Annenberg School for Communication

Departmental Papers (ASC)

University of Pennsylvania *Year* 2001

Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse

Bo Lu*
Elaine Zanutto†
Robert C. Hornik‡
Paul R. Rosenbaum**

*University of Pennsylvania
†University of Pennsylvania, zanutto@wharton.upenn.edu
‡University of Pennsylvania, rhornik@asc.upenn.edu
**University of Pennsylvania, rosenbap@wharton.upenn.edu

This paper is posted at ScholarlyCommons.
http://repository.upenn.edu/asc_papers/177

Matching With Doses in an Observational Study of a Media Campaign Against Drug Abuse

Bo LU, Elaine ZANUTTO, Robert HORNIK, Paul R. ROSENBAUM

Multivariate matching with doses of treatment differs from the treatment-control matching in three ways. First, pairs must not only balance covariates, but also must differ markedly in dose. Second, any two subjects may be paired, so that the matching is nonbipartite, and different algorithms are required. Finally, a propensity score with doses must be used in place of the conventional propensity score. We illustrate multivariate matching with doses using pilot data from a media campaign against drug abuse. The media campaign is intended to change attitudes and intentions related to illegal drugs, and the evaluation compares stated intentions among ostensibly comparable teens who reported markedly different exposures to the media campaign.

KEY WORDS: Coherent signed rank test; Equal percent bias reducing; Matching with doses; Nonbipartite matching; Observational studies; Ordinal logit model; Optimal matching; Propensity score.

1. INTRODUCTION: BALANCE WITH DIFFERENT DOSES

1.1 Varied Exposure to Antidrug Messages

The United States Office of National Drug Control Policy (ONDCP) recently launched a media campaign intended to reduce illegal drug use by the young Americans. Because the campaign was implemented throughout the United States, there is no unexposed or control group available for use in evaluating the effects of the campaign. An experiment that divided the nation into a checkerboard of media markets, with some markets exposed to the campaign and others not, was judged impractical.

The National Institute on Drug Abuse (NIDA) proposed that the evaluation plan will compare teens with varied degrees of exposure to media campaign, that is, with varied doses of the treatment. One would like to compare teens who received different exposures to the campaign, but who were similar in terms of baseline characteristics. Matching with doses means forming pairs with very different doses of treatment in such a way that the final high- and low-dose groups have similar or balanced distributions of observed covariates.

Here, we illustrate new methods for matching with doses using data on 521 teens who participated in the pilot for the evaluation. Matched pairs are formed in which 22 covariates are balanced, but the doses are very different. Because our goal is to illustrate matching techniques and not to evaluate the media campaign, our approach resembles the actual evaluation plan in a few respects, but diverges from it in many others. For instance, the pilot data are a convenient piece of the population, whereas the evaluation will use a national sample.

1.2 Outline: Data, Method, Results, Theory

This article is organized as follows. Sections 1.3 and 1.4 discuss aspects of the data, and Section 1.5 discusses the differences between matching with doses and matching treated subjects to controls. The proposed method is discussed in Section 2, including propensity scores with doses, optimal matching in a nonbipartite graph, and a distance measure for use with doses. Section 3 examines the performance of the method using the pilot data for the media campaign. Finally, the appendix presents some theory showing that the method has a desirable standard property namely equal percent bias reduction.

1.3 Doses

The dose of exposure to the media campaign was defined using the answers to the following three questions. In recent months, about how often have you seen such antidrug commercials on TV, or heard them on the radio? In recent months, about how often have you seen such antidrug ads in newspapers or magazines? In recent months, about how often have you seen such antidrug ads in movie theatres or on rental videos? We ignored the ordinal nature of the response categories and added the scaled responses to the three questions, and then formed five dose groups. The three lowest dose groups each contained about a quarter of the children, with the rest evenly divided between the two highest dose groups. This division reflected, in part, the discreteness of the underlying data and, in part, a right skewed distribution with tighter spacing at lower doses. The use of *five* groups, as distinct from three or ten, was motivated by the somewhat related results of Cox (1957) and Cochran (1968).

To give a sense of what the dose categories mean, Table 1 describes the responses of five particular teens, one from each of the categories. Obviously, a variety of patterns of responses are found in each category, and only one is listed in the Table 1. Nonetheless, the dose categories are quite different.

The use of five dose categories truncates extreme responses. Is this wise? After inspecting the data, we thought perhaps it was. For instance, one question asked: On the weekend, about how many hours of TV do you usually watch? Please include both Saturday and Sunday. One teen responded with "85 hours," another with "68 hours," and several more with "50 hours." So, in the end, we did feel comfortable truncating extreme responses.

Bo Lu is a PhD candidate, Elaine Zanutto is assistant professor, and Paul Rosenbaum is Robert G. Putzel Professor in the Department of Statistics of the Wharton School of the University of Pennsylvania, Philadelphia, PA 19104-6302 (E-mail: *rosenbaum@stat.wharton.upenn.edu*). Robert Hornik is Wilbur Schramm Professor of Communication and Health Policy at the Annenberg School for Communication of the University of Pennsylvania. The hospitality and support of the Center for Advanced Study in the Behavioral Sciences are gratefully acknowledged. Rosenbaum's work was supported by a grant from the National Science Foundation. The authors thank Monique Guignard-Spielberg for advice about optimal nonbipartite matching.

© 2001 American Statistical Association
Journal of the American Statistical Association
December 2001, Vol. 96, No. 456, Applications and Case Studies

Table 1. Understanding the Dose Categories: Responses of One Typical Teen in Each of the Five Dose Categories

Dose	*TV*	*Print*	*Movie*
5 = *high*	>1 per day	>1 per day	≥4 per month
4	daily	daily	≥4 per month
3	1 to 3 per week	1 to 3 per week	≥4 per month
2	1 to 3 per month	1 to 3 per week	1 to 3 per month
1 = *low*	1 to 3 per month	<1 per month	<1 per month

NOTE: Responses to "How often have you seen such antidrug commercials (on TV or radio, in newspapers or magazines, in movie theatres or rental videos)?"

1.4 Covariates

In the current illustration of matching technology, we used the 22 covariates in Table 2. The variables describe demographics, TV habits, organized activities, and self-reported drug use. Table 2 also reports the rank correlation with dose, together with the two-sided P-value testing no association. Notice that a missing mother's education is counted in the middle category, high school graduate. An alternative strategy with propensity scores is to code "missing" as a separate variable and include it in the propensity score (Rosenbaum and Rubin 1984, app B). Also, due to extreme skewness of a few covariates, we took square root transformations of some variables.

