Propensity Scores as a Basis for Equating Groups: Basic Principles and Application in Clinical Treatment Outcome Research

Stephen G. West
Arizona State University

Heining Cham
Fordham University

Felix Thoemmes
Cornell University

Babette Renneberg, Julian Schulze, and Matthias Weiler
Freie Universität Berlin

A propensity score is the probability that a participant is assigned to the treatment group based on a set of baseline covariates. Propensity scores provide an excellent basis for equating treatment groups on a large set of covariates when randomization is not possible. This article provides a nontechnical introduction to propensity scores for clinical researchers. If all important covariates are measured, then methods that equate on propensity scores can achieve balance on a large set of covariates that mimics that achieved by a randomized experiment. We present an illustration of the steps in the construction and checking of propensity scores in a study of the effectiveness of a health coach versus treatment as usual on the well-being of seriously ill individuals. We then consider alternative methods of equating groups on propensity scores and estimating treatment effects including matching, stratification, weighting, and analysis of covariance. We illustrate a sensitivity analysis that can probe for the potential effects of omitted covariates on the estimate of the causal effect. Finally, we briefly consider several practical and theoretical issues in the use of propensity scores in applied settings. Propensity score methods have advantages over alternative approaches to equating groups particularly when the treatment and control groups do not fully overlap, and there are nonlinear relationships between covariates and the outcome.

Keywords: nonrandomized study, propensity scores, matching, group equating

Propensity scores provide a basis for equating groups on a large number of covariates measured at baseline. When participants cannot be randomized to treatment conditions, familiar exact matching procedures permit equating of groups on only a small number of covariates, given the difficulties of finding good matches. Propensity scores provide a basis for overcoming this limitation. Using matching, stratification, weighting, or analysis of covariance procedures described later, researchers can equate the treatment (t) and control (c) groups at baseline on a large number of important covariates and compare the equated groups on the outcome variable Y. In clinical treatment outcome studies in which randomization is not possible or breaks down, such comparisons enhance internal validity in the estimation of treatment effects by ruling out potential confounding variables (Shadish, Cook, & Campbell, 2002; West, Cham, & Liu, 2014). Propensity scores can also provide a basis for ruling out the potential influence of confounding factors in the estimation of mediation effects (Jo, Stuart, MacKinnon, & Vinokur, 2011); here, the researcher’s interest is in understanding whether the effect of t on the outcome Y operates through an intermediate variable, providing evidence about the processes underlying the observed effect of treatment on the outcome. In terms of external validity, propensity scores can be used to construct weighting schemes for generalizing treatment effects to the population of interest when participants from a known population choose to enroll or not enroll in a study (Stuart, Cole, Bradhaw, & Leaf, 2011). Given their usefulness and versatility, there has been an exponential increase in applications of propensity scores in the health sciences in the 1990s (Thoemmes & Kim, 2011). A similar increase appears to have begun more recently in psychology. However, propensity scores are not yet well understood by most clinical scientists.

The goal in this article is to provide clinical scientists with an introduction to propensity scores, their strengths, and limitations. Our focus throughout the article is on a typical study in which a t and a c group are compared with the goal of estimating a causal effect of the treatment on the outcome variable Y. We use the terms...
treatment and intervention interchangeably in this article. We initially provide a brief overview of the basic ideas of causal inference from the perspective of the potential outcomes model (Holland, 1986; Rubin, 1978, 2005). We consider how we can mimic the randomized experiment, the gold standard design, in a nonrandomized observational study (or nonequivalent control group design; Shadish et al., 2002). Propensity scores provide a vehicle through which the t and c groups can be equated at baseline on a large number of measured covariates. We then consider the theory of propensity scores and its assumptions. Once constructed, propensity scores can be used to equate groups in several ways. We present an illustration of the construction of propensity scores, equating groups, and estimation of treatment effects from a study evaluating the effectiveness of a telephone health counseling intervention for seriously ill patients (Renneberg, Böhme, Schulze, & Weiler, 2013). Finally, we briefly consider practical and theoretical issues that can arise in the use of propensity scores in practice. We believe that propensity scores provide a promising method of equating groups, but one that requires care and a critical scientific perspective to achieve the best possible (least biased) estimate of the treatment effect and an appropriate understanding of the uncertainty associated with that estimate.

### The Potential Outcomes Model

Propensity scores were initially developed (Rosenbaum & Rubin, 1983) and are frequently interpreted in the context of the potential outcomes model (Holland, 1986; Rubin, 1978, 2005; West & Thoemmes, 2010), a formal statistical model of causal inference. The key feature of the potential outcomes model is that, prior to a study, each participant has two potential outcomes that could occur: one outcome, $Y_t(i)$, that would be observed if individual $i$ were given $t$, and one outcome, $Y_c(i)$, that would be observed if individual $i$ were given $c$. $Y_t(i) - Y_c(i)$ represents the hypothetical causal effect for individual $i$; each individual may have a different causal effect. The hypothetical comparison of the outcomes $Y_t(i)$ and $Y_c(i)$ for a single participant would ideally occur at the identical time, in the identical context, and with the participant in the identical pretreatment state. These conditions cannot be realized. In practice, researchers often attempt to equate a group of participants given $t$ with a group of participants given $c$ at baseline prior to delivery of any treatment. If the equating of groups is successful, this strategy permits the estimation of a causal effect. As described later, one of two different causal effects is commonly estimated: the average treatment effect (ATE), or the average treatment effect on the treated participants (ATT). One of two general approaches is typically taken to equating the $t$ and $c$ groups at baseline. The preferred "gold standard" approach is randomization. In randomization, a sample of participants is selected and then assigned to the $t$ and $c$ conditions with a specified a priori probability. Most commonly, an equal number of participants are assigned to the $t$ and $c$ conditions, $P(T = t) = 0.5; P(T = c) = 0.5$, where $P$ represents probability and $T$ represents treatment assignment. As developed in statistical theory, randomization guarantees that the means of the $t$ and $c$ conditions are equal in expectation (in large samples, technically when sample size $N = \infty$) on all possible baseline covariates, whether measured or unmeasured. This theoretical result means that the $t$ and $c$ groups can be considered to be equated, on average, on all covariates at baseline. Given the theoretical absence of confounding, the average treatment effect, $ATE = \bar{Y}_t - \bar{Y}_c$, can be used as the estimate of the average causal effect. In finite samples, baseline differences may exist in practice, but if $N$ is of reasonable size, these differences typically are relatively small, and these differences would not be systematically replicated in future experiments.

