Constructing a Control Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score

PAUL R. ROSENBAUM and DONALD B. RUBIN*

Matched sampling is a method for selecting units from a large reservoir of potential controls to produce a control group of modest size that is similar to a treated group with respect to the distribution of observed covariates. We illustrate the use of multivariate matching methods in an observational study of the effects of prenatal exposure to barbiturates on subsequent psychological development. A key idea is the use of the propensity score as a distinct matching variable.

KEY WORDS: Observational studies; Bias reduction; Propensity scores; Mahalanobis metric matching; Nearest available matching.

1. INTRODUCTION: BACKGROUND; WHY MATCH?

Matched Sampling in Observational Studies. In many observational studies, there is a relatively small group of subjects exposed to a treatment and a much larger group of control subjects not exposed. When the costs associated with obtaining outcome or response data from subjects are high, some sampling of the control reservoir is often necessary. Matched sampling attempts to choose the controls for further study so that they are similar to the treated subjects with respect to background variables measured on all subjects.

The Danish Cohort. We examine multivariate matched sampling using initial data from a proposed study of the effects on psychological development of prenatal exposure to barbiturates. The children under study were born between 1959 and 1961 and have been the object of other studies (e.g., Mednick et al. 1971; Zachau-Christiansen and Ross 1975). Prenatal and perinatal information is available for 221 barbiturate-exposed children and 7,027 unexposed children. A battery of measures of subsequent psychological development are to be obtained from all 221 exposed children, but cost considerations require sampling of the unexposed children to create a control group in which the measures will be obtained. The cost of the study will be approximately linear in the number of children studied—with basic costs that are largely independent of the number of children, and other costs associated with locating and examining the children that are approximately proportional to the number of children studied.

Approximate Efficiency Considerations. To obtain a rough idea of the loss in efficiency involved in not including all unexposed children as controls, suppose for the moment that there is no concern with biases between exposed and unexposed groups, in the sense that the mean difference between the groups can be regarded as an unbiased estimate of the effect of prenatal exposure to barbiturates. If control children are randomly sampled from among the 7,027 unexposed children, and if the variance, σ^2, of a particular psychological response is the same in the treated and control groups, then the standard error of the treated versus control difference in means will be $\sigma(1/221 + 1/N_c)^{1/2}$, where N_c is the number of control children studied. For N_c = 100, 221, 442 (= 2 × 221), 663 (= 3 × 221), 884 (= 4 × 221), 2,210 (= 10 × 221), and 7,027 (= 31.8 × 221), the multipliers $(1/221 + 1/N_c)^{1/2}$ are, respectively, .121, .095, .082, .078, .075, .071, and .068. The cost of studying all 7,027 control children would be substantially greater than the cost of a modest sample, and the gain in precision would not be commensurate with the increase in cost. Cost considerations in this study led to a sample of 221 matched controls.

Distributions of Background Variables Before Matching. In fact, forming a control group by random sampling of the unexposed children is not a good idea. Many of the unexposed children may not be good controls because they are quite different from all exposed children with respect to background variables. Consequently, controls will not be selected by random sampling, but rather by matched sampling on the basis of the covariates listed in Table 1. From the t statistics and standardized differences in Table 1, we see that the exposed and unexposed children differ considerably. The hope is that matched sampling will produce a control group that is similar to the treated group with respect to these covariates.

For Nontechnical Audiences, Matched Sampling Is Often a Persuasive Method of Adjustment. One virtue, not the least important, of matched sampling is that nontechnical audiences often find that matching, when successful, is a persuasive method of adjusting for imbalances in observed covariates. Although matching algorithms can be complex,

*Paul R. Rosenbaum is Research Statistician, Research Statistics Group, Educational Testing Service, Princeton, NJ 08541. Donald B. Rubin is Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. This work was partially supported by National Institutes of Health Grant HD17574-01A2 to the Kinsey Institute for Research at Indiana University, by Datametrics Research, Inc., and by Educational Testing Service. This work was completed while the second author was Professor in the Departments of Statistics and Education at the University of Chicago. The analyses presented are preliminary and intended only to explore methodological options; none of the matched samples are the actual ones to be used for study of in utero exposure to barbiturates. The authors are grateful to Robert T. Patrick for extensive assistance in computing.

