Statistical Science
2010, Vol. 25, No. 1, 1–21
DOI: 10.1214/09-STS313

Matching Methods for Causal Inference: A Review and a Look Forward

Elizabeth A. Stuart

Abstract. When estimating causal effects using observational data, it is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distributions. This goal can often be achieved by choosing well-matched samples of the original treated and control groups, thereby reducing bias due to the covariates. Since the 1970s, work on matching methods has examined how to best choose treated and control subjects for comparison. Matching methods are gaining popularity in fields such as economics, epidemiology, medicine and political science. However, until now the literature and related advice has been scattered across disciplines. Researchers who are interested in using matching methods—or developing methods related to matching—do not have a single place to turn to learn about past and current research. This paper provides a structure for thinking about matching methods and guidance on their use, coalescing the existing research (both old and new) and providing a summary of where the literature on matching methods is now and where it should be headed.

Key words and phrases: Observational study, propensity scores, subclassification, weighting.

1. INTRODUCTION

One of the key benefits of randomized experiments for estimating causal effects is that the treated and control groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved. Work on matching methods has examined how to replicate this as much as possible for observed covariates with observational (nonrandomized) data. Since early work in matching, which began in the 1940s, the methods have increased in both complexity and use. However, while the field is expanding, there has been no single source of information for researchers interested in an overview of the methods and techniques available, nor a summary of advice for applied researchers interested in implementing these methods. In contrast, the research and resources have been scattered across disciplines such as statistics (Rosenbaum, 2002; Rubin, 2006), epidemiology (Brookhart et al., 2006), sociology (Morgan and Harding, 2006), economics (Imbens, 2004) and political science (Ho et al., 2007). This paper coalesces the diverse literature on matching methods, bringing together the original work on matching methods—of which many current researchers are not aware—and tying together ideas across disciplines. In addition to providing guidance on the use of matching methods, the paper provides a view of where research on matching methods should be headed.

We define "matching" broadly to be any method that aims to equate (or "balance") the distribution of covariates in the treated and control groups. This may involve 1 : 1 matching, weighting or subclassification. The use of matching methods is in the broader context of the careful design of nonexperimental studies (Rosenbaum, 1999, 2002; Rubin, 2007). While extensive time and effort is put into the careful design of randomized experiments, relatively little effort is put into the corresponding "design" of nonexperimental studies. In fact, precisely because nonexperimental studies do not have the benefit of randomization, they require

Elizabeth A. Stuart is Assistant Professor, Departments of Mental Health and Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, USA (e-mail: estuart@jhsph.edu).

even more careful design. In this spirit of design, we can think of any study aiming to estimate the effect of some intervention as having two key stages: (1) design, and (2) outcome analysis. Stage (1) uses only background information on the individuals in the study, designing the nonexperimental study as would be a randomized experiment, without access to the outcome values. Matching methods are a key tool for stage (1). Only after stage (1) is finished does stage (2) begin, comparing the outcomes of the treated and control individuals. While matching is generally used to estimate causal effects, it is also sometimes used for noncausal questions, for example, to investigate racial disparities (Schneider, Zaslavsky and Epstein, 2004).

Alternatives to matching methods include adjusting for background variables in a regression model, instrumental variables, structural equation modeling or selection models. Matching methods have a few key advantages over those other approaches. First, matching methods should not be seen in conflict with regression adjustment and, in fact, the two methods are complementary and best used in combination. Second, matching methods highlight areas of the covariate distribution where there is not sufficient overlap between the treatment and control groups, such that the resulting treatment effect estimates would rely heavily on extrapolation. Selection models and regression models have been shown to perform poorly in situations where there is insufficient overlap, but their standard diagnostics do not involve checking this overlap (Dehejia and Wahba, 1999, 2002; Glazerman, Levy and Myers, 2003). Matching methods in part serve to make researchers aware of the quality of resulting inferences. Third, matching methods have straightforward diagnostics by which their performance can be assessed.

The paper proceeds as follows. The remainder of Section 1 provides an introduction to matching methods and the scenarios considered, including some of the history and theory underlying matching methods. Sections 2–5 provide details on each of the steps involved in implementing matching: defining a distance measure, doing the matching, diagnosing the matching, and then estimating the treatment effect after matching. The paper concludes with suggestions for future research and practical guidance in Section 6.

1.1 Two Settings

Matching methods are commonly used in two types of settings. The first is one in which the outcome values are not yet available and matching is used to select subjects for follow-up (e.g., Reinisch et al., 1995; Stuart and Ialongo, 2009). It is particularly relevant for studies with cost considerations that prohibit the collection of outcome data for the full control group. This was the setting for most of the original work in matching methods, particularly the theoretical developments, which compared the benefits of selecting matched versus random samples of the control group (Althauser and Rubin, 1970; Rubin, 1973a, 1973b). The second setting is one in which all of the outcome data is already available, and the goal of the matching is to reduce bias in the estimation of the treatment effect.

A common feature of matching methods, which is automatic in the first setting but not the second, is that the outcome values are not used in the matching process. Even if the outcome values are available at the time of the matching, the outcome values should not be used in the matching process. This precludes the selection of a matched sample that leads to a desired result, or even the appearance of doing so (Rubin, 2007). The matching can thus be done multiple times and the matched samples with the best balance—the most similar treated and control groups—are chosen as the final matched samples; this is similar to the design of a randomized experiment where a particular randomization may be rejected if it yields poor covariate balance (Hill, Rubin and Thomas, 1999; Greevy et al., 2004).

This paper focuses on settings with a treatment defined at some particular point in time, covariates measured at (or relevant to) some period of time before the treatment, and outcomes measured after the treatment. It does not consider more complex longitudinal settings where individuals may go in and out of the treatment group, or where treatment assignment date is undefined for the control group. Methods such as marginal structural models (Robins, Hernan and Brumback, 2000) or balanced risk set matching (Li, Propert and Rosenbaum, 2001) are useful in those settings.

1.2 Notation and Background: Estimating Causal Effects

As first formalized in Rubin (1974), the estimation of causal effects, whether from a randomized experiment or a nonexperimental study, is inherently a comparison of potential outcomes. In particular, the causal effect for individual i is the comparison of individual i's outcome if individual i receives the treatment (the potential outcome under treatment), $Y_i(1)$, and individual i's outcome if individual i receives the control (the potential outcome under control), $Y_i(0)$. For simplicity, we use the term "individual" to refer to the units that

receive the treatment of interest, but the formulation would stay the same if the units were schools or communities. The "fundamental problem of causal inference" (Holland, 1986) is that, for each individual, we can observe only one of these potential outcomes, because each unit (each individual at a particular point in time) will receive either treatment or control, not both. The estimation of causal effects can thus be thought of as a missing data problem (Rubin, 1976a), where we are interested in predicting the unobserved potential outcomes.

For efficient causal inference and good estimation of the unobserved potential outcomes, we would like to compare treated and control groups that are as similar as possible. If the groups are very different, the prediction of $Y(1)$ for the control group will be made using information from individuals who look very different from themselves, and likewise for the prediction of $Y(0)$ for the treated group. A number of authors, including Cochran and Rubin (1973), Rubin (1973a, 1973b), Rubin (1979), Heckman, Ichimura and Todd (1998), Rubin and Thomas (2000) and Rubin (2001), have shown that methods such as linear regression adjustment can actually increase bias in the estimated treatment effect when the true relationship between the covariate and outcome is even moderately nonlinear, especially when there are large differences in the means and variances of the covariates in the treated and control groups.

Randomized experiments use a known randomized assignment mechanism to ensure "balance" of the covariates between the treated and control groups: The groups will be only randomly different from one another on all covariates, observed and unobserved. In nonexperimental studies, we must posit an assignment mechanism, which determines which individuals receive treatment and which receive control. A key assumption in nonexperimental studies is that of a strongly ignorable treatment assignment (Rosenbaum and Rubin, 1983b) which implies that (1) treatment assignment (T) is independent of the potential outcomes ($Y(0), Y(1)$) given the covariates (X): $T \perp (Y(0), Y(1))|X$, and (2) there is a positive probability of receiving each treatment for all values of X: $0 < P(T = 1|X) < 1$ for all X. The first component of the definition of strong ignorability is sometimes termed "ignorable," "no hidden bias" or "unconfounded." Weaker versions of the ignorability assumption are sufficient for some quantities of interest, as discussed further in Imbens (2004). This assumption is often more reasonable than it may sound at first since matching on or controlling for the observed covariates also matches on or controls for the unobserved covariates, in so much as they are correlated with those that are observed. Thus, the only unobserved covariates of concern are those unrelated to the observed covariates. Analyses can be done to assess sensitivity of the results to the existence of an unobserved confounder related to both treatment assignment and the outcome (see Section 6.1.2). Heller, Rosenbaum and Small (2009) also discuss how matching can make effect estimates less sensitive to an unobserved confounder, using a concept called "design sensitivity." An additional assumption is the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980), which states that the outcomes of one individual are not affected by treatment assignment of any other individuals. While not always plausible—for example, in school settings where treatment and control children may interact, leading to "spillover" effects—the plausibility of SUTVA can often be improved by design, such as by reducing interactions between the treated and control groups. Recent work has also begun thinking about how to relax this assumption in analyses (Hong and Raudenbush, 2006; Sobel, 2006; Hudgens and Halloran, 2008).

To formalize, using notation similar to that in Rubin (1976b), we consider two populations, P_t and P_c, where the subscript t refers to a group exposed to the treatment and c refers to a group exposed to the control. Covariate data on p pre-treatment covariates is available on random samples of sizes N_t and N_c from P_t and P_c. The means and variance covariance matrix of the p covariates in group i are given by μ_i and Σ_i, respectively ($i = t, c$). For individual j, the p covariates are denoted by X_j, treatment assignment by T_j ($T_j = 0$ or 1), and the observed outcome by Y_j. Without loss of generality, we assume $N_t < N_c$.

