Published in final edited form as:

Med Care Res Rev. 2010 October ; 67(5): 528–554. doi:10.1177/1077558710361486.

Applying Propensity Score Methods in Medical Research: Pitfalls and Prospects

Zhehui Luo¹, Joseph C. Gardiner¹, and Cathy J. Bradley²

¹Michigan State University, East Lansing

²Virginia Commonwealth University, Richmond

Abstract

The authors review experimental and nonexperimental causal inference methods, focusing on assumptions for the validity of instrumental variables and propensity score (PS) methods. They provide guidance in four areas for the analysis and reporting of PS methods in medical research and selectively evaluate mainstream medical journal articles from 2000 to 2005 in the four areas, namely, examination of balance, overlapping support description, use of estimated PS for evaluation of treatment effect, and sensitivity analyses. In spite of the many pitfalls, when appropriately evaluated and applied, PS methods can be powerful tools in assessing average treatment effects in observational studies. Appropriate PS applications can create experimental conditions using observational data when randomized controlled trials are not feasible and, thus, lead researchers to an efficient estimator of the average treatment effect.

Keywords

propensity score matching; treatment effect; evaluation methods; observational study

Use of propensity score (PS) methods in medical research to estimate causal effects from nonexperimental data has grown considerably over the past decade. Introduced by Rosenbaum and Rubin (1983b), the main aim of PS methods in nonrandomized studies is to balance the treatment (exposure) and comparison groups on observed preintervention characteristics. An individual's PS is the probability of being treated (exposed) conditional on observed characteristics. In general, there are three ways of using the estimated PS. First, the predicted probability is used as a covariate in addition to the treatment indicator in a multivariable regression for the outcome of interest. The most commonly used models are logistic regression for binary outcomes and the Cox model for time-to-event data. Second, a treated subject is matched to one or more comparison subjects based on the closeness in their estimated PS. Differences in the outcome of interest between the treated subject and the matched comparison subject(s) are summarized to derive an average treatment effect. In the third way, all subjects are stratified into bins (quintiles or quartiles) of the estimated PS. The treatment effect is estimated by a weighted difference in the matched pairs (or clusters), or stratified samples. Henceforth, we use "treatment" to represent the objective of an empirical investigation that may be a specific medical intervention, an exposure to a risk factor, a policy initiative, or a characteristic or trait of a population.

Corresponding Author: Zhehui Luo, Department of Epidemiology, Michigan State University, B627 West Fee Hall, East Lansing, MI 48824, zluo@msu.edu.

Declaration of Conflicting Interests: The author(s) declared no potential conflicts of interest with respect to the authorship and/or publication of this article.

Notes

One recent review of applications of PS methods (Weitzen, Lapane, Toledano, Hume, & Mor, 2004) lamented on the lack of procedural guidance in appropriately using and clearly reporting the methods, and the paucity of sensitivity analyses of results. Stürmer et al. (2006) abstracted 177 published studies from 1998 to 2003 that used PS methods and found that estimates based on PS methods and estimates obtained by conventional methods of covariate adjustment differed for only 9 out of 69 studies in which such comparisons were possible. This led some researchers to advocate joint consideration of PS estimates and multivariate analysis for a check of robustness (Baser, 2006). A more statistically sophisticated review (Austin, 2008a) focused on matching methods and the need to account for the matched design when assessing balance between treated and untreated subjects and estimating treatment effects. The author found that many PS applications in medical research were not appropriate and listed common errors (e.g., using logistic regression, t tests, and chi-squared tests in the matched sample) in employing matching methods.

New Contribution

We review experimental and nonexperimental data analysis methods and provide guidelines for the analysis and reporting of various PS methods in medical research. Evaluation of PS methods in a given study should be based on four factors. First, the estimated PS must satisfy the balancing property (Rosenbaum & Rubin, 1983b). Second, when we apply the PS on an irrelevant (or preintervention) measure that is not affected by the treatment, the estimated effect should be null (Rosenbaum, 2002). This is similar to the evaluation test proposed by Heckman, Ichimura, and Todd (1997) in the context of evaluating training programs. Third, one should explicitly compare the matched cases with the population of interest when a substantial number of subjects are eliminated due to trimming or lack of matched controls (Smith & Todd, 2005). Fourth, sensitivity analyses with respect to the specification of the PS and matching techniques are crucial steps in any evaluation of PS methods (Rosenbaum & Rubin, 1983a). We provide guidelines in implementing PS methods and selectively evaluate mainstream medical journal articles from 2000 to 2005 against our proposed guidelines.

Experimental and Nonexperimental Data Analysis Methods

Experimental Data

In a randomized controlled trial (RCT), the treatment effect can be estimated from the simple difference in outcomes between the treated and the control groups, assuming the assignment of treatment is well randomized with respect to both observed and unobserved subject characteristics, and the treatment effect is homogeneous across the study population. The strongest argument supporting randomized experiments is that under certain assumptions they solve the fundamental evaluation problem of unobservable counterfactuals. A counterfactual outcome refers to what would have happened hypothetically if a person could be observed in a state where she or he was not, that is, the would-be outcome of a person in treatment if she or he was not treated or the would-be outcome of a person in control if she or he was treated. In well-designed and well-conducted RCTs, the counterfactual outcomes of the treated are estimated by the observed outcomes of the control group.

The list of criticisms associated with randomization is long. Most RCTs are designed to demonstrate efficacy of a treatment in a selective group of patients. External validity is threatened when the sample or practice settings are not representative of the general population (Cook & Campbell, 1979). Treatment effects may vary by subgroups of the selected patients. As discussed in Kravitz, Duan, and Braslow (2004), individuals may depart from the group average because of differences in susceptibility, responsiveness to the

treatment, vulnerability to adverse side effects, and utilities for different outcomes. Many of these variables are not available to the analyst and render the internal validity of an RCT suspect. If the treatment and control groups are unbalanced, theoretical and practical solutions have not been adequately addressed. Deleting the unbalanced subgroup will lose the basis justifying randomization. If we could use regression methods to correct the imbalance, we could have done so without randomization.¹ Finally, ethical issues and high costs of randomized trials prohibit some experiments.

Nonexperimental Data Method: Propensity Score

There have been many important developments in nonexperimental causal inference methods in past decades (see Heckman & Vytlačil, 2007, and the references therein; see also Heckman, 2005a, 2005b; Heckman & Smith, 1995; Sobel, 2005). To the best of our knowledge, however, there is no consensus on which method is superior to the others, nor is an approach deemed a panacea against all potential causes of bias in analyses. All methods have underlying assumptions on which the validity of the estimates depends. See Imbens and Wooldridge (2009) for a technical review of the convergence of the statistical and econometric literatures on the evaluation of treatment.

In nonexperimental analyses, the main threat to validity is selection bias. Selection bias consists of the difference between the adjusted outcomes of the “control” subjects and the desired counterfactual. There are two types of selection bias: overt bias due to measured but not appropriately adjusted confounders and hidden bias due to unmeasured confounders. In nonexperimental data, individuals are not randomly assigned but rather “self-selected” into the treatment or control groups. As a result, the standard statistical techniques may yield biased estimates of true effects if the subjects in the two groups differ in unobserved ways that relate to both the outcome and the treatment. For example, suppose more severely ill patients are more likely to take a new (or higher dose) drug, then the effect of the drug may be overestimated due to the phenomenon of regression to the mean, or it may be underestimated due to poor prognosis. An analyst is forced to construct a “control” sample through poststratification matching or other statistical/econometric techniques to achieve as much balance as possible ad hoc. PS methods and the instrumental variables (IV) approach are two commonly used techniques, with the former developed and used more often in medical, biostatistics, and epidemiologic research, and the latter in social science and economic research.

