Abstract

Propensity scores represent the probability that an individual is assigned to a treatment group given his scores on a set of covariates. Propensity scores are used in nonrandomized studies to equate treatment and control groups at baseline prior to any treatment. Propensity scores concisely summarize in a single number the key information about baseline differences between the groups. Randomized experiments provide the gold standard to which other nonrandomized research designs are compared. Randomized experiments are unsurpassed for inferring that a treatment produced an effect on the outcome variable. Consider an experiment in which individuals are assigned to one of two treatment groups $T$: psychotherapy ($T = t$) or a minimal treatment control ($T = c$). In a randomized experiment, individuals are assigned to $t$ and $c$ using a random process (e.g., coin flip, random number) to guarantee that, on average, the two groups will be comparable on all possible baseline covariates. The two groups will be approximately equal at baseline on depression, proportion of females, income, or any other measured or unmeasured covariates, a condition referred to as balance. Given successful randomization, the difference between the mean of the $t$ and $c$ groups, $\bar{Y}_t - \bar{Y}_c$, provides an unbiased estimate of the average causal effect of the treatment on the outcome variable $Y$ across individuals.

Randomization often cannot be conducted for ethical or practical reasons. A study of major childhood stressors cannot randomly assign children to experience or not experience Hurricane Sandy or parental death. In such cases the process through which individuals are assigned to $T$ is unknown. For example, communities that receive the “treatment” (e.g., Hurricane Sandy) are compared with those that did not. Or individuals with depression who choose to receive may be compared to those who choose not to receive psychotherapy treatment. Possible differences on covariates between the $t$ and $c$ groups at baseline must be presumed to exist. Methods of equating the $t$ and $c$ groups prior to treatment are needed to reduce the possibility that pre-existing differences rather than the treatment account for the observed $\bar{Y}_t - \bar{Y}_c$ difference. Each of these methods of
equating groups attempts to measure key baseline covariates known as *confounders* that are related to both the treatment assignment $T$ and the outcome variable and to remove the influence of these confounders as plausible causes of the observed treatment effect. One historically popular method of equating $t$ and $c$ groups on baseline covariates is matching (Rubin, 2006). In its simplest form, baseline covariates are measured and an individual in the $t$ group is paired with an individual in the $c$ group who has the same set of scores. For example, a 30-year-old married male in the $t$ group would be matched with another 30-year-old married male in the $c$ group. The process is carried out until all individuals who can be paired in the two groups are identified and matched. There are technical and conceptual limitations of this approach. The technical limitation is that it is nearly impossible in many data sets to find good matches on more than a few covariates. For example, Chapin (1947) attempted to match 671 $t$ group and 523 $c$ group individuals on six covariates, identifying only 23 pairs that matched exactly. The conceptual limitation is that individuals in the $t$ and $c$ groups need to be equated on all potential confounders. Matching on a small number of demographic covariates that are convenient to assess will almost never be sufficient to equate the groups (Steiner, Cook, Shadish, & Clark, 2010). Rather, a comprehensive assessment of individuals at baseline on an extensive set of potential confounders suggested by theory and prior empirical work is needed. For example, West et al. (2014) used 98 covariates in a study of the effectiveness of a telemedicine intervention on the quality of life and wellbeing of patients with serious chronic diseases; Moser, West, and Hughes (2012) used 72 covariates in a study of the effects of retention (holding back in grade) on the achievement and psychosocial outcomes of low-achieving school children.

Propensity scores provide a vehicle for potentially overcoming the limitations of traditional matching. The propensity score is the probability that the individual will be assigned to the treatment group based of the full set of covariates, $P(T = t | X_1, X_2, \ldots, X_k)$, where $P$ represents the probability, $|$ means ‘conditional on,’ and the $X$s represent different covariates. The propensity score is a single score that summarizes all information contained in the full set of covariates. Careful equating of the $t$ and $c$ groups on the propensity scores can achieve balance on all of the baseline covariates from which the propensity scores were constructed, potentially a large number of covariates. If a comprehensive set of baseline covariates has been carefully selected, then potential bias of the average treatment effect estimate across individuals will be greatly reduced. Any remaining bias will be due to *hidden covariates*; that is, potential confounders that are related to both $T$ and the outcome that were *not* assessed at baseline. This bias is a function of the unique effect of the hidden covariates *over and above* the bias removed by the propensity scores.

The theory underlying propensity scores (Rosenbaum & Rubin, 1983) requires that three assumptions be met to achieve an unbiased estimate of the causal effect of a treatment.