From the perspective of the basic potential outcomes model (Rubin, 2005), the interpretation of the causal effect as the ATE in a randomized experiment rests on four additional assumptions: (a) participants adhere fully with the treatment assignment, (b) all participants are successfully measured on the outcome at posttest, (c) there is no interference in which the outcome of one participant affects that of another participant (each participant can have only one potential outcome), and (d) there are no hidden variations of the treatment conditions. If assumption (a) is violated, the interpretation of the nature of the causal effect may be changed (see later discussion). If assumption (b) is violated, the initial equating of the $t$ and $c$ groups may be undermined. If assumption (c) or (d) is violated, there may be more than one potential outcome associated with $t$ and more than one with $c$ for each participant. Randomized experiments in which one or more of the assumptions are violated have been termed broken randomized experiments (Barnard, Du, Hill, & Rubin, 1998). Many studies planned as randomized clinical trials turn into broken randomized experiments, requiring additional adjustment to produce an unbiased estimate of the average treatment effect.

Although random assignment typically represents the gold standard, it cannot always be implemented. Random assignment may not be possible because of practical or ethical concerns (West et al., 2008). Participants cannot be randomly assigned to exposure to a major hurricane; children cannot be randomly assigned to be exposed to the death of a parent. In addition, sometimes participants are willing to be randomized but are so unrepresentative of the desired clinical population that the estimate of the causal effect may be of little interest (e.g., evaluations of Alcoholics Anonymous programs; Kownacki & Shadish, 1999). In practice, randomization is more commonly implemented in clinical efficacy trials than in evaluations of the effectiveness of actual clinical programs.

An alternative, less preferred, but commonly used approach to equating the $t$ and $c$ groups at baseline is matching. In matching, the strategy is to attempt to identify all baseline covariates that the researcher believes might cause both treatment assignment ($t$ vs. $c$) and the outcome variable. Such baseline covariates are termed confounders (see Figure 1A). Covariates that are only related to the outcome (independent cause, Figure 1B) or are only related to treatment assignment (instrument, Figure 1C) do not confound the estimate of the causal effect. If all confounders can be identified

---

1. Another approach is the application of the within-subject design in both the $t$ and $c$ conditions to each individual (Holland, 1986). This approach assumes (a) temporal stability, that the value of $Y_t(i)$ or $Y_c(i)$ will be identical regardless of when it is measured (e.g., no history or maturation effects) and (b) causal transience, the value of $Y_t(i)$ or $Y_c(i)$ is not affected by exposure to the other condition (no carryover effects).

2. The traditional intention to treat analysis estimates the average effect of treatment assignment. The potential outcomes model emphasizes the average treatment effect on the treated (described later).

3. Methods for addressing violations of each of these assumptions exist (see West et al., 2014). Assumptions (c) and (d) are collectively referred to as the stable unit treatment value assumption.
This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of individual users and is not to be disseminated publicly.

Figure 1. Three forms of relationship among the baseline covariate X, the treatment condition T (t vs. c), and the outcome variable Y. In Panel A, X is a confounder that causes both T and Y. X must be controlled to achieve an unbiased estimate of the causal effect. In Panel B, X is an independent cause of the outcome Y but has no direct relationship to T. X does not need to be controlled to achieve an unbiased estimate of the causal effect. In Panel C, X directly causes T but has no direct relationship to Y. X is an instrument and does not need to be controlled to achieve an unbiased estimate of the causal effect.

and successfully equated, then an unbiased estimate of the average causal effect can be produced. The key challenge is that the researcher must now specify all of the important confounding variables. Traditional matching approaches that equate groups on a few convenient demographic variables such as gender, age, and ethnicity will not suffice. Rubin (2001) has recommended an inclusive strategy in which as many potential confounders as possible are measured at baseline. This strategy gives rise to a practical problem: How can good matches be achieved when there are many baseline covariates? For example, in attempting to use traditional exact matching procedures using six variables (age, sex, nationality of parents, father’s occupation, neighborhood status, and high school grades) on a sample of 1,194 participants (671 treatment and 523 control), Chapin (1947) found only 23 pairs that matched exactly. Exact matching methods quickly become impractical with more than a few covariates given sample sizes typically achievable in clinical research (median N = 208 for a sample of studies reported in three issues of the 2010 volume of Journal of Consulting and Counseling Psychology; West, Aiken, Cham, & Liu, 2013).

Propensity scores provide an alternative approach that permits matching on large numbers of baseline covariates. The key idea is that the researcher can match on a single score that represents a summary of all of the information contained in the full set of covariates. The goal is to create balance on each of the covariates in the two groups in a manner that parallels the balance of covariates that would be achieved in a randomized experiment rather than to create exactly matched pairs of observations.

The propensity score is the predicted probability that the person will be assigned to the treatment group based on his or her scores on each of the full set of covariates, \( P(T = t | X_1, X_2, \ldots, X_p) \), where \( T = t \) represents assignment to the treatment condition, and each of the \( X \)s represents a baseline covariate. Propensity scores range strictly between 0 and 1. Propensity scores are typically estimated using linear logistic regression where the outcome is treatment assignment \( T (= t \ or \ c) \), and the predictors are the full set of covariates; more complex propensity score models will be briefly considered later. Rosenbaum and Rubin (1983) developed large sample (\( N = \infty \)) statistical theory concerning propensity scores. There are two remarkable implications of this theory: (a) If propensity scores can be successfully constructed, and (b) if treatment and control groups can be successfully equated on propensity scores, then theoretically the groups can be expected to be equated on all measured covariates that enter into the estimation of the propensity scores. Given this result, propensity scores provide an attractive basis on which to equate groups.

For propensity scores to be successful in producing unbiased estimates of the average causal effect, three additional assumptions must be met in addition to the assumptions for the randomized experiment. Collectively, Assumptions (1) and (2) are termed strong ignorability.

1. All potential confounders are included. All confounders that might bias the treatment effect estimate must be included in the propensity score model. There can be no hidden variables (Rosenbaum, 1992), unmeasured confounders that are related to both treatment assignment and outcome.

2. Common support region. No individuals may be included that have a propensity score of 0 or 1. For these cases, no alternative potential outcome exists as they can only be assigned to the \( c \) or to the \( t \) condition, respectively. More generally, there must be an area of overlap between the \( t \) and \( c \) groups on the propensity scores, an area that is termed the common support region. If there are regions of nonoverlap, no interpretable causal effect can be estimated in the nonoverlapping regions.

3. Correct estimation of propensity scores. Rosenbaum and Rubin’s (1983) theory applies to true propensity scores. If the propensity score model is not correctly specified, then systematic errors of prediction will be introduced over part of the domain of the covariates. This systematic error will lead to bias in the estimate of the average causal effect (Drake, 1993).