To define the treatment effect, let $E(Y(1)|X) = R_1(X)$ and $E(Y(0)|X) = R_0(X)$. In the matching context effects are usually defined as the difference in potential outcomes, $\tau(x) = R_1(x) - R_0(x)$, although other quantities, such as odds ratios, are also sometimes of interest. It is often assumed that the response surfaces, $R_0(x)$ and $R_1(x)$, are parallel, so that $\tau(x) = \tau$ for all x. If the response surfaces are not parallel (i.e., the effect varies), an average effect over some population is generally estimated. Variation in effects is particularly relevant when the estimands of interest are not difference in means, but rather odds ratios or relative risks, for which the conditional and marginal effects are not necessarily equal (Austin, 2007; Lunt

et al., 2009). The most common estimands in nonexperimental studies are the "average effect of the treatment on the treated" (ATT), which is the effect for those in the treatment group, and the "average treatment effect" (ATE), which is the effect on all individuals (treatment and control). See Imbens (2004), Kurth et al. (2006) and Imai, King and Stuart (2008) for further discussion of these distinctions. The choice between these estimands will likely involve both substantive reasons and data availability, as further discussed in Section 6.2.

1.3 History and Theoretical Development of Matching Methods

Matching methods have been in use since the first half of the 20th Century (e.g., Greenwood, 1945; Chapin, 1947), however, a theoretical basis for these methods was not developed until the 1970s. This development began with papers by Cochran and Rubin (1973) and Rubin (1973a, 1973b) for situations with one covariate and an implicit focus on estimating the ATT. Althauser and Rubin (1970) provide an early and excellent discussion of some practical issues associated with matching: how large the control "reservoir" should be to get good matches, how to define the quality of matches, how to define a "close-enough" match. Many of the issues identified in that work are topics of continuing debate and discussion. The early papers showed that when estimating the ATT, better matching scenarios include situations with many more control than treated individuals, small initial bias between the groups, and smaller variance in the treatment group than the control group.

Dealing with multiple covariates was a challenge due to both computational and data problems. With more than just a few covariates, it becomes very difficult to find matches with close or exact values of all covariates. For example, Chapin (1947) finds that with initial pools of 671 treated and 523 controls there are only 23 pairs that match exactly on six categorical covariates. An important advance was made in 1983 with the introduction of the propensity score, defined as the probability of receiving the treatment given the observed covariates (Rosenbaum and Rubin, 1983b). The propensity score facilitates the construction of matched sets with similar distributions of the covariates, without requiring close or exact matches on all of the individual variables.

In a series of papers in the 1990s, Rubin and Thomas (1992a, 1992b, 1996) provided a theoretical basis for multivariate settings with affinely invariant matching methods and ellipsoidally symmetric covariate distributions (such as the normal or t-distribution), again focusing on estimating the ATT. Affinely invariant matching methods, such as propensity score or Mahalanobis metric matching, are those that yield the same matches following an affine (linear) transformation of the data. Matching in this general setting is shown to be Equal Percent Bias Reducing (EPBR; Rubin, 1976b). Rubin and Stuart (2006) later showed that the EPBR feature also holds under much more general settings, in which the covariate distributions are discriminant mixtures of ellipsoidally symmetric distributions. EPBR methods reduce bias in all covariate directions (i.e., makes the covariate means closer) by the same amount, ensuring that if close matches are obtained in some direction (such as the propensity score), then the matching is also reducing bias in all other directions. The matching thus cannot be increasing bias in an outcome that is a linear combination of the covariates. In addition, matching yields the same percent bias reduction in bias for any linear function of X if and only if the matching is EPBR.

Rubin and Thomas (1992b) and Rubin and Thomas (1996) obtain analytic approximations for the reduction in bias on an arbitrary linear combination of the covariates (e.g., the outcome) that can be obtained when matching on the true or estimated discriminant (or propensity score) with normally distributed covariates. In fact, the approximations hold remarkably well even when the distributional assumptions are not satisfied (Rubin and Thomas, 1996). The approximations in Rubin and Thomas (1996) can be used to determine in advance the bias reduction that will be possible from matching, based on the covariate distributions in the treated and control groups, the size of the initial difference in the covariates between the groups, the original sample sizes, the number of matches desired and the correlation between the covariates and the outcome. Unfortunately these approximations are rarely used in practice, despite their ability to help researchers quickly assess whether their data will be useful for estimating the causal effect of interest.

1.4 Steps in Implementing Matching Methods

Matching methods have four key steps, with the first three representing the "design" and the fourth the "analysis":

1. Defining "closeness": the distance measure used to determine whether an individual is a good match for another.

2. Implementing a matching method, given that measure of closeness.
3. Assessing the quality of the resulting matched samples, and perhaps iterating with steps 1 and 2 until well-matched samples result.
4. Analysis of the outcome and estimation of the treatment effect, given the matching done in step 3.

The next four sections go through these steps one at a time, providing an overview of approaches and advice on the most appropriate methods.

2. DEFINING CLOSENESS

There are two main aspects to determining the measure of distance (or "closeness") to use in matching. The first involves which covariates to include, and the second involves combining those covariates into one distance measure.

2.1 Variables to Include

The key concept in determining which covariates to include in the matching process is that of strong ignorability. As discussed above, matching methods, and in fact most nonexperimental study methods, rely on ignorability, which assumes that there are no unobserved differences between the treatment and control groups, conditional on the observed covariates. To satisfy the assumption of ignorable treatment assignment, it is important to include in the matching procedure all variables known to be related to both treatment assignment and the outcome (Rubin and Thomas, 1996; Heckman, Ichimura and Todd, 1998; Glazerman, Levy and Myers, 2003; Hill, Reiter and Zanutto, 2004). Generally poor performance is found of methods that use a relatively small set of "predictors of convenience," such as demographics only (Shadish, Clark and Steiner, 2008). When matching using propensity scores, detailed below, there is little cost to including variables that are actually unassociated with treatment assignment, as they will be of little influence in the propensity score model. Including variables that are actually unassociated with the outcome can yield slight increases in variance. However, excluding a potentially important confounder can be very costly in terms of increased bias. Researchers should thus be liberal in terms of including variables that may be associated with treatment assignment and/or the outcomes. Some examples of matching have 50 or even 100 covariates included in the procedure (e.g., Rubin, 2001). However, in small samples it may not be possible to include a very large set of variables. In that case priority should be given to variables believed to be related to the outcome, as there is a higher cost in terms of increased variance of including variables unrelated to the outcome but highly related to treatment assignment (Brookhart et al., 2006). Another effective strategy is to include a small set of covariates known to be related to the outcomes of interest, do the matching, and then check the balance on all of the available covariates, including any additional variables that remain particularly unbalanced after the matching. To avoid allegations of variable selection based on estimated effects, it is best if the variable selection process is done without using the observed outcomes, and instead is based on previous research and scientific understanding (Rubin, 2001).

One type of variable that should not be included in the matching process is any variable that may have been affected by the treatment of interest (Rosenbaum, 1984; Frangakis and Rubin, 2002; Greenland, 2003). This is especially important when the covariates, treatment indicator and outcomes are all collected at the same point in time. If it is deemed to be critical to control for a variable potentially affected by treatment assignment, it is better to exclude that variable in the matching procedure and include it in the analysis model for the outcome (as in Reinisch et al., 1995).[1]

Another challenge that potentially arises is when variables are fully (or nearly fully) predictive of treatment assignment. Excluding such a variable should be done only with great care, with the belief that the problematic variable is completely unassociated with the outcomes of interest and that the ignorability assumption will still hold. More commonly, such a variable indicates a fundamental problem in estimating the effect of interest, whereby it may not be possible to separate out the effect of the treatment of interest from this problematic variable using the data at hand. For example, if all adolescent heavy drug users are also heavy drinkers, it will be impossible to separate out the effect of heavy drug use from the effect of heavy drinking.

2.2 Distance Measures

The next step is to define the "distance": a measure of the similarity between two individuals. There

[1]The method is misstated in the footnote in Table 1 of that paper. In fact, the potential confounding variables were not used in the matching procedure, but were utilized in the outcome analysis (D. B. Rubin, personal communication).

are four primary ways to define the distance D_{ij} between individuals i and j for matching, all of which are affinely invariant:

1. Exact:
$$D_{ij} = \begin{cases} 0, & \text{if } X_i = X_j, \\ \infty, & \text{if } X_i \neq X_j. \end{cases}$$
2. Mahalanobis:
$$D_{ij} = (X_i - X_j)'\Sigma^{-1}(X_i - X_j).$$
If interest is in the ATT, Σ is the variance covariance matrix of X in the full control group; if interest is in the ATE, then Σ is the variance covariance matrix of X in the pooled treatment and full control groups. If X contains categorical variables, they should be converted to a series of binary indicators, although the distance works best with continuous variables.
3. Propensity score:
$$D_{ij} = |e_i - e_j|,$$
where e_k is the propensity score for individual k, defined in detail below.
4. Linear propensity score:
$$D_{ij} = |\operatorname{logit}(e_i) - \operatorname{logit}(e_j)|.$$
Rosenbaum and Rubin (1985b), Rubin and Thomas (1996) and Rubin (2001) have found that matching on the linear propensity score can be particularly effective in terms of reducing bias.

Below we use "propensity score" to refer to either the propensity score itself or the linear version.

Although exact matching is in many ways the ideal (Imai, King and Stuart, 2008), the primary difficulty with the exact and Mahalanobis distance measures is that neither works very well when X is high dimensional. Requiring exact matches often leads to many individuals not being matched, which can result in larger bias than if the matches are inexact but more individuals remain in the analysis (Rosenbaum and Rubin, 1985b). A recent advance, coarsened exact matching (CEM), can be used to do exact matching on broader ranges of the variables; for example, using income categories rather than a continuous measure (Iacus, King and Porro, 2009). The Mahalanobis distance can work quite well when there are relatively few covariates (fewer than 8; Rubin, 1979; Zhao, 2004), but it does not perform as well when the covariates are not normally distributed or there are many covariates (Gu and Rosenbaum, 1993). This is likely because Mahalanobis metric matching essentially regards all interactions among the elements of X as equally important; with more covariates, Mahalanobis matching thus tries to match more and more of these multi-way interactions.