1. Inclusion of all confounders. All confounders that are related to both treatment assignment $T$ and the outcome must be included in the propensity score model.
2. Common support region. In mathematical terms, no participants may be included that have a propensity score of 0.0 or 1.0. More generally, there must be an area of overlap between the $t$ and $c$ groups on the propensity scores, termed the common support.
region. Outside this common support region, individuals do not exist in one of the groups with whom to properly equate individuals in the other group.

3. **Proper estimation of propensity scores.** Rosenbaum and Rubin’s (1983) theory applies to *true* propensity scores, but participants are equated based on estimated propensity scores. If the propensity score model is not properly specified, systematic errors in estimating the propensity score may occur. In those parts of the domain of the covariates in which estimation errors occur, *t* and *c* groups will not be properly equated.

The strength of the randomized experiment for causal inference derives from its ability to rule out all possible measured and unmeasured confounders even without conditioning on any covariates. Propensity score analysis allows nonrandomized studies to mimic randomized experiments for the set of covariates measured at baseline after appropriate conditioning on the propensity score.

The assumptions are not easy to meet and present challenges for propensity score analysis; however, they also point to some of its advantages. (a) Propensity scores are constructed based solely on the baseline covariates, independent of the outcome variable. This feature separates design (equating) from decisions about how to properly specify the outcome model, minimizing the possibility that the researcher’s hypotheses could inappropriately influence the estimate of the causal effect (Rubin, 2001). (b) Matching on propensity scores clearly identifies the common support region over which *t* and *c* comparisons can legitimately be made. Extrapolation of causal effects beyond the data is minimized. (c) Matching on propensity scores implies that the *t* and *c* groups should be balanced on all baseline covariates that contribute to the estimation of the propensity score. The means and standard deviations of the measured baseline covariates should be equal within sampling error at baseline, just as in a randomized experiment. This property permits checks on the success on the success of the propensity score matching.

Given that these three assumptions are met, participants can be properly equated at baseline using propensity score methods. Of course, as in the randomized experiment, other issues may occur following initial equating of the groups that need to be addressed: Participants may not adhere to their treatment assignment, they may drop out and not be available for outcome measurement, or interference may occur between participants that affects their outcomes. These issues are not considered here (see West, Cham, & Liu, 2014 for a discussion of these issues).

**Steps in a Propensity Score Analysis**

**Step 1, Selection of Covariates**

The selection of a comprehensive set of covariates is the most important step in a propensity score analysis. All confounders that are related to both *T* and the outcome must be included in the construction of the propensity score. This step is largely a scientific rather than statistical matter: Clinical research and theory will often provide a strong basis for selection of covariates. At the same time, variables that are strongly related to *T* but that are incidentally related to the outcome are often less understood. Consider the example of a nonrandomized study comparing cognitive behavioral therapy versus bibliotherapy for depression. If depressed participants were given the option of choosing between the two therapies, those living further from the therapy treatment site may be more likely to choose bibliotherapy given the effort required to attend the therapy sessions. If these neighborhoods were also associated with other variables (e.g., anomie...
of neighborhood) that are related to the outcome variable of depression, then distance to
treatment site could be a potential confounder. Covariates that are only related to the
outcome (but not T) do not bias the estimate of the causal effect and may improve
statistical power. Propensity scores that include covariates that are only related to T (but
not the outcome) can under some conditions lead to a decrease in bias reduction.

Step 2, Estimating Propensity Scores
Propensity scores can be estimated using a wide variety of statistical models. Most
typically used is a linear logistic regression model (equation 1):
\[ \text{ln} \left( \frac{PS}{1 - PS} \right) = b_0 + b_1 X_1 + b_2 X_2 + ... + b_k X_k + e \]  
where \( \text{ln} \) represents the natural logarithm function, the \( X \)s represent different
covariates, the \( b \)s represent the regression coefficients, \( e \) represents the error of
prediction, \( k \) is the total number of covariates, and \( PS \) is the estimated propensity score
(i.e., the predicted probability that the participant is in the \( t \) condition given his scores on
all \( k \) covariates). The predicted logit of the propensity score, \( \text{ln} \left( \frac{PS}{1 - PS} \right) \), is typically
used instead of the predicted \( PS \) because the logits have a more normal distribution and
do not have a minimum value of 0 and a maximum value of 1. The linear logistic
regression equation will often yield a solution in which good balance between the \( t \) and \( c \) groups is achieved on the propensity scores and all \( k \) covariates that go into its calculation
(see step 4, balance diagnostics).