Table 1. Covariate Imbalance Prior to Matching

Covariate	Description of original covariate	As used for estimating the propensity score	Differences in covariate means prior to matching: Two-sample t statistic	Differences in covariate means prior to matching: Standardized difference in %*
Child characteristics				
Sex	Female/male	0, 1	−1.02	−7
Twin	Single/multiple birth	0, 1	−1.28	−10
Sibpos	Oldest child (no, yes)	0, 1	−2.33	−16
C-age	Age at start of study	Months	.46	3
Mother characteristics				
SES	Socioeconomic status (9 ordered categories)	Integers 1–9	3.66	26
Education	Mother's education (4 ordered categories)	Integers 1–4	2.09	15
Single	Unmarried (no, yes)	0, 1	−5.70	−43
M-age	Age (years)	Years	8.99	59
Height	Mother's height (5 ordered categories)	Integers 1–5	2.55	18
Characteristics of the pregnancy				
WGTHGT3	(Weight gain)/height^3 (30 values based on category midpoints)	30 values	−.00	−0
PBC415	Pregnancy complications (an index)	Index value and its square	2.61	17
PRECLAM	Preeclamsia (no, yes)	0, 1	1.82	9
RESPILL	Respiratory illness (no, yes)	0, 1	1.73	10
LENGEST	Length of gestation (10 ordered categories)	$(10 - i)^{1/2}$ and i for $i = 1, 2, \ldots, 10$	.72	6
Cigarette	Cigarette consumption, last trimester (0 = none, plus 4 ordered categories)	Integers 0–4 and their squares	−.48	−3
Other Drugs				
Antihistamine	No. of exposures to antihistamines (0–6)	Integers 0–6 and their squares	1.76	10
Hormone	No. of exposures to hormones (0–6)	Integers 0–6 and their squares	8.41	28
HRMG1	Exposed to hormone type 1 (no, yes)	0, 1	2.67	15
HRMG2	Exposed to hormone type 2 (no, yes)	0, 1	3.75	19
HRMG3	Exposed to hormone type 3 (no, yes)	0, 1	3.46	18

*The standardized difference in percent is the mean difference as a percentage of the average standard deviation: $100(\bar{x}_1 - \bar{x}_{0R})/[(s_1^2 + s_{0R}^2)/2]^{1/2}$, where for each covariate, $\bar{x}_1$ and $\bar{x}_{0R}$ are the sample means in the treated group and the control reservoir and s_1^2 and s_{0R}^2 are the corresponding sample variances.

the simplest methods, such as comparisons of sample moments, often suffice to indicate whether treated and matched control groups can be directly compared without bias due to observed covariates.

The Limitations of Incomplete Categorical Matching. Perhaps the most obvious matching method involves categorizing each of the 20 variables in Table 1 and considering a treated child and control child as a suitable matched pair only if they fall in the same category on each variable. Unfortunately, even if each of the 20 variables is divided into just two categories, the 7,027 potential controls will be distributed among $2^{20} \doteq 1$ million matching categories, so exact matches may be hard to find for some treated children. If a version of categorical matching described by Rosenbaum and Rubin (in press) is applied to the current data, only 126 of the 221 treated children have exact matches, so 95 (43%) of the treated children are discarded as unmatchable. Discarding treated children in this way can lead to serious biases, since the unmatched treated children differ systematically from the matched treated children. We consider only methods that match all 221 exposed children.