If these assumptions are met, propensity score approaches to equating groups minimize bias in the estimate of the causal effect. Consideration of these assumptions highlights the challenges of the propensity score approach as well as some of its conceptual advantages over common alternatives such as analysis of covariance. First, propensity scores only represent the model of selection into the \( t \) and \( c \) groups; they can and should be constructed only on the basis of knowledge of the baseline covariates and treatment assignment, a feature used in our example. No information about...
outcomes is involved. This procedure of estimating the propensity scores and equating the groups prior to any consideration of the outcomes greatly reduces the possibility that the researchers’ hypotheses can unintentionally introduce bias into estimates of causal effects (Rubin, 2001). Second, comparison of the distributions of the propensity scores in the $t$ and $c$ groups can identify the common support region over which proper equating of units can be achieved. Matching on propensity score avoids extrapolation of causal effects beyond the region of common support, which can introduce bias. Third, achieving matches on properly estimated propensity scores theoretically leads to balance between the $t$ and $c$ groups on every one of the covariates that enter into the construction of the propensity scores. This result permits checks on the success of the construction of the propensity scores via comparison of the mean and variance of the matched groups on each of the measured covariates as shown in our illustrative example.

**Illustrative Example**

We present an evaluation of a telephone counseling intervention (health coach) on the perceived well-being of individuals suffering from one or more of the following serious chronic diseases: diabetes, coronary heart disease [CHD], chronic heart failure [CHF], and high-blood pressure (see Renneberg et al., 2013). These chronic diseases have a highly negative impact on patients’ well-being and are associated with increased risk of mortality. These diseases are also associated with high utilization of medical care and frequent hospitalization. Prior evaluations of telephone counseling interventions with several chronic diseases have shown improvements in outcomes such as medication adherence, dietary habits, quality of life, and individual health outcomes (e.g., Buhrman, Fältenhag, Ström, & Anderson, 2004) and decreased impairments in everyday life (Böhme, Geiser, Mühlenhoff, Holtmann, & Renneberg, 2012). Our outcome variable here is the reported well-being of the patients. We focus initially on the simplest propensity score procedures for pedagogical reasons, later introducing options and potential complications.

**Method**

**Design.** The planned design was a randomized encouragement (invitation) design (Holland, 1988) in which insured individuals with one or more of the three chronic diseases were randomly assigned to receive an offer of intervention ($t$) or to the control ($c$) condition. In the $t$ condition, participants were invited to participate in the telephone counseling program plus treatment as usual (25% agreed). In the $c$ condition, participants received treatment as usual as covered by the insurer under the regulations of the German national health care system. The invitation process in the $t$ group might undermine the success of the original randomization in equating the groups at baseline. Our assumption was that the subgroup that accepted treatment might differ at baseline from those who refused treatment as well as from the control group. Data could not be collected from individuals who declined the intervention in the treatment group, precluding the possibility of intention to treat analyses or instrumental variable analyses that directly address treatment nonadherence (Angrist, Imbens, & Rubin, 1996; Sagarin et al., in press). We illustrate here the use of propensity scores to construct a subgroup of patients from the $c$ group who would be closely matched on a large number of important health variables measured at baseline to those participants who received $t$.

**Intervention.** Counselors were health practitioners with a professional background in counseling, such as psychologists and nurses, all trained in motivational interviewing. A manual for the telephone intervention was developed with four main modules related to essential behaviors that influence the conditions of chronic diseases and health: (a) physical activity, (b) nutrition, (c) fluid intake, and (d) adherence to medication intake. Each module included (a) a detailed analysis of the participant’s current behavior, (b) education about the relevant disease and specific health behavior, (c) the identification and setting of goals, and (d) the planning of health-enhancing behavior in order to achieve these goals. Counselors had the task of providing action and coping planning as well as goal setting. Psychoeducational strategies were applied to improve risk perception and self-efficacy and to develop an adequate outcome expectancy. Participants were contacted once every 2 weeks until both the participant and the coach decided active coaching on relevant modules was completed. Participants were contacted for one or two booster sessions 4–6 weeks after program completion to facilitate maintenance of behavior change.

**Baseline measures.** Participants reported on sociodemographic characteristics (e.g., gender, year of birth); a variety of health-related measures including level of disability, nutrition, activity behavior, and medical parameters (e.g., kind of diagnosis, high blood pressure, severity of chronic heart failure); and psychological variables (e.g., motivation to participate, outcome expectancy, and self-efficacy). From self-report inventories and patient records, a total of 98 baseline variables that might potentially be related to both actual treatment condition (health coach intervention vs. treatment as usual) and the outcome measure were assessed.

**Outcome measure.** The primary outcome variable for this illustration is the World Health Organization (WHO) Well-Being Index (Bonsignore, Barkow, Jessen, & Heun, 2001), a five-item brief screening instrument originally developed for the World Health Organization for use with general populations of older adults. Potential scores on the Well-Being Index range between 0 and 100, where high scores indicate positive well-being. The Well-Being Index was collected at baseline and at 4 months, which immediately followed completion of the booster session for intervention participants (immediate posttest). Sample items include “I have felt cheerful and in good spirits” and “My daily life has been filled with things that interested me.” Participants responded on a 6-point scale anchored by the descriptors at no time and all of the time. Bonsignore et al. (2001) reported the measure has shown adequate evidence of internal consistency and unidimensionality (Loevinger’s coefficient of homogeneity = 0.47; Mokken’s coefficient of item homogeneity ranges from 0.42 to 0.50). The WHO Well-Being Index has shown sensitivity of 1.00 and specificity of 0.68 with respect to current diagnosis of depression as measured by the Composite International Diagnostic Interview (World Health Organization, 1990).

**Participants.** Potential participants were selected from the database of individuals insured by the Techniker Krankenkasse, Hamburg, a large German health insurance company. Participants were selected who met criteria for diabetes, CHF, or CHD, and who had a high value on a statistical indicator of the insurance
company that reflects the predicted likelihood of hospitalization. Details of the selection and randomization procedure in the main evaluation design are available in Renneberg et al. (2013). For the purposes of our illustrative example, there was a total of 3,467 participants with complete data at baseline, with 1,126 participants actually receiving telephone counseling in addition to their usual treatment (t) and 2,341 participants receiving treatment as usual (c).

**Propensity Score Procedure and Estimation of Treatment Effect**

Although the design was initially planned to be a randomized encouragement design, the combination of nonadherence to the t condition (not all participants assigned to t accepted the invitation to be in the intervention) and the failure to collect data on those participants who refused treatment turned the design into a broken randomized experiment. If individuals who decline to participate differ from individuals who choose to participate in the intervention on baseline covariates, the treated participants will differ from control participants, potentially introducing bias into estimates of treatment effects. Thus, it was important to equate participants in the t and c groups prior to data analysis. We selected 98 baseline measures that we believed might possibly be related to the actual treatment condition and the outcome. These 98 measures were used in the construction of the propensity scores. To avoid clutter, we focused in our graphs on 15 key covariates that are known to be very important in health outcome research. For the full sample of 3,467 participants, treatment and control group participants showed some appreciable differences at baseline. On the 15 variables, the standardized effect sizes ranged in magnitude between 0.01 and 0.50. These observed differences suggested that the two groups initially differed in important ways on some key baseline variables.