When good balance between the \( t \) and \( c \) groups cannot be achieved, it often
indicates that there is a nonlinear relationship between at least some of the covariates and
the \( PS \). Two approaches may be taken to this problem. One approach is to add quadratic
\( (X_i^2) \) and two-way interaction \( (X_i X_j) \) terms to equation 1 to represent potentially nonlinear
effects. Candidate quadratic and interaction terms may be suggested by careful
examination of the balance diagnostics in step 4, identifying specific covariates on which
balance was not achieved. An alternative approach is to use any of a variety of machine
learning approaches that attempt to approximate the optimal nonparametric function
relating the set of covariates to \( T \). Several variants exist, of which “random forests” and
“generalized boosting” currently appear to be the most promising (Lee, Lesser, & Stuart,
2010). Preliminary evaluations suggest that these methods perform better than linear
logistic regression when the true propensity score model is nonlinear, particularly when
the sample size is large. Imai and Ratkovic (2014) have recently proposed an alternative
estimation method, called the covariate-balancing propensity score, in which a single
model estimates the relationship between covariates and treatment assignment, but with
regression weights chosen that minimize the imbalance of covariates.

Step 3, Equating Groups: Conditioning on the Propensity Scores
Several methods are commonly used to equate the \( t \) and \( c \) groups by conditioning on
the propensity score.

Matching.
Conceptually, the simplest form of matching is 1:1 nearest-neighbor matching. In
nearest-neighbor matching individuals in the smaller group (here assumed to be \( t \)
become the targets for matching with individuals from the larger group (here \( c \)). A target individual is selected from the \( t \) group and the individual in the \( c \) group whose \( PS \) is closest to the \( PS \) of the target is identified. The \( t \) and \( c \) individuals are then matched and removed from the sample. The matching process continues until all \( t \) individuals are matched.

There are several improvements that can be made on this simple form of matching. First, the success of the matching process depends on the order in which individuals in the \( t \) group are considered for matching. Alternative “optimal matching” algorithms allow matches to be broken if an alternative match could later be created that reduces the overall distance between the propensity scores in the two groups. Second, when there are individuals in the \( t \) group who are outside of the region of common support, these \( t \) individuals will not have a corresponding \( c \) individual with whom a close match can be formed. To avoid bad matches, the matching algorithm can be constrained to accept only those matches that fall within a specified caliper width defining good matches, often chosen to be <0.25 SD of the logit of \( PS \). Third, there may be multiple potential matches for specific individuals. Full matching (Rosenbaum, 1991) divides 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. Allowing a variable number of matches for each target can produce good results. For example, in areas of the common support region in which there is only a single \( c \) individual whose logit of the propensity score close to that of the target \( t \) individual, a single match will be created. In areas in which many \( c \) individuals have logits of \( PS \)s close to that of the target \( t \) individual, many individuals will be matched with the target \( t \) individual. The full matching procedure retains a larger portion of the original sample and thereby potentially increases statistical power without materially degrading the quality of the matches.

Subclassification.

Rather than matching individual participants, the \( PS \) can be divided into strata and comparisons can be made within each of the strata (e.g., regions including 0–20th percentile, 21st–40th percentile, …, and 81st–100th percentile on the \( PS \)s). In subclassification at least five strata are normally used, based on earlier work showing five strata are sufficient to reduce bias of the average treatment effect by 90%; the means in the \( t \) and \( c \) groups on the propensity scores and the covariates from which they are constructed will often be close to those achieved by matching. If the \( t \) and \( c \) groups are not closely equated within one or more of the strata, a finer-grained subclassification scheme may be used to improve balance. Relative to matching, subclassification normally retains a larger proportion of cases from the original sample, but at a potential cost of less closely equating the \( t \) and \( c \) groups.

Analysis of covariance.

The groups may also be equated by using the \( PS \) as the covariate in an analysis of covariance. Typically, all individuals are included in the analysis and the following model is estimated (equation 2):

\[
\hat{Y} = b_0 + b_1 PS + b_2 T \quad (2)
\]

where \( \hat{Y} \) is the predicted outcome, \( T \) is the treatment group (\( t=1 \), \( c=0 \)), \( b_0 \) is the mean in the \( c \) group, \( b_1 \) is the linear relationship between \( PS \) and \( Y \), and \( b_2 \) is the estimate
of the causal effect of treatment, adjusted for the PS. In standard analysis of covariance, the adjustment assumes that the relationship of the PS and the outcome is linear. If this relationship is nonlinear, then standard analysis of covariance can lead to biased estimate of the causal effect. Analysis of covariance can be extended to estimate nonlinear relationships between the PS and the outcome, but bias in the estimate of the causal effect can result if the true relationship does not match the estimated relationship. A second problem with analysis of covariance is that it can also potentially extrapolate far beyond the data, producing an estimate of the causal effect even when there is no overlap between the PS of the t and c individuals.