2. THEORY RELEVANT TO THE CHOICE OF A MATCHING METHOD

2.1 The Most Important Scalar Matching Variable: The Propensity Score

Let $\mathbf{x}$ denote the vector of covariates for a particular child, and let the binary variable z indicate whether the child was exposed ($z = 1$) or unexposed ($z = 0$). The propensity score, $e(\mathbf{x})$, is the conditional probability of exposure given the covariates; that is, $e(\mathbf{x}) = \Pr(z = 1|\mathbf{x})$. Treated children and control children selected to have the same value of $e(\mathbf{x})$ will have the same distributions of $\mathbf{x}$; formally, z and $\mathbf{x}$ are conditionally independent given $e(\mathbf{x})$. Exact matching on $e(\mathbf{x})$ will, therefore, tend to balance the $\mathbf{x}$ distributions in the treated and control groups. Moreover, matching on $e(\mathbf{x})$ and any function of $\mathbf{x}$, such as selected coordinates of $\mathbf{x}$, will also balance $\mathbf{x}$. (For proofs of these balancing properties of propensity scores, see Rosenbaum and Rubin 1983a, Theorems 1 and 2. For related discussions of propensity scores, see Rubin 1983, 1984; Rosenbaum 1984b; and Rosenbaum and Rubin 1984.) The propensity score is a po-

tential matching variable because it does not depend on response information that will be collected after matching. Since exact adjustment for a known propensity score will, on average, remove all of the bias in $\mathbf{x}$, the propensity score $e(\mathbf{x})$ is in a sense the most important scalar matching variable.

Matching on $e(\mathbf{x})$ balances the *observed* covariates $\mathbf{x}$; however, unlike randomization, matching on $e(\mathbf{x})$ does not balance *unobserved* covariates except to the extent that they are correlated with $\mathbf{x}$. For discussion of methods for addressing the possible effects of unobserved covariates in observational studies, see Rosenbaum and Rubin (1983b) and Rosenbaum (1984a and in press).

In practice, several issues need to be addressed before the propensity score can be used as a matching variable. First, the functional form of $e(\mathbf{x})$ is rarely if ever known, at least in observational studies such as the one we describe, and therefore $e(\mathbf{x})$ must be estimated from the available data. Second, exact matches will rarely be available, and so issues of closeness on $e(\mathbf{x})$ must be addressed. Third, adjustment for $e(\mathbf{x})$ balances $\mathbf{x}$ only in expectation, that is, averaging over repeated studies. In any particular study, further adjustment for $\mathbf{x}$ may be required to control chance imbalances in $\mathbf{x}$. Such adjustments, for example by covariance analysis, are often used in randomized experiments to control chance imbalances in observed covariates.

As noted by Rosenbaum and Rubin (1983a, sec. 2.3), matching on the propensity score is generalization to arbitrary $\mathbf{x}$ distributions of discriminant matching for multivariate normal $\mathbf{x}$ as proposed by Rubin (1970) and discussed by Cochran and Rubin (1973) and Rubin (1976a,b; 1979; 1980). Propensity matching is not, however, the same as any of the several procedures proposed by Miettinen (1976): the propensity score is not generally a confounder score (see Rosenbaum and Rubin 1983a, sec. 3.3, for discussion).

2.2 Estimating the Propensity Score

We estimated the propensity score in the Danish cohort using a logit model (Cox 1970):

$$q(\mathbf{x}) \equiv \log[(1 - e(\mathbf{x}))/e(\mathbf{x})] = \alpha + \boldsymbol{\beta}^T\mathbf{f}(\mathbf{x}),$$

where α and $\boldsymbol{\beta}$ are parameters to be estimated, $q(\mathbf{x})$ is the log odds against exposure, and $\mathbf{f}(\mathbf{x})$ is a specified function, which in this instance included quadratic terms and transforms (see Table 1 for details). A logit model for $e(\mathbf{x})$ can be formally derived from $\Pr(\mathbf{x}|z = t)$ if the latter has any of a variety of exponential family distributions, such as the multivariate normal $N(\boldsymbol{\mu}_t, \Sigma)$, the multivariate logit model for binary data in Cox (1972), the quadratic exponential family of Dempster (1971), or the multinomial/multivariate normal distribution of Dempster (1973); see Rosenbaum and Rubin [1983a, sec. 2.3(ii)] for details.