Older children reported less exposure to the media campaign than the younger children, but gender and race showed little or no relationship with exposure. Not surprisingly, TV habits were related to exposure, but self-reported past drug use showed little or no relationship. Participation in sports and "other" activities were associated with higher exposures, but most activities showed little or no relationship.

A few technical decisions about the covariates require brief mention. For the purpose of evaluating the media campaign, it is unclear whether adjustments should be made for TV habits. On the one hand, much of the variation in exposure to the media campaign is created by varied TV habits, and this argues against adjustments. On the other hand, different types of TV attract different types of viewers, and this argues in favor of adjustments. For instance, the so-called "sensation-seeking" is said to be related to drug abuse (e.g., Ball, Carroll, and Rounsaville 1994; Kosten, Ball, and Rounsaville 1994; Bardo, Donohew, and Harrington 1996), and certain types of TV, such as MTV, might possibly be more popular with sensation-seekers. A thorough evaluation would synthesize the results of many different analyses. Here, however, we wanted to illustrate the matching techniques and so we picked the more challenging matching problem. Because TV habits are related to exposure, matching on TV habits makes the matching task more challenging. If our algorithm can balance all the 22 covariates, then it can also balance any subset of the 22 covariates, so including all the 22 covariates provide a better test of what our algorithm can accomplish.

The pilot data are not longitudinal, but for our current purpose of illustrating matching techniques, we ignore certain

Table 2. Covariates Before Matching and Their Rank Correlations With Dose of Exposure to the Media Campaign

Covariate	*Kendall's* τ	*P-Value*
Age	−.10	.0013
Black (1 = yes, 0 = no)	.00	.9500
Other Nonwhite Race (1 = yes, 0 = no)	−.02	.6300
Gender (1 = female, 0 = male)	−.01	.8300
Mother graduated high school (0 = no, 1 = yes or missing)	.00	.9300
Mother graduated college (0 = no or missing, 1 = yes)	.07	.0600
Cable or satellite TV at home (1 = yes, 0 = no)	.01	.7400
$\sqrt{\text{hours of TV on a weekday}}$	.07	.0300
$\sqrt{\text{hours of TV on a weekend}}$	.09	.0056
Music TV per month (e.g., MTV)		
0 = *never*, . . . , 3 = 15 to 30 days/month	.10	.0048
Sports channel per month (e.g., ESPN)		
0 = *never*, . . . , 3 = 15 to 30 days/month	.10	.0048
Music, dance, or theater (1 = yes, 0 = no)	.04	.2500
Athletic teams or sports (1 = yes, 0 = no)	.11	.0047
Clubs (e.g., Girl Scouts) (1 = yes, 0 = no)	−.01	.8200
Religious youth groups (1 = yes, 0 = no)	−.05	.2300
Other activities (1 = yes, 0 = no)	.10	.0098
Ever smoked cigarettes		
(0 = never, . . . , 4 = yes in the last 30 days)	.01	.7600
Ever used marijuana (1 = yes, 0 = no)	.00	.9900
Marijuana, last 12 months		
(0 = never, . . . , 6 = 40 or more times)	.01	.8100
Ever used inhalants (1 = yes, 0 = no)	.03	.4300
Inhalants, last 12 months		
(0 = never, . . . , 6 = 40 or more times)	.02	.6400
Any friends use drugs (1 = yes, 0 = no)	.02	.6900

related problems that will be addressed in the actual evaluation study. Specifically, respondents answer questions at a single moment, but some questions ask about the past whereas others about intentions for the future. For the current purpose of evaluating matching techniques, we accept responses to questions naively, accepting answers about the past as facts about the past, and answers about the future as facts about the future. Obviously, there is more to be considered in this respect (Rosenbaum 1984b), but not within the limited scope of the current article.

1.5 How is Matching-With-Doses Different?

In three ways, matching-with-doses differs from matching treated subjects to untreated controls. When treated subjects are matched to untreated controls, the individuals to be matched divide into two disjoint groups—treated and control—and individuals in one group are matched to individuals in the other. In contrast, when all the subjects are exposed to treatment but the doses vary, there are no longer two disjoint groups; rather there is a single group, and any individual can, in principle, be matched to any other individual. This difference affects three aspects of matching: the definition of the propensity score, the definition of distance, and the choice of optimization algorithm.

In matching treated subjects to untreated controls, the propensity score is the conditional probability of treatment given observed covariates, and matching on an estimate of the propensity score tends to balance observed covariates (Rosenbaum and Rubin 1983, 1985). A simulation study suggested that, when there are 20 covariates, matching on propensity scores is much better than other methods considered in the simulation (Gu and Rosenbaum 1993). Dehejia and Wahba (1998, 1999) reach a similar conclusion by comparing methods on an empirical example. In practice, propensity scores are unknown and must be estimated. Various consequences of estimating the propensity score are discussed by Rosenbaum (1984, 1987), Heckman, Ichimura, and Todd (1998), and Hirano, Imbens, and Ridder (2000). This definition of propensity scores is not applicable when there are doses.

When the treatment is not binary, and instead comes in doses, Joffe and Rosenbaum (1999, p. 331) showed that, under certain circumstances, a scalar balancing score exists such that matching subjects with different doses but the same balancing score tends to balance covariates. For example, if the conditional distribution of doses given covariates is correctly described by McCullagh's (1980) ordinal logit model, then the linear portion of that model is a scalar balancing score. Balancing scores with doses will be discussed and used at several points later.

When treated subjects are matched to controls, there is a distance between each treated subject and each possible choice of control describing how similar they are in terms of observed covariates. For instance, Rubin (1980) suggests using the Mahalanobis distance. In contrast, when all the subjects are exposed to treatment but at varied doses, the goal is to identify pairs that are similar in terms of observed covariates but very different in terms of dose. The distance must measure both the similarity in terms of covariates and the difference in dose. We propose new distances for matching with doses. These new distances reproduce familiar distances when there are just two doses, e.g., treated versus untreated.

In the network optimization literature, matching one group to another disjoint group is called a 'bipartite' matching, where bipartite signifies two disjoint parts; see Papadimitriou and Steiglitz (1998), Rosenbaum (1989), and Bergstralh, Kosanke, and Jacobsen (1996). Matching within a single group is called 'nonbipartite' matching, and the required algorithms are quite different.