A major advance was made in 1983 with the introduction of propensity scores (Rosenbaum and Rubin, 1983b). Propensity scores summarize all of the covariates into one scalar: the probability of being treated. The propensity score for individual i is defined as the probability of receiving the treatment given the observed covariates: $e_i(X_i) = P(T_i = 1|X_i)$. There are two key properties of propensity scores. The first is that propensity scores are balancing scores: At each value of the propensity score, the distribution of the covariates X defining the propensity score is the same in the treated and control groups. Thus, grouping individuals with similar propensity scores replicates a mini-randomized experiment, at least with respect to the observed covariates. Second, if treatment assignment is ignorable given the covariates, then treatment assignment is also ignorable given the propensity score. This justifies matching based on the propensity score rather than on the full multivariate set of covariates. Thus, when treatment assignment is ignorable, the difference in means in the outcome between treated and control individuals with a particular propensity score value is an unbiased estimate of the treatment effect at that propensity score value. While most of the propensity score results are in the context of finite samples and the settings considered by Rubin and Thomas (1992a, 1996), Abadie and Imbens (2009a) discuss the asymptotic properties of propensity score matching.

The distance measures described above can also be combined, for example, doing exact matching on key covariates such as race or gender followed by propensity score matching within those groups. When exact matching on even a few variables is not possible because of sample size limitations, methods that yield "fine balance" (e.g., the same proportion of African American males in the matched treated and control groups) may be a good alternative (Rosenbaum, Ross and Silber, 2007). If the key covariates of interest are continuous, Mahalanobis matching within propensity score calipers (Rubin and Thomas, 2000) defines the distance between individuals i and j as
$$D_{ij} = \begin{cases} (Z_i - Z_j)'\Sigma^{-1}(Z_i - Z_j), \\ \qquad \text{if } |\operatorname{logit}(e_i) - \operatorname{logit}(e_j)| \leq c, \\ \infty, \quad \text{if } |\operatorname{logit}(e_i) - \operatorname{logit}(e_j)| > c, \end{cases}$$
where c is the caliper, Z is the set of "key covariates," and Σ is the variance covariance matrix of Z. This will yield matches that are relatively well matched

on the propensity score and particularly well matched on Z. Z often consists of pre-treatment measures of the outcome, such as baseline test scores in educational evaluations. Rosenbaum and Rubin (1985b) discuss the choice of caliper size, generalizing results from Table 2.3.1 of Cochran and Rubin (1973). When the variance of the linear propensity score in the treatment group is twice as large as that in the control group, a caliper of 0.2 standard deviations removes 98% of the bias in a normally distributed covariate. If the variance in the treatment group is much larger than that in the control group, smaller calipers are necessary. Rosenbaum and Rubin (1985b) generally suggest a caliper of 0.25 standard deviations of the linear propensity score.

A more recently developed distance measure is the "prognosis score" (Hansen, 2008). Prognosis scores are essentially the predicted outcome each individual would have under the control condition. The benefit of prognosis scores is that they take into account the relationship between the covariates and the outcome; the drawback is that it requires a model for that relationship. Since it thus does not have the clear separation of the design and analysis stages that we advocate here, we focus instead on other approaches, but it is a potentially important advance in the matching literature.

2.2.1 *Propensity score estimation and model specification.* In practice, the true propensity scores are rarely known outside of randomized experiments and thus must be estimated. Any model relating a binary variable to a set of predictors can be used. The most common for propensity score estimation is logistic regression, although nonparametric methods such as boosted CART and generalized boosted models (gbm) often show very good performance (McCaffrey, Ridgeway and Morral, 2004; Setoguchi et al., 2008; Lee, Lessler and Stuart, 2009).

The model diagnostics when estimating propensity scores are not the standard model diagnostics for logistic regression or CART. With propensity score estimation, concern is not with the parameter estimates of the model, but rather with the resulting balance of the covariates (Augurzky and Schmidt, 2001). Because of this, standard concerns about collinearity do not apply. Similarly, since they do not use covariate balance as a criterion, model fit statistics identifying classification ability (such as the c-statistic) or stepwise selection models are not helpful for variable selection (Rubin, 2004; Brookhart et al., 2006; Setoguchi et al., 2008). One strategy that is helpful is to examine the balance of covariates (including those not originally included in the propensity score model), their squares and interactions in the matched samples. If imbalance is found on particular variables or functions of variables, those terms can be included in a re-estimated propensity score model, which should improve their balance in the subsequent matched samples (Rosenbaum and Rubin, 1984; Dehejia and Wahba, 2002).

Research indicates that misestimation of the propensity score (e.g., excluding a squared term that is in the true model) is not a large problem, and that treatment effect estimates are more biased when the outcome model is misspecified than when the propensity score model is misspecified (Drake, 1993; Dehejia and Wahba, 1999, 2002; Zhao, 2004). This may in part be because the propensity score is used only as a tool to get covariate balance—the accuracy of the model is less important as long as balance is obtained. Thus, the exclusion of a squared term, for example, may have less severe consequences for a propensity score model than it does for the outcome model, where interest is in interpreting a particular regression coefficient (that on the treatment indicator). However, these evaluations are fairly limited; for example, Drake (1993) considers only two covariates. Future research should involve more systematic evaluations of propensity score estimation, perhaps through more sophisticated simulations as well as analytic work, and consideration should include how the propensity scores will be used, for example, in weighting versus subclassification.

3. MATCHING METHODS

Once a distance measure has been selected, the next step is to use that distance in doing the matching. In this section we provide an overview of the spectrum of matching methods available. The methods primarily vary in terms of the number of individuals that remain after matching and in the relative weights that different individuals receive. One way in which propensity scores are commonly used is as a predictor in the outcome model, where the set of individual covariates is replaced by the propensity score and the outcome models run in the full treated and control groups (Weitzen et al., 2004). Unfortunately the simple use of this method is not an optimal use of propensity scores, as it does not take advantage of the balancing property of propensity scores: If there is imbalance on the original covariates, there will also be imbalance on the propensity score, resulting in the same degree of model

extrapolation as with the full set of covariates. However, if the model regressing the outcome on the treatment indicator and the propensity score is correctly specified or if it includes nonlinear functions of the propensity score (such as quantiles or splines) and their interaction with the treatment indicator, then this can be an effective approach, with links to subclassification (Schafer and Kang, 2008). Since this method does not have the clear "design" aspect of matching, we do not discuss it further.

3.1 Nearest Neighbor Matching

One of the most common, and easiest to implement and understand, methods is $k:1$ nearest neighbor matching (Rubin, 1973a). This is generally the most effective method for settings where the goal is to select individuals for follow-up. Nearest neighbor matching nearly always estimates the ATT, as it matches control individuals to the treated group and discards controls who are not selected as matches.

In its simplest form, $1:1$ nearest neighbor matching selects for each treated individual i the control individual with the smallest distance from individual i. A common complaint regarding $1:1$ matching is that it can discard a large number of observations and thus would apparently lead to reduced power. However, the reduction in power is often minimal, for two main reasons. First, in a two-sample comparison of means, the precision is largely driven by the smaller group size (Cohen, 1988). So if the treatment group stays the same size, and only the control group decreases in size, the overall power may not actually be reduced very much (Ho et al., 2007). Second, the power increases when the groups are more similar because of the reduced extrapolation and higher precision that is obtained when comparing groups that are similar versus groups that are quite different (Snedecor and Cochran, 1980). This is also what yields the increased power of using matched pairs in randomized experiments (Wacholder and Weinberg, 1982). Smith (1997) provides an illustration where estimates from $1:1$ matching have lower standard deviations than estimates from a linear regression, even though thousands of observations were discarded in the matching. An additional concern is that, without any restrictions, $k:1$ matching can lead to some poor matches, if, for example, there are no control individuals with propensity scores similar to a given treated individual. One strategy to avoid poor matches is to impose a caliper and only select a match if it is within the caliper. This can lead to difficulties in interpreting effects if many treated individuals do not receive a match, but can help avoid poor matches. Rosenbaum and Rubin (1985a) discuss those trade-offs.

3.1.1 *Optimal matching.* One complication of simple ("greedy") nearest neighbor matching is that the order in which the treated subjects are matched may change the quality of the matches. Optimal matching avoids this issue by taking into account the overall set of matches when choosing individual matches, minimizing a global distance measure (Rosenbaum, 2002). Generally, greedy matching performs poorly when there is intense competition for controls, and performs well when there is little competition (Gu and Rosenbaum, 1993). Gu and Rosenbaum (1993) find that optimal matching does not in general perform any better than greedy matching in terms of creating groups with good balance, but does do better at reducing the distance within pairs (page 413): "...optimal matching picks about the same controls [as greedy matching] but does a better job of assigning them to treated units." Thus, if the goal is simply to find well-matched groups, greedy matching may be sufficient. However, if the goal is well-matched pairs, then optimal matching may be preferable.

3.1.2 *Selecting the number of matches: Ratio matching.* When there are large numbers of control individuals, it is sometimes possible to get multiple good matches for each treated individual, called ratio matching (Smith, 1997; Rubin and Thomas, 2000). Selecting the number of matches involves a bias:variance trade-off. Selecting multiple controls for each treated individual will generally increase bias since the 2nd, 3rd and 4th closest matches are, by definition, further away from the treated individual than is the 1st closest match. On the other hand, utilizing multiple matches can decrease variance due to the larger matched sample size. Approximations in Rubin and Thomas (1996) can help determine the best ratio. In settings where the outcome data has yet to be collected and there are cost constraints, researchers must also balance cost considerations. More methodological work needs to be done to more formally quantify the trade-offs involved. In addition, $k:1$ matching is not optimal since it does not account for the fact that some treated individuals may have many close matches while others have very few. A more advanced form of ratio matching, variable ratio matching, allows the ratio to vary, with different treated individuals receiving differing numbers of matches (Ming and Rosenbaum, 2001). Variable ratio matching is related to full matching, described below.

3.1.3 *With or without replacement.* Another key issue is whether controls can be used as matches for more than one treated individual: whether the matching should be done "with replacement" or "without replacement." Matching with replacement can often decrease bias because controls that look similar to many treated individuals can be used multiple times. This is particularly helpful in settings where there are few control individuals comparable to the treated individuals (e.g., Dehejia and Wahba, 1999). Additionally, when matching with replacement, the order in which the treated individuals are matched does not matter. However, inference becomes more complex when matching with replacement, because the matched controls are no longer independent—some are in the matched sample more than once and this needs to be accounted for in the outcome analysis, for example, by using frequency weights. When matching with replacement, it is also possible that the treatment effect estimate will be based on just a small number of controls; the number of times each control is matched should be monitored.