PS methods can correct overt bias if all pertinent variables are observed and measured without error. When a confounder to the treatment is not observed or is measured with error in an observational study (or in a clinical trial with noncompliance), the PS methods correct the hidden bias only to the extent that the unobserved confounders are correlated with the observed covariates. Rosenbaum and Rubin (1983b) clearly stated the assumption under which the PS estimates are valid, namely, the strong ignorability of treatment assignment. This assumption has two components: first, among individuals with the same observed characteristics before treatment, their treatment assignment does not depend on the potential benefit associated with the treatment or control condition; and second, no individual is guaranteed to be in the treatment or control group. In other words, given the set of covariates used for matching, treatment assignment is independent of potential outcomes (also referred to as “conditional independence”). Because only observed characteristics are used for matching (also referred to as “selection on observables”), unlike randomization, the potential for selection bias still exists. For a more technical discussion of the estimation of average treatment effects under these assumptions, see Imbens (2004).

¹We thank one of the reviewers for pointing out the limitations in randomized control trials.

Wooldridge (2002) showed that under the strong ignorability assumption the linear regression approach (also called the model-based approach)—regressing outcome on treatment indicator and a consistently estimated PS—leads to a consistent estimator of the average treatment effect for a continuous outcome. The corresponding results for some outcomes modeled by a nonlinear regression approach do not generally hold (Austin, Grootendorst, Normand, & Anderson, 2007). Austin, Grootendorst, Normand, et al. (2007) suggest that this is because the model-based approach estimates a marginal effect using the matched pairs. For some nonlinear models, the distinction between marginal estimates obtained by generalized estimating equation method and conditional estimates by the subject-specific random effects models is important, both in terms of magnitude and interpretation (Gardiner, Luo, & Roman, 2009).

Nonexperimental Data Method: Instrumental Variables

The IV methods are hallmark techniques in econometrics that can correct for hidden bias under certain assumptions. If the treatment variable (or any explanatory variable of interest) is correlated with unmeasured confounder (and thus with the error term in the equation of interest), then the treatment variable is endogenous. The usual regression estimates are biased when one or more endogenous variables are included as explanatory variables. The IV approach provides a solution to the endogeneity problem under very strong assumptions. We restrict our description of the approach to the case where the outcome is continuous and the instrument and the endogenous regressor are binary. Many-valued discrete endogenous variables (i.e., variables that have several discrete levels), multiple endogenous variables, and nonlinear problems are beyond the scope of this review.

In standard textbook linear IV methods, the set up is as follows. Assume a causal relationship is linear between the outcome of interest Y_i and the treatment indicator D_i : $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$. In the absence of randomization or the presence of systematic non-compliance, D_i may be correlated with the unobserved component of the error, ε_i . Suppose we have a third variable Z_i , an instrument (or IV), that satisfies two conditions: First, it is uncorrelated with ε_i , and second, it is correlated with the endogenous D_i . The first assumption requires that the instrument, similar to random assignment, does not affect the outcome directly, that is, uncorrelated with the error term ε_i . This is called the exclusion restriction. Examples of IVs could be encouragement to participate in a health education program or the distance from a patient's residence to facilities that offer one type of treatment versus another for emergency medical conditions. It has been shown that the IV estimator of the average treatment effect is a local average treatment effect for subpopulations whose endogenous variable changes value induced by the instrument (Imbens & Angrist, 1994). Without the constant treatment effect assumption, the IV estimator may not coincide with the population average. Under certain assumptions, the two-stage least squares estimator is the most efficient IV estimator.

When the number of instruments equals the number of endogenous variables, the problem is called “exactly identified” or “just identified.” Because there is no empirical test that the instrument is uncorrelated with the error term ε_i , there is no test for IV analysis of an exactly identified model. A model with p endogenous variables is called “overidentified” when the number of instruments, M , is greater than p , in which case we say that there are $M - p$ overidentification restrictions. When we have more instruments than necessary, we can test whether the additional instruments are valid in the sense that they are uncorrelated with ε_i , that is, the overidentification test. Notice that this is not a direct test of the first assumption either. For the testing procedure, see Wooldridge (2002).

A good instrument is strongly correlated with the endogenous variable. The two-stage least-squares estimator is biased when the instruments are weak, that is, the correlation with the

endogenous regressor is low (Bound, Jaeger, & Baker, 1995). In practice, convincing instruments are hard to come by, and not all assumptions for the validity of the IV approach can be tested, just as not all assumptions for the validity of the PS methods or RCTs can be tested. Nevertheless, the strength of the instruments can and should be tested (Staiger & Stock, 1997). Baser (2009) illustrated how weak instruments may produce inconsistent and inefficient results. For a highly readable explanation of the danger of many weak instruments, see Angrist and Pischke (2009). Greenland (2000) provides an excellent introduction of the IV approach for medical researchers.

Implementation of PS Methods

The PS methods have gained great popularity because they are easy to implement empirically. However, to appropriately use various PS methods, there is a plethora of choices a researcher has to make (Figure 1). To facilitate full understanding of these choices, we summarize the basic framework of estimating an average treatment effect. For each subject i , we observe the following: a binary treatment indicator variable D_i , equal to 1 if the subject is treated and 0 if not; a vector of observed pretreatment covariates $\mathbf{x}_i$; and the outcome variable of interest y_i . The PS (p_i) for subject i represents the probability of being treated given the pretreatment factors: $p_i = \Pr(D_i = 1 \mid \mathbf{x}_i)$. The regression adjustment (or model-based) method typically estimates the average treatment effect through a regression of y_i on D_i , p_i , and interaction of D_i and p_i . Although it is the most widely used and easily implemented method, it is not recommended by Rubin (2004) who stated that this method

can be effective when nonlinear terms in the propensity score are included in the model, but in general the method must be used cautiously, and the user must be confident that the propensity scores are well estimated, at least up to a monotone transformation. (p. 856)

It is easy to construct examples where a misspecified equation of interest leads to bias in estimating treatment effect. The model-based adjustment is not robust to misspecifications in either the propensity equation or the equation of interest without further corrections.

Common PS matching estimators are the following: (a) The single nearest-neighbor matching without replacement, where an untreated subject with the value of p_j that is closest to p_i is selected as the match. (b) The multiple nearest-neighbors matching without replacement, where each treated subject has more than one matched neighbor with equal weight. This may reduce variance of the estimator, but may increase its bias due to more poor matches. (c) The caliper or radius matching that imposes a tolerance on the maximum distance between p_i and p_j . Participants for whom no match is found within the caliper c are excluded from the analysis. (d) In the stratification or interval matching or binning method, the common support of p_i is partitioned into intervals and treatment effects are estimated within each interval and then averaged. (e) The kernel matching method gives a kernel weight for the i th treated and j th control pair. The kernel matching differs from the conventional matching methods in (a) to (d) in that the former is matching with replacement. Although methods (a) to (c) can be modified to match with replacement, when the number of untreated subjects is small, doing so may result in one untreated subject being matched repeatedly with multiple treated subjects and lead to complication in statistical tests for balance and estimating standard errors of the treatment effect. Kernel matching with replacement essentially is a weighted regression of y_i .

These matching methods assume that after conditioning on a set of observed characteristics, potential outcomes are conditionally mean independent of treatment. If unobserved systematic differences exist between treated and untreated groups, for example, if more sickly patients are more likely to select a treatment, then the above PS estimates are subject

to hidden bias. Alternatively, in a longitudinal study we can employ the difference-in-differences matching method that allows for temporally invariant differences in unobserved or unmeasured covariates.

