Propensity Scores in Cardiovascular Research

Ralph B. D’Agostino, Jr, PhD

Propensity scores have been used to reduce bias in observational studies in many fields and are becoming more widely used in cardiovascular research. The goal of this statistical primer is to present the definition of propensity scores and to illustrate their use by describing some recent examples found in the cardiovascular disease research literature.

Large-scale epidemiological cohort studies such as the Multi-Ethnic Study of Atherosclerosis (MESA) are designed to follow a large sample of participants over time without active administration of any interventions. Within MESA, lack of randomization can complicate potential treatment comparisons such as the impact of β-blocker versus angiotensin-converting enzyme inhibitor usage. Nonrandomized comparisons may also arise from within a randomized clinical trial. For instance, the Clopidogrel as Adjunctive Reperfusion Therapy - Thrombolysis in Myocardial Infarction 28 (CLARITY-TIMI 28) trial is a randomized study that compares clopidogrel with placebo in 3491 ST-elevation myocardial infarction patients aged 18 to 75 years who have undergone fibrinolysis. In addition to the primary end points, investigators wished to compare the effects of low molecular weight heparin with unfractionated heparin on angiographic and clinical outcomes in participants. These treatments were not randomly assigned.

In studies such as these, the treatment groups may markedly differ with respect to the observed pretreatment covariates measured on participants. These differences could lead to biased estimates of treatment effects. The propensity score for an individual, defined as the conditional probability of being treated given the individual’s covariate values, can be used to balance the covariates in the 2 groups and thus reduce this bias.

In a randomized experiment, the randomization of participants to different treatments minimizes the chance of differences on observed or unobserved covariates. However, in nonrandomized studies, systematic differences can exist between treatment groups. To control for this potential bias, information on measured covariates can be incorporated into the study design (eg, through matched sampling) or into estimation of the treatment effect (eg, through stratification or covariance adjustment). However, such methods of adjustment can often use only a limited number of covariates, whereas adjustments that use propensity scores do not have this limitation.

A simple illustration of how an imbalance on covariates could influence a treatment effect estimate is as follows. Consider a nonrandomized study with 2 groups and a binary outcome with the data shown in the Table.

We can clearly see in the Table that gender is not balanced between the 2 groups (80% males in Group A versus 10% in Group B). The apparent treatment difference between groups would not be significant if we adjusted for gender. In other words, if we created balance between the 2 groups on the basis of gender, we would recognize that the apparent treatment effect is caused by gender and not group. In most observational studies, there may be several variables that are imbalanced at the same time between groups and the propensity score methodology allows one to simultaneously balance all the covariates and make more valid inferences about treatment effects.

Definition

The propensity score for an individual is the probability of being treated conditionally on (or only on the basis of) the individual’s covariate values. Intuitively, the propensity score is a measure of the likelihood that a person would have been treated on the basis of only his or her covariate scores. The propensity score is the probability that a participant is in the “treated” group given his/her background (pretreatment) characteristics. This score is frequently estimated by logistic regression where the treatment variable (treated yes/no) is the outcome and the covariates are the predictor variables in the model.

The propensity score method can be used if conditional independence exists between the treatment assignment and potential outcomes given the covariates (referred to as strongly ignorable treatment assignment). In other words, the treatment assignment can be associated with covariate values but not be related to outcome values once the covariates are controlled for. The above is a description of the relationship of treatment assignment (eg, being placed in the β-blocker group) to covariates and outcomes, not a description of the relationship of the treatment effect (eg, the impact of taking β-blockers) to the covariates or outcomes. When the treatment assignment is strongly ignorable, as is most often the case, one can estimate the propensity score and use this score.
as a balancing score to "balance" the distribution of the covariates in the treated and control groups. Matching, stratification, or regression (covariance) adjustment with the propensity score can be used to produce unbiased estimates of the treatment effects and create covariate balance between groups. In some of these methods, the propensity score itself is used in the analyses as a weight or factor (regression adjustment), whereas in others it is used to construct the appropriate comparisons (stratification or matching) but not in the analyses directly.

In practice, the success of propensity score modeling is judged by whether balance on covariate values is achieved between the treatment groups after its use. Because of this, one can be more liberal with inclusion of covariates in the model than in most traditional settings. For instance, covariates with \( P \) values larger than 0.05 can be included in the propensity score model. One limitation that concerns the number of covariates that can be included in the model is that there needs to be a sufficient number of participants in each treatment group for each covariate that is included. For instance, if a study includes 30 treated and 50 untreated individuals, the propensity score model should have much less than 30 covariates included. Once the model is fit, one method to evaluate the success of a particular propensity score model is to compare the amount of bias (or imbalance) that existed on observed covariates in the treated and control groups before and after adjustment for propensity scores.

One advantage of propensity scores is that if 2 subjects are found, 1 subject in the treated group and 1 subject in the control, with the same propensity score, then one could imagine that these 2 subjects were "randomly" assigned to each group in the sense of being equally likely to be treated or control. Because propensity scores are estimated with only observed covariates, one has to assume that unobserved covariates would not have changed the model had they been measured. When this assumption is true, one can be fairly confident that approximately unbiased estimates for the treatment effect can be obtained.