3.2 Subclassification, Full Matching and Weighting

For settings where the outcome data is already available, one apparent drawback of $k:1$ nearest neighbor matching is that it does not necessarily use all the data, in that some control individuals, even some of those with propensity scores in the range of the treatment groups' scores, are discarded and not used in the analysis. Weighting, full matching and subclassification methods instead use all individuals. These methods can be thought of as giving all individuals (either implicit or explicit) weights between 0 and 1, in contrast with nearest neighbor matching, in which individuals essentially receive a weight of either 0 or 1 (depending on whether or not they are selected as a match). The three methods discussed here represent a continuum in terms of the number of groupings formed, with weighting as the limit of subclassification as the number of observations and subclasses go to infinity (Rubin, 2001) and full matching in between.

3.2.1 *Subclassification.* Subclassification forms groups of individuals who are similar, for example, as defined by quintiles of the propensity score distribution. It can estimate either the ATE or the ATT, as discussed further in Section 5. One of the first uses of subclassification was Cochran (1968), which examined subclassification on a single covariate (age) in investigating the link between lung cancer and smoking. Cochran (1968) provides analytic expressions for the bias reduction possible using subclassification on a univariate continuous covariate; using just five subclasses removes at least 90% of the initial bias due to that covariate. Rosenbaum and Rubin (1985b) extended that to show that creating five propensity score subclasses removes at least 90% of the bias in the estimated treatment effect due to all of the covariates that went into the propensity score. Based on those results, the current convention is to use 5–10 subclasses. However, with larger sample sizes more subclasses (e.g., 10–20) may be feasible and appropriate (Lunceford and Davidian, 2004). More work needs to be done to help determine the optimal number of subclasses: enough to get adequate bias reduction but not too many that the within-subclass effect estimates become unstable.

3.2.2 *Full matching.* A more sophisticated form of subclassification, full matching, selects the number of subclasses automatically (Rosenbaum, 1991; Hansen, 2004; Stuart and Green, 2008). Full matching creates a series of matched sets, where each matched set contains at least one treated individual and at least one control individual (and each matched set may have many from either group). Like subclassification, full matching can estimate either the ATE or the ATT. Full matching is optimal in terms of minimizing the average of the distances between each treated individual and each control individual within each matched set. Hansen (2004) demonstrates the method in the context of estimating the effect of SAT coaching. In that example the original treated and control groups had propensity score differences of 1.1 standard deviations, but the matched sets from full matching differed by only 0.01 to 0.02 standard deviations. Full matching may thus have appeal for researchers who are reluctant to discard some of the control individuals but who want to obtain optimal balance on the propensity score. To achieve efficiency gains, Hansen (2004) also introduces restricted ratios of the number of treated individuals to the number of control individuals in each matched set.

3.2.3 *Weighting adjustments.* Propensity scores can also be used directly as inverse weights in estimates of the ATE, known as inverse probability of treatment weighting (IPTW; Czajka et al., 1992; Robins, Hernan and Brumback, 2000; Lunceford and Davidian, 2004). Formally, the weight $w_i = \frac{T_i}{\hat{e}_i} + \frac{1-T_i}{1-\hat{e}_i}$, where $\hat{e}_k$ is the estimated propensity score for individual k. This weighting serves to weight both the treated and control groups up to the full sample, in the same way that

survey sampling weights weight a sample up to a population (Horvitz and Thompson, 1952).

An alternative weighting technique, weighting by the odds, can be used to estimate the ATT (Hirano, Imbens and Ridder, 2003). Formally, $w_i = T_i + (1 - T_i)\frac{\hat{e}_i}{1-\hat{e}_i}$. With this weight, treated individuals receive a weight of 1. Control individuals are weighted up to the full sample using the $\frac{1}{1-\hat{e}_i}$ term, and then weighted to the treated group using the $\hat{e}_i$ term. In this way both groups are weighted to represent the treatment group.

A third weighting technique, used primarily in economics, is kernel weighting, which averages over multiple individuals in the control group for each treated individual, with weights defined by their distance (Imbens, 2000). Heckman, Hidehiko and Todd (1997), Heckman et al. (1998) and Heckman, Ichimura and Todd (1998) describe a local linear matching estimator that requires specifying a bandwidth parameter. Generally, larger bandwidths increase bias but reduce variance by putting weight on individuals that are further away from the treated individual of interest. A complication with these methods is this need to define a bandwidth or smoothing parameter, which does not generally have an intuitive meaning; Imbens (2004) provides some guidance on that choice.

A potential drawback of the weighting approaches is that, as with Horvitz–Thompson estimation, the variance can be very large if the weights are extreme (i.e., if the estimated propensity scores are close to 0 or 1). If the model is correctly specified and thus the weights are correct, then the large variance is appropriate. However, a worry is that some of the extreme weights may be related more to the estimation procedure than to the true underlying probabilities. Weight trimming, which sets weights above some maximum to that maximum, has been proposed as one solution to this problem (Potter, 1993; Scharfstein, Rotnitzky and Robins, 1999). However, there is relatively little guidance regarding the trimming level. Because of this sensitivity to the size of the weights and potential model misspecification, more attention should be paid to the accuracy of propensity score estimates when the propensity scores will be used for weighting vs. matching (Kang and Schafer, 2007). Another effective strategy is doubly-robust methods (Bang and Robins, 2005), which yield accurate effect estimates if either the propensity score model or the outcome model are correctly specified, as discussed further in Section 5.

3.3 Assessing Common Support

One issue that comes up for all matching methods is that of "common support." To this point, we have assumed that there is substantial overlap of the propensity score distributions in the two groups, but potentially density differences. However, in some situations there may not be complete overlap in the distributions. For example, many of the control individuals may be very different from all of the treatment group members, making them inappropriate as points of comparison when estimating the ATT (Austin and Mamdani, 2006). Nearest neighbor matching with calipers automatically only uses individuals in (or close to) the area of common support. In contrast, the subclassification and weighting methods generally use all individuals, regardless of the overlap of the distributions. When using those methods it may be beneficial to explicitly restrict the analysis to those individuals in the region of common support (as in Heckman, Hidehiko and Todd, 1997; Dehejia and Wahba, 1999).

Most analyses define common support using the propensity score, discarding individuals with propensity score values outside the range of the other group. A second method involves examining the "convex hull" of the covariates, identifying the multidimensional space that allows interpolation rather than extrapolation (King and Zeng, 2006). While these procedures can help identify who needs to be discarded, when many subjects are discarded it can help the interpretation of results if it is possible to define the discard rule using one or two covariates rather than the propensity score itself.

It is also important to consider the implications of common support for the estimand of interest. Examining the common support may indicate that it is not possible to reliably estimate the ATE. This could happen, for example, if there are controls outside the range of the treated individuals and thus no way to estimate $Y(1)$ for the controls without extensive extrapolation. When estimating the ATT it may be fine (and in fact beneficial) to discard controls outside the range of the treated individuals, but discarding treated individuals may change the group for which the results apply (Crump et al., 2009).

4. DIAGNOSING MATCHES

Perhaps the most important step in using matching methods is to diagnose the quality of the resulting matched samples. All matching should be followed by an assessment of the covariate balance in the matched

groups, where balance is defined as the similarity of the empirical distributions of the full set of covariates in the matched treated and control groups. In other words, we would like the treatment to be unrelated to the covariates, such that $\tilde{p}(X|T=1) = \tilde{p}(X|T=0)$, where $\tilde{p}$ denotes the empirical distribution. A matching method that results in highly imbalanced samples should be rejected, and alternative methods should be attempted until a well-balanced sample is attained. In some situations the diagnostics may indicate that the treated and control groups are too far apart to provide reliable estimates without heroic modeling assumptions (e.g., Rubin, 2001; Agodini and Dynarski, 2004). In contrast to traditional regression models, which do not examine the joint distribution of the predictors (and, in particular, of treatment assignment and the covariates), matching methods will make it clear when it is not possible to separate the effect of the treatment from other differences between the groups. A well-specified regression model of the outcome with many interactions would show this imbalance and may be an effective method for estimating treatment effects (Schafer and Kang, 2008), but complex models like that are only rarely used.

When assessing balance we would ideally compare the multidimensional histograms of the covariates in the matched treated and control groups. However, multidimensional histograms are very coarse and/or will have many zero cells. We thus are left examining the balance of lower-dimensional summaries of that joint distribution, such as the marginal distributions of each covariate. Since we are attempting to examine different features of the multidimensional distribution, though, it is helpful to do a number of different types of balance checks, to obtain a more complete picture.

All balance metrics should be calculated in ways similar to how the outcome analyses will be run, as discussed further in Section 5. For example, if subclassification was done, the balance measures should be calculated within each subclass and then aggregated. If weights will be used in analyses (either as IPTW or because of variable ratio or full matching), they should also be used in calculating the balance measures (Joffe et al., 2004).

4.1 Numerical Diagnostics

One of the most common numerical balance diagnostics is the difference in means of each covariate, divided by the standard deviation in the full treated group: $\frac{\overline{X}_t - \overline{X}_c}{\sigma_t}$. This measure, sometimes referred to as the "standardized bias" or "standardized difference in means," is similar to an effect size and is compared before and after matching (Rosenbaum and Rubin, 1985b). The same standard deviation should be used in the standardization before and after matching. The standardized difference of means should be computed for each covariate, as well as two-way interactions and squares. For binary covariates, either this same formula can be used (treating them as if they were continuous), or a simple difference in proportions can be calculated (Austin, 2009).

Rubin (2001) presents three balance measures based on the theory in Rubin and Thomas (1996) that provide a comprehensive view of covariate balance:

1. The standardized difference of means of the propensity score.
2. The ratio of the variances of the propensity score in the treated and control groups.
3. For each covariate, the ratio of the variance of the residuals orthogonal to the propensity score in the treated and control groups.

Rubin (2001) illustrates these diagnostics in an example with 146 covariates. For regression adjustment to be trustworthy, the absolute standardized differences of means should be less than 0.25 and the variance ratios should be between 0.5 and 2 (Rubin, 2001). These guidelines are based both on the assumptions underlying regression adjustment as well as on results in Rubin (1973b) and Cochran and Rubin (1973), which used simulations to estimate the bias resulting from a number of treatment effect estimation procedures when the true relationship between the covariates and outcome is even moderately nonlinear.