We systematically outline the many choices faced by a researcher when she or he uses the various matching methods (Figure 1). First, although not universally accepted, some researchers (Heckman, Ichimura, Smith, & Todd, 1998) have advocated choosing nonexperimental comparison samples as close to the experimental sample as possible before estimating PSs, for example, restricting the comparison sample to be in the same age groups or geographic area as the treated sample. However, it can be argued that it is the role of the PS to remove any preintervention differences based on observed variables. The main advantage of PS methods over direct adjustment by these variables in the outcome equation of interest is reduction of dimensionality. One should avoid including posttreatment variables in estimating PS using some form of stepwise regressions. When simply guided by statistical tests for “best”-fit models we are tempted to “over” control for variables that may be affected by and measured after the treatment and hence risk the violation of the ignorability of treatment assignment assumption (Wooldridge, 2005).²

Second, all PS methods require $0 < p_i(\mathbf{x}_i) < 1$ and sufficient overlap of the estimated PS between the treated and control groups, that is, the common support problem. The common support of the estimated PS refers to the interval of PSs outside of which there are either treated subjects or comparison subjects but not both. For example, at extremely high (low) values of the PS there is no observation from the comparison (treatment) sample. Without this restriction on the support, any treated subject with very high $\hat{p}_i$ will have poorly matched controls in that the difference in PSs between the treated and comparison observations might be large, especially in single nearest-neighbor matching without replacement. Researchers have proposed different ways of constructing the common support, the simplest of which is to discard comparison observations with estimated PS below the minimum or above the maximum of the estimated PS in the treatment group. Smith and Todd (2005) demonstrated that variations in common support regions influenced estimated treatment effects.

Third, once the PSs are estimated and the matched sample is constructed, the covariate balance between the treated group and the comparison group must be assessed. There is some debate on whether the use of significance testing is appropriate for assessing balance (Austin, 2008a; Hansen, 2008). However, the assessment of imbalance using some form of significance testing has a long tradition in the field (D'Agostino, 1998; Rosenbaum & Rubin, 1984). Our view is that the test, if used, should be consistent with the matching method used to construct the matched sample. If single nearest-neighbor matching or caliper matching without replacement is used, then a paired *t* test or signed rank test can be used to compare the means or medians of continuous variables, McNemar's chi-square test can be used for binary variables, and more general test for marginal homogeneity in square tables can be used for multinomial responses (Agresti, 1990). When the number of covariates is large there is a multiple-testing problem causing inflation of Type I error. If multiple matches are selected for each treated subject, one may rely on the random intercept model for continuous variables, conditional logistic regression for binary variables, or the generalized estimating equation methods to account for the correlation between matched subjects when assessing imbalance. Simple two-sample *t* tests and chi-square tests assuming independence between

²Austin, Grootendorst, and Anderson (2007) found that, when forming propensity score estimation, including covariates that are strongly associated with the treatment exposure but not associated with the outcome will lead to inefficiency. We emphasize that their observation is under the assumption that there is no hidden bias. Indeed, these covariates are precisely instrumental variables (IV) when hidden bias exists, and the validity of IV hinges on strong association with the treatment and no association with the outcome.

observations are not suitable. When matching is done with replacement, it leads to dependency of observations that is beyond the matched pairs (or clusters). Some researchers use the reduction in standardized difference (Austin, 2008a)—measured by the mean differences as a percentage of the average standard deviation between the treated group and the comparison group—before and after matching as a gauge for the quality of matching. However, even with large reductions in standardized differences, balance in covariates may not be achieved. In the stratification or binning method, *F* tests are used for assessing differences between treatment groups by PS strata. Rosenbaum and Rubin (1984), following the seminal article by Cochran (1968), suggest that using five strata could remove about 90% of the bias, although the *F* tests may be sensitive to the number of strata used. Thus, some researchers use regression-based tests adjusted by the estimated PSs. This is the weakest form of all tests for the balancing property. Clearly, as stated by Rubin (2006), there is still substantial work to be done in carrying out diagnostic checks of balance.

Fourth, assuming that the above steps are satisfactory (perhaps after several iterations of estimating PS and assessing balance), many alternative matching estimators of average treatment effect exist as we described earlier. When using the nearest-neighbor matching with replacement, the seed for the random number generator might affect how the tied scores are handled. The most frequently used matching algorithm is greedy matching where the treated units are ordered and the nearest available match from the controls is selected after setting aside previously selected controls. Greedy matching without replacement has two defects. First, if there are many treated subjects with high values of p_i and few untreated comparisons with such values, matching without replacement will result in many poor matches. Second, the estimate depends on the order in which the observations are matched. Matching with replacement may increase precision but it may also increase variance. An optimal matching algorithm, proposed by Gu and Rosenbaum (1993), overcomes these problems by minimizing the total distance (measured in various ways) within matched sets. The choices of the radius in the caliper matching method and the bandwidth in the kernel matching are all subject to judgment.

Fifth, our review of the medical literature found that PS methods have been applied widely when the main outcome is binary or time-to-event. Logistic regression or Cox regression is typically used on the matched sample or on the unmatched sample while adjusting for the estimated PS. Recent studies have shown that PS methods result in biased estimates of conditional odds ratios for binary outcomes, biased estimates of conditional hazard ratios for time-to-event outcomes, unbiased estimates of rate ratios for count outcomes, and unbiased estimates of relative risks for binary outcomes (Austin, 2008b; Austin, Grootendorst, & Anderson, 2007). One solution is to use a collapsible estimand as an alternative where the marginal association between treatment and outcome remains invariant to the omission of another variable (Greenland, Robins, & Pearl, 1999; Pearl, 2009). Instead of an odds ratio, one can use risk difference to examine the effect of treatment on a binary outcome through linear probability models. These findings are consistent with Rosenbaum and Rubin's (1983a) original conclusion that the PS methods lead to unbiased estimates of treatment effects for linear models.

Finally, when using various PS methods, researchers may have to trade off between bias and variance of the estimates. Eliminating bias should be the primary goal of the analysis, albeit inflation in the variance could lead to incorrect conclusions about the efficacy of the intervention. Sensitivity analysis is a crucial part of assessment for bias. Assessment for overt bias can be carried out by exploring different specifications for the estimation of PS. One way to gauge the specification of PS is to use an outcome or a preintervention variable that is known to be unaffected by the treatment. When we apply PS methods using such variables as outcomes, the estimated treatment effect should be null.

Sensitivity Analysis Methods: Assessing Hidden Bias

When crucial variables are omitted from the estimation of PSs or when important confounders are, for any reason, not measured, the estimated treatment effects may exhibit positive or negative hidden biases. To assess hidden bias, various methods for sensitivity analysis have been proposed. To paraphrase Rosenbaum (2002), a sensitivity analysis examines how inferences about treatment effects would be altered by hidden biases of various magnitudes. We briefly summarize three most frequently used methods for binary outcomes.