An alternative to the propensity score for doses proposed by Joffe and Rosenbaum (1999) and used here is a useful, interesting method proposed by Imbens (2000). This paragraph discusses the relationship between our approach and that of Imbens (2000). Whereas, the approach we use entails a single scalar propensity score for all dose levels, Imbens (2000) uses a different propensity score for each dose level. The score Imbens uses for dose level z is the probability that an individual will receive dose z given observed covariates, so that if there are five dose levels, then there are five propensity scores. In contrast, under certain models, such as McCullagh's (1980) ordinal logit model, the single score we use characterizes the entire distribution of dose z given covariates. An attraction of Imbens' approach is that it does not require any model for the dose, at least conceptually, and moreover it could be used with several unordered treatments. In effect, he adjusts the marginal distribution of response separately one dose level at a time. However, in that approach, because propensity scores for different doses are different functions of covariates, individuals with different doses and the "same" propensity score are not similar in any particular respect, and it is not natural to match them (Imbens 2000, p. 709). As a result, in Imben's approach, the investigator must apply direct adjustment to propensity strata to estimate a causal effect—stratum specific results are not interpretable and do not estimate causal effects even when there is no hidden bias due to unobserved covariates (Imbens 2000, p. 709). In contrast, with a single scalar propensity score for all doses, a stratum or matched pair is defined by a single function of covariates, the same for all dose levels, so individuals in the same stratum or pair have similar values of this one function of covariates, and stratum or pair specific results do estimate causal effects when there is no hidden bias due to unobserved covariates (Joffe and Rosenbaum 1999, p. 331).

2. MATCHING WITH DOSES

2.1 Any Two Subjects Can be Matched

There are K subjects available, $k = 1, \ldots, K$, and I matched pairs of two subjects, $j = 1, 2$, will be formed, $i = 1, \ldots, I$, where $K \geq 2I$. In the example, $K = 521$ and we form $I = 260$ pairs, discarding one subject; however, when K is very large and costly, additional data will be collected from matched subjects, one might set I well below $K/2$.

Subject k has a vector $\mathbf{x}_k$ of observed covariates, received the treatment at dose Z_k, and exhibits a response, R_k. Formally, a *matched sample* is a collection $\mathcal{C}$ of I disjoint, unordered subsets of two distinct elements selected from $\{1, \ldots, K\}$, so $\mathcal{C} = \{P_1, \ldots, P_I\}$ where each P_i is a subset of two distinct elements from $\{1, \ldots, K\}$ and $P_i \cap P_{i'} = \varnothing$ for $i, i' = 1, \ldots, I$.

Associated with each possible pairing of two distinct subjects, $\{k, k'\}$ is a nonnegative, possibly infinite, distance, $d_{\{k,k'\}}$ computed from the covariates $\mathbf{x}_k$ and doses Z_k. Later sections will discuss the choice of distance. In particular, we will set $d_{\{k,k'\}} = \infty$ if $Z_k = Z_{k'}$, and this will effectively forbid matching subjects with the same dose. This is analogous to forbidding, with binary doses, the matching of controls to controls or treated subjects to treated subjects. The *total distance* associated with a matched sample $\mathcal{C} = \{P_1, \ldots, P_I\}$ is the sum of the I distances between paired subjects, $D_{\mathcal{C}} = \sum_{i=1}^{I} d_{P_i}$. A matched sample $\mathcal{C} = \{P_1, \ldots, P_I\}$ is *optimal* if it minimizes the total distance $D_{\mathcal{C}}$ among all possible pairings of $2I$ distinct subjects into I pairs. An optimal matched sample need not be unique. If $D_{\mathcal{C}} = \infty$ for an optimal matching, then the problem is called *infeasible*; otherwise, it is *feasible*.

2.2 Optimal Matching: Minimize Total Distance

Within the field of operations research, there is a large literature on optimal nonbipartite matching; see, for example, Papadimitriou and Steiglitz (1998, sec. 11.3). The alternative to optimal matching is some form of greedy algorithm, in which a best available pair is formed and removed, a best pair is formed from the remaining data and also removed, and so on. Greedy algorithms can be very poor when compared to the optimal algorithms. As an illustration, suppose we had four individuals with ages 30, 34, 36, 40. A greedy algorithm would pick the closest pair, (34, 36), remove them, and then pair (30, 40) for a total distance of $|34-36|+|30-40| = 12$. The optimal match would pick (30, 34) and (36, 40) for a total distance of $|30-34|+|36-40| = 8$. The total distance for the greedy match is 50% larger, $12/8 = 1.5$, and one pair is ten years apart whereas the optimal match has both pairs matched within five years. The same issue arises with caliper matching, say matching within five years, in which the distance is defined to be zero for subjects whose ages differ by at most five years and is defined to be infinite if the ages differ by more than five years. In the example just given, a five-year caliper match exists, and optimal matching will find it whenever it exists, but greedy matching will not find it if it starts by matching 34 to 36. In large problems, greedy matching can be extremely poor when compared to optimal matching (Snyder and Steele 1990).

We use an algorithm and Fortran code due to Derigs (1988) to find an optimal match. Derigs' algorithm is for nonbipartite matching, which is somewhat more complex than bipartite matching. He gives various computational comparisons on computers available in 1988 and finding the algorithm solved large problems reasonably quickly. His algorithm quickly solved the current matching problem with $K = 521$, so it is useful for problems of practical size.

Actually, Derigs' algorithm only works in the case of a "perfect match" with $K = 2I$. When $K > 2I$, we may use Derigs' algorithm anyway by making use of the following device. For $K \geq k \geq 1$, $K \geq k' \geq 1$, $k \neq k'$, add the same positive constant to every $d_{\{k,k'\}}$, so that all the distances are strictly positive. The constant does not affect the optimal match. Now introduce $K-2I$ sinks, or extra phantom units, $s_1, \ldots, s_{K-2I}$, and set $d_{\{k,s_i\}} = 0$, $d_{\{s_i,s_j\}} = \infty$ for $k = 1, \ldots, K$, $i, j = 1, \ldots, K-2I$, and construct an optimal perfect matching with these $K + K - 2I$ units. In a feasible optimal perfect match, sinks will not be matched to other sinks as this would cost $d_{\{s_i,s_j\}} = \infty$, and $K-2I$ subjects among $\{1, \ldots, K\}$ will be matched to sinks as these cost $d_{\{k,s_i\}} = 0$, so $2I$ subjects will be matched to each other, producing I pairs of two subjects. In the example, there were $K = 521$ subjects, $I = 260$ pairs, and $K - 2I = 1$ sink was used. A related strategy is used in a different context by Ming and Rosenbaum (2001).