Although common, hypothesis tests and p-values that incorporate information on the sample size (e.g., t-tests) should not be used as measures of balance, for two main reasons (Austin, 2007; Imai, King and Stuart, 2008). First, balance is inherently an in-sample property, without reference to any broader population or super-population. Second, hypothesis tests can be misleading as measures of balance, because they often conflate changes in balance with changes in statistical power. Imai, King and Stuart (2008) show an example where randomly discarding control individuals seemingly leads to increased balance, simply because of the reduced power. In particular, hypothesis tests should not be used as part of a stopping rule to select a matched sample when those samples have varying sizes (or effective sample sizes). Some researchers argue that hypothesis tests are okay for testing balance since the outcome analysis will also have reduced

power for estimating the treatment effect (Hansen, 2008), but that argument requires trading off Type I and Type II errors. The cost of those two types of errors may differ for balance checking and treatment effect estimation.

4.2 Graphical Diagnostics

With many covariates it can be difficult to carefully examine numeric diagnostics for each; graphical diagnostics can be helpful for getting a quick assessment of the covariate balance. A first step is to examine the distribution of the propensity scores in the original and matched groups; this is also useful for assessing common support. Figure 1 shows an example with adequate overlap of the propensity scores, with a good control match for each treated individual. For weighting or subclassification, plots such as this can show the dots with their size proportional to their weight.

For continuous covariates, we can also examine quantile–quantile (QQ) plots, which compare the empirical distributions of each variable in the treated and control groups (this could also be done for the variables squared or two-way interactions, getting at second moments). QQ plots compare the quantiles of a variable in the treatment group against the corresponding quantiles in the control group. If the two groups have identical empirical distributions, all points would lie on the 45 degree line. For weighting methods, weighted boxplots can provide similar information (Joffe et al., 2004).

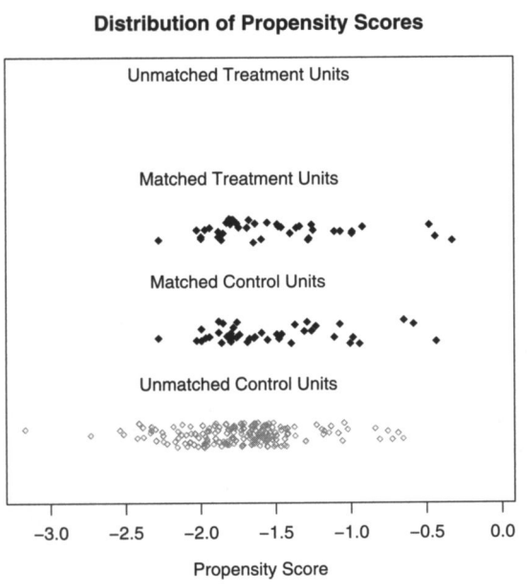

FIG. 1. *Matches chosen using* 1 : 1 *nearest neighbor matching on propensity score. Black dots indicate matched individuals; grey unmatched individuals. Data from Stuart and Green* (2008).

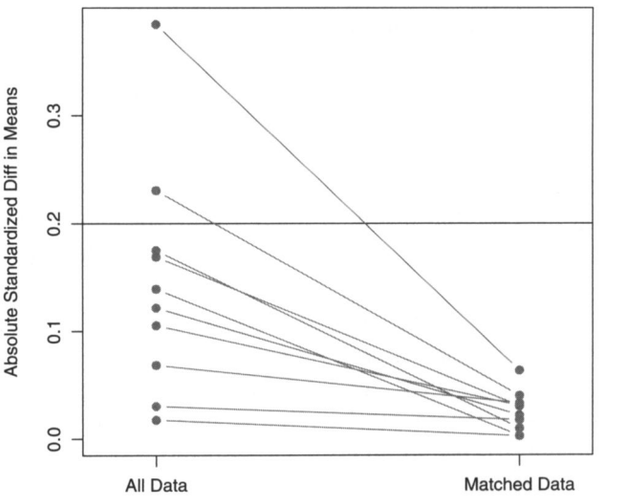

FIG. 2. *Plot of standardized difference of means of* 10 *covariates before and after matching. Data from Stuart and Green* (2008).

Finally, a plot of the standardized differences of means, as in Figure 2, gives us a quick overview of whether balance has improved for individual covariates (Ridgeway, McCaffrey and Morral, 2006). In this example the standardized difference of means of each covariate has decreased after matching. In some situations researchers may find that the standardized difference of means of a few covariates will increase. This may be particularly true of covariates with small differences before matching, since they will not factor heavily into the propensity score model (since they are not predictive of treatment assignment). In these cases researchers should consider whether the increase in bias on those covariates is problematic, which it may be if those covariates are strongly related to the outcome, and modify the matching accordingly (Ho et al., 2007). One solution for that may be to do Mahalanobis matching on those covariates within propensity score calipers.

5. ANALYSIS OF THE OUTCOME

Matching methods are not themselves methods for estimating causal effects. After the matching has created treated and control groups with adequate balance (and the observational study thus "designed"), researchers can move to the outcome analysis stage. This stage will generally involve regression adjustments using the matched samples, with the details of the analysis depending on the structure of the matching. A key point is that matching methods are not designed to

"compete" with modeling adjustments such as linear regression, and, in fact, the two methods have been shown to work best in combination (Rubin, 1973b; Carpenter, 1977; Rubin, 1979; Robins and Rotnitzky, 1995; Heckman, Hidehiko and Todd, 1997; Rubin and Thomas, 2000; Glazerman, Levy and Myers, 2003; Abadie and Imbens, 2006). This is similar to the idea of "double robustness," and the intuition is the same as that behind regression adjustment in randomized experiments, where the regression adjustment is used to "clean up" small residual covariate imbalance between the groups. Matching methods should also make the treatment effect estimates less sensitive to particular outcome model specifications (Ho et al., 2007).

The following sections describe how outcome analyses should proceed after each of the major types of matching methods described above. When weighting methods are used, the weights are used directly in regression models, for example, using weighted least squares. We focus on parametric modeling approaches since those are the most commonly used, however, nonparametric permutation-based tests, such as Fisher's exact test, are also appropriate, as detailed in Rosenbaum (2002, 2010). The best results are found when estimating marginal treatment effects, such as differences in means or differences in proportions. Greenland, Robins and Pearl (1999) and Austin (2007) discuss some of the challenges in estimating noncollapsible conditional treatment effects and which matching methods perform best for those situations.

5.1 After $k:1$ Matching

When each treated individual has received k matches, the outcome analysis proceeds using the matched samples, as if those samples had been generated through randomization. There is debate about whether the analysis needs to account for the matched pair nature of the data (Austin, 2007). However, there are at least two reasons why it is not necessary to account for the matched pairs (Schafer and Kang, 2008; Stuart, 2008). First, conditioning on the variables that were used in the matching process (such as through a regression model) is sufficient. Second, propensity score matching, in fact, does not guarantee that the individual pairs will be well-matched on the full set of covariates, only that groups of individuals with similar propensity scores will have similar covariate distributions. Thus, it is more common to simply pool all the matches into matched treated and control groups and run analyses using the groups as a whole, rather than using the individual matched pairs. In essence, researchers can do the exact same analysis they would have done using the original data, but using the matched data instead (Ho et al., 2007).

Weights need to be incorporated into the analysis for matching with replacement or variable ratio matching (Dehejia and Wahba, 1999; Hill, Reiter and Zanutto, 2004). When matching with replacement, control group individuals receive a frequency weight that reflects the number of times they were selected as a match. When using variable ratio matching, control group members receive a weight that is proportional to the number of controls matched to "their" treated individual. For example, if 1 treated individual was matched to 3 controls, each of those controls receives a weight of $1/3$. If another treated individual was matched to just 1 control, that control receives a weight of 1.

5.2 After Subclassification or Full Matching

With standard subclassification (e.g., the formation of 5 subclasses), effects are generally estimated within each subclass and then aggregated across subclasses (Rosenbaum and Rubin, 1984). Weighting the subclass estimates by the number of treated individuals in each subclass estimates the ATT; weighting by the overall number of individuals in each subclass estimates the ATE. There may be fairly substantial imbalance remaining in each subclass and, thus, it is important to do regression adjustment within each subclass, with the treatment indicator and covariates as predictors (Lunceford and Davidian, 2004). When the number of subclasses is too large—and the number of individuals within each subclass too small—to estimate separate regression models within each subclass, a joint model can be fit, with subclass and subclass by treatment indicators (fixed effects). This is especially useful for full matching. This estimates a separate effect for each subclass, but assumes that the relationship between the covariates X and the outcome is constant across subclasses. Specifically, models such as $Y_{ij} = \beta_{0j} + \beta_{1j}T_{ij} + \gamma X_{ij} + e_{ij}$ are fit, where i indexes individuals and j indexes subclasses. In this model, β_{1j} is the treatment effect for subclass j, and these effects are aggregated across subclasses to obtain an overall treatment effect: $\beta = \frac{N_j}{N}\sum_{j=1}^{J}\beta_{1j}$, where J is the number of subclasses, N_j is the number of individuals in subclass j, and N is the total number of individuals. (This formula weights subclasses by their total size, and so estimates the ATE, but could be modified to estimate the ATT.) This procedure is somewhat more complicated for noncontinuous outcomes when

the estimand of interest, for example, an odds ratio, is noncollapsible. In that case the outcome proportions in each treatment group should be aggregated and then combined.

5.3 Variance Estimation

One of the most debated topics in the literature on matching is variance estimation. Researchers disagree on whether uncertainty in the propensity score estimation or the matching procedure needs to be taken into account, and, if so, how. Some researchers (e.g., Ho et al., 2007) adopt an approach similar to randomized experiments, where the models are run conditional on the covariates, which are treated as fixed and exogenous. Uncertainty regarding the matching process is not taken into account. Other researchers argue that uncertainty in the propensity score model needs to be accounted for in any analysis. However, in fact, under fairly general conditions (Rubin and Thomas, 1996; Rubin and Stuart, 2006), using estimated rather than true propensity scores leads to an overestimate of variance, implying that not accounting for the uncertainty in using estimated rather than true values will be conservative in the sense of yielding confidence intervals that are wider than necessary. Robins, Mark and Newey (1992) also show the benefit of using estimated rather than true propensity scores. Analytic expressions for the bias and variance reduction possible for these situations are given in Rubin and Thomas (1992b). Specifically, Rubin and Thomas (1992b) states that "... with large pools of controls, matching using estimated linear propensity scores results in approximately half the variance for the difference in the matched sample means as in corresponding random samples for all covariates uncorrelated with the population discriminant." This finding has been confirmed in simulations (Rubin and Thomas, 1996) and an empirical example (Hill, Rubin and Thomas, 1999). Thus, when it is possible to obtain 100% or nearly 100% bias reduction by matching on true or estimated propensity scores, using the estimated propensity scores will result in more precise estimates of the average treatment effect. The intuition is that the estimated propensity score accounts for chance imbalances between the groups, in addition to the systematic differences—a situation where overfitting is good. When researchers want to account for the uncertainty in the matching, a bootstrap procedure has been found to outperform other methods (Lechner, 2002; Hill and Reiter, 2006). There are also some empirical formulas for variance estimation for particular matching scenarios (e.g., Abadie and Imbens, 2006, 2009b; Schafer and Kang, 2008), but this is an area for future research.