The first method was proposed by Rosenbaum and Rubin (1983a). Recall the condition under which the PS estimates are valid requires the treatment assignment to be strongly ignorable given all observed variables $\mathbf{x}$. Suppose this assumption is violated given $\mathbf{x}$, but strong ignorability of treatment assignment holds given $\mathbf{x}$ and u , an unobserved binary covariate such as genetic predisposition for the disease of interest. Under the new assumption, PS estimates are valid if u is included in the analysis. If conclusions of treatment effects are insensitive over a range of plausible associations between u and the outcome and treatment assignment, the qualitative causal inferences are more defensible. For example, assume an unobserved variable u exists and it would double the odds of receiving the treatment and halve the odds of improvement in outcome. Now assume u is observed and added in the analysis. If this does not change the qualitative conclusion of treatment effects, then we can be more confident with the original results. Ichino, Mealli, and Nannicini (2008) proposed a simulation-based technique for this sensitivity analysis; the technique is implemented by Nannicini (2007).³

The second method for sensitivity analysis was put forth by Rosenbaum (1987). This approach involves only one sensitivity parameter, that is, the association of u with treatment assignment, as opposed to the Rosenbaum and Rubin approach (1983a) that considers the full spectrum of the distribution of u and its association with treatment and outcome. This second approach provides bounds for the significance level of the null hypothesis of no treatment effect under the assumption of strong ignorability given $\mathbf{x}$. For example, in a one-to-one matched study, without loss of generalizability, assume u is binary and denote by Γ the odds ratio of being treated for u equals to 1 versus u equals to 0. For various values of Γ the Rosenbaum (1987) approach provides the corresponding upper and lower bounds for the significance level of the original inference. A study is insensitive to hidden bias if extreme values of Γ are required to alter the conclusion. Becker and Caliendo (2007) implement this approach for binary outcomes, and DiPrete and Gangl (2004) implement it for binary as well as continuous outcomes.

The above two sensitivity methods examine the consequences of weakening the strong ignorability assumption. The third approach, proposed by Manski (1990), consists of constructing bounds for treatment effects that rely on either the outcome being bounded or on alternative identifying assumptions, such as selection rules. We use an example to clarify the relationship between the Manski bounds and the Rosenbaum and Rubin (1983a) approach. Manski (1990) shows that when outcomes are bounded, treatment effects are also bounded. In the special case of a binary outcome, the upper bound for the treatment effect among those who were treated is the expected value of the potential outcome when treated,

³The procedure is implemented in Stata by Nannicini (2007). Stata programs to estimate treatment effects include `att*` (Becker & Ichino, 2002), `psmatch2` (Leuven & Sianesi, 2003), and `nnmatch` (Abadie, Drukker, Herr, & Imbens, 2004). We have found that not all programs gave the same results for some matching estimators. The SAS macro for greedy matching is not evaluated against these programs. DiPrete and Gangl (2004) provide a program `rbounds` that tests sensitivity for continuous outcomes. Becker and Caliendo (2007) implement Rosenbaum bounds in their program `mhbounds`. Nannicini (2007) presents `sensatt` that implements the simulation-based sensitivity analysis proposed by Ichino et al. (2008).

and the lower bound is 1 less than the upper bound. It can be shown that the upper bound is achieved when all treated individuals have $u = 1$ and all controls with $u = 1$ have outcome equal to 0. The lower bound is achieved when all treated individuals have $u = 1$ and all controls with $u = 1$ have outcome equal to 1. These two sets of circumstances are the extreme cases in the Rosenbaum and Rubin (1983a) approach and are highly implausible. Because the Rosenbaum and Rubin and Rosenbaum (1987) methods are built on reasonable assumptions on associations between unobserved confounders, treatment assignment, and/or potential outcomes, these two approaches are more informative than Manski (1990) bounds. These two methods are related to the control function approach in Heckman and Navarro-Lozano (2004) where the unobserved factors are strictly continuous, specific to potential outcomes, and modeled structurally. Lin, Psaty, and Kronmal (1998) generalized Rosenbaum and Rubin's sensitivity analysis to work with both binary and censored survival time data.

Review of Recent Applications

Next, we evaluate selected applications from 2000 to 2005 in mainstream medical journals against the four criteria we proposed in the New Contribution section, namely, examination of balance, overlapping support description, use of estimated PS for evaluation of treatment effect, and sensitivity analyses in the sense of Rosenbaum and Rubin (1983a). Table 1 summarizes 27 articles in medical journals by the four criteria. Most of these articles used PS methods as a supplementary analysis and conclusions therein were seldom based on PS methods alone. Our purpose is to compare the use of PS methods with our recommended criteria, not to challenge the findings of any article.

Balancing Test

As seen in Table 2, after estimating PS, 13 studies used matched samples. These studies would have required balancing tests of observed covariates between matched pairs. One study using one-to-one matched sample examined balance appropriately by standardized differences and another study by signed rank test. Three studies did not present balancing tests. The other eight studies did not take into account the matched nature of the selected sample and carry out a t test, z test, χ^2 test, Kruskal–Wallis test, Wilcoxon rank sum test, or Fisher exact test, all assuming independent sampling. Among the four studies that used stratification methods (by quintiles, quartiles, or deciles), which should have required balancing tests within strata, only one study reported the appropriate F test, and the other three did not present balancing tests.

Overlapping Support

Among the 27 studies reviewed, only one explicitly trimmed their sample based on the common support. Additionally, three studies reported the number of matched cases and controls within each quartile or quintile of PS. Reporting the range of common support and the number of matches should be a routine practice.

Estimation of Treatment Effects

When matched designs were used (one-to-one or one-to- m matching), eight studies employed appropriate analytic methods, such as weighted average of paired differences, McNemar test, and conditional logistic regression (Table 3). Three of the eight studies and additional five studies used inappropriate analytic methods to estimate treatment effects, such as log-rank test of Kaplan–Meier curves, logistic and Cox regressions. All four studies using stratification methods appropriately derived estimates of treatment effects within each stratum. Without matching, 10 studies used the estimated PS as covariate(s) in regressions, either linearly or in dummy variables format of the PS quintiles, quartiles, or deciles. None

of these studies took into consideration that the PS was an estimated regressor. When using fitted PS, from a first-stage binary model, such as probit or logit, in the second-stage regression model, one needs to at least make the second-stage inference robust to heteroscedasticity using the Huber–White standard errors (Wooldridge, 2002). A more efficient estimator is the full information maximum likelihood method applied to the equations of both stages (Wooldridge, 2002).

Sensitivity Analysis

Only 1 out of the 27 studies carried out the sensitivity analysis of Rosenbaum and Rubin (1983a). Wang et al. (2005) found that to explain fully the increased risk of death associated with conventional antipsychotic agents compared with atypical antipsychotic medications, a hypothetical confounder would need to have very large relative risk.

It may be of interest to some readers that Wang et al. (2005) used both PS and IV methods in addition to the traditional multivariable regressions. They found that all three approaches led to the same conclusions about the association of the risk of death and conventional antipsychotic agents. For a more thorough empirical comparison of the two approaches, Landrum and Ayanian (2001) studied 18-month mortality reduction associated with ambulatory cardiology care following myocardial infarction. They demonstrated the importance of understanding each model's assumptions, carrying out sensitivity analyses, and paying attention to interpretations of results.

An Example

None of the studies reviewed above reported more than one PS method in their analyses. However, many alternative matching estimators of average treatment effect exist. In the example below, we demonstrate that different matching estimators may give substantially different results.

In a previous study, we examined the impact of being diagnosed and treated for prostate cancer on employment status for men aged between 30 and 65 years (Bradley, Neumark, Luo, Bednarek, & Schenk, 2005), using multivariable regressions where PS adjustment was made. A control group was constructed from the Current Population Survey (CPS) in the same geographic area and of the same age group. We estimated PSs for having cancer, using age, race, marital status, education, number of children, income, and occupation with four specifications. In the first specification (Model 1), we included the above covariates. In the second specification (Model 2), we added to the first specification higher order terms of the baseline continuous variables. In the third specification (Model 3), we added to the first specification interactions of age with other covariates. In the fourth specification (Model 4), we included baseline variables, higher order terms, and interactions. All four models had a high *c* statistic (.81-.88), and Model 4 out performed the other models based on the likelihood ratio test.