When building the propensity score model, only covariates that occur pretreatment should be included. If one includes covariates that are measured posttreatment, then the propensity score model may explain part of the treatment effect itself. For example, if one wished to compare in an observational study the impact of a \( \beta \)-blocker versus an angiotensin-converting enzyme inhibitor, the propensity score model could include age, smoking status, and prior medical history. However, patient characteristics measured after the treatment began, such as an ejection fraction measurement taken posttreatment (e.g., after \( \beta \)-blocker initiation) should not be included. Indeed, ejection fraction may indeed be imbalanced between the treatment groups; however, this imbalance may be caused by the treatment and therefore is part of the outcome.

### Uses of Propensity Scores

The 3 most common techniques that use the propensity score are matching, stratification, and regression adjustment. Each of these techniques is a way to make an adjustment for covariates before calculation of the treatment effect (matching and stratification) or during calculation of treatment effect (stratification and regression adjustment). With all 3 techniques, the propensity score is calculated in the same way, but once estimated it is applied differently. Propensity scores are useful for these techniques because by definition the propensity score is the conditional probability of treatment given the observed covariates; thus, subjects in treatment and control groups with equal (or nearly equal) propensity scores will tend to have the same (or nearly the same) distributions on their background covariates. Exact adjustments made with the propensity score will, on average, remove all the bias in the background covariates. Therefore, bias-removing adjustments can be made with the propensity scores rather than all the background covariates individually.

### Matching

An example of propensity score matching was described in a recently published study that examined whether hypothermic circulatory arrest (HCA) is a risk factor for neurologic morbidity in aortic surgery. More than 500 individuals (238 HCA+ and 273 HCA−) participated in this study. When the investigators first compared the groups, they determined that many characteristics were different between the 2 groups, such as gender, age, smoking history, and hypertension. Thus, they were concerned that if differences between HCA+ and HCA− were found, these could be the result of differences in pretreatment conditions only. To handle this, the investigators estimated propensity scores for all participants with 9 covariates and then matched HCA+ and HCA− participants with the propensity score. The maximum difference in the propensity score allowed for a match was 0.015, and with this criterion 220 closely matched patients (110 in each group) were identified for their final analyses. The investigators demonstrated that all baseline characteristics that had been significantly different (unbalanced) between groups in the overall study were balanced on the propensity-matched pairs. Thus, HCA comparisons could be made on this subgroup of participants.

Matching is a common technique used to select control subjects who are “matched” with the treated subjects on background covariates that the investigator believes need to be controlled. Although the idea of finding matches seems straightforward, it is often difficult to find subjects who are similar on all covariates, even when only a few background covariates of interest exist. The investigators for the HCA example above would have confronted this problem as they had identified 9 variables on which they wished to match subjects.

Propensity score matching solves this problem by allowing an investigator to control for many background covariates.
simultaneously by matching on a single variable, the propensity score. Propensity scores can be calculated with many covariates, and the result for each participant is a scalar summary (single number) of his/her covariates.

To evaluate the success of propensity score matching, a common technique is to compare covariates in the treated and control groups before and after matching. For continuous variables one can compare means or \( t \) statistics pre- and postmatching, and for categorical variables one can compare frequencies/percent and \( \chi^2 \) statistics pre- and postmatching. Estimates of the percent reduction in bias from propensity score matching can be found by calculation of an initial bias (as the difference in covariate mean values between the treated and control groups before matching, \( b_m \)) and the postmatching bias (as the difference in covariate mean values after matching, \( b_n \)) and then calculation of the percent reduction in bias as \( 100(1 - b_n/b_m) \).

In many settings, propensity score matching can also be very cost-effective. In particular, if an investigator has access to a large database or patient population where the treatment indicator and background covariates have been measured, but outcomes of interest have not been measured yet, propensity score matching can be used to identify the appropriate subset of individuals from which to gather additional outcome measures rather than have data collected on all individuals.

**Stratification**

Stratification is also commonly used in observational studies to control for systematic differences between the control and treated groups. This technique consists of grouping subjects into strata determined by observed background characteristics. Once the strata are defined, treated and control subjects who are in the same strata are compared directly. Many of the same problems occur in stratification as in matching when the number of covariates increases. As the number of covariates increases, the number of strata grows exponentially. For instance, if there were \( k \) dichotomous covariates to be controlled for, then there would need to be \( 2^k \) strata created. If \( k \) is large, then some strata might contain subjects from only 1 group, which would make it impossible to estimate a treatment effect in that stratum. Here again, the propensity score is very useful. Because the propensity score is a scalar summary of all the observed background covariates, stratification on it alone can improve the overall balance of the distributions of the covariates in the treated and control groups without the exponential increase in number of strata.

It has been shown that stratification based on the propensity score will produce strata where the average treatment effect within strata is an unbiased estimate of the true treatment effect. In addition, research has shown that creation of 5 strata (ie, quintiles) can in general remove approximately 90% of the bias caused by strata variables (propensity score). In fact, stratification on the propensity score balances all covariates that are used to estimate the propensity score, and often 5 subclasses based on the propensity score will remove >90% of the bias in each of these covariates.