<txt>Weighting. In weighting each individual is included in the analysis. Depending on the individual’s PS and the type of causal effect to be estimated, the individual’s outcome is given an appropriate weight in estimating the causal effect. This approach is discussed below in step 5.

<b>Step 4, Checking Balance: Assessing the Adequacy of the Propensity Score Model</b> Balance checking is conceptually simplest to understand when 1:1 matching has been used to equate the participants in the t and c groups. The matching process leads to close matches in the t and c groups on the estimated propensity scores. If close matches are achieved and the estimated propensity score is a good approximation of the true propensity score, then the t and c groups should be balanced on all covariates that were used to construct the propensity scores. The means and variances of the t and c groups on each can be statistically tested for equality; however, authors (e.g., Stuart, 2010) have criticized this practice, preferring to focus on standardized mean differences and variance ratios of each of the covariates, indices that are not affected by sample size.

The standardized mean difference (smd) is the difference between the means of the t and c groups on the covariate following matching, divided by the SD of the treatment group before matching, \[ \text{smd} = \frac{\bar{X}_t - \bar{X}_c}{SD_{t\text{-prematchin}}} \] (Stuart, 2010). Typically, values of \text{smd} < 0.25 in magnitude are taken as evidence that the two groups were successfully matched on the covariate. The variance ratio, \[ \frac{SD_t^2}{SD_c^2}, \] provides a second, important comparison of the groups. Typically, values of the variance ratio between 0.5 and 2.0 are taken as evidence that the two groups were successfully matched. Compact graphical depictions such as boxplots, kernel density plots, or dotplots can be used to visually compare the t and c groups both before and after matching on the smds and variance ratios.

Figures 1, 2, and 3 provide an illustration of this balance-checking process based on data from Moser et al. (2012). Moser et al. identified school children who were at risk for retention at the end of first grade based on below-median scores in reading at school entry. They collected data from the child, the teacher, peers, parents, and school records in first grade prior to the decision to retain the child in first grade or to promote the child to second grade the following year. Seventy-two covariates were selected that were believed to be possibly related to both the retention decision and school achievement and psychosocial outcomes.
For our illustration, imagine a clinical researcher wished to study the effects of early grade retention on measures of child psychopathology in early adolescence. The researcher would identify covariates believed to be related both to the grade retention decision and later childhood psychopathology. For ease of presentation (and to avoid clutter in the graphs) we focus here on 10 covariates in the data set that would be likely to be chosen for inclusion in the construction of the PS. In practice, additional covariates would likely be chosen based on subject-matter expertise.

Figure 1 shows superimposed kernel density plots of the PS for the t and c groups. Figure 1a shows the smoothed distribution of PS in the two groups. The average score in the retained group is lower than in the promoted group. Figure 1b shows that the distributions of the two groups are much more similar following matching and that there is no longer a difference in propensity scores. The common support region appears to cover a substantial portion (approximately 0.0–0.6) of the potential range (0.0–1.0) of the propensity scores. Figure 2 shows a dotplot of the smds comparing the retained to promoted groups for each of the 10 covariates before and after matching. Although the mean differences in this example prior to matching are not large, they are reduced following matching. None of the smds exceeded the guideline of 0.25 SDs following matching. Figure 3 shows the variance ratios comparing the retained to promoted groups for each of the 10 covariates before and after matching. Again, none of the variance ratios were outside the range of acceptable matches range of 0.5 to 2.0.

Two supplemental methods provide additional evidence concerning the success of the equating procedure and are particularly useful in detecting problems arising from nonlinear relationships between the covariates and propensity score. First, the t and c groups can be compared within each stratum of the PS, essentially checking for a group × stratum interaction in standardized effect sizes. Second, smds and variance ratios for nonlinear terms such as quadratic ($X_i^2$) and two-way interaction ($X_iX_j$) terms may be compared between the t and c groups.

If this series of checks detects a failure to achieve balance, then one of two approaches is normally taken. The preferred approach is to re-estimate the propensity score using a method that captures nonlinear relationships between T and the covariates. As noted above, quadratic terms, two-way interaction terms, or both may be added to the propensity score model to capture the nonlinearity. Alternatively, machine learning approaches may be investigated. The revised propensity scores need to be checked to verify that balance has in fact been attained. Alternatively, the problematic covariates can be included in the outcome analysis, providing an adjustment for the failure to attain balance.