The sample means of the maximum likelihood estimates $\hat{q}(\mathbf{x})$ of $q(\mathbf{x})$ are 3.06 and 3.76 in the treated and control groups, respectively. The sample variance of $\hat{q}(\mathbf{x})$ is 2.3 times greater in the treated group than in the control group. The standardized difference for $\hat{q}(\mathbf{x})$ is .77 (calculated as in the footnote to Table 1), which, as one would expect, is larger than the standardized difference for any single variable in Table 1. In the treated and control groups, respectively, the minimum values of $\hat{q}(\mathbf{x})$ are -3.9 and $-.6$; the lower quartiles, 2.5 and 3.2; the medians, 3.1 and 3.7; the upper quartiles, 3.8 and 4.2; and the maximums, 5.1 and 7.5. Three treated children have $\hat{q}(\mathbf{x})$ values lower than any control child: their $\hat{q}(\mathbf{x})$ values are -3.9, -1.3, and -1.2.

There is, then, a substantial difference along the propensity score. The larger variance in the treated groups suggests that finding appropriate matches will be relatively more difficult than if the variances were equal (Cochran and Rubin 1973, Table 2.3.1; Rubin 1973a, Table 5.1; Rubin 1980, Table 1). The reason for the difficulty is the concentration of the $\hat{q}(\mathbf{x})$'s around 3.76 in the control group and the wider dispersion of the $\hat{q}(\mathbf{x})$'s around 3.06 in the treated group: for matching, controls are required with low values of $\hat{q}(\mathbf{x})$, which are relatively uncommon in the control group.

2.4 Matching Methods That Are Equal-Percent Bias Reducing

The mean bias or expected difference in $\mathbf{x}$ prior to matching is $E(\mathbf{x}|z = 1) - E(\mathbf{x}|z = 0)$, whereas the mean bias in $\mathbf{x}$ after matching is $E(\mathbf{x}|z = 1) - \boldsymbol{\mu}_{0M}$, where $\boldsymbol{\mu}_{0M}$ is the expected value of $\mathbf{x}$ in the matched control group. Generally, $\boldsymbol{\mu}_{0M}$ depends on the matching method used, whereas $E(\mathbf{x}|z = 1)$ and $E(\mathbf{x}|z = 0)$ depend only on population characteristics. As defined by Rubin (1976a,b), a matching method is equal-percent bias reducing (EPBR) if the reduction in bias is the same for each coordinate of $\mathbf{x}$, that is, if

$$E(\mathbf{x}|z = 1) - \boldsymbol{\mu}_{0M} = \gamma\{E(\mathbf{x}|z = 1) - E(\mathbf{x}|z = 0)\}$$

for some scalar $0 \leqslant \gamma \leqslant 1$. If a matching method is not EPBR, then matching actually increases the bias for some linear functions of $\mathbf{x}$. If little is known about the relationship between $\mathbf{x}$ and the response variables that will be collected after matching, then EPBR matching methods are attractive, since they are the only methods that reduce bias in all variables having linear regression on $\mathbf{x}$. Rosenbaum and Rubin (1983a, sec. 3.2) show that matching on the population propensity score alone is EPBR whenever $\mathbf{x}$ has a linear regression on some scalar function of e; that is, whenever $E(\mathbf{x}|e) = \boldsymbol{\alpha} + \boldsymbol{\gamma}^T g(e)$ for some scalar function $g(\cdot)$.

3. CONSTRUCTING A MATCHED SAMPLE: AN EMPIRICAL COMPARISON OF THREE MULTIVARIATE METHODS

3.1 Overview: How Much Importance Should Be Given to the Propensity Score?

Matched samples were constructed by using three different methods that matched every treated child to one control child. By design, all three methods required exact matches on sex. The three methods differed in the importance given to the estimated propensity score relative to the other variables in $\mathbf{x}$.