We stratified the data into five bins for each of the above models and used the *F* test to test group differences in all covariates between the cancer patients versus the control subjects in each model. Model 3 yielded estimated PS that balanced the two groups where all *F* statistics for baseline covariates were below the critical values. Thus, we used this specification in the analysis that followed.⁴

⁴The sample sizes in this article differ slightly from the published work because here we dropped the few observations where income was missing. The results were not substantially affected. To avoid the complication of imputation, we assumed missing completely at random here.

Figure 2 displays the predicted probability of having prostate cancer. A lack of overlap implied that there were combinations of covariate values that predicted PS in only one of the groups so that cancer patients cannot be matched to controls. The existence of unobserved variables related to cancer places the assumption of ignorability in doubt. We consider sensitivity to possible unobserved confounders later.

Table 4 presents the estimates of the effect of cancer on percent reduction in 6-month employment using the regression-based method, single nearest-neighbor matching with or without replacement, three nearest-neighbors matching, caliper matching, and kernel density matching methods. Matching estimates differed from each other, sometimes substantially (ranging from 8% to 16%). This could be because different matching methods may retain different subgroups of patients. Furthermore, single nearest neighbor with replacement and multiple nearest neighbors increased the standard errors compared with single nearest neighbor without replacement. The two kernel matching methods had the smallest standard errors. In addition, caliper matching with a radius of estimated PSs of 0.1 and stratification with five bins led to insignificant effects. Thus, simply comparing different PS methods will not give conclusive results. Additional sensitivity analysis is needed.

PS methods lead to unbiased estimates under the assumption that having cancer was independent of employment conditional on the set of observed characteristics. This assumption would be violated if unobserved variables that affected employment independently affected the likelihood of having cancer. Potential unobserved confounders in this study could be genetic disposition for cancer or having access to health insurance that offers cancer screening. Unobserved confounders can bias the results in either direction. For example, if patients with the genetic disposition for cancer were less likely to be continuously employed, the PS estimates of the average causal effect will be larger than the true effect of cancer. On the other hand, if patients having health insurance through employment had increased chance of getting screened for cancer, the PS estimates will be smaller than the true effect. We examined the robustness of the causal estimates to unobserved confounders using the Rosenbaum and Rubin (1983a) and Rosenbaum (1987) methods described above.

Following the Rosenbaum and Rubin (1983a) approach, we considered situations in which the hypothetical unobserved variable has varying degrees of associations with cancer and employment. For example, for the single nearest neighbor without replacement estimate, the estimated treatment effect is 10.5% under the assumption of no unobserved confounder ($p = .05$). Assume a binary confounder, u , exists and let the conditional expectations of u given cancer status, employment outcome, and other covariates vary between 0.1 and 0.9. We found the range of estimates was from 3.4% to 15.3%. However, to alter the inference to 3.4%, the odds ratio of u for employment would have to be 2.1 and the odds ratio of u for cancer would have to be 0.02, extremely small to be likely. To alter the inference to 15.3%, the odds ratio of u for employment would have to be 3.5 and the odds ratio of u for cancer would have to be 60.2, extremely large to be likely. In addition, when u is associated with employment at the odds ratio of 3.7 and with cancer at the odds ratio of 4.8, the simulated treatment effect is 9.9%, which is very close to the estimated effect of 10.5%. This evidence suggests that the estimated effect is robust to likely potential unobserved confounders.

Following Rosenbaum (1987), we made the matched individuals differ in their odds of having cancer as a result of a hidden confounder u and let the odds ratio go from 1 to 5, and we found that for various matching estimates to be overestimates of the effect of cancer, the odds ratio of u for cancer between matched pairs needs to be at least as large as 2 to 2.5. Note that this approach makes fewer assumptions about u and thus is less informative than the first approach.

The data above were not ideal in that the comparison sample derived from CPS was small, which violates a commonly recommended rule that the comparison sample should be at least 10 times as large as the treated sample. However, this example demonstrated that not all PS estimates were equivalent, and one should always provide sensitivity analyses. Because we do not have good instruments for cancer diagnosis and treatment, we cannot demonstrate the use of IV approach or sample selection models. But the sensitivity analysis provides partial evidence that it may not be necessary to use IV or sample selection correction in this example. Sensitivity analyses of the Rosenbaum and Rubin (1983a) and the Rosenbaum (1987) forms are very informative to the plausibility of the inference.

Conclusions

We summarized the differences between RCTs, PS methods, and IV methods; described some criteria for appropriately using PS methods; and reviewed some articles in leading medical journals against these criteria. Most applications presented only certain aspects of the PS results, such as using *c* statistics to gauge the discrimination of the logistic regression of estimating PS. We reiterate the emphasis in Austin (2008a) that the matched nature of the sample should be taken into account when performing balancing tests and estimating final treatment effects. The abundance of choices in using PS matching methods may lead to divergent results, suggesting that researchers should examine several matching methods for sensitivity. After each matching method, one should make explicit who the matched subjects were in relation to the total sample. Single nearest-neighbor matching method is especially sensitive. Often one has to eliminate many observations without a match and there is no clear guideline as to which matching methods to use when they produce different estimates.

For the purpose of evaluating treatment effects, RCT and nonexperimental data analyses rely on different assumptions to estimate counterfactual outcomes. RCT assumes randomization balances both observed and unobserved factors. PS methods assume strong ignorability of treatment assignment, or selection on observables. IV methods assume valid IVs exist. No matter which method a researcher employs, the internal validity of the method needs to be carefully examined. In RCT, this involves meticulous implementation of randomization, strict adherence to treatment protocol by providers, compliance by patients, and no differential attrition between groups. In PS methods, to ensure validity involves careful consideration of the research question that can be answered with the selected treatment and control population, diligent check of balance in observed variables, and scrupulous examination of the plausibility of strong ignorability of treatment assignment, and lucid presentation and interpretation of the results. In IV methods, not all assumptions concerning the validity of instruments are testable, just as in PS the strong ignorability assumption is not testable. Whenever possible, the overidentification test should be performed and weak instruments should be avoided.

In a nonexperimental study, a sensitivity analysis assessing the degree to which the qualitative conclusions are affected due to hidden bias is crucial. Researchers should carry out the Rosenbaum and Rubin (1983a) type of sensitivity analyses and examine the plausibility of all assumptions in their study settings. When a sensitivity analysis indicates that an unobserved confounder may influence the substantive conclusion, IV methods have the advantage of being able to resolve some hidden bias issues if good instruments can be found, whereas PS methods may or may not. Pearl (2009), in his reflection on the controversy surrounding PS methods (Smith & Todd, 2005), noted that there is a tendency among investigators to play down the cautionary note concerning the required conditions for PS estimates to be valid and to interpret the mathematical proof of PS methods as a guarantee to eliminate confounding. Investigators are tempted to assume that strong ignorability is satisfied or likely to be satisfied if one includes as many covariates as

possible. Pearl points out that it is not enough to warn practitioners against the dangers they cannot recognize, but rather, they must be given “eyeglasses to spot the threats and a meaningful language to reason about them” (Pearl, 2009, p. 352). In spite of the many pitfalls, when appropriately evaluated and applied, PS methods can be powerful tools in assessing average treatment effects in observational studies. Appropriate PS applications can create experimental conditions using observational data when RCTs are not feasible and, thus, lead researchers to an efficient estimator of the average treatment effect.

Acknowledgments

We thank the two anonymous reviewers and the editor for providing extensive and valuable comments and suggestions that have led to the present version.

Funding: The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: Agency for Healthcare Research & Quality (Grant R01 HS14206) and The National Cancer Institute (Grant R01 CA86045-01).