6. DISCUSSION

6.1 Additional Issues

This section raises additional issues that arise when using any matching method, and also provides suggestions for future research.

6.1.1 *Missing covariate values.* Most of the literature on matching and propensity scores assume fully observed covariates, but of course most studies have at least some missing data. One possibility is to use generalized boosted models to estimate propensity scores, as they do not require fully observed covariates. Another recommended approach is to do a simple single imputation of the missing covariates and include missing data indicators in the propensity score model. This essentially matches based both on the observed values and on the missing data patterns. Although this is generally not an appropriate strategy for dealing with missing data (Greenland and Finkle, 1995), it is an effective approach in the propensity score context. Although it cannot balance the missing values themselves, this method will yield balance on the observed covariates and the missing data patterns (Rosenbaum and Rubin, 1984). A more flexible method is to use multiple imputation to impute the missing covariates, run the matching and effect estimation separately within each "complete" data set, and then use the multiple imputation combining rules to obtain final effect estimates (Rubin, 1987; Song et al., 2001). Qu and Lipkovich (2009) illustrate this method and show good results for an adaptation that also includes indicators of missing data patterns in the propensity score model.

In addition to development and investigation of matching methods that account for missing data, one particular area needing development is balance diagnostics for settings with missing covariate values, including dignostics that allow for nonignorable missing data mechanisms. D'Agostino, Jr. and Rubin (2000) suggests a few simple diagnostics such as assessing available-case means and standard deviations of the continuous variables, and comparing available-case cell proportions for the categorical variables and missing-data indicators, but diagnostics should be developed that explicitly consider the interactions between the missing data and treatment assignment mechanisms.

6.1.2 *Violation of ignorable treatment assignment.* A critique of any nonexperimental study is that there may be unobserved variables related to both treatment assignment and the outcome, violating the assumption of ignorable treatment assignment and biasing the treatment effect estimates. Since ignorability can never be directly tested, researchers have instead developed sensitivity analyses to assess its plausibility, and how violations of ignorability may affect study conclusions. One type of plausibility test estimates an effect on a variable that is known to be unrelated to the treatment, such as a pre-treatment measure of the outcome variable (as in Imbens, 2004), or the difference in outcomes between multiple control groups (as in Rosenbaum, 1987b). If the test indicates that the effect is not equal to zero, then the assumption of ignorable treatment assignment is deemed to be less plausible.

A second approach is to perform analyses of sensitivity to an unobserved variable. Rosenbaum and Rubin (1983a) extends the ideas of Cornfield (1959), examining how strong the correlations would have to be between a hypothetical unobserved covariate and both treatment assignment and the outcome to make the observed treatment effect go away. Similarly, bounds can be created for the treatment effect, given a range of potential correlations of the unobserved covariate with treatment assignment and the outcome (Rosenbaum, 2002). Although sensitivity analysis methods are becoming more and more developed, they are still used relatively infrequently. Newly available software (McCaffrey, Ridgeway and Morral, 2004; Keele, 2009) will hopefully help facilitate their adoption by more researchers.

6.1.3 *Choosing between methods.* There are a wide variety of matching methods available, and little guidance to help applied researchers select between them (Section 6.2 makes an attempt). The primary advice to this point has been to select the method that yields the best balance (e.g., Harder, Stuart and Anthony, 2010; Ho et al., 2007; Rubin, 2007). But defining the best balance is complex, as it involves trading off balance on multiple covariates. Possible ways to choose a method include the following: (1) the method that yields the smallest standardized difference of means across the largest number of covariates, (2) the method that minimizes the standardized difference of means of a few particularly prognostic covariates, and (3) the method that results in the fewest number of "large" standardized differences of means (greater than 0.25). Another promising direction is work by Diamond and Sekhon (2006), which automates the matching procedure, finding the best matches according to a set of balance measures. Further research needs to compare the performance of treatment effect estimates from methods using criteria such as those in Diamond and Sekhon (2006) and Harder, Stuart and Anthony (2010), to determine what the proper criteria should be and examine issues such as potential overfitting to particular measures.

6.1.4 *Multiple treatment doses.* Throughout this discussion of matching, it has been assumed that there are just two groups: treated and control. However, in many studies there are actually multiple levels of the treatment (e.g., doses of a drug). Rosenbaum (2002) summarizes two methods for dealing with doses of treatment. In the first method, the propensity score is still a scalar function of the covariates (e.g., Joffe and Rosenbaum, 1999; Lu et al., 2001). In the second method, each of the levels of treatment has its own propensity score (e.g., Rosenbaum, 1987a; Imbens, 2000) and each propensity score is used one at a time to estimate the distribution of responses that would have been observed if all individuals had received that dose.

Encompassing these two approaches, Imai and van Dyk (2004) generalizes the propensity score to arbitrary treatment regimes (including ordinal, categorical and multidimensional). They provide theorems for the properties of this generalized propensity score (the propensity function), showing that it has properties similar to that of the propensity score in that adjusting for the low-dimensional (not always scalar, but always low-dimensional) propensity function balances the covariates. They advocate subclassification rather than matching, and provide two examples as well as simulations showing the performance of adjustment based on the propensity function. Diagnostics are also complicated in this setting, as it becomes more difficult to assess the balance of the resulting samples when there are multiple treatment levels. Future work is needed to examine these issues.

6.2 Guidance for Practice

So what are the take-away points and advice regarding when to use each of the many methods discussed? While more work is needed to definitively answer that question, this section attempts to pull together the current literature to provide advice for researchers interested in estimating causal effects using matching methods. The lessons can be summarized as follows:

1. Think carefully about the set of covariates to include in the matching procedure, and err on the side of

including more rather than fewer. Is the ignorability assumption reasonable given that set of covariates? If not, consider in advance whether there are other data sets that may be more appropriate, or if there are sensitivity analyses that can be done to strengthen the inferences.

2. Estimate the distance measure that will be used in the matching. Linear propensity scores estimated using logistic regression, or propensity scores estimated using generalized boosted models or boosted CART, are good choices. If there are a few covariates on which particularly close balance is desired (e.g., pretreatment measures of the outcome), consider using the Mahalanobis distance within propensity score calipers.

3. Examine the common support and implications for the estimand. If the ATE is of substantive interest, is there enough overlap of the treated and control groups' propensity scores to estimate the ATE? If not, could the ATT be estimated more reliably? If the ATT is of interest, are there controls across the full range of the treated group, or will it be difficult to estimate the effect for some treated individuals?

4. Implement a matching method.

- If estimating the ATE, good choices are generally IPTW or full matching.
- If estimating the ATT and there are many more control than treated individuals (e.g., more than 3 times as many), $k:1$ nearest neighbor matching without replacement is a good choice for its simplicity and good performance.
- If estimating the ATT and there are not (or not many) more control than treated individuals, appropriate choices are generally subclassification, full matching and weighting by the odds.

5. Examine the balance on covariates resulting from that matching method.

- If adequate, move forward with treatment effect estimation, using regression adjustment on the matched samples.
- If imbalance on just a few covariates, consider incorporating exact or Mahalanobis matching on those variables.
- If imbalance on quite a few covariates, try another matching method (e.g., move to $k:1$ matching with replacement) or consider changing the estimand or the data.

Even if for some reason effect estimates will not be obtained using matching methods, it is worthwhile to go through the steps outlined here to assess the adequacy of the data for answering the question of interest. Standard regression diagnostics will not warn researchers when there is insufficient overlap to reliably estimate causal effects; going through the process of estimating propensity scores and assessing balance before and after matching can be invaluable in terms of helping researchers move forward with causal inference with confidence.

Matching methods are important tools for applied researchers and also have many open research questions for statistical development. This paper has provided an overview of the current literature on matching methods, guidance for practice and a road map for future research. Much research has been done in the past 30 years on this topic, however, there are still a number of open areas and questions to be answered. We hope that this paper, combining results from a variety of disciplines, will promote awareness of and interest in matching methods as an important and interesting area for future research.

7. SOFTWARE APPENDIX

In previous years software limitations made it difficult to implement many of the more advanced matching methods. However, recent advances have made these methods more and more accessible. This section lists some of the major matching procedures available. A continuously updated version is also available at http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html.

- Matching software for R

– **cem**, http://gking.harvard.edu/cem/
Iacus, S. M., King, G. and Porro, G. (2009). cem: Coarsened exact matching software. Can also be implemented through MatchIt.
– **Matching**, http://sekhon.berkeley.edu/matching
Sekhon, J. S. (in press). Matching: Multivariate and propensity score matching with balance optimization. Forthcoming, *Journal of Statistical Software.* Uses automated procedure to select matches, based on univariate and multivariate balance diagnostics. Primarily $k:1$ matching, allows matching with or without replacement, caliper, exact. Includes built-in effect and variance estimation procedures.
– **MatchIt**, http://gking.harvard.edu/matchit
Ho, D. E., Imai, K., King, G. and Stuart, E. A. (in press). MatchIt: Nonparametric preprocessing for parameteric causal inference. Forthcoming, *Journal of Statistical Software.* Two-step process: does matching, then user does outcome analysis. Wide array of estimation procedures and matching methods

available: nearest neighbor, Mahalanobis, caliper, exact, full, optimal, subclassification. Built-in numeric and graphical diagnostics.