3.2 Nearest Available Matching on the Estimated Propensity Score

With nearest available propensity score matching, (a) treated and control children are randomly ordered; (b) the

first treated child is matched with the control child of the same sex having the nearest $\hat{q}(\mathbf{x})$, and both children are removed from the lists of treated and control children; (c) step (b) is repeated for the remaining unmatched treated children. The decision to define distance in terms of $\hat{q}(\mathbf{x})$ rather than $\hat{e}(\mathbf{x})$ was somewhat but not entirely arbitrary and probably had negligible effect. It did, however, avoid the compression of the $\hat{e}(\mathbf{x})$ scale near 0 and 1, and moreover, $\hat{q}(\mathbf{x})$ was more nearly normally distributed, which is relevant in the context of Section 3.4. Nearest available matching on a scalar covariate x was studied by Rubin (1970), reviewed by Cochran and Rubin (1973), and extended by Rubin (1973a,b); its application to matching on linear discriminant scores with bivariate normal $\mathbf{x}$ was studied, using Monte Carlo, by Rubin (1979, 1980). The effects of random ordering in step (a) rather than ordering by $q(\hat{\mathbf{x}})$ are discussed by Rubin (1973a).

In Tables 2 and 3, column 1 describes the balance obtained in the samples matched by nearest available propensity matching. Note that the standardized differences in Table 2 have the same denominator as the standardized differences in Table 1, whereas the t statistics indicated by the footnotes to Table 2 have denominators that are affected by the matching and so are not directly comparable. (For discussion of the relationship between t statistics on covariate means and the coverage of confidence intervals for treatment effects formed by ignoring the variable, see, e.g., Cochran 1965, sec. 3.1.) The *two-sample t* statistics (values indicated by pluses in Table 2) are relevant for comparing the distributions of the covariates in the treated and matched control groups. The *paired t* statistics (values indicated by asterisks) are relevant for assessing the effects of residual biases in the covariates in analyses of outcome variables based on matched pair differences.

Table 2. Covariate Imbalance in Matched Samples: Standardized Differences (%)

Factor	Nearest available matching on the propensity score	Mahalanobis metric matching: Including the propensity score	Mahalanobis metric matching: Within propensity score calipers
Child characteristics			
Sex	0	0	0
Twin	−3	0	0
SIBPOS	−5	5*	0
C-age	7	6	−6
Mother characteristics			
SES	−10*	5	−1
Education	−17+	3	−7*
Single	−7	−3*	−2
M-age	−8*	5	−1
Height	−8	3	−9+**
Characteristics of the pregnancy			
WGTHGT3	−0	−3	1
PBC415	−14+*	6*	1
PRECLAM	0	0	0
RESPILL	−7	0	0
LENGEST	−12+	−3*	−4
Cigarette	0	10+***	9***
Drugs			
Antihistamine	−3	4	9*
Hormone	8*	6***	6*
HRMG1	−2	−7**	−6
HRMG2	−2	−5**	−9*
HRMG3	−3	−8**	−11*
$\hat{q}(\mathbf{x})$	−3**	−20+***	−3**

NOTE: The standardized difference in percent is $100(\bar{x}_1 - \bar{x}_{0M})/\{(s_1^2 + s_{0R}^2)/2\}^{1/2}$, where for each covariate, $\bar{x}_1$ and $\bar{x}_{0M}$ are the sample means in the treated group and matched control group and s_1^2 and s_{0R}^2 are the sample variances in the treated group and control reservoir. Note that the denominator of the standardized difference is the same for all three matching methods. The values of paired (*) and two-sample (+) t-statistics are indicated as follows: * and +, between 1.0 and 1.5 in absolute value; ** and ++, between 1.5 and 2.0 in absolute value; ***, between 2 and 3 in absolute value; ****, above 3 in absolute value.

In Table 3, the sample percent reduction in bias for a covariate is $100(1 - b_M/b_I)$, where b_I and b_M are the treated versus control differences in covariate means initially and after matching, respectively. When the initial mean sample bias, b_I, is small, the sample percent reduction in bias, $100(1 - b_M/b_I)$, is quite unstable; therefore, we report percent reductions only for variables with large initial biases (i.e., standardized differences above 20% in Table 1).

Tables 2 and 3 both suggest that nearest available matching on the propensity score has removed almost all of the mean difference along the propensity score—arguably the most important variable—and that there has been substantial reduction in the standardized differences for most variables. Still, the residual differences on several variables (Education, PBC415, LENGEST) are bothersome; further analytical adjustments for these variables might be required, for example, using analysis of covariance on matched-pair differences (Rubin 1973b, 1979).