2.3 Propensity Scores With Doses

In observational studies comparing a treated and a control group, propensity scores and related balancing scores were proposed by Rosenbaum and Rubin (1983) to aid in constructing matched pairs or sets or strata that balance many covariates at once. In that context, the propensity score is the probability of treatment given observed covariates. Two subjects with the same propensity score have the same distribution of covariates, that is, the multivariate covariate distribution is balanced at each value of the scalar propensity score (Rosenbaum and Rubin 1983, thm 1). Theoretical results, practical experience, and simulation results all suggest that estimated propensity scores perform slightly better than true propensity scores, because they cannot distinguish chance imbalances in covariates from systematic imbalances, and so the estimated scores tend to remove both to some degree (Rosenbaum and Rubin 1983, Rosenbaum 1987; Gu and Rosenbaum 1993).

When doses of treatment replace treated and control groups, Joffe and Rosenbaum (1999) showed that scalar balancing scores exist for certain models describing the distribution of dose, given covariates. The key issue is whether the distribution of doses given covariates depends on the covariates via a scalar function of the covariates. This happens, for example, in McCullagh's (1980) ordinal logit model, and also in a conventional Gaussian multiple linear regression model *with errors of constant variance*. However, it does not happen in a multiple linear regression model with unequal variances, in which the expectation of dose varies with one scalar function of covariates and the variance varies with another. As a practical matter, the propensity score is a device for approximately balancing observed covariates, and one can straightforwardly check whether or not a particular model for the propensity score has, in fact, approximately balanced the observed covariates.

In McCullagh's (1980) ordinal logit model, which we use here, the distribution of doses, Z_k, given observed covariates, $\mathbf{x}_k$, is modeled as:

$$\log\left\{\frac{\Pr(Z_k \geq d)}{\Pr(Z_k < d)}\right\} = \theta_d + \beta^{\mathrm{T}}\mathbf{x}_k, \quad \text{for } d = 2, 3, 4, 5.$$

Under this model, the distribution of doses given covariates depends on the observed covariates only through $e(\mathbf{x}_k) = \beta^{\mathrm{T}}\mathbf{x}_k$, with the consequence that the observed covariates $\mathbf{x}$ and the doses Z are conditionally independent given the scalar $e(\mathbf{x}_k)$, so $e(\mathbf{x}_k)$ is a balancing score. The maximum likelihood estimate, $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$, is used in the matching.

We fit the ordinal logit model using the covariates in Table 2. As is often true when many variables measure different aspects of a few attributes or behaviors, the partial associations represented by the coefficients in the model are related

but distinct from the marginal associations represented by the rank correlations in Table 2. A covariate may have a statistically insignificant coefficient in the model not because it is unrelated to dose, but because that relationship can be captured by other covariates in the model. For example, both the hours of TV on a weekday and the hours of TV on a weekend have significant marginal associations with dose in Table 2, but neither had a significant partial association with dose in the model, perhaps because either variable could be used in place of the other. For use as a balancing score, what is important is $\beta^{\mathrm{T}}\mathbf{x}_k$, not individual coefficients in β. For this reason, we did not discard covariates with statistically insignificant coefficients, nor did we search for a parsimonious model. The coefficients with statistically significant coefficients at the .05 level were "age," "sports channel," and "other activities," whereas "music channel" and "religious youth groups" were marginally significant at the .1 level.

2.4 Distances: Close on Covariates; Far Apart on Doses

In matching treated subjects to untreated controls, various distances $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) \geq 0$ have been proposed between the values of the observed covariates $\mathbf{x}_k$ and $\mathbf{x}_{k'}$ of subjects k and k'. These include the Mahalanobis distance, $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = (\mathbf{x}_k - \mathbf{x}_{k'})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_k - \mathbf{x}_{k'})$, where $\mathbf{S}$ is an estimate of the covariance matrix of the observed covariates, the propensity distance, $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \{e(\mathbf{x}_k) - e(\mathbf{x}_{k'})\}^2$, and the Mahalanobis distance within propensity score calipers, namely $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \infty$ if $\{e(\mathbf{x}_k) - e(\mathbf{x}_{k'})\}^2 \geq c$ and $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = (\mathbf{x}_k - \mathbf{x}_{k'})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{x}_k - \mathbf{x}_{k'})$ if $\{e(\mathbf{x}_k) - e(\mathbf{x}_{k'})\}^2 < c$. See Rosenbaum and Rubin (1985) for an empirical comparison of these three distances. A simulation suggests that matching on the propensity score is best when there are many, say 20, covariates (Gu and Rosenbaum 1993).

In matching with doses, the goal is not only to balance the observed covariates, but also to produce pairs with very different doses. To this end, consider the following distance:

$$\Delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \frac{\delta(\mathbf{x}_k, \mathbf{x}_{k'}) + \epsilon}{(Z_k - Z_{k'})^2}$$

where $\epsilon > 0$ is a vanishingly small but strictly positive number. The constant ϵ is a formal device, and it is not actually used in computations; rather, it signifies how perfect matches on covariates or doses will be handled. Specifically, ϵ serves two functions. First, if subjects k and k' have the same dose, $Z_k = Z_{k'}$, then the distance between them is ∞ even if they have identical observed covariates $\mathbf{x}_k = \mathbf{x}_{k'}$ and $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = 0$. Second, when two subjects have identical observed covariates and $\delta(\mathbf{x}_k, \mathbf{x}_{k'}) = 0$, the dose distance $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ will be smaller as the difference in doses $(Z_k - Z_{k'})^2$ increases. The optimal match is the same for all sufficiently small $\epsilon > 0$, so it is just a formal device.

In a sense, $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ generalizes $\delta(\mathbf{x}_k, \mathbf{x}_{k'})$ to the case of doses. Specifically, if there are just two doses, say 1 for treated and 0 for control, then $\Delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \infty$ if $Z_k = Z_{k'}$, that is if both k and k' are treated or both are control, and otherwise $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ is essentially just $\delta(\mathbf{x}_k, \mathbf{x}_{k'})$. In this case, an optimal feasible matching will never match two treated subjects together, or two controls together, because of the infinite distance, and so an optimal feasible matching with two doses using $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ will be the same as an optimal treated/control matching using $\delta(\mathbf{x}_k, \mathbf{x}_{k'})$.

Notice that if the doses Z_k were linearly transformed by adding a constant and by multiplying by a strictly positive constant, then $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ would be divided by the square of the multiplier, but the optimal match would not change. However, nonlinear transformations of the dose, such as a log or rank transformation, would change the distance, and such a transformation may be useful depending upon the initial distribution of doses.