- **optmatch**, http://cran.r-project.org/web/packages/optmatch/index.html
  Hansen, B. B. and Fredrickson, M. (2009). optmatch: Functions for optimal matching. Variable ratio, optimal and full matching. Can also be implemented through MatchIt.
- **PSAgraphics**, http://cran.r-project.org/web/packages/PSAgraphics/index.html
  Helmreich, J. E. and Pruzek, R. M. (2009). PSAgraphics: Propensity score analysis graphics. *Journal of Statistical Software* **29**. Package to do graphical diagnostics of propensity score methods.
- **rbounds**, http://cran.r-project.org/web/packages/rbounds/index.html
  Keele, L. J. (2009). rbounds: An R package for sensitivity analysis with matched data. Does analysis of sensitivity to assumption of ignorable treatment assignment.
- **twang**, http://cran.r-project.org/web/packages/twang/index.html
  Ridgeway, G., McCaffrey, D. and Morral, A. (2006). twang: Toolkit for weighting and analysis of nonequivalent groups. Functions for propensity score estimating and weighting, nonresponse weighting, and diagnosis of the weights. Primarily uses generalized boosted regression to estimate the propensity scores.

• Matching software for Stata

- **cem**, http://gking.harvard.edu/cem/
  Iacus, S. M., King, G. and Porro, G. (2009). cem: Coarsened exact matching software.
- **match**, http://www.economics.harvard.edu/faculty/imbens/software_imbens
  Abadie, A., Drukker, D., Herr, J. L. and Imbens, G. W. (2004). Implementing matching estimators for average treatment effects in Stata. *The Stata Journal* **4** 290–311. Primarily $k:1$ matching (with replacement). Allows estimation of ATT or ATE, including robust variance estimators.
- **pscore**, http://www.lrz-muenchen.de/~sobecker/pscore.html
  Becker, S. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal* **2** 358–377. Does $k:1$ nearest neighbor matching, radius (caliper) matching and subclassification.
- **psmatch2**, http://econpapers.repec.org/software/bocbocode/s432001.htm
  Leuven, E. and Sianesi, B. (2003). psmatch2. Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Allows $k:1$ matching, kernel weighting, Mahalanobis matching. Includes built-in diagnostics and procedures for estimating ATT or ATE.
- **Note**: 3 procedures for analysis of sensitivity to the ignorability assumption are also available: rbounds (for continuous outcomes), mhbounds (for categorical outcomes), and sensatt (to be used after the pscore procedures).
  rbounds, http://econpapers.repec.org/software/bocbocode/s438301.htm;
  mhbounds, http://ideas.repec.org/p/diw/diwwpp/dp659.html;
  sensatt, http://ideas.repec.org/c/boc/bocode/s456747.html.

• Matching software for SAS

- **SAS usage note**: http://support.sas.com/kb/30/971.html
- **Greedy 1:1 matching**, http://www2.sas.com/proceedings/sugi25/25/po/25p225.pdf
  Parsons, L. S. (2005). Using SAS software to perform a case-control match on propensity score in an observational study. In *SAS SUGI 30*, Paper 225-25.
- **gmatch macro**, http://mayoresearch.mayo.edu/mayo/research/biostat/upload/gmatch.sas
  Kosanke, J. and Bergstralh, E. (2004). gmatch: Match 1 or more controls to cases using the GREEDY algorithm.
- **Proc assign**, http://pubs.amstat.org/doi/abs/10.1198/106186001317114938
  Can be used to perform optimal matching.
- **1:1 Mahalanobis matching within propensity score calipers**, www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf
  Feng, W. W., Jun, Y. and Xu, R. (2005). A method/macro based on propensity score and Mahalanobis distance to reduce bias in treatment comparison in observational study.
- **vmatch macro**, http://mayoresearch.mayo.edu/mayo/research/biostat/upload/vmatch.sas
  Kosanke, J. and Bergstralh, E. (2004). Match cases to controls using variable optimal matching. Variable ratio matching (optimal algorithm).
- **Weighting**, http://www.lexjansen.com/wuss/2006/Analytics/ANL-Leslie.pdf

Leslie, S. and Thiebaud, P. (2006). Using propensity scores to adjust for treatment selection bias.

ACKNOWLEDGMENTS

Supported in part by Award K25MH083846 from the National Institute of Mental Health. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health. The work for this paper was partially done while the author was a graduate student at Harvard University, Department of Statistics, and a Researcher at Mathematica Policy Research. The author particularly thanks Jennifer Hill, Daniel Ho, Kosuke Imai, Gary King and Donald Rubin, as well as the reviewers and Editor, for helpful comments.

REFERENCES

ABADIE, A. and IMBENS, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica* **74** 235–267. MR2194325

ABADIE, A. and IMBENS, G. W. (2009a). Bias corrected matching estimators for average treatment effects. *Journal of Educational and Behavioral Statistics*. To appear. Available at http://www.hks.harvard.edu/fs/aabadie/bcm.pdf.

ABADIE, A. and IMBENS, G. W. (2009b). Matching on the estimated propensity score. Working Paper 15301, National Bureau of Economic Research, Cambridge, MA.

AGODINI, R. and DYNARSKI, M. (2004). Are experiments the only option? A look at dropout prevention programs. *Review of Economics and Statistics* **86** 180–194.

ALTHAUSER, R. and RUBIN, D. (1970). The computerized construction of a matched sample. *American Journal of Sociology* **76** 325–346.

AUGURZKY, B. and SCHMIDT, C. (2001). The propensity score: A means to an end. Discussion Paper 271, Institute for the Study of Labor (IZA).

AUSTIN, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Stat. Med.* **26** 3078–3094. MR2380505

AUSTIN, P. (2009). Using the standardized difference to compare the prevalence of a binary variable between two groups in observational research. *Comm. Statist. Simulation Comput.* **38** 1228–1234.

AUSTIN, P. C. and MAMDANI, M. M. (2006). A comparison of propensity score methods: A case-study illustrating the effectiveness of post-ami statin use. *Stat. Med.* **25** 2084–2106. MR2239634

BANG, H. and ROBINS, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61** 962–972. MR2216189

BROOKHART, M. A., SCHNEEWEISS, S., ROTHMAN, K. J., GLYNN, R. J., AVORN, J. and STURMER, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology* **163** 1149–1156.

CARPENTER, R. (1977). Matching when covariables are normally distributed. *Biometrika* **64** 299–307.

CHAPIN, F. (1947). *Experimental Designs in Sociological Research*. Harper, New York.

COCHRAN, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24** 295–313. MR0228136

COCHRAN, W. G. and RUBIN, D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā Ser. A* **35** 417–446.

COHEN, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. Earlbaum, Hillsdale, NJ.

CORNFIELD, J. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173–200.

CRUMP, R., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199.

CZAJKA, J. C., HIRABAYASHI, S., LITTLE, R. and RUBIN, D. B. (1992). Projecting from advance data using propensity modeling. *J. Bus. Econom. Statist.* **10** 117–131.

D'AGOSTINO, JR., R. B. and RUBIN, D. B. (2000). Estimating and using propensity scores with partially missing data. *J. Amer. Statist. Assoc.* **95** 749–759.

DEHEJIA, R. H. and WAHBA, S. (1999). Causal effects in nonexperimental studies: Re-evaluating the evaluation of training programs. *J. Amer. Statist. Assoc.* **94** 1053–1062.

DEHEJIA, R. H. and WAHBA, S. (2002). Propensity score matching methods for non-experimental causal studies. *Review of Economics and Statistics* **84** 151–161.

DIAMOND, A. and SEKHON, J. S. (2006). Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. Working paper. Univ. California, Berkeley. Available at http://sekhon.berkeley.edu/papers/GenMatch.pdf.

DRAKE, C. (1993). Effects of misspecification of the propensity score on estimators of treatment effects. *Biometrics* **49** 1231–1236.

FRANGAKIS, C. E. and RUBIN, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58** 21–29. MR1891039

GLAZERMAN, S., LEVY, D. M. and MYERS, D. (2003). Nonexperimental versus experimental estimates of earnings impacts. *Annals of the American Academy of Political and Social Science* **589** 63–93.

GREENLAND, S. (2003). Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology* **14** 300–306.

GREENLAND, S. and FINKLE, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology* **142** 1255–1264.

GREENLAND, S., ROBINS, J. M. and PEARL, J. (1999). Confounding and collapsibility in causal inference. *Statist. Sci.* **14** 29–46.

GREENWOOD, E. (1945). *Experimental Sociology: A Study in Method*. King's Crown Press, New York.

GREEVY, R., LU, B., SILBER, J. H. and ROSENBAUM, P. (2004). Optimal multivariate matching before randomization. *Biostatistics* **5** 263–275.

GU, X. and ROSENBAUM, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *J. Comput. Graph. Statist.* **2** 405–420.

HANSEN, B. B. (2004). Full matching in an observational study of coaching for the SAT. *J. Amer. Statist. Assoc.* **99** 609–618. MR2086387

HANSEN, B. B. (2008). The essential role of balance tests in propensity-matched observational studies: Comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. *Stat. Med.* **27** 2050–2054.

HANSEN, B. B. (2008). The prognostic analogue of the propensity score. *Biometrika* **95** 481–488. MR2521594

HARDER, V. S., STUART, E. A. and ANTHONY, J. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*. To appear.

HECKMAN, J. J., HIDEHIKO, H. and TODD, P. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econom. Stud.* **64** 605–654.

HECKMAN, J. J., ICHIMURA, H., SMITH, J. and TODD, P. (1998). Characterizing selection bias using experimental data. *Econometrica* **66** 1017–1098. MR1639419

HECKMAN, J. J., ICHIMURA, H. and TODD, P. (1998). Matching as an econometric evaluation estimator. *Rev. Econom. Stud.* **65** 261–294. MR1623713

HELLER, R., ROSENBAUM, P. and SMALL, D. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101.

HILL, J. L. and REITER, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Stat. Med.* **25** 2230–2256. MR2240098

HILL, J., REITER, J. and ZANUTTO, E. (2004). A comparison of experimental and observational data analyses. In *Applied Bayesian Modeling and Causal Inference From an Incomplete-Data Perspective* (A. Gelman and X.-L. Meng, eds.). Wiley, Hoboken, NJ. MR2134801

HILL, J., RUBIN, D. B. and THOMAS, N. (1999). The design of the New York School Choice Scholarship Program evaluation. In *Research Designs: Inspired by the Work of Donald Campbell*, (L. Bickman, ed.) 155–180. Sage, Thousand Oaks, CA.