3.3 Mahalanobis Metric Matching Including the Propensity Score

Mahalanobis metric matching has been described by Cochran and Rubin (1973) and Rubin (1976a) and studied in detail by Carpenter (1977) and Rubin (1979, 1980). With nearest available Mahalanobis metric matching, treated and control children are randomly ordered. The first treated child is matched with the closest control child of the same sex, where distance is defined by the Mahalanobis distance:

$$d(\mathbf{u}, \mathbf{v}) = (\mathbf{u} - \mathbf{v})^T \mathbf{C}_{0R}^{-1} (\mathbf{u} - \mathbf{v}), \qquad (1)$$

where $\mathbf{u}$ and $\mathbf{v}$ are values of $\{\mathbf{x}^T, \hat{q}(\mathbf{x})\}^T$ and $\mathbf{C}_{0R}$ is the sample covariance matrix of $\{\mathbf{x}^T, \hat{q}(\mathbf{x})\}$ in the control reservoir. The two matched children are then removed from the treated

Table 3. Covariate Imbalance in Matched Samples: Percent Reductions in Bias for Variables With Substantial Initial Bias (standardized absolute bias of 20% or greater)

	Nearest available matching on the propensity score	Mahalanobis metric matching: Including the propensity score	Mahalanobis metric matching: Within propensity score calipers
Single	84	93	95
Hormone	71	79	79
SES	140	81	105
M-age	114	91	102
$\hat{q}(\mathbf{x})$	96	74	96

and control lists, and the process is repeated. In the case of multivariate normal covariates with common covariance matrix in treated and control groups, Rubin (1976a, Theorem 2) has shown that Mahalanobis metric matching is EPBR.

The results of applying Mahalanobis metric matching are given in column 2 of Tables 2 and 3. As one might expect, Mahalanobis metric matching is somewhat more successful than propensity matching in reducing the standardized differences for individual coordinates of $\mathbf{x}$, but it is far less successful in reducing the standardized difference along the propensity score. The standardized difference of 20% and two-sample t statistic of -1.99 for $\hat{q}(\mathbf{x})$ are disturbing.

Mahalanobis metric matching produces several large matched-pair t statistics, as indicated in the footnotes to Table 2. The standard deviations of the within-pair differences in covariate values are smaller than under nearest available propensity score matching, so the matched-pair t statistics are larger. As noted previously (Sec. 3.2), the t statistics for different methods are not directly comparable. Nevertheless, the large values of the matched-pair t statistics indicate that analyses based on matched-pair differences can be misleading unless analysis of covariance is used to control within-pair differences due to $\mathbf{x}$.

3.4 Nearest Available Mahalanobis Metric Matching Within Calipers Defined by the Propensity Score

In an effort to obtain the best features of both previous methods, we now consider a hybrid matching method that first defines a subset of potential controls who are close to each treated child on the propensity score (i.e., within "calipers," Althauser and Rubin 1971) and then selects the control child from this subset by using nearest available Mahalanobis metric matching (for variables $\{\mathbf{x}, \hat{q}(\mathbf{x})\}$). The details of the procedure are given in Figure 1. With multivariate normal covariates having common covariance matrices in treated and control groups, and with $\hat{q}(\mathbf{x})$ replaced by its population value (the linear discriminant), this matching method would be EPBR, since each of the two stages would reduce bias in $\mathbf{x}$ by a constant percentage. A computational advantage of this method is a substantial reduction in the number of Mahalanobis distances that need to be computed. The method in Section 3.3 required the computation of about 1.5 million Mahalanobis distances.

1. Randomly order the treated children.
2. *Caliper Matching on the Propensity Score:* For the first treated child, find all available untreated children of the same sex with $\hat{q}(\mathbf{x})$ values that differ from the $\hat{q}(\mathbf{x})$ value for the treated child by less than a specified constant c. If there is no such untreated child, match the treated child to the untreated child of the same sex with the nearest value of $\hat{q}(\mathbf{x})$.
3. *Nearest Available Mahalanobis Metric Matching Within Calipers:* From the subset of children defined in step 2, select as a match the untreated child of the same sex who is closest in the sense of the Mahalanobis distance for the variables $\{\mathbf{x}, \hat{q}(\mathbf{x})\}$.
4. Remove the treated child and the matched control child from the lists of treated and untreated children. Go to step 2 for the next treated child.