The particular distance used here is

$$\Delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \frac{\left(\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k - \widehat{\beta}^{\mathrm{T}}\mathbf{x}_{k'}\right)^2 + \epsilon}{(Z_k - Z_{k'})^2} \tag{1}$$

so the distance is small when, based on the observed covariates $\mathbf{x}$, we would have expected k and k' to have received similar doses Z, but in fact, their actual doses are very different. The appendix discusses some theoretical motivation for the form of (1). In fact, both $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$ and Z_k were replaced by their ranks before using formula (1), with average ranks for ties.

Although the pilot data are limited, the larger actual evaluation will estimate program effects separately for teens who had previously used drugs and for those who had not. If one wanted to do this using matching techniques, then one might slightly alter the distance, $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$. Specifically, if both teen k and teen k' had not previously used drugs, or if both had previously used drugs, then $\Delta(\mathbf{x}_k, \mathbf{x}_{k'})$ would be given by (1), but if only one teen had previously used drugs then $\Delta(\mathbf{x}_k, \mathbf{x}_{k'}) = \infty$. This would force matched pairs to agree on previous drug use, so the pairs could be divided into two disjoint groups—prior users and nonusers—for some analyses. Similarly, one could require pairs to agree on the type or extent of prior drug use.

Once an optimal matching has been selected, it is convenient to order the two subjects in each pair so that the first subject received the higher dose. For the chosen feasible optimal matching, $\mathcal{C} = \{P_1, \ldots, P_I\}$, assign subscripts $(i, 1)$ and $(i, 2)$ to the two subjects in pair P_i so that $Z_{i1} > Z_{i2}$.

2.5 Sampling Potential Controls in Research Design

In the present study, the $K = 521$ subjects are divided into $I = 260$ pairs, with one subject discarded. When costly additional data are to be collected from matched subjects after matching, one might decide that the number of matched subjects, $2I$, should be much less than the reservoir size, K, so that the matching serves as both an analytical device and a sampling technique. In this case, use of a distance such as (1) will sample pairs with similar estimated distributions of doses given observed covariates, measured by $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$, but with very different realized doses, measured by Z_k. This is a reasonable research design if the goal is to build an observational study resembling an efficient experiment, in which similar subjects receive very different doses. It is less attractive as a research

design if the goal is to estimate the typical effect in a population, because high and low doses may be comparatively uncommon in the population, and oversampling high and low doses may increase the power to detect a treatment effect at the expense of decreasing the precision of estimates of typical population responses.

In many settings, including the media campaign here, the typical doses are not unalterable features of policy. If it were found that high doses of the media campaign were highly effective, but typical doses were not, then one policy option is to intensify the media campaign so that the high doses become typical. On the other hand, if all doses, even high doses, have negligible effects, then optimizing the dose level is futile. In experimental design, early studies of treatment efficacy often compare quite distinct doses, to determine whether the treatment has sufficient efficacy to merit refinement. For example, in designing clinical trials, Peto et al. (1976, p. 590) write:

> A positive result is more likely, and a null result is more informative, if the main comparison is of only 2 treatments, these being as different as possible. . . . it is the mark of good trial design that a null result, if it occurs, will be of interest.

In short, sampling using a distance such as (1) is useful in building an observational study that resembles, to some extent, an experiment in which dramatically different doses are assigned in an ostensibly haphazard manner. Obviously, randomization in an experiment balances both the observed and the unobserved covariates, whereas a distance computed from observed covariates cannot be expected to address imbalances in unobserved covariates, which must be addressed by other means in an observational study (e.g., Rosenbaum 1995, sec. 4). Sampling using (1) is not useful for describing typical individuals in a population.

3. RESULTS: COMPARABILITY AND INTENTIONS

3.1 Doses in Matched Pairs

The algorithm constructs 260 matched pairs consisting of a teen who experienced a relatively higher dose of the media campaign and a teen who experienced a relatively lower dose. Table 3 describes the dose distributions within the 260 matched pairs. Review Table 1 for the meaning of the five dose categories.

Table 3 describes the joint distribution of doses within the 260 pairs, together with the two marginal distributions for high-dose teens and low-dose teens. Each count in Table 3 is one pair of two people. For example, in 24 pairs, the high dose was 4 and the low dose was 1. In every pair, the dose difference is at least one level, in $154/260 = 59\%$ of pairs, the dose difference is at least two levels, and in $62/260 = 24\%$, the dose difference is at least three levels. The mean dose for the high-dose teens was 3.6 and for the low-dose teens was 1.7. In aggregate, the high-dose teens reported much higher exposure to the media campaign.

Table 3. Doses of Exposure to the Media Campaign: Joint Distributions (High, Low) in Matched Pairs

Dose	*1*	*2*	*3*	*4*	*5*	*Total*
1	0	0	0	0	0	0
2	37	0	0	0	0	37
3	43	49	0	0	0	92
4	24	28	14	0	0	66
5	14	24	21	6	0	65
Total	118	101	35	6	0	260

NOTE: Rows are high dose, columns are low dose.

In short, the algorithm has constructed matched pairs of teens reporting very different doses of exposure to the media campaign.

3.2 Balance Obtained by Matching

Table 4 shows the balance on the 22 covariates after matching. Table 4 gives means and percentages for the high- and low-dose subjects, averaging over the 260 pairs. For instance, at both the high and low dose, the mean age was 15.9, and 24% of the high dose-teens were black, as opposed to 25% of the low-dose teens. Overall, the high- and low-dose teens look fairly comparable. If one compares the means using the two sample t-test, then none of the 22 t-statistics is larger than one in absolute value. In a completely randomized experiment, 1 in 20 t-statistics comparing covariate balance should be larger than about 2 in absolute value. There is more balance on observed covariates than one would expect in a randomized experiment. Of course, randomization also balances unobserved covariates, whereas matching generally does not.

The comparisons in Table 4 provide a check on the fitted propensity score. In theory, the correct propensity score should balance the covariates (Joffe and Rosenbaum 1999, p. 331), so that the failure to balance the covariates indicates a failure of the model. Specifically, failure to balance the covariates indicates that there is information in the covariates that is useful in predicting the dose but which is not accurately reflected in the fitted propensity score. Of course, the propensity score was fitted with the single goal of matching to balance many covariates and so checking covariate balance is the most relevant way to appraise the fit.