**Construction of propensity scores.** To estimate the propensity scores, we used the most common procedure, logistic regression. We entered all 98 covariates into a logistic regression equation for the sample of 3,467 participants. Conceptually, in terms of the following equation

$$\text{Propensity score} = \text{Predicted logit} (T=t) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_98X_{98}, \quad (1)$$

where \(T = \) treatment condition received. In the analysis, \(t = \) telephone counseling + treatment as usual was coded as 1, and \(c = \) treatment as usual was coded as 0. \(X_i\) designates a covariate. The predicted logit is the logarithm of the odds \(\frac{P}{1-P}\), where \(P\) is the probability of the participant being in the intervention group, given his or her scores on each of the 98 covariates. By substituting each participant’s scores on the 98 covariates into the equation, the predicted value of the logit, can be calculated. The logits are used in the analyses, but for easier interpretation, the logit values can be transformed to probabilities using equations given in Cohen, Cohen, West, and Aiken (2003, p. 493). All of these calculations and graphs in the present illustration were performed in R using the MatchIt package (Ho, Imai, King, & Stuart, 2011). Thoemmes (2012) provided a similar package that performs the calculations in SPSS.

In Figure 2A, we provide kernel density estimates of propensity scores for the full intervention and control group samples. Kernel density estimates (Cohen et al., 2003, pp. 105–108) provide an estimate of the underlying distribution in each group. As can be seen, the distribution for the intervention group is shifted to the right, indicating that these participants have higher propensity scores than control participants, meaning intervention group participants have a higher average probability of receiving treatment. This lack of comparability undermines the usefulness of the simple prima facie difference between the means of the \(t\) and \(c\) groups, \(\bar{Y}_t - \bar{Y}_c\), as representing the causal effect. Given that the two groups differ on baseline covariates, a method of equating the groups must be chosen to achieve balance.

**Equating groups: Matching.** We initially present the conceptually simplest method of equating groups, 1:1 nearest neighbor matching on propensity scores. This method of matching involves choosing a unit in the \(t\) condition and matching its propensity score with the unit in the \(c\) condition that has the closest propensity score. The matched pair is removed from the data set, and the process continues until all pairs are matched.

One common problem with nearest neighbor matches is that “bad matches” can be formed if there is little overlap between the distributions of propensity scores in the treatment and control groups. A simple method of avoiding this problem is to specify a caliper, a maximum distance between the propensity scores in the two groups that is allowable for a match to be made. In the present case, we used a caliper of .025 standard deviation difference in the propensity scores: the smaller the width of the caliper, the closer the match of the participants in the two groups. However, the choice of a very small caliper width may exclude a large proportion of participants from the matched sample. Through trial and error, a caliper width can normally be identified that yields close matches with minimal loss of participants. Using this algorithm, we were able to successfully match 993 pairs of patients in the intervention and control groups. The success of this procedure in equating the propensity scores is shown in Figure 2B. The kernel density plots in the intervention and control groups overlap nearly perfectly. These plots provide the first indication that we have successfully equated the two groups on the propensity scores in the present example. Alternative methods of matching and equating groups (discussed later) exist that have advantages and disadvantages relative to nearest neighbor matching.

**Checks on covariate balance.** Our matching procedure was successful in producing almost identical distributions on the propensity scores. An even more important check on the equating of the groups is whether we have been able to achieve balance on each of the baseline covariates that go into the calculation of the propensity scores (see Figure 3). Panel A presents a dot plot comparing the standardized mean difference between the \(t\) and \(c\) conditions in the full sample for the 15 key covariates before (open circles) and after matching (dark filled circles) on the propensity scores. The standardized mean difference (sdmd) is closely related to Cohen’s \(d\), except that the same denominator (typically the SD of the treatment group before adjustment) is used to standardize

---

5 For simplicity, we only used participants with complete data at pretest. Estimation of propensity scores when participants have missing data on the baseline measures is considered briefly later in the article.
Panel B provides another check on the equating procedure, comparing the ratio of the variances in the treatment and control groups. If the groups have been successfully equated, the ratios should be not too discrepant from 1.0 (Rubin, 2001, considers 0.50 and 2.0 “far too extreme,” p. 174), the value that represents equal variances in the treatment and control groups at baseline. This criterion is met here. As another check, a histogram of the standardized differences before and after matching can be constructed for all first-order, quadratic, and two-way interaction terms involving the 98 covariates (see Figure 4). This criterion is met here.

In summary, the present results supported the success of the present propensity score and matching procedure on equating the groups on the 15 key baseline covariates. Although we highlighted only 15 key covariates, the results on the 83 other covariates were similar. These results indicate that we have successfully equated the intervention and control groups on the set of 98 covariates. In essence, we have mimicked the results of randomization for achieving balance on the set of measured covariates. No claims can be made about the treatment and control groups being equated on unmeasured (hidden) covariates. We also note that the means and standard deviations of the pretest measure of the outcome were well equated: As intended, the treatment and control groups did not differ in well-being at baseline: intervention: $N = 993$, $M = 60.61$, $SD = 22.85$; intervention: $N = 993$, $M = 60.38$, $SD = 22.92$; smd = −0.01). Of importance, all of the standardized differences on the baseline covariates were small, and none exceeded 0.25, a value suggested by Stuart (2010) as a rule of thumb guideline for a failure to successfully equate the treatment and control groups on a covariate.

If we had not been successful in achieving balance in the distributions of the covariates, more complicated propensity score models would need to be considered. Within the logistic regression framework, nonlinear terms such as quadratic $X_i^2$, $X_i X_j$ interactions, or both may be added to the propensity score model. Alternatively, computer-intensive procedures such as generalized boosted modeling and random forests (see Lee, Lesser, & Stuart, 2010) that attempt to approximate nonlinear relationships may be considered. The treatment and control groups would be matched on the re-estimated propensity scores and compared to see if better balance has been achieved (Ho, Imai, King, & Stuart, 2007). Finally, a key point—if there is not a common support region on the propensity scores, appropriate comparison of the groups is not possible.

**Outcome analysis.** Our matched procedure resulted in two equated groups. If there were no missing data on the outcome, the groups could be compared with a simple t test or regression analysis. However, given that there were missing data (54%) on the outcome variable in the matched data set, we identified 13 candidate variables measured at pretest that were moderately to highly correlated with both treatment and missingness (whether the outcome was measured or not). These candidates served as auxiliary variables. We performed our regression analysis using the Mplus program (Muthén & Muthén, 1998–2012) that uses full-information maximum likelihood in conjunction with the auxiliary variables to estimate the average causal effect (Enders, 2010). This procedure adjusts the results, reducing any bias due to missing data, under an assumption of missing at random (meaning that no unobserved causes of both missingness and outcome variables were left out). These results indicated that participation in the intervention produced a small positive effect for well-being: $\bar{Y}_t - \bar{Y}_c = 63.44 - 59.78 = 3.66$, Cohen’s $d = 0.08$, $z = 3.66$, $p < .001$. Participation in the intervention produced the largest effect for the
group of patients with a diagnosis of CHD ($d = 0.20$ for 318 matched pairs of patients and their controls with CHD; see Renneberg et al., 2013).