References

- Abadie A, Drukker D, Leber Herr J, Imbens GW. Implementing matching estimators for average treatment effects in Stata. *Stata Journal*. 2004; 4:290–311.
- Abidov A, Rozanski A, Hachamovitch R, Hayes SW, Aboul-Enein F, Cohen I, et al. Prognostic significance of dyspnea in patients referred for cardiac stress testing. *New England Journal of Medicine*. 2005; 353:1889–1898. [PubMed: 16267320]
- Agresti, A. *Categorical data analysis*. New York: Wiley; 1990.
- Angrist, JD.; Pischke, J. *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press; 2009.
- Aronow HD, Topol EJ, Roe MT, Houghtaling PL, Wolski KE, Lincoff AM, et al. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: An observational study. *Lancet*. 2001; 357:1063–1068. [PubMed: 11297956]
- Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*. 2008a; 27:2037–2049. [PubMed: 18038446]
- Austin PC. The performance of different propensity-score methods for estimating relative risks. *Journal of Clinical Epidemiology*. 2008b; 61:537–545. [PubMed: 18471657]
- Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*. 2007; 26:734–753. [PubMed: 16708349]
- Austin PC, Grootendorst P, Normand SL, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Statistics in Medicine*. 2007; 26:754–768. [PubMed: 16783757]
- Ayanian JZ, Landrum MB, Guadagnoli E, Gaccione P. Specialty of ambulatory care physicians and mortality among elderly patients after myocardial infarction. *New England Journal of Medicine*. 2002; 347:1678–1686. [PubMed: 12444183]
- Babaev A, Frederick PD, Pasta DJ, Every N, Sichrovsky T, Hochman JS. Trends in management and outcomes of patients with acute myocardial infarction complicated by cardiogenic shock. *Journal of the American Medical Association*. 2005; 294:448–454. [PubMed: 16046651]
- Baser O. Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*. 2006; 9:377–385. [PubMed: 17076868]
- Baser O. Too much ado about instrumental variable approach: Is the cure worse than the disease? *Value in Health*. 2009; 12:1201–1209. [PubMed: 19497084]
- Becker SO, Caliendo M. Sensitivity analysis for average treatment effects. *Stata Journal*. 2007; 7:71–83.
- Becker SO, Ichino A. Estimation of average treatment effects based on propensity scores. *Stata Journal*. 2002; 2:358–377.

- Bhatt DL, Roe MT, Peterson ED, Li Y, Chen AY, Harrington RA, et al. Utilization of early invasive management strategies for high-risk patients with non-ST-segment elevation acute coronary syndromes: Results from the CRUSADE quality improvement initiative. *Journal of the American Medical Association*. 2004; 292:2096–2104. [PubMed: 15523070]
- Bound J, Jaeger DA, Baker RM. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*. 1995; 90:443–450.
- Bradley CJ, Neumark D, Luo Z, Bednarek H, Schenk M. Employment outcomes of men treated for prostate cancer. *Journal of National Cancer Institute*. 2005; 97:958–965.
- Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968; 24:295–313. [PubMed: 5683871]
- Cook, TD.; Campbell, DT. *Quasi-experimentation: Design and analysis issues for field settings*. Chicago: Rand McNally; 1979.
- D'Agostino RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine*. 1998; 17:2265–2281. [PubMed: 9802183]
- DiPrete T, Gangl M. Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*. 2004; 34:271–310.
- Ferguson TB, Coombs LP, Peterson ED. Preoperative beta-blocker use and mortality and morbidity following CABG surgery in North America. *Journal of the American Medical Association*. 2002; 287:2221–2227. [PubMed: 11980522]
- Frolkis JP, Pothier CE, Blackstone EH, Lauer MS. Frequent ventricular ectopy after exercise as a predictor of death. *New England Journal of Medicine*. 2003; 348:781–790. [PubMed: 12606732]
- Gardiner JC, Luo ZH, Roman LA. Fixed effects, random effects and GEE: What are the differences? *Statistics in Medicine*. 2009; 28:221–239. [PubMed: 19012297]
- Girou E, Brun-Buisson C, Taille S, Lemaire F, Brochard L. Secular trends in nosocomial infections and mortality associated with noninvasive ventilation in patients with exacerbation of COPD and pulmonary edema. *Journal of the American Medical Association*. 2003; 290:2985–2991. [PubMed: 14665660]
- Greenland S. An introduction to instrumental variables for epidemiologists. *International Journal of Epidemiology*. 2000; 29:722–729. [PubMed: 10922351]
- Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science*. 1999; 14:29–46.
- Grzybowski M, Clements EA, Parsons L, Welch R, Tintinalli AT, Ross MA, et al. Mortality benefit of immediate revascularization of acute ST-segment elevation myocardial infarction in patients with contraindications to thrombolytic therapy: A propensity analysis. *Journal of the American Medical Association*. 2003; 290:1891–1898. [PubMed: 14532318]
- Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Computational and Graphical Statistics*. 1993; 2:405–420.
- Gum PA, Thamaras M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: A propensity analysis. *Journal of the American Medical Association*. 2001; 286:1187–1194. [PubMed: 11559263]
- Hansen, BB. *Statistics in Medicine*. Vol. 27. 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*; p. 2050-2054. p. 2066-2069.
- Hannan EL, Racz MJ, Walford G, Jones RH, Ryan TJ, Bennett E, et al. Long-term outcomes of coronary-artery bypass grafting versus stent implantation. *New England Journal of Medicine*. 2005; 352:2174–2183. [PubMed: 15917382]
- Heckman JJ. Rejoinder: Response to Sobel. *Sociological Methodology*. 2005a; 35:135–162.
- Heckman JJ. The scientific model of causality. *Sociological Methodology*. 2005b; 35:1–97.
- Heckman JJ, Ichimura H, Smith J, Todd P. Characterizing selection bias using experimental data. *Econometrica*. 1998; 66:1017–1098.

- Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: Evidence from evaluating a job training program. *Review of Economic Studies*. 1997; 64:605–654.
- Heckman JJ, Navarro-Lozano S. Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics*. 2004; 86:30–57.
- Heckman JJ, Smith JA. Assessing the case for social experiments. *Journal of Economic Perspectives*. 1995; 9:85–110.
- Heckman, JJ.; Vytlačil, EJ. Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman, JJ.; Leamer, E., editors. *Handbook of econometrics*. Vol. 6B, chap. 71. Amsterdam: Elsevier; 2007.
- Hsueh EC, Essner R, Foshag LJ, Ye X, Wang HJ, Morton DL. Prolonged survival after complete resection of metastases from intraocular melanoma. *Cancer*. 2004; 100:122–129. [PubMed: 14692032]
- Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Quarterly*. 2004; 82:661–687. [PubMed: 15595946]
- Ichino A, Mealli F, Nannicini T. From temporary help jobs to permanent employment: What can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*. 2008; 23:305–327.
- Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*. 2004; 86:4–29.
- Imbens GW, Angrist JD. Identification and estimation of local average treatment effects. *Econometrica*. 1994; 62:467–475.
- Imbens GW, Wooldridge JM. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*. 2009; 47:5–86.
- Landrum MB, Ayanian JZ. Causal effect of ambulatory specialty care on mortality following myocardial infarction: A comparison of propensity score and instrumental variable analyses. *Health Services and Outcomes Research Methodology*. 2001; 2:221–245.
- Leuven, E.; Sianesi, B. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. Chestnut Hill, MA: Boston College, Department of Economics; 2003.
- Lin DY, Psaty BM, Kronmal RA. Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*. 1998; 54:948–963. [PubMed: 9750244]
- Lindenauer PK, Pekow P, Wang KJ, Mamidi DK, Gutierrez B, Benjamin EM. Perioperative beta-blocker therapy and mortality after major noncardiac surgery. *New England Journal of Medicine*. 2005; 353:349–361. [PubMed: 16049209]
- Manski CF. Nonparametric bounds on treatment effects. *American Economic Review*. 1990; 80:319–323.
- Mehta RL, Pascual MT, Soroko S, Chertow GM. Diuretics, mortality, and nonrecovery of renal function in acute renal. *Journal of the American Medical Association*. 2002; 288:2547–2553. [PubMed: 12444861]
- Mitra N, Schnabel FR, Neugut AI, Heitjan DF. Estimating the effect of an intensive surveillance program on stage of breast carcinoma at diagnosis: A propensity score analysis. *Cancer*. 2001; 91:1709–1715. [PubMed: 11335895]
- Mukamal KJ, Maclure M, Muller JE, Sherwood JB, Mittleman MA. Prior alcohol consumption and mortality following acute myocardial infarction. *Journal of the American Medical Association*. 2001; 285:1965–1970. [PubMed: 11308432]
- Nannicini T. Simulation-based sensitivity analysis for matching estimators. *Stata Journal*. 2007; 7:334–3350.
- Newby LK, Kristinsson A, Bhapkar MV, Aylward PE, Dimas AP, Klein WW, et al. Early statin initiation and outcomes in patients with acute coronary syndromes. *Journal of the American Medical Association*. 2002; 287:3087–3095. [PubMed: 12069671]
- Orosz GM, Magaziner J, Hannan EL, Morrison RS, Koval K, Gilbert M, et al. Association of timing of surgery for hip fracture and patient outcomes. *Journal of the American Medical Association*. 2004; 291:1738–1743. [PubMed: 15082701]