By averaging the two columns in Table 4, means or percentages for the $2I = 520$ teens are obtained. For instance, the mean age for all 520 teens is $(15.9 + 15.9)/2 = 15.9$.

The matching was based on the score $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$. The mean score for the high-dose teens was .98 and for the low-dose teens was .97 with a two sample t-statistic of $t = .0088$. Evidently, the matching on $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$ was quite close. In every pair, the high-dose teen received a dose at least 1 unit greater than the low-dose teen, so the t-statistic for dose is, by construction, enormous. In other words, within pairs, the fitted distribution of doses based on $\widehat{\beta}^{\mathrm{T}}\mathbf{x}_k$ is quite balanced, but the actual doses are very different. The matching has produced groups with similar distributions of observed covariates, $\mathbf{x}$, but very different doses, Z.

3.3 Will You Use Marijuana or Inhalants Next Year?

The pilot data contain four questions about intentions for drug use over the next year, two about marijuana and hashish and two about inhalants. Might high exposures to the media campaign reduce stated intentions about future drug use? As an illustration of methodology, these intentions are compared for matched pairs. Keep in mind that the pilot data are adequate to try out methodology, but too limited in scope and

Table 4. Covariate Balance in 260 Matched Pairs of a High-Dose Teen and a Low-Dose Teen

Covariate	*High Dose*	*Low Dose*
Age	15.9	15.9
Black (1 = yes, 0 = no)	24%	25%
Other Nonwhite Race (1 = yes, 0 = no)	3%	3%
Gender (1 = female, 0 = male)	53%	52%
Mother graduated high school (0 = no, 1 = yes or missing)	80%	79%
Mother graduated college (0 = no or missing, 1 = yes)	27%	30%
Cable or satellite TV at home (1 = yes, 0 = no)	77%	77%
$\sqrt{\text{Hours of TV on a weekday}}$	2.5	2.5
$\sqrt{\text{Hours of TV on a weekend}}$	2.6	2.6
Music TV per month (e.g., MTV)		
0 = *never*, . . . , 3 = 15 to 30 days/month	2.1	2.0
Sports channel per month (e.g., ESPN)		
0 = *never*, . . . , 3 = 15 to 30 days/month	1.0	1.1
Music, dance, or theater (1 = yes, 0 = no)	72%	68%
Athletic teams or sports (1 = yes, 0 = no)	79%	77%
Clubs (e.g., Girl Scouts) (1 = yes, 0 = no)	46%	46%
Religious youth groups (1 = yes, 0 = no)	51%	48%
Other activities (1 = yes, 0 = no)	65%	65%
Ever smoked cigarettes		
(0 = never, . . . , 4 = yes in the last 30 days)	1.3	1.2
Ever used marijuana (1 = yes, 0 = no)	40%	39%
Marijuana, last 12 months		
(0 = never, . . . , 6 = 40 or more times)	.83	.78
Ever used inhalants (1 = yes, 0 = no)	10%	10%
Inhalants, last 12 months		
(0 = never, . . . , 6 = 40 or more times)	.096	0.092
Any friends use drugs (1 = yes, 0 = no)	68%	68%

NOTE: Each mean or percentage is based on 260 teens.

too unrepresentative of the nation as a whole to be a basis for evaluating the media campaign.

All the four questions use the same four point scale: (1) I definitely will not, (2) I probably will not, (3) I probably will, and (4) I definitely will. One question asks: "How likely is it that you will use marijuana, even once or twice, in the next 12 months?" A second asks: "How likely is it that you will use marijuana nearly every month for the next 12 months?" A (1) response to the first question is understood to imply a (1) response to the second. There are two parallel questions about inhalants. The same four point scale is used for all questions, but the questions themselves are very different, so the scale points mean different things for different questions. Therefore, in our analysis, scale points for one question are not compared to scale points for another question; rather, the responses of different subjects to the same question are compared.

For instance, in the fourth matched pair, the high-dose teen said "I definitely will not" to all four questions, that is, $(1, 1, 1, 1)$. The low-dose teen in this pair said "I probably will not" use marijuana once or twice, "I probably will not" use marijuana nearly every month, "I probably will" use inhalants once or twice, "I probably will not" use inhalants nearly every month, that is, $(2, 2, 3, 2)$. The four-variate matched pair difference for this pair is: $(1, 1, 1, 1) - (2, 2, 3, 2) = (-1, -1, -2, -1)$, so this one pair exhibits the pattern of intentions hoped for by the designers of the media campaign.

Table 5 describes the matched pair differences in intentions for the 251 pairs (of 260) with complete data on the intentions. For instance, a zero difference in a pair for "Marijuana use once or twice" would mean that the high- and low-dose teen in that pair gave the same response to this question. If the high dose teen said "I definitely will not" or 1 and the low dose teen said "I definitely will" or 4, then the difference would be $1 - 4 = -3$. Generally, negative values signify greater intentions to use drugs by the low-dose teen in a pair, and positive values signify the opposite. The median matched pair difference is zero for all the four questions, and the four sets of quartiles are symmetric about zero, with most equal to zero. Roughly speaking, in half the pairs, the high-dose teen expressed greater intentions to use drugs than the low-dose teen and in half the pairs, the opposite was true. There is no sign that dose is associated with stated intentions. When the

Table 5. Do Matched High- and-Low-Dose Teens Have Different Intentions About Future Drug Use?

	Marijuana 1-2	*Marijuana-E*	*Inhalants 1-2*	*Inhalants-E*
Minimum	−3	−3	−3	−3
Quartile 1	−1	0	0	0
Median	0	0	0	0
Quartile 3	1	0	0	0
Maximum	3	3	3	3
Deviate	.45	−1.06	−.18	1.00

NOTE: Quantiles of 260 matched pair differences, high-minus-low, for four questions, with deviates for the Wilcoxon-signed rank test. A negative difference in a pair indicates the high-dose teen stated lower intentions to use drugs than did the low-dose teen. One question (1-2) asks about use once or twice in the next 12 months, whereas another (every) asks about use in nearly every month in the next 12 months.

Wilcoxon signed rank test is applied, using average ranks for ties, the standardized deviates are all less than 1.1 in absolute value, so none are significant at the .05 level, and two are positive, whereas a significant negative relationship between exposure and intentions—that is, a one-sided .05 P-value—would correspond to a deviate of -1.65.