Figure 1. Nearest Available Mahalanobis Metric Matching Within Calipers Defined by the Propensity Score.

The caliper width, c, used in step 2 of Figure 1 was determined by using results from Cochran and Rubin (1973) concerning the performance of caliper matching. Write σ_1^2 and σ_{0R}^2 for the variances of $\hat{q}(\mathbf{x})$ in the treated and untreated groups, and let $\sigma = [(\sigma_1^2 + \sigma_{0R}^2)/2]^{1/2}$. Table 2.3.1 of Cochran and Rubin (1973) suggests that when $\sigma_1^2/\sigma_{0R}^2 = 2$, a caliper width of $c = .2\sigma$ would remove 98% of the bias in a normally distributed covariate, that $c = .4\sigma$ would remove 93%, and that $c = .6\sigma$ would remove 86%. That table also suggests that narrower caliper widths (i.e., smaller values of c) are required as σ_1^2/σ_{0R}^2 increases. The point estimate of σ_1^2/σ_{0R}^2 for $\hat{q}(\mathbf{x})$ is $s_1^2/s_{0R}^2 = 2.3$. Therefore, in the hope of removing at least 90% of the bias along $\hat{q}(\mathbf{x})$ by caliper matching, we took $c = .25s = (.25)(.930) = .232$, where $s = [(s_1^2 + s_{0R}^2)/2]^{1/2}$. (For further discussion of caliper matching, see Cochran 1972 and Raynor 1983.) There were four treated children who had no available matches within the calipers; following step 2 of Figure 1, they were matched with the nearest available control on $\hat{q}(\mathbf{x})$.

Some results of this matching appear in column 3 of Tables 2 and 3. Mahalanobis metric matching within calipers defined by the propensity score appears superior to the two other methods: it is better than matching on the propensity score in that it yields fewer standardized differences above 10% in absolute value, and it is better than Mahalanobis metric matching in controlling the difference along the propensity score.

3.5 Nonlinear Response Surfaces

Tables 1–3 compare the three matching methods in terms of the means of $\mathbf{x}$ in the treated and matched control groups. If, however, the response has a nonlinear regression on $\mathbf{x}$ in the treated and control groups, then equal $\mathbf{x}$ means in matched samples do not necessarily indicate the absence of bias due to $\mathbf{x}$. For example, if the regressions on $\mathbf{x}$ are quadratic, then the means on $\mathbf{x}$ and $\mathbf{x}\mathbf{x}^T$ are both relevant. Table 4 summarizes the standardized biases for $230 = \binom{20}{2} + 2 \times 20$ variables: the 20 coordinates of $(\mathbf{x}, \hat{q}(\mathbf{x}))$ with sex excluded because exact matches for sex were obtained, the squares of these 20 variables, and the cross products of pairs of these variables. The third matching method—Mahalanobis metric matching within propensity score calipers—appears clearly superior.

Table 4. Summarized Standardized Differences (in %) for Covariates, Squares of Covariates, and Cross Products of Covariates

	Root mean square*	Maximum absolute
Prior to matching	24	78
Nearest available matching on the propensity score	9	49
Mahalanobis metric matching		
Including the propensity score	9	76
Using propensity score calipers	7	27

*$100(\bar{b}^2 + s_b^2)^{1/2}$, where $\bar{b}$ and s_b^2 are the sample mean and variance of standardized differences for the $\binom{20}{2} + 2\binom{20}{1} = 230$ variables.