**Step 5, Estimating the Causal Effect**

Researchers typically estimate one of two causal effects of interest. (a) The average treatment effect on the treated (ATT) is the average causal effect for the population corresponding to the full set of individuals who are assigned to the t group. When everyone in the t group can be matched and matching is 1:1, the ATT is the mean difference between the t and c groups, $\bar{Y}_t - \bar{Y}_c$. When not all t individuals can be matched, weighting by the odds is used in which each t individual receives a weight $W = 1$, and each c individual a weight $W = PS/(1 - PS)$. When multiple c individuals are matched to each t individual or subclassification is used, these factors also need to be considered in
the weighting scheme. (b) The *average treatment effect* (ATE) estimates the average causal effect for the full population. This quantity can be directly estimated as $\bar{Y}_t - \bar{Y}_c$ when all individuals in both the $t$ and $c$ groups can be matched. Alternatively, a variant of the weighting scheme can be used, again potentially taking into account the possibility that multiple $c$ individuals are matched to each $t$ individual or subclassification is used in constructing the weighting scheme (see Stuart, 2010). In randomized experiments, ATT and ATE represent identical estimates. In nonrandomized designs the ATT and ATE estimates can be substantially different, since not all participants would agree to receive both the $t$ and $c$ treatments.

Some Complications and Extensions

Designs with More than Two Groups

Nearly all of the work on propensity score analysis has been on two-group designs comparing one $t$ and one $c$ group. With more than three treatment groups, groups may be compared on a pairwise basis or a set of propensity scores corresponding to each of the groups may be constructed using multinomial logistic regression.

Covariates with Missing Values

There is a variety of approaches to addressing missing values on the covariates (Stuart, 2010). These approaches include using multiple imputation, using single imputation with an indicator variable to denote the extent to which a covariate is incomplete, and using the same propensity model to estimate the propensity score separately for each pattern of incomplete covariates. The propensity score is redefined as the probability that the individual is assigned to treatment group given the scores and the patterns of missing data on the covariates.

Multilevel Designs

Some clinical designs such as group psychotherapy involve delivery of treatment to individual participants within a group. Equating the groups now involves consideration of covariates at two levels: level 1, the individual participant level; level 2, the group level. The covariates at level 1 (e.g., age, level of depressive symptoms, previous clinical diagnosis of depression) will often be strikingly different from those at level 2 (e.g., group size, therapist experience). Thoemmes and West (2011) discuss multilevel propensity score models that take both levels 1 and 2 simultaneously into account using either fixed effects (in which indicator variables represent each group) or random effects (in which each group has a different estimated intercept and slope). A variety of clinical research contexts have this multilevel structure (e.g., each therapist has multiple clients, treatments are administered at different community centers).

Sensitivity Analyses

Although propensity scores equate groups at baseline on measured covariates, they do not address *unmeasured* covariates. In sensitivity analysis, the magnitude of the difference on a hypothetical unmeasured covariate at baseline that is necessary to yield a nonsignificant estimate of the causal effect is examined. In some cases, the causal effect of the treatment may be very robust to the presence of an unmeasured confounder, further buttressing the conclusion.
Evaluation of Propensity Scores

Propensity scores provide a conceptually strong method of equating groups in nonrandomized studies. Propensity scores often lead to similar estimates of the causal effect to other approaches; they are likely to lead to less biased estimates than other approaches (e.g., analysis of covariance) when there are many covariates, there are nonlinear relationships between the covariates and the outcome, and there is a limited common support region (Schafer & Kang, 2008). Available studies comparing estimates of causal effects from randomized and nonrandomized studies highlight the importance of careful measurement of a comprehensive set of covariates, the selection of samples that are as similar as possible for the comparison, and the use of all available information that permits accurate modeling of the process of selection into the \( t \) and \( c \) groups (Cook, Shadish, & Wong, 2008).

SEE ALSO: Conditional Probabilities; Missing Data; Models of Specific and Nonspecific Change Mechanisms in Mental Health Treatments

References


*Further Reading*


*Figure 1* Plots of the propensity scores for low-achieving students to be retained in grade. The solid line represents the group retained in first grade; the dotted line represents the group promoted to second grade. (a) Before matching the distribution of the propensity scores of the retained and promoted groups differ. (b) After matching the distributions are virtually identical.

*Figure 2* A dotplot of the smd between the retained and promoted groups on 10 baseline covariates. The smds before matching (denoted by ○) are in almost all cases farther from 0.0 than the smd after matching (denoted by ×) on the propensity scores. P represents peer rating, PA represents parent rating, T represents teacher rating; ADHD, attention deficit hyperactivity disorder.

*Figure 3* A dotplot of the variance ratio of the retained to the promoted groups on 10 baseline covariates. The variance ratios before matching (denoted by ○) are in general farther from 1.0 than those after matching (denoted by ×) on the propensity scores. Abbreviations as in Figure 2.