HIRANO, K., IMBENS, G. W. and RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71** 1161–1189. MR1995826

HO, D. E., IMAI, K., KING, G. and STUART, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15** 199–236.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 945–960. MR0867618

HONG, G. and RAUDENBUSH, S. W. (2006). Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *J. Amer. Statist. Assoc.* **101** 901–910. MR2324091

HORVITZ, D. and THOMPSON, D. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *J. Amer. Statist. Assoc.* **103** 832–842. MR2435472

IACUS, S. M., KING, G. and PORRO, G. (2009). CEM: Software for coarsened exact matching. *J. Statist. Software* **30** 9. Available at http://gking.harvard.edu/files/abs/cemR-abs.shtml.

IMAI, K. and VAN DYK, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *J. Amer. Statist. Assoc.* **99** 854–866. MR2090918

IMAI, K., KING, G. and STUART, E. A. (2008). Misunderstandings among experimentalists and observationalists in causal inference. *J. Roy. Statist. Soc. Ser. A* **171** 481–502. MR2427345

IMBENS, G. W. (2000). The role of the propensity score in estimating dose–response functions. *Biometrika* **87** 706–710. MR1789821

IMBENS, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86** 4–29.

JOFFE, M. M. and ROSENBAUM, P. R. (1999). Propensity scores. *American Journal of Epidemiology* **150** 327–333.

JOFFE, M. M., TEN HAVE, T. R., FELDMAN, H. I. and KIMMEL, S. E. (2004). Model selection, confounder control, and marginal structural models. *Amer. Statist.* **58** 272–279. MR2109415

KANG, J. D. and SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statist. Sci.* **22** 523–539. MR2420458

KEELE, L. (2009). rbounds: An R package for sensitivity analysis with matched data. R package. Available at http://www.polisci.ohio-state.edu/faculty/lkeele/rbounds.html.

KING, G. and ZENG, L. (2006). The dangers of extreme counterfactuals. *Political Analysis* **14** 131–159.

KURTH, T., WALKER, A. M., GLYNN, R. J., CHAN, K. A., GAZIANO, J. M., BERGER, K. and ROBINS, J. M. (2006). Results of multivariable logistic regresion, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology* **163** 262–270.

LECHNER, M. (2002). Some practical issues in the evaluation of heterogeneous labour market programmes by matching methods. *J. Roy. Statist. Soc. Ser. A* **165** 59–82. MR1888488

LEE, B., LESSLER, J. and STUART, E. A. (2009). Improving propensity score weighting using machine learning. *Stat. Med.* **29** 337–346.

LI, Y. P., PROPERT, K. J. and ROSENBAUM, P. R. (2001). Balanced risk set matching. *J. Amer. Statist. Assoc.* **96** 455, 870–882. MR1946360

LU, B., ZANUTTO, E., HORNIK, R. and ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Amer. Statist. Assoc.* **96** 1245–1253. MR1973668

LUNCEFORD, J. K. and DAVIDIAN, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat. Med.* **23** 2937–2960.

LUNT, M., SOLOMON, D., ROTHMAN, K., GLYNN, R., HYRICH, K., SYMMONS, D. P., STURMER, T., THE BRITISH SOCIETY FOR RHEUMATOLOGY BIOLOGICS REGISTER and THE BRITISH SOCIETY FOR RHEUMATOLOGY BIOLOGICS REGISTER CONTRL CENTRE CONSORTIUM (2009). Different methods of balancing covariates leading to different effect estimates in the presence of effect modification. *American Journal of Epidemiology* **169** 909–917.

MCCAFFREY, D. F., RIDGEWAY, G. and MORRAL, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods* **9** 403–425.

MING, K. and ROSENBAUM, P. R. (2001). A note on optimal matching with variable controls using the assignment algorithm. *J. Comput. Graph. Statist.* **10** 455–463. MR1939035

MORGAN, S. L. and HARDING, D. J. (2006). Matching estimators of causal effects: Prospects and pitfalls in theory and practice. *Sociological Methods & Research* **35** 3–60. MR2247150

POTTER, F. J. (1993). The effect of weight trimming on nonlinear survey estimates. In *Proceedings of the Section on Survey Research Methods of American Statistical Association*. Amer. Statist. Assoc., San Francisco, CA.

QU, Y. and LIPKOVICH, I. (2009). Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* **28** 1402–1414.

REINISCH, J., SANDERS, S., MORTENSEN, E. and RUBIN, D. B. (1995). In utero exposure to phenobarbital and intelligence deficits in adult men. *Journal of the American Medical Association* **274** 1518–1525.

RIDGEWAY, G., MCCAFFREY, D. and MORRAL, A. (2006). twang: Toolkit for weighting and analysis of nonequivalent groups. Software for using matching methods in R. Available at http://cran.r-project.org/web/packages/twang/index.html.

ROBINS, J. and ROTNITZKY, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *J. Amer. Statist. Assoc.* **90** 122–129. MR1325119

ROBINS, J. M., HERNAN, M. A. and BRUMBACK, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology* **11** 550–560.

ROBINS, J. M., MARK, S. and NEWEY, W. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48** 479–495. MR1173493

ROSENBAUM, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.

ROSENBAUM, P. R. (1987a). Model-based direct adjustment. *J. Amer. Statist. Assoc.* **82** 387–394.

ROSENBAUM, P. R. (1987b). The role of a second control group in an observational study (with discussion). *Statist. Sci.* **2** 292–316.

ROSENBAUM, P. R. (1991). A characterization of optimal designs for observational studies. *J. Roy. Statist. Soc. Ser. B* **53** 597–610. MR1125717

ROSENBAUM, P. R. (1999). Choice as an alternative to control in observational studies (with discussion). *Statist. Sci.* **14** 259–304.

ROSENBAUM, P. R. (2002). *Observational Studies*, 2nd ed. Springer, New York. MR1899138

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York.

ROSENBAUM, P. R. and RUBIN, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

ROSENBAUM, P. R. and RUBIN, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROSENBAUM, P. R. and RUBIN, D. B. (1985a). The bias due to incomplete matching. *Biometrics* **41** 103–116. MR0793436

ROSENBAUM, P. R. and RUBIN, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Amer. Statist.* **39** 33–38.

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534

RUBIN, D. B. (1973a). Matching to remove bias in observational studies. *Biometrics* **29** 159–184.

RUBIN, D. B. (1973b). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* **29** 185–203.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688–701.

RUBIN, D. B. (1976a). Inference and missing data (with discussion). *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1976b). Multivariate matching methods that are equal percent bias reducing, I: Some examples. *Biometrics* **32** 109–120. MR0400555

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

RUBIN, D. B. (1980). Bias reduction using Mahalanobis metric matching. *Biometrics* **36** 293–298.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

RUBIN, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology* **2** 169–188.

RUBIN, D. B. (2004). On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety* **13** 855–857.

RUBIN, D. B. (2006). *Matched Sampling for Causal Inference*. Cambridge Univ. Press, Cambridge. MR2307965

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Stat. Med.* **26** 20–36. MR2312697

RUBIN, D. B. and STUART, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *Ann. Statist.* **34** 1814–1826. MR2283718

RUBIN, D. B. and THOMAS, N. (1992a). Affinely invariant matching methods with ellipsoidal distributions. *Ann. Statist.* **20** 1079–1093. MR1165607

RUBIN, D. B. and THOMAS, N. (1992b). Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika* **79** 797–809. MR1209479

RUBIN, D. B. and THOMAS, N. (1996). Matching using estimated propensity scores, relating theory to practice. *Biometrics* **52** 249–264.

RUBIN, D. B. and THOMAS, N. (2000). Combining propensity score matching with additional adjustments for prognostic covariates. *J. Amer. Statist. Assoc.* **95** 573–585.

SCHAFER, J. L. and KANG, J. D. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated case study. *Psychological Methods* **13** 279–313.

SCHARFSTEIN, D. O., ROTNITZKY, A. and ROBINS, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models. *J. Amer. Statist. Assoc.* **94** 1096–1120. MR1731478

SCHNEIDER, E. C., ZASLAVSKY, A. M. and EPSTEIN, A. M. (2004). Use of high-cost operative procedures by Medicare beneficiaries enrolled in for-profit and not-for-profit health plans. *The New England Journal of Medicine* **350** 143–150.

SETOGUCHI, S., SCHNEEWEISS, S., BROOKHART, M. A., GLYNN, R. J. and COOK, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: A simulation study. *Pharmacoepidemiology and Drug Safety* **17** 546–555.

SHADISH, W. R., CLARK, M. and STEINER, P. M. (2008). Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *J. Amer. Statist. Assoc.* **103** 1334–1344.

SMITH, H. (1997). Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27** 325–353.

SNEDECOR, G. W. and COCHRAN, W. G. (1980). *Statistical Methods*, 7th ed. Iowa State Univ. Press, Ames, IA. MR0614143

SOBEL, M. E. (2006). What do randomized studies of housing mobility demonstrate?: Causal inference in the face of interference. *J. Amer. Statist. Assoc.* **101** 1398–1407. MR2307573

SONG, J., BELIN, T. R., LEE, M. B., GAO, X. and ROTHERAM-BORUS, M. J. (2001). Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths. *Health Services & Outcomes Research Methodology* **2** 317–329.

STUART, E. A. (2008). Developing practical recommendations for the use of propensity scores: Discussion of "A critical appraisal of propensity score matching in the medical literature between 1996 and 2003" by P. Austin. *Stat. Med.* **27** 2062–2065. MR2439885

STUART, E. A. and GREEN, K. M. (2008). Using full matching to estimate causal effects in non-experimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology* **44** 395–406.

STUART, E. A. and IALONGO, N. S. (2009). Matching methods for selection of subjects for follow-up. *Multivariate Behavioral Research*. To appear.

WACHOLDER, S. and WEINBERG, C. R. (1982). Paired versus two-sample design for a clinical trial of treatments with dichotomous outcome: Power considerations. *Biometrics* **38** 801–812.

WEITZEN, S., LAPANE, K. L., TOLEDANO, A. Y., HUME, A. L. and MOR, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety* **13** 841–853.

ZHAO, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *Review of Economics and Statistics* **86** 91–107.