- Pearl, J. Causality: Models, reasoning, and inference. New York: Cambridge University Press; 2009.
- Petersen LA, Normand ST, Daley J, McNeil BJ. Outcome of myocardial infarction in Veterans Health Administration patients as compared with Medicare patients. *New England Journal of Medicine*. 2000; 343:1934–1941. [PubMed: 11136265]
- Polanczyk CA, Rohde LE, Goldman L, Cook EF, Thomas EJ, Marcantonio ER, et al. Right heart catheterization and cardiac complications in patients undergoing noncardiac surgery: An observational study. *Journal of the American Medical Association*. 2001; 286:309–314. [PubMed: 11466096]
- Rao SV, Jollis JG, Harrington RA, Granger CB, Newby LK, Armstrong PW, et al. Relationship of blood transfusion and clinical outcomes in patients with acute coronary syndromes. *Journal of the American Medical Association*. 2004; 292:1555–1562. [PubMed: 15467057]
- Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*. 1987; 74:13–26.
- Rosenbaum, PR. *Observational Studies*. New York: Springer; 2002.
- Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 1983a; 45:212–218.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983b; 70:41–55.
- Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*. 1984; 79:516–524.
- Rubin DB. On principles for modeling propensity scores in medical research. *Pharmacoepidemiology and Drug Safety*. 2004; 13:855–857.
- Rubin, DB. *Matched sampling for causal effects*. New York: Cambridge University Press; 2006.
- Schneeweiss S, Walker AM, Glynn RJ, Maclure M, Dormuth C, Soumerai SB. Outcomes of reference pricing for angiotensin-converting-enzyme inhibitors. *New England Journal of Medicine*. 2002; 346:822–829. [PubMed: 11893794]
- Smith JA, Todd PE. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*. 2005; 125:305–353.
- Sobel ME. Discussion: the scientific model of causality. *Sociological Methodology*. 2005; 35:99–133.
- Staiger D, Stock JH. Instrumental variables regression with weak instruments. *Econometrica*. 1997; 65:557–586.
- Stenestrand U, Wallentin L. Early revascularisation and 1-year survival in 14-day survivors of acute myocardial infarction: A prospective cohort study. *Lancet*. 2002; 359:1805–1811. [PubMed: 12044375]
- Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the applications of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. *Journal of Clinical Epidemiology*. 2006; 59:437–447. [PubMed: 16632131]
- Vikram HR, Buenconsejo J, Hasbun R, Quagliarello VJ. Impact of valve surgery on 6-month mortality in adults with complicated, left-sided native valve endocarditis: A propensity analysis. *Journal of the American Medical Association*. 2003; 290:3207–3214. [PubMed: 14693873]
- Vincent JL, Baron JF, Reinhart K, Gattinoni L, Thijs L, Webb A, et al. Anemia and blood transfusion in critically ill patients. *Journal of the American Medical Association*. 2002; 288:1499–1507. [PubMed: 12243637]
- Wang PS, Schneeweiss S, Avorn J, Fischer MA, Mogun H, Solomon DH, et al. Risk of death in elderly users of conventional vs. atypical antipsychotic medications. *New England Journal of Medicine*. 2005; 353:2335–2341. [PubMed: 16319382]
- Wassertheil-Smoller S, Psaty B, Greenland P, Oberman A, Kotchen T, Mouton C, et al. Association between cardiovascular outcomes and antihypertensive drug treatment in older women. *Journal of the American Medical Association*. 2004; 292:2849–2859. [PubMed: 15598916]
- Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmaco-epidemiology & Drug Safety*. 2004; 13:841–853.

- Wooldridge, JM. Econometric analysis of cross section and panel data. Cambridge: MIT Press; 2002.
- Wooldridge JM. Violating ignorability of treatment by controlling for too many factors. *Econometric Theory*. 2005; 21:1026–1028.

Figure 1.
The process of evaluating nonexperimental estimates

Figure 2.

Boxplot of estimated propensity scores after trimming

Note: The box encloses the interquartile range (IQR) from the 25th to 75th percentile of the estimated propensity scores (PS) in each group; the line in the middle of the box represents the median; the ends of the whiskers are the most extreme PS within one and half IQR of the nearest quartile, that is, the top whisker = the largest available PS that is less than the 75th percentile plus 1.5 IQR, and the bottom whisker = the smallest available PS that is greater than the 25th percentile minus 1.5 IQR; and the dots are outliers.

Table 1
Evaluation of Selected Applications of Propensity Score (PS) Methods in Leading Medical Journals