4. SUMMARY

When combined with covariance adjustments of matched-pair differences, multivariate matched sampling is known to be one of the most robust methods for reducing bias due to imbalances in observed covariates (Rubin 1973b, 1979). Three methods for multivariate matched sampling have been illustrated and compared on data concerning the effects of prenatal barbiturate exposure. The first method was nearest available matching on the estimated propensity score; this method required less computation than the others and was fairly successful in reducing bias. The second method was nearest available Mahalanobis metric matching using all variables and the estimated propensity score; this method produced smaller standardized differences for individual variables but left a substantial difference along the propensity score. The third method—Mahalanobis metric matching within calipers defined by the estimated propensity score—appeared superior to the others with respect to balancing the covariates, their squares, and their cross products. Since the current study has examined just three possible methods on a single set of data, additional work on multivariate matching is needed in several areas: (a) theory concerning the effects on the best choice of matching method of (i) reservoir size, (ii) the magnitude of initial biases, (iii) dimensionality of $\mathbf{x}$, and (iv) covariate distributions; (b) theory concerning multivariate measures of the quality of matched samples involving non-Gaussian covariates; and (c) further empirical studies of multivariate matching methods.

[*Received December 1983. Revised September 1984.*]

REFERENCES

Althauser, R. P., and Rubin, D. B. (1971), "The Computerized Construction of a Matched Sample," *American Journal of Sociology*, 76, 325–346.

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

Cochran, W. G. (1965), "The Planning of Observational Studies of Human Populations" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 128, 234–255.

——— (1972), "Observational Studies," in *Statistical Papers in Honor of George W. Snedecor*, Ames: Iowa State University Press, pp. 70–90.

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

Cox, D. R. (1970), *The Analysis of Binary Data*, London: Methuen.

——— (1972), "The Analysis of Multivariate Binary Data," *Applied Statistics*, 21, 113–120.

Dempster, A. P. (1971), "An Overview of Multivariate Analysis," *Journal of Multivariate Analysis*, 1, 316–346.

——— (1973), "Aspects of Multinomial Logit Model," in *Multivariate Analysis III*, ed. P. R. Krishnaiah, New York: Academic Press, pp. 129–142.

Mednick, S. A., Mura, E., Schulsinger, F., and Mednick, B. (1971), "Prenatal Conditions and Infant Development in Children With Schizophrenic Parents," *Social Biology*, 18, 5103–5113.

Miettinen, O. (1976), "Stratification on a Multivariate Confounder Score," *American Journal of Epidemiology*, 104, 609–620.

Raynor, W. J. (1983), "Caliper Pair-Matching on a Continuous Variable in Case-Control Studies," *Communications in Statistics: Theory and Methods*, 12, 1499–1509.

Rosenbaum, P. R. (1984a), "From Association to Causation in Observational Studies: The Role of Tests of Strongly Ignorable Treatment Assignment," *Journal of the American Statistical Association*, 79, 41–48.

——— (1984b), "Conditional Permutation Tests and the Propensity Score in Observational Studies," *Journal of the American Statistical Association*, 79, 565–574.

——— (in press), "The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment," *Journal of the Royal Statistical Society*, Ser. A, 147.

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

——— (1983b), "Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study With Binary Outcome," *Journal of the Royal Statistical Society*, Ser. B, 45, 212–218.

——— (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score," *Journal of the American Statistical Association*, 79, 516–524.

——— (in press), "The Bias Due to Incomplete Matching," *Biometrics*.

Rubin, D. B. (1970), "The Use of Matched Sampling and Regression Adjustment in Observational Studies," unpublished Ph.D. dissertation, Harvard University, Dept. of Statistics.

——— (1973a), "Matching to Remove Bias in Observational Studies," *Biometrics*, 29, 159–183; Printer's correction (1974), 30, 728.

——— (1973b), "The Use of Matching and Regression Adjustment to Remove Bias in Observational Studies," *Biometrics*, 29, 185–203.

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

——— (1976b), "Matching Methods That Are Equal Percent Bias Reducing: Maximums on Bias Reduction Fixed Sample Sizes," *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.

——— (1983), "Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys" (discussion of Hansen, Madow, and Tepping), *Journal of the American Statistical Association*, 78, 803–805.

——— (1984), "Comment: Assessing the Fit of Logistic Regressions Using the Implied Discriminant Analysis" (discussion of Landwehr, Pregibon, and Shoemaker), *Journal of the American Statistical Association*, 79, 79–80.

Zachau-Christiansen, B. G., and Ross, E. M. (1975), *Babies: Human Development During the First Year*, New York: John Wiley.