The technique used to determine strata is straightforward. Once the propensity score is estimated, the investigator must decide how many strata should be used. As stated above, 5 strata (ie, quintiles) are usually sufficient; however, the number of strata used depends on how many participants are available in the overall study. The strata boundaries are normally based on the values of the propensity score for both groups combined rather than on the treated or control group alone. A recent publication that used propensity scores for stratification examined whether excessive variation exists in providing coronary angiography to patients after acute myocardial infarction on the basis of chronic kidney disease and whether an association exists between angiography and mortality. The investigators estimated propensity scores for the probability of undergoing coronary angiography during hospitalization among 6794 chronic kidney disease patients who were rated appropriate for the procedure. Here the dependent variable (ie, treatment indicator) was provision of angiography, and the covariates used in the model included both patient level and hospital characteristics. Once propensity scores were estimated for all participants, the investigators ranked all appropriate chronic kidney disease patients by their estimated propensity scores and created quintiles based on these propensity scores. Analyses were then performed within each of the 5 strata to compare odds ratios and 95% confidence intervals for 1-year mortality for those who underwent coronary angiography versus those who did not. With this approach, all quintiles except the lowest (where the likelihood of angiography was <6%) showed that the odds of death were higher for those with no angiography. Although these results were similar to those found with an overall logistic regression, the investigators concluded, “Given that the propensity score approach requires fewer assumptions and tends to balance differences between treated and untreated groups, we prefer these results to those of the logistic regression model.”

**Regression (Covariance) Adjustment**

Propensity scores can also be used in regression (covariance) adjustment. In regression adjustment, the treatment effect is estimated by adjustment for the impact of background covariates in a regression model. In general, covariance-adjusted models can contain 1 or more covariates. The propensity score is a useful variable in regression adjustments, because one can first fit a propensity score model that includes many potential covariates, and then the final treatment effect model only has to include the propensity score as a covariate to derive adjusted estimates. Another approach to regression adjustment is to use a large set of background covariates to estimate the propensity score and then use a subset of these covariates and the propensity score in the regression adjustment. A recent article in the cardiovascular research literature examined whether mitral valve annuloplasty (MVA) improves long-term mortality in patients with mitral regurgitation and left ventricular systolic dysfunction in 419 patients felt to be candidates for MVA. To examine this question the investigators estimated propensity scores that predicted whether a patient would undergo MVA on the basis of demographics, physical examination findings, electrocardiography and echocardiography measurements, and medications that clinically would likely affect...
the probability of undergoing MVA. Once the propensity scores were estimated for each participant, Cox proportional hazards models were fit to examine the impact of MVA on event-free survival where the propensity score was forced into the model as a covariate. Additional models were fit that included the propensity score and other covariates, and the investigators found that final predicted values remained consistent with or without the propensity score as long as a subset of important covariates were included.

One question that may arise when regression adjustment with propensity scores is used is whether any gain results from the use of the propensity score rather than performance of a regression adjustment with all the covariates used to estimate the propensity score included in the model. Rubin12 showed that the results from both methods should often lead to the same conclusions as in the case in the MVA example above. However, one advantage to the 2-step procedure (with propensity scores) is that one can fit a very complicated propensity score model with interactions and higher order terms first. Because the goal of this propensity score model is to obtain the best estimated probability of treatment assignment, one is not concerned with over-parameterizing this model. Then when the model for the treatment effect estimation is fit, the investigator can include only a subset of the most important variables, such as the propensity score, in the model. This smaller model may allow the investigator to perform diagnostic checks on the fit of the model more reliably than if many covariates were included in the model.

One can combine the previous 2 techniques, stratification and regression adjustment, by first stratifying the data on the basis of the propensity score and then using regression adjustment with a subset of important covariates within each stratum. It has been suggested that this estimator of the treatment effect may be better than deriving the treatment effect with any of the 3 methods (matching, stratification, or regression adjustment) alone.

Summary

Propensity scores are being widely used in statistical analyses, particularly in the area of cardiovascular disease research. Their use is likely to continue to increase as the cost for randomized clinical trials rises and more investigators turn to observational studies as a method of research. The propensity score methodology appears to produce the greatest benefits when it can be incorporated into the design stages of studies (through matching or stratification). These benefits include providing more precise estimates of the true treatment effects as well as saving time and money. This savings results from an ability to avoid recruitment of subjects who may not be appropriate for particular studies. The propensity score is not the only tool that can be used in analysis of data from observational studies; rather, it should be thought of as an additional tool available to investigators as they try to estimate the effects of treatments in studies where potential bias may exist.

Acknowledgments

The author would like to thank his wife Carey and his family for their support.

Source of Funding

This work was supported in part by National Cancer Institute Grant 1 RO1 CA79934.

Disclosures

None.

References


Key Words: cardiovascular diseases ▪ epidemiology ▪ risk factors ▪ statistics

Downloaded from http://ahajournals.org by on September 17, 2019