Study	Balance Test for Covariates and Residual Imbalance Controlled if Necessary	Use of Estimated PS	Overlap of PS Presented	Unobserved Confounder Sensitivity Analysis
Petersen, Normand, Daley, and McNeil (2000)	Standardized difference. Balanced, all covariates with standard difference less than 5%	Single nearest neighbor within strata by age and hospital types. Average difference equals precision weighted paired differences	No	No
Ayanian, Landrum, Guadagnoli, and Gaccione (2002)	t test and Pearson χ^2 test; balanced based on p values	Single nearest neighbor within 0.6 standard deviation of estimated logits. McNemar test for mortality and log-rank test for Kaplan-Meier curves among matched pairs	No	No
Schneeweiss et al. (2002)	Balancing test not available	Quintile dummies of PS in log-linear regressions	No	No
Frolikis, Pothier, Blackstone, and Lauer (2003)	Kruskal-Wallis and χ^2 tests; balanced, based on p values	Single nearest neighbor. Log-rank test for Kaplan(x2013)Meier curves among matched pairs	No	No
Abidov et al. (2005)	Balancing test not available	One-to-one match. Cox regression on matched pairs	No	No
Hannan et al. (2005)	Balancing test not available	Hazard ratio in each quintile of PS	No	No
Lindenauner et al. (2005)	z test and χ^2 test; unbalanced, residual unbalance controlled	Greedy matching (up to five digits) with up to two controls, conditional logistic regression using matched sample	No	No
Wang et al. (2005)	Balancing test not available, covariates controlled	Stratified Cox regression by deciles of PS, Cox regression with deciles of PS as covariate	No	Yes
Aronow et al. (2001)	Within decile of PS compared, three unbalanced deciles excluded, covariates controlled	PS linearly entered in Cox regression	No	No
Stenestrand and Wallentin (2002)	Balancing test not available, covariates controlled	Cox regression within quartiles of PS	Number of T and C in each quartile	No
Babaev et al. (2005)	Balancing test not available, covariates controlled	PS linearly entered in logistic regression	No	No
Bhatt et al. (2004)	Wilcoxon rank-sum test and χ^2 test, 1 out of 31 variables unbalanced	Greedy matching, difference in mortality calculated in matched sample	No	No
Ferguson, Coombs, and Peterson (2002)	Signed rank test and χ^2 test, 7 out of 21 variables unbalanced	One-to-one match, conditional logistic regression using matched pairs	No	No
Girou, Brun-Buisson, Taille, Lemaire, and Brochard (2003)	Within quintile of PS compared, balance stated but not available	PS linearly entered in logistic regression	Number of T and C in each quartile	No
Grzybowski et al. (2003)	Wilcoxon rank-sum test and χ^2 test, balanced based on p values	Greedy matching on eight digits, PS linearly entered in logistic regression on matched sample, difference in mortality calculated by PS quintiles	No	No
Gum, Thamilarasan, Watanabe, Blackstone, and Lauer (2001)	Wilcoxon rank-sum test and χ^2 test, 1 out of 31 covariates unbalanced, covariates controlled	Greedy matching up to five digits; Cox regression using matched sample with or without PS entered linearly	No	No

Study	Balance Test for Covariates and Residual Imbalance Controlled if Necessary	Use of Estimated PS	Overlap of PS Presented	Unobserved Confounder Sensitivity Analysis
Mehta, Pascual, Soroko, and Chertow (2002)	Balancing test not available, covariates controlled	PS entered linearly in logistic regression	No	No
Mukamal, MacLure, Muller, Sherwood, and Mittleman(2001)	Balancing test not available, covariates controlled	PS entered linearly or quintile or decile dummies in Cox regression	No	No
Newby et al. (2002)	Balancing test not available, covariates controlled	PS entered linearly in logistic regression	Number of T and C in each quintile	No
Orosz et al. (2004)	Statement of balance declared	Single nearest neighbor within 0.1 PS and closest age. Difference or odds ratio of matched pairs	No	No
Polanczyk et al.(2001)	Wilcoxon signed rank test, balanced based on p values, covariates controlled	Single nearest neighbor within 0.03 PS. Conditional logistic regression of matched pairs	No	No
Rao et al. (2004)	Balancing test not available, covariates controlled	PS entered linearly in Cox regression	No	No
Vikram, Buenconsejo, Hasbun, and Quagliariello (2003)	t test, Wilcoxon rank sum test, χ^2 and Fisher exact test, balanced based on p values, covariates controlled	Greedy matching up to five digits, PS entered linearly in Cox regression	No	No
Vincent et al. (2002)	t test, χ^2 test, balanced based on p values	Greedy matching up to six digits, mortality difference, and Kaplan-Meier of matched pairs	No	No
Wassertheil-Smolter et al. (2004)	Balancing test not available, covariates controlled	PS quintile dummies in Cox regression	No	No
Hsueh et al. (2004)	t test, χ^2 and Wilcoxon rank sum test, one out of six covariates unbalanced	PS entered linearly in Cox regression	Overlapping PS used	No
Mitra, Schnabel, Neugut, and Heitjan (2001)	F test in two-way analysis of variance of PS quintiles and treatment, balanced	Logistic regression in each quintile	No	No

Table 2
Balancing Tests With Matched or Stratified Sample

Study	Matched or Stratified Sample	Balancing Test	Congruent to Design
Petersen et al. (2000), Ferguson et al. (2002)	One-to-one match	Standardized differences, signed rank test	Yes
Ayanian et al. (2002), Frolkis et al. (2003), Bhatt et al. (2004), Grzybowski et al. (2003), Gum et al. (2001), Polanczyk et al. (2001), Vikram et al. (2003), Vincent et al. (2002)	One-to-one match	<i>t</i> test, <i>z</i> test, χ^2 test, Kruskal–Wallis test, Wilcoxon rank sum test, Fisher exact test	No
Abidov et al. (2005), Orosz et al. (2004) ^a	One-to-one match	Not presented	Unknown
Lindenauer et al. (2005)	One-to- <i>m</i> match	<i>z</i> test, χ^2 test	No
Mitra et al. (2001)	Stratified	<i>F</i> test	Yes
Hannan et al. (2005), Wang et al. (2005), Stenestrand and Wallentin (2002)	Stratified	Not presented	Unknown

^aStatement about balance was declared.

Table 3
Use of Estimated Propensity Score (PS) in Analyses for Outcomes of Interest

Study	Sample	Estimation of Treatment Effects	Congruent to Design
Petersen et al. (2000), Ayanian et al. (2002), Lindenauer et al. (2005), Bhatt et al. (2004), Ferguson et al. (2002), Orosz et al. (2004), Polanczyk et al. (2001), Vincent et al. (2002)	Matched	Weighted average of paired differences, McNemar test, conditional logistic regression	Yes
Ayanian et al. (2002), Frolkis et al. (2003), Abidov et al. (2005), Grzybowski et al. (2003), Gum et al. (2001), Vikram et al. (2003), Orosz et al. (2004), Vincent et al. (2002)	Matched	Log-rank test of Kaplan–Meier curves, logistic or Cox regressions	No
Schneeweiss et al. (2002), Aronow et al. (2001), Bahaev et al. (2005), Girou et al. (2003), Mehta et al. (2002), Mukamal et al. (2001), Newby et al. (2002), Rao et al. (2004), Wassertheil-Smoller et al. (2004), Hsueh et al. (2004)	Unmatched	Log-linear, logistic, or Cox regressions	Yes
Hannan et al. (2005), Wang et al. (2005), Stenestrand and Wallentin (2002), Mitra et al. (2001)	Stratified	Estimates within strata	Yes

Table 4
Effect of Prostate Cancer on Employment Probability at 6 Months After Diagnosis
(Expressed as Percentage Points)

Method	Estimate	Bootstrap Standard Error	Bootstrap 95% Confidence Interval	<i>p</i> Value
Regression	-8.95	4.17	(-17.15, -0.75)	.032
Single nearest neighbor, without replacement ^a	-10.48	5.31	(-16.81, 0.93)	.050
Single nearest neighbor, with replacement ^b	-12.05	5.86	(-23.53, -0.56)	.040
Three nearest neighbors ^b	-11.38	5.64	(-22.43, -0.32)	.044
Caliper matching (0.1) ^a	-10.83	4.70	(-23.92, -2.25)	.022
Caliper matching (0.2) ^b	-12.05	5.66	(-23.15, -0.95)	.033
Kernel matching (0.06) ^a	-8.70	5.31	(-16.84, 8.58)	.103
Kernel matching (0.4) ^b	-14.74	3.49	(-21.58, -7.90)	<.01
Stratified (five bins) ^a	-8.21	5.94	(-19.93, 3.51)	.167

^a Stata program to estimate treatment effects used is *att** (Becker & Ichino, 2002) with bias-corrected bootstrap confidence interval. The *p* value is based on *t* test with bootstrap standard error.

^b Stata program to estimate treatment effects used is *psmatch2* (Leuven & Sianesi, 2003) with bootstrap confidence interval.