**Sensitivity analysis.** A final step in propensity score analysis is to probe how the obtained results might change if there were one or more unmeasured confounders (hidden variables). In our illustrative example, we used logistic regression to estimate the propensity scores, constructed matched pairs, and estimated the treatment effect. We used Cohen’s $d = \frac{Y_t - Y_c}{SD}$ to represent the estimated standardized treatment effect. We illustrate here a procedure for matched pairs suggested by Rosenbaum (1986) and Hong (2004).

Suppose there is a hidden variable $U$ (which may represent a composite of several hidden variables). The $t$ and $c$ participants have been equated on all 98 measured variables through propensity score matching. We believe this set is a good representation of potential covariates suggested by prior theory and research. However, the $t$ and $c$ participants have not been equated on $U$. We do not know $U$, so we have to assume a value. As a worst case value, Hong (2004) suggested the $smd$ with the largest magnitude observed on any baseline covariate other than the pretest. In the present example, that was 0.50 for the 5-point Likert item “I am motivated to participate in health programs to improve and maintain my health status.” A scaling problem arises because we would like to have the proper variance for the standardized difference between the $t$ and $c$ groups rather than the variance for the original means of the separate $t$ and $c$ groups (see Rosenbaum, 1986). To achieve this, we must rescale $smd$ to $smd'$:

$$smd' = \frac{smd}{\sqrt{2}} = 0.35.$$  

This equation has the proper standardization for the sensitivity analysis. The magnitude of the confounding is also determined by

---

**Figure 3.** Dot plot showing the success of propensity score matching for the means of 15 key covariates. Panel A: The open circles represent the mean difference between the unmatched intervention and control groups at baseline. The dark-filled circles represent the mean difference between the matched intervention and control groups at baseline. The initial large differences are reduced to near 0 through propensity score matching. $SMD = standardized$ mean difference. Panel B: The open circles represent the ratio of the variance of the treatment to the control group at baseline. The dark-filled circles represent the ratio of the variances of the matched intervention and control groups at baseline. The initial differences in the variance ratios are reduced to near 1 (equal variances) through propensity score matching.
of this magnitude actually exists? Should the effects of larger or smaller values of smd and γ be explored? Rosenbaum (2002, chapter 4) considers sensitivity analyses for binary outcomes and a variety of nonparametric tests of treatment effects.

A Key Issue: What Causal Effect Is Being Estimated?

In randomized experiments with full treatment adherence, the average treatment effect (ATE), \( Y_t - Y_c \), represents the estimate of the gain from receiving the treatment rather than the control for an individual randomly selected from the population. When the assumption of full treatment adherence is violated, the interpretation of the treatment effect becomes more complicated. When treatment adherence is less than perfect in the \( t \) and \( c \) conditions, \( Y_t - Y_c \) becomes the intention to treat (ITT) estimate. The ITT estimate represents the causal effect of treatment assignment on the outcome. It does not take into account whether the participant actually received treatment. For randomized experiments with treatment nonadherence, Angrist et al. (1996) proposed an alternative known as the average treatment effect on the treated (ATT). Following the logic of the potential outcomes model, this estimate compares the average outcome of the participants in the \( t \) condition who actually received the treatment with those in the \( c \) condition who would have accepted the treatment if offered. With full treatment adherence, the ATE and ATT estimates are identical.

In nonequivalent control group designs and broken randomized experiments, the ATE and ATT estimates nearly always differ. The simple 1:1 matching procedure described earlier produces an estimate of the ATT. Given that (a) the assumptions of propensity score analysis are met and (b) all participants in the treatment group can be successfully matched, this estimate is unbiased. In cases with a very large number of \( c \) participants, criterion (b) can often be met. In the present example, we were only able to match 993 (88%) of the 1,126 intervention participants. Few participants in the \( t \) group could be matched who had a propensity score greater than 0.5. Consequently, the obtained estimate might be a somewhat biased estimate of the ATT; its generalization is limited to participants with propensity scores below 0.5. Although it was not a material concern here, in studies with smaller \( Ns \), the loss of participants from the final analysis due to 1:1 matching may reduce the statistical power of the test of the treatment effect. Finally, rather than the ATT, the researcher may wish to estimate the ATE that applies to the entire population regardless of whether the \( t \) participants actually received the treatment. Such estimates are important when the goal is to compare the results of nonrandomized studies and randomized clinical trials are analyzed using ITT analyses.

Alternative Methods of Equating Groups and Estimating Treatment Effects

We now present alternative methods of equating groups and estimating treatment effects. We briefly consider known strengths, limitations, areas of optimal application, and complexities. Several of these methods are more suitable for clinical investigations with limited sample sizes.
Alternative Matching Algorithms Allowing Multiple Matches

One potential improvement is to use a matching procedure that permits the retention of a larger proportion of the cases. In many contexts, there is an excess of \( c \) relative to \( t \) participants or of \( t \) relative to \( c \) participants either overall or for certain ranges of the propensity score. In such cases, we can retain more participants if we use a procedure that permits a larger number of matches for each \( t \) (or \( c \)) participant to be selected from the larger group. Full matching (Rosenbaum, 1991; Stuart & Green, 2008) partitions the sample into matched sets with one \( t \) participant and any number of \( c \) participants, or one \( c \) participant and any number of \( t \) participants. Full matching leads to optimal balance of the covariates between the \( t \) and \( c \) groups. However, optimal balance comes at a potential cost of increased standard errors in tests of the treatment effect, and consequently decreased statistical power, given that there is no fixed limit on the number of cases that can be matched to a single \( t \) or \( c \) participant. Consequently, restrictions such as a maximum number of possible matches from the larger subgroup and the use of calipers that define the maximum permissible difference between the propensity scores are frequently imposed. Often a maximum of five comparison participants is suggested as the maximum number of matches from the larger subgroup permitted with each target participant. However, different guidelines have been suggested for the maximum number of permitted matches (see Stuart & Green, 2008); their relative performance is unknown.