One might hope to strengthen the test, combining the results from the four questions by adding the corresponding signed rank statistics (Rosenbaum 1997). Notice that this combined test ranks separately for each question and combines the separately ranked results; it does not assume a common scale for different questions. It is possible to show that if the treatment affects each outcome in the intended direction, but the outcomes are imperfectly correlated, then the combined test may have much more power than each of the individual tests. Here, however, the standardized deviate for the combined statistic is .20, far from -1.65. These pilot data are preliminary in form, unrepresentative of the nation as a whole, and are used only for illustration of methodology. In this illustrative use of the data, there is no indication that higher doses of the media campaign were associated with lower intentions for future drug use when comparable teens are compared.

4. DISCUSSION

A single group of 521 teens receiving varied doses of exposure to a media campaign was matched to form 260 high dose—low dose pairs. The pairs balanced 22 covariates, but within the pairs the doses were quite different. The approach combined a recent proposal of Joffe and Rosenbaum (1999) for propensity scores with doses and an algorithm for optimal nonbipartite matching due to Derigs (1988) in which a single group is optimally divided into pairs.

In this study, there was little or no indication of a treatment effect. Had evidence of an effect been found, additional analyses would be interesting. For instance, in an observational or nonrandomized comparison, such as this one, there is typically the concern that an ostensible treatment effect may, in fact, reflect a hidden bias due to an unobserved covariate that was not controlled by matching. This possibility is partially clarified by sensitivity analysis, and a method of sensitivity analysis for matched data with doses is discussed by Gastwirth, Krieger, and Rosenbaum (1998). The signed rank statistic may be altered to incorporate dose information, (see Rosenbaum 1997). If there were evidence of an effect, then one might wish to model its relationship with dose, possibly with effect proportional to dose for each outcome, leading to estimates and confidence intervals for the proportionality constants. For instance, matched pair differences in responses might be regressed on matched pair differences in doses, perhaps with additional adjustments for imperfectly matched observed covariates. Alternatively, one could, of course, take a less structured approach, estimating separate effects for the five dose categories. The multivariate signed rank statistic in Section 3.3 and its associated sensitivity analysis may be used in equivalence testing; in this case, the null hypothesis asserts a substantial treatment effect and in the alternative asserts there is little or no effect, (see e.g., Li, Propert, and Rosenbaum 2001).

The matching procedure may be used in several other ways. As discussed in Section 2.5, the procedure may be used to select matched high-dose/low-dose pairs from a large reservoir when costly additional information is to be obtained from the selected individuals. Unlike most other methods of adjustment, matching permits quantitative and ethnographic methods to work in mutual support (Rosenbaum and Silber 2001). Matching may be combined with covariance adjustment of matched pair differences to improve robustness to misspecification of the covariance model, in parallel with the method of Rubin (1979).

APPENDIX: EQUAL PERCENT BIAS REDUCTION WHEN MATCHING WITH DOSES

This appendix provides some theoretical motivation for the form of the distance (1) using the concept of equal percent bias reduction. Matched sampling is often used when covariate and treatment information is available for a large reservoir of subjects, but responses will be obtained later at significant cost only for matched subjects. In this situation, response information is not available when the matching is performed, so one intends the matching to reduce bias due to observed covariates $\mathbf{x}$ no matter what relationship is later found between the responses and the covariates. Consider a simple case, in which the response R_k has a linear regression on the dose, Z_k, and the covariates, $\mathbf{x}_k$

$$E\left(R_k \middle| Z_k, \mathbf{x}_k\right) = \alpha + \beta Z_k + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_k,$$

where β describes the relationship between dose of treatment and response, and $\boldsymbol{\theta}$ is an unknown vector parameter. One would like to match on covariates $\mathbf{x}_k$ in such a way that the matching tends to reduce bias in estimating β no matter what $\boldsymbol{\theta}$ turns out to be, because there is no information on responses R_k to estimate $\boldsymbol{\theta}$ at the time the matching is performed. Obviously, exact matching on $\mathbf{x}$ would remove all of the bias due to $\mathbf{x}$ no matter what $\boldsymbol{\theta}$ is, but exact matching is often impossible when $\mathbf{x}$ is of high dimension. The ith imperfectly matched pair formed using Z and $\mathbf{x}$ yields an estimate of β, namely the slope $\left(R_{i1} - R_{i2}\right)/\left(Z_{i1} - Z_{i2}\right)$, with $Z_{i1} > Z_{i2}$, and this estimate has bias

$$E\left(\frac{R_{i1} - R_{i2}}{Z_{i1} - Z_{i2}} \middle| Z_{i1}, Z_{i2}, \mathbf{x}_{i1}, \mathbf{x}_{i2}\right) - \beta = \boldsymbol{\theta}^{\mathrm{T}}\left(\frac{\mathbf{x}_{i1} - \mathbf{x}_{i2}}{Z_{i1} - Z_{i2}}\right).$$

Rubin (1976) considered the case of treatment/control bipartite matching, so $Z_{i1} = 1$ and $Z_{i2} = 0$, and the bias is simply $\boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{x}_{i1} - \mathbf{x}_{i2}\right)$. He defined a matching method to be an *equal percent bias reducing* if the matching reduced bias in every coordinate of the vector $\mathbf{x}$ by the same percentage, say ψ, so that the percent bias reduction in $\boldsymbol{\theta}^{\mathrm{T}}\left(\mathbf{x}_{i1} - \mathbf{x}_{i2}\right)$ is also ψ *for every* $\boldsymbol{\theta}$. Moreover, he observed that if a matching method is not an equal percent bias reducing, then there exists a possible value of $\boldsymbol{\theta}$ such that the bias is actually increased. In other words, if a matching method is not an equal percent bias reducing, then we may be making the bias smaller for each coordinate of $\mathbf{x}$, and yet the bias in $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$ will increase for some $\boldsymbol{\theta}$. If $b(\mathbf{x})$ is a scalar balancing score, and if the expectation of $\mathbf{x}$ given $b(\mathbf{x})$ is linear, so that $E\{\mathbf{x} \mid b(\mathbf{x})\} = \boldsymbol{\pi} + \boldsymbol{\lambda} b(\mathbf{x})$ for some vectors $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, then Rosenbaum and Rubin (1983) showed that treated/control matching on the balancing score $b(\mathbf{x})$ is an equal percent bias reducing: making pairs more similar in $b(\mathbf{x})$ ensures they are more similar in $\boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}$ *for every* $\boldsymbol{\theta}$. A somewhat similar result holds for matching with doses, as we now demonstrate.

