z}^*)\} + \{1 - \mathrm{E}(a_*^2)\} \\ &\approx \mathrm{E}\{a_*^2 v_t(\mathcal{Z}^*)\} + \{1 - \mathrm{E}(a_*^2)\}.\end{aligned}$$

Substituting (A.10) into $\mathrm{E}\{a_*^2 v_t(\mathcal{Z}^*)\}$ yields

$$\mathrm{E}\{v_{mc}(\mathcal{Z})\} - \mathrm{E}\{v_t(\mathcal{Z})\} = (1-\sigma^2)\{1 - \mathrm{E}(a_*^2)\},$$

and using (A.12),

$$\begin{aligned}\left|\frac{\mathrm{E}\{v_{mc}(\mathcal{Z})\}}{\mathrm{E}\{v_t(\mathcal{Z})\}} - 1\right| &\approx |(1-\sigma^2)\{1 - \mathrm{E}(a_*^2)\}|/\mathrm{E}\{v_t(\mathcal{Z})\} \\ &\approx \left|(1-\sigma^2)\frac{r-1}{r + B_{r|s}^2/V}\right| \Big/ \mathrm{E}\{v_t(\mathcal{Z})\}.\end{aligned}$$

A similar derivation using (A.13) and the fact that $\mathrm{E}\{v_t(\mathcal{W})\} = \sigma^2$ establishes the approximation for $\mathrm{E}\{v_{mc}(\mathcal{W})\}$.

*[Received April 1997. Revised August 1999.]*

## REFERENCES


Ashikaga, T., and Chang, P. (1981), "Robustness of Fisher's Linear Discriminant Function Under Two-Component Mixed Normal Models," *Journal of the American Statistical Association*, 76, 676–680.

Carpenter, R. G. (1977), "Matching When Covariates are Normally Distributed," *Biometrika*, 64, 299–307.

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 295–313.

Cochran, W. G., and Rubin, D. B. (1973), "Controlling Bias in Observational Studies: A Review," *Sankhya*, Ser. A, 35, 417–446.

Daudin, J. J. (1986), "Selection of Variables in Mixed-Variable Discriminant Analysis," *Biometrics*, 42, 473–481.

Drake, C. (1993), "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect," *Biometrics*, 49, 1231–1236.

Efron, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis," *Journal of the American Statistical Association*, 70, 892–898.

Gu, X., and Rosenbaum, P. (1993), "Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms," *Journal of Computational and Graphical Statistics*, 2, 405–420.

Hill, J., Rubin, D. B., and Thomas, N. (1999), "The Design of the New York School Choice Scholarship Program Evaluation," in *Research Design: Donald Campbell's Legacy*, ed. L. Bickman, London: Sage.

Krzanowski, W. (1975), "Discrimination and Classification Using Both Binary and Continuous Variables," *Journal of the American Statistical Association*, 70, 782–790.

——— (1980), "Mixtures of Continuous and Categorical Variables in Discriminant Analysis," *Biometrics*, 36, 493–499.

Olkin, I., and Tate, R. F. (1961), "Multivariate Correlation Models With Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, 32, 448–465.

Reinisch, J. M., Sanders, S. A., Lykke-Mortensen, E., and Rubin, D. B. (1995), "Prenatal Exposure to Phenobarbitol and Intelligence Deficits in Adult Human Males," *The Journal of the American Medical Association*, 274, 1518–1525.

Rosenbaum, P. (1989), "Optimal Matching for Observational Studies," *Journal of the American Statistical Association*, 84, 1024–1032.

——— (1991), "A Characterization of Optimal Designs for Observational Studies," *Journal of the Royal Statistical Society*, Ser. B, 53, 597–610.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55.

——— (1985a), "The Bias Due to Incomplete Matching," *Biometrics*, 41, 103–116.

——— (1985b), "Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1973a), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203.

——— (1973b), "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183.

——— (1976a), "Multivariate Matching Methods That are Equal Percent Bias Reducing, I: Some Examples," *Biometrics*, 32, 109–120.

——— (1976b), "Multivariate Matching Methods That are Equal Percent Bias Reducing, II: Maximums on Bias Reduction," *Biometrics*, 32, 121–132.

——— (1979), "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies," *Journal of the American Statistical Association*, 74, 318–328.

——— (1980), "Bias Reduction Using Mahalanobis-Metric Matching," *Biometrics*, 36, 293–298.

Rubin, D. B., and Thomas, N. (1992a), "Affinely Invariant Matching Methods With Ellipsoidal Distributions," *The Annals of Statistics*, 20, 1079–1093.

——— (1992b), "Characterizing the Effect of Matching Using Linear Propensity Score Methods With Normal Distributions," *Biometrika*, 79, 797–809.

——— (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 254–268.