Among the advantages of the full matching with restrictions is that this procedure offers a potential gain in the statistical power of the test of the treatment effect. Full matching requires the use of weighting methods commonly used in the analysis of survey research data that adjust the treatment estimate for multiple matches. For example, if one \( t \) participant were matched with five \( c \) participants, the \( t \) participant would be given a weight of 1 and the \( c \) participants would each be given a weight of 0.2. Such procedures require software that accommodates survey-weighting procedures, such as Mplus, SAS, and Stata (StataCorp, College Station, TX). In the present large \( N \) example, gains in statistical power from the use of these more complex matching procedures would be expected to be minimal. With smaller sample sizes, the increase in statistical power could be substantial.

Stratification (Subclassification)

In stratification, the data set is divided into five or more strata that represent specific ranges on the propensity score (e.g., in percentiles, the lowest 20%, 21%–40%, . . . 81%–100%). On the positive side, stratification has the advantage that close matches need not be found; the mean of the propensity scores in each stratum needs only be very similar. Stratification is flexible in that additional strata may be added to achieve balance; the strata do not need to be of equal width. Another advantage of the stratification approach is that both the ATE and the ATT can be estimated using different schemes for weighting each stratum (Schafer & Kang, 2008).

\[
\text{ATE} = \frac{\sum_{s=1}^{S} N_s \left( \bar{Y}_c - \bar{Y}_t \right)_s}{N}, \text{ with standard error} \\
\sqrt{\frac{\sum_{s=1}^{S} \left( N_s \text{SE} \left( \bar{Y}_c - \bar{Y}_t \right)_s \right)}{N}}
\]

\[
\text{ATT} = \frac{\sum_{s=1}^{S} N_s \left( \bar{Y}_c - \bar{Y}_t \right)_s}{N_t}, \text{ with standard error} \\
\sqrt{\frac{\sum_{s=1}^{S} \left( N_s \text{SE} \left( \bar{Y}_c - \bar{Y}_t \right)_s \right)}{N_t}}
\]

where subscript \( s \) is the stratum (1 to \( S \)), \( N \) is the total sample size and \( N_t \) is the treatment group size. \( \text{SE} \) is the standard error. Stratification often retains more participants than matching when participants who are not matched are discarded. On the negative side, stratification may approach but will not likely equal the bias reduction produced by matching. Imbalances between the \( t \) and \( c \) groups may remain in some of the strata and bias the treatment effect estimate. The covariates on which balance cannot be achieved may differ across the strata, making it difficult to determine the optimal stratification for a particular data set. Austin, Grotendorst, and Anderson (2007) showed in a simulation study that matching is often preferred because the larger residual imbalances from stratification can potentially introduce substantial bias.

Weighting

In weighting approaches, each individual’s propensity score \( P_i \) is used to represent the probability that the individual \( i \) will be in the \( t \) group and \( 1 - P_i \) represents the probability that the individual is in the \( c \) group. The full sample of \( N \) participants is considered. Bringing together ideas from survey sampling and the potential outcomes model, the contribution of the observed outcome is weighted by the inverse of the probability that individual \( i \) would be in the \( t \) group versus the \( c \) group, given his or her propensity score.

One advantage of the propensity score weighting approach is that both the ATE and the ATT can be estimated using slightly different schemes (Schafer & Kang, 2008). For the ATE, the following equation is used:

\[
\text{ATE} = \frac{\sum_{i=1}^{n} T_i Y_i \left( \frac{1}{P_i} \right)}{\sum_{i=1}^{n} \left( 1 - T_i \right) Y_i \left( \frac{1}{1 - P_i} \right)} - \frac{\sum_{i=1}^{n} \left( 1 - T_i \right) Y_i \left( \frac{1}{P_i} \right)}{\sum_{i=1}^{n} \left( 1 - T_i \right) \left( \frac{1}{1 - P_i} \right)}.
\]

Here, \( T_i \) represents the treatment condition (coded \( t = 1, c = 0 \)) and \( Y_i \) represents the observed outcome for individual \( i \). When individual \( i \) is in the treatment group, \( T_i = 1 \) so that the first term

\footnote{Propensity score matching reduces the standard error of statistical tests of \( t \) versus \( c \) on the outcome variable because of the closer balance on the covariates. Despite the smaller sample size often associated with propensity-score-matched \( t \) and \( c \) groups relative to the full sample used in ANCOVA, this feature often leads to higher statistical power for the \( t \) vs. \( c \) comparisons on the outcome for the matched samples (Smith, 1997).}
PROPENSITY SCORES FOR EQUATING GROUPS

of 1: Equation 3 is modified by giving each treated individual a weight (PS) as the only covariate.

\[ ATT = \frac{\sum_{i=1}^{n} T_i Y_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} (1 - T_i) Y_i \left( \frac{P_i}{1 - P_i} \right)}{\sum_{i=1}^{n} (1 - T_i) \left( \frac{P_i}{1 - P_i} \right)}. \] (4)

We used Mplus to implement these weighting procedures using the missing data procedures described earlier for our illustrative example of the health coach intervention study with the complete sample \((N_t = 1,126; N_c = 2,341)\). The propensity-score-weighted ATT theoretically estimates a statistical quantity identical to that estimated by the matched sample analysis if all treatment cases can be matched in large samples. Here, we found \(ATT = 3.66, d = 0.08, z = 3.42, p < .001\), identical to results presented earlier. Recall that the ATE estimates the treatment effect on the full population including those who have a near 0 probability of accepting \(t\) if offered (nonadherers). The ATE estimate = 2.41, \(d = 0.05, z = 2.02, p < .05\), is, as expected, smaller than the ATT estimate. Given that the assumptions of propensity score analysis are met, the ATE estimate would theoretically equal the ITT estimate in a randomized experiment.

Weighting methods have the advantages that they use all available participants and can be used to produce estimates of both the ATE and ATT effects. On the negative side, they require specialized software that incorporates survey sampling weighting. The standard errors may be too large, producing conservative hypothesis tests, particularly in smaller samples (see Schaefer & Kang, 2008, appendix). Recall that propensity scores are estimates; they have an associated standard error. In cases in which the sample includes propensity scores close to 0 or 1, the inverse-weighting scheme may become very unstable. Procedures to stabilize the weights have been proposed, but there is not yet consensus about their use. Problems with unstable weights are more common in estimates of the ATE than the ATT since the former procedure is more likely to include participants with propensity scores near 0 or 1 in the probability metric (i.e., individuals with propensity scores outside the common support region).

Analysis of Covariance (ANCOVA)

Two variants of ANCOVA may be utilized: (a) Traditional ANCOVA forms a linear composite of the full set of covariates and \(T\) to predict the outcome. For our example,

\[ \hat{Y} = b_0 + b_1X_1 + b_2 + \ldots + b_{98}X_{98} + b_9T. \]

(b) ANCOVA may also be conducted using the propensity score (PS) as the only covariate.

\[ \hat{Y} = b_0 + b_1PS + b_2T. \]

The goal of traditional ANCOVA is to estimate optimal regression weights to adjust the outcome for the effect of the set of covariates and the treatment. In contrast, the goal of propensity score models is to optimize the prediction of selection into treatment groups. The use of Variant (b) typically is less satisfactory than the use of Variant (a). Propensity scores were not developed for this purpose.