If $b(\mathbf{x})$ is a balancing score, so that $\Pr\{\mathbf{x} \mid Z, b(\mathbf{x})\} = \Pr\{\mathbf{x} \mid b(\mathbf{x})\}$, and the pair is imperfectly matched using just Z and $b(\mathbf{x})$, then the

expected bias is

$$
\begin{aligned}
&E\left(\frac{R_{i1}-R_{i2}}{Z_{i1}-Z_{i2}}\,\middle|\, Z_{i1}, Z_{i2}, b(\mathbf{x}_{i1}), b(\mathbf{x}_{i2})\right)-\beta \\
&\quad = E\left\{\boldsymbol{\theta}^{\mathrm{T}}\left(\frac{\mathbf{x}_{i1}-\mathbf{x}_{i2}}{Z_{i1}-Z_{i2}}\right)\,\middle|\, Z_{i1}, Z_{i2}, b(\mathbf{x}_{i1}), b(\mathbf{x}_{i2})\right\} \\
&\quad = \boldsymbol{\theta}^{\mathrm{T}}\left[\frac{E\{\mathbf{x}_{i1} \mid Z_{i1}, b(\mathbf{x}_{i1})\}-E\{\mathbf{x}_{i2} \mid Z_{i2}, b(\mathbf{x}_{i2})\}}{Z_{i1}-Z_{i2}}\right] \\
&\quad = \boldsymbol{\theta}^{\mathrm{T}}\left[\frac{E\{\mathbf{x}_{i1} \mid b(\mathbf{x}_{i1})\}-E\{\mathbf{x}_{i2} \mid b(\mathbf{x}_{i2})\}}{Z_{i1}-Z_{i2}}\right], \qquad \text{(A.1)}
\end{aligned}
$$

providing the relevant expectations exist. Now if $E\{\mathbf{x} \mid b(\mathbf{x})\} = \boldsymbol{\pi} + \boldsymbol{\lambda} b(\mathbf{x})$ where $b(\mathbf{x})$ is a scalar, then (A.1) equals:

$$
\boldsymbol{\theta}^{\mathrm{T}}\left[\frac{\{\boldsymbol{\pi}+\boldsymbol{\lambda} b(\mathbf{x}_{i1})\}-\{\boldsymbol{\pi}+\boldsymbol{\lambda} b(\mathbf{x}_{i2})\}}{Z_{i1}-Z_{i2}}\right]=\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\lambda}\right)\left[\frac{b(\mathbf{x}_{i1})-b(\mathbf{x}_{i2})}{Z_{i1}-Z_{i2}}\right]
$$

so the absolute bias is made smaller *for every* $\boldsymbol{\theta}$ by making $|\{b(\mathbf{x}_{i1}) - b(\mathbf{x}_{i2})\}/(Z_{i1} - Z_{i2})|$ smaller. This motivates the distance (1).

[Received October 2000. Revised August 2001.]

REFERENCES

Ball, S. A., Carroll, K. M., and Rounsaville, B. J. (1994), "Sensation Seeking, Substance Abuse, and Pychopathology in Treatment-Seeking and Community Cocaine Abusers," *Journal of Consulting and Clinical Pyschology*, 62, 1053–1057.

Bardo, M. T., Donohew, R. L., and Harrington, N. G. (1996), "Psychobiology of Novelty Seeking and Drug Seeking Behavior," *Behavioural Brain Research*, 77, 23–43.

Bergstralh, E. J., Kosanke, J. L., and Jacobsen, S. L. (1996), "Software for Optimal Matching in Observational Studies," *Epidemiology*, 7, 331–332. http://www.mayo.edu/hsr/sasmac.html

Cochran, W. G. (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies," *Biometrics*, 24, 205–213.

Cox, D. R. (1957), "Note on Grouping," *Journal of the American Statistical Association*, 52, 543–547.

Dehejia, R. H., and Wahba, S. (1998), "Propensity Score Matching Methods for Non-Experimental Causal Studies," Working Paper 6829, National Bureau of Economic Research, http://www.nber.org/.

Dehejia, R. H., and Wahba, S. (1999), "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association,* 94, 1053–1062.

Derigs, U. (1988), "Solving Non-Bipartite Matching Problems via Shortest Path Techniques," *Annals of Operations Research*, 13, 225–261.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998), "Dual and Simultaneous Sensitivity Analysis for Matched Pairs," *Biometrika*, 85, 907–920.

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

Heckman, J. J., Ichimura, H., and Todd, P. (1998), "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65, 261–294.

Hirano, K., Imbens, G. W., and Ridder, G. (2000), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," Working Paper T0251, National Bureau of Economic Research, http://www.nber.org/.

Imbens, G. W. (2000), "The Role of The Propensity Score in Estimating Dose-Response Functions," *Biometrika*, 87, 706–710.

Joffe, M. M., and Rosenbaum, P. R. (1999), "Propensity Scores," *American Journal of Epidemiology*, 150, 327–333.

Kosten, T. A., Ball, S. A., and Rounsaville, B. J. (1994), "A Sibling Study of Sensation Seeking and Opiate Addiction," *Journal of Nervous and Mental Disease*, 182, 284–289.

Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001) Balanced Risk set Matching, *Journal of the American Statistical Association*, 96, 870–882.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society*, Ser. B, 42, 109–142.

Ming, K., and Rosenbaum, P. R. (2001), "A Note on Optimal Matching With Variable Controls Using the Assignment Algorithm," *Journal of Computational and Graphical Statistics*, 10, 455–463.

Papadimitriou, C. H., and Steiglitz, K. (1998), *Combinatorial Optimization: Algorithms and Complexity*, New York: Dover.

Peto, R., Pike, M., Armitage, P., Breslow, N., Cox, D., Howard, S., Mantel, N., McPherson, K., Peto, J., and Smith, P. (1976), "Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patient, I: Introduction and Design," *British Journal of Cancer*, 34, 585–612.

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

———. (1987), "Model-Based Direct Adjustment," *Journal of the American Statistical Association*, 82, 387–394.

———. (1989), "Optimal matching in observational studies," *Journal of the American Statistical Association*, 84, 1024–32.

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

———. (1995), *Observational Studies,* New York: Springer-Verlag.

———. (1997), "Signed Rank Statistics for Coherent Predictions," *Biometrics*, 53, 556–566.

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.

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

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

Rosenbaum, P. R., and Silber, J. H. (2001), "Matching and Thick Description in an Observational Study of Mortality After Surgery," *Biostatistics*, 2, 217–232.

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

———. (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.

Snyder, T. L., and Steele, J. M. (1990), "Worst Case Greedy Matching in the Unit d-cube," *Networks*, 20, 779–800.