Given trivial differences between the \(t\) and \(c\) groups at baseline and that the standard assumptions (linearity, homoscedasticity, No Treatment X Covariate interactions) of ANCOVA are met, ANCOVA produces an unbiased estimate of the ATE. ANCOVA relies on the regression model to equate the \(t\) and \(c\) groups over the full domain of the covariates. In contrast, propensity score methods identify a region of common support for the \(t\) and \(c\) groups and closely equate the groups in that region. When two conditions both occur, ANCOVA may lead to a seriously biased estimate of the ATE (Schaefer & Kang, 2008): (a) the regression model is not correctly specified (e.g., nonlinear effects or interactions are omitted), and (b) the \(t\) and \(c\) groups do not fully overlap in their common support region. In addition, under Condition (b), the ATT and the ATE typically differ, with ATT being the statistical quantity that is of interest to the researcher. ANCOVA only estimates the ATE, whereas propensity scores can estimate both the ATT and ATE. Finally, the inclusion of the outcome variable may encourage the researcher to unconsciously select the model most favorable to the researcher’s hypothesis if the ANCOVA model needs to be respecified to meet its assumptions.

Many studies have compared ATEs resulting from propensity score approaches with ATEs from the traditional linear ANCOVA model using actual empirical data sets (e.g., Qin, Titler, Shever, & Kim, 2008). The majority of these studies have shown that the two methods yield identical conclusions regarding statistical significance but often yield differences in the magnitude of the treatment effect estimates. An important characteristic of these studies is that the true treatment effect is unknown so no conclusion can be reached about the potential bias of either approach. Cook, Steiner, and Pohl (2009) conducted a study in which participants were randomly assigned to a randomized experiment or nonrandomized study in which identical \(t\) and \(c\) conditions were compared. This four group design provided an estimate of the true treatment effect. These authors found little difference between estimates of the propensity score and ANCOVA approaches. However, their study was restricted to college students, and the interventions involved very brief training in vocabulary or math. Schaefer and Kang (2008) conducted a simulation study based on a large national data set on the effect of dieting. They created a known rule for selection into dieting versus nondieting. With their large sample size, they found a substantially larger magnitude of bias for ANCOVA than for propensity score approaches in terms of their statistical criteria; however, the absolute magnitude of the difference in the treatment effect estimates was not of clinical significance. As noted earlier, propensity score approaches can be expected to provide substantially less biased estimates of the true ATE than ANCOVA when the two conditions are both met. It is unknown how often the combination of the two conditions occurs in practice.

Selecting Covariates: The Most Critical Issue in Propensity Score Analysis

Selecting a comprehensive set of covariates to be measured at baseline is the most critical issue in propensity score analysis. Recall that the goal in selecting covariates is to identify all vari-
ables that might be confounders—potentially causing both treatment assignment $T(t vs. c)$ and the outcome (see Figure 1A). Prior research and theory in clinical psychology provide a strong basis for identifying many confounders, but this foundation can often be incomplete. Imagine a study of depressed patients in which cognitive behavioral therapy delivered in a clinic is compared with a bibliotherapy control condition in which written therapeutic materials are mailed to participants. Baseline level of depression is a clear candidate for a confounder—it is related to the level of depression at posttest and almost certainly to the likelihood of choosing to participate in the active versus control treatment. Pretests or proxy pretests of the outcome variable are always strong candidates for inclusion. Beyond this, clinical scientists have a systematic empirical base of variables known to predict level of depression from which they can select candidate covariates. In contrast, clinical scientists often have a far weaker empirical and theoretical understanding of the determinants of nonrandom assignment (or self-selection) to treatment conditions. In our current example, suppose distance to the clinic, a variable of no scientific interest, is negatively associated with choosing the active treatment given the effort involved in traveling to the clinic versus receiving the therapeutic materials in the mail. But if distance to the clinic were also associated with other variables (e.g., anomic of neighborhood) that are related to depression, it would be a potential confounder. Researchers are encouraged to think through all possible variables that might be related to both treatment assignment and outcome and to include as many as possible in the covariates measured at baseline. Input here also needs to come from staff, like recruiters and administrators, who often have important craft knowledge with respect to the recruitment process. This input can be supplemented by input (e.g., focus groups) from potential participants and, when participants self-select into treatment, by baseline measures of the participants’ preference for each treatment (Imbens, 2010; see earlier section on sensitivity analysis for an example).

Some studies have approximated these desiderata. In an educational context, Wu, West, and Hughes (2008) had content experts and school administrators identify variables (total covariates = 72) potentially related to the “treatment” assignment at the end of first grade (retention in first grade vs. promotion to second grade) and scores later in elementary school on standardized tests of reading and math. In the German health coach study presented here, we identified 98 covariates that might be potential confounders. All too often in practice, this comprehensive strategy of covariate selection contrasts with a nonoptimal default practice of selecting a few convenient demographic variables such as age, income, gender, and ethnic group on which to match participants. The comprehensive strategy addresses the first assumption underlying strong ignorability: All potential confounders that might potentially bias the treatment effect estimate must be included in the propensity score equation. The comprehensive strategy attempts to achieve this goal by identifying and correcting for a comprehensive set of covariates that may be potential confounders. Then any unmeasured covariate can bias the estimate of the treatment effect only to the extent it provides an unaccounted source of confounding that operates over and above (i.e., a unique effect) the effects of all of the baseline variables included in the propensity score. Studies by Cook et al. (2009) and Steiner, Cook, Shadish, and Clark (2010) have compared estimates of treatment effects obtained in studies in which participants are randomly assigned to $t$ and $c$ groups in either a randomized experiment or a nonrandomized observational study in which participants self-select into treatment conditions. The results showed that correction for a comprehensive set of covariates minimizes bias in the estimate of the treatment effect in the observational study compared with the gold standard estimate of the treatment effect from the randomized experiment. Correction for convenient demographic covariates did not lead to substantial bias reduction. A reliable measure of the pretest of the outcome variable typically leads to substantial bias reduction and often is the most important covariate.

Despite the attractiveness of the comprehensive covariate selection strategy, theoretical considerations identify some potential issues that can arise. Earlier we considered Figure 1, which presents three types of potential relationships among the covariate, treatment assignment, and the outcome. Panel A illustrates confounding; correction for a confounder always reduces bias in the estimate of the treatment. We seek to include all variables of this type in the propensity score model to minimize bias in our estimate of the causal effect. Panel B illustrates a covariate that is related to the outcome but not to treatment assignment. Inclusion of this type of variable has no effect on the estimate of the unstandardized treatment effect. It can have the beneficial effect of accounting for error variance in the outcome variable, increasing the power of the statistical test of the treatment effect. Panel C illustrates an instrumental variable, a covariate that is related only to treatment assignment but not to the outcome (Morgan & Winship, 2007). Instrumental variables also do not bias the estimate of the treatment effect if they are omitted from the estimation of the propensity score. In contrast, controlling for an instrumental variable can theoretically have one of two effects. If an unmeasured variable exists that has a residual confounding effect after the propensity score has been controlled, then controlling for the instrumental variable can increase bias (Wooldridge, 2009). If no unmeasured confounder exists, controlling for the instrumental variable increases the standard error, leading to a lower power statistical test. However, failing to control for a potential covariate appears to be a far more likely problem than inadvertently controlling for an instrumental variable (see Rubin, 2009).

Other Issues in the Propensity Score Approach

Other practical and theoretical issues commonly arise when researchers implement the propensity score approach. Space limitations permit only a brief introduction to these issues here. Fuller treatment can be found in the references cited.

Designs With More Than Two Treatment Groups

Our focus has been on constructing propensity scores to compare two groups ($t$ vs. $c$). Propensity score methods can also be employed in situations in which more than two groups are being compared. Rubin (1997) suggested constructing a separate set of propensity scores for each pairwise comparison and focusing separately on differences between each pair of treatment conditions. The estimate of the causal effect for each pairwise comparison is based on a unique subsample, making it difficult to compare causal effects obtained from different pairwise comparisons. Imbens (2000) has proposed an alternative approach using propensity
score models that estimate the selection probability for each of the treatment groups simultaneously using multinomial logistic regression. This method leads to more straightforward interpretation of the results but can easily fail if the sample size is not large, the common support region is small, or both. Spreewenberg et al. (2010) presented a clinical illustration of the latter approach.

**Covariates With Missing Values**

Baseline covariates will often have missing values in practice. Rosenbaum and Rubin (1984) proposed a general procedure of using a separate binary indicator for each covariate with missing data. The indicator is given a value equal to 1 when the covariate is observed for a participant and 0 when missing. The definition of the propensity score now changes to reflect potential missingness: In this case, it is defined as the logit of the probability that the participant is assigned to the treatment condition, given his or her observed values on the measured covariates and pattern of missing data on the measured covariates. The key idea of strong ignorability now also includes the additional assumption that the participant’s outcome is independent of the participant’s pattern of missing data (D’Agostino & Rubin, 2000).

Several different related approaches have been used in the literature to implement these theoretical ideas (see Harder, Stuart, & Anthony, 2010, and Stuart, 2010, for reviews). (a) The propensity score may be estimated separately for each pattern of missingness of the covariates but with the same propensity score model (Rosenbaum & Rubin, 1984). (b) Using logistic regression, an arbitrary constant is imputed for the missing values, and the propensity score is estimated with single imputed data set and missingness indicators (D’Agostino, Lang, Walkup, Morgan, & Karter, 2001). (c) Multiple imputation may be used (Hughes, Chen, Thoemmes, & Kwok, 2010; Qu & Lipkovich, 2009). Whichever approach to missing data is employed, balance checks need to focus on investigating the balance of both the observed values of the measured covariates and the participants’ missingness patterns for the measured covariates after propensity score matching.

**Multilevel Designs (Clustered Data)**

Propensity score matching becomes more challenging when researchers are conducting studies that yield clustered data (i.e., data that are nested). Such data are commonly found in clinical studies in which treatments are administered to a group (e.g., group psychotherapy), each therapist has multiple clients, or treatments are administered at different sites. In simulation studies, Thoemmes and West (2011) systematically examined different ways to estimate the propensity score with clustered data and ways to condition on this estimated propensity score taking the clustering into account. They found that the propensity score is ideally modeled using either a multilevel model with random effects (random intercepts and slopes across clusters for the relationship between covariates and the outcome) or so-called fixed effects models in which an indicator variable for each cluster (subscripted intercept) is entered into the prediction of the propensity score. This estimated propensity score can then be used to match participants either within clusters or across different clusters. The former approach simply means that any treated participant can only be matched to a participant in the same cluster, whereas the latter approach refers to allowing matches of treated and control participants regardless of cluster membership. Matching within clusters is theoretically preferred, but if treatment assignment is highly imbalanced, it can be very difficult to find matches within clusters (Hughes et al., 2010). Applications in medical science (e.g., Griswold, Localio, & Mulrow, 2010) have used propensity scores in multilevel design contexts.

**Conclusion**

This article has provided an overview of the basics of propensity scores for clinical researchers. Propensity scores can serve as a vehicle for equating nonrandomized groups on a large number of measured baseline covariates, mimicking the balance achieved by random assignment on those covariates. Propensity score methods also provide clear methods for checking the balance that is achieved on each measured covariate. Although we have focused on the comparison of treatment and control groups in this article, we pointed to other applications in our introduction such as minimizing the role of confounders in mediational analysis and generalizing the results of self-selected samples to the full population. Although propensity scores have important potential applications in clinical research, they are not a panacea. Very careful attention needs to be given to selection of covariates; all potential covariates that might influence both treatment assignment and the outcome must be included and reliably measured. The assumption of strong ignorability, that there are no hidden covariates that influence both the treatment assignment and outcome, is a critical one that is challenging to meet in practice. Careful thinking is needed about what potential covariates should be included in the design of the study for propensity score approaches to be effective. It is clear that selection of a few convenient demographic covariates (e.g., gender, age) almost never adequately represents the key variables that characterize selection into treatment conditions; the construction of propensity scores on this basis almost always fails to substantially reduce bias. The use of a comprehensive set of covariates, selected on the basis of prior theory and research, is far more likely to successfully minimize bias in the estimate of the causal effect.

Except in rare special cases (e.g., Diaz & Handa, 2006), the covariates that actually determine selection into treatment conditions are unknown in nonrandomized observational studies. Consequently, the estimate of the average causal effect associated with propensity scores (or any other method of equating groups) is always associated with some uncertainty. Cook, Shadish, and Wong’s (2008) review of designs comparing randomized and nonrandomized designs sharing the same treatment condition indicates that similar effect sizes are obtained when treatment and control participants have been sampled to be very similar. They emphasized comparisons of treatment groups with what they term “focal, local” comparison groups of similar participants rather than convenient comparison groups such as those from national surveys or other potentially different populations. Rosenbaum (1986, 2010) has advocated the use of sensitivity analyses in which the causal effect is estimated under different plausible assumptions about the magnitude of hidden confounders. To the extent that the estimate of the causal effect remains statistically significant as the
magnitude of the confounding effect increases, uncertainty about the causal influence of the treatment is reduced.

References


Received December 20, 2012
Revision received January 13, 2014
Accepted February 3, 2014