A brief guide to propensity score analysis

Ameneh Ebrahim Valojerdi1, Leila Janani*2

Received: 6 May 2017 Published: 7 Dec 2018

Abstract

In the statistical analysis of observational data, propensity score is a technique that attempts to estimate the effect of a treatment (exposure) by accounting for the covariates that predict receiving the treatment (exposure). The aim of this paper is to provide a brief guide for clinicians and researchers who are applying propensity score analysis as a tool for analyzing observational data. We reviewed literature about how, when and why propensity score is used and then we discussed some important practical issues in using propensity score in observational studies. Applying propensity score as a method for analyzing observational studies is very useful but, we should know when and how we can use this method. Moreover, new methods of propensity score analysis such as Bayesian and doubly robust approaches were established in recent years, and these methods could be more useful for researchers in estimating causal effect from observational studies.

Keywords: Propensity score, Observational study, Causal inference

Introduction

Randomized controlled trials (RCTs) are considered the “gold standard” for assessing intervention effects because of their random allocation in the assignment of units to groups (1). But there are some limitations for using this type of design. For example, cost or ethics may imply that an RCT is impossible. In these cases, the researcher can use observational studies; e.g. investigating the causal relationship between insulin therapy in diabetic patients and incidence of cardiovascular disease (CVD). We know that RCT is the best option in this situation, although it might be unethical because of random allocation of patients in two groups (insulin user and insulin naïve). However, depending on the clinical situation, doctors decide to prescribe oral medication or injectable insulin. In this situation, we need to design an observational study. But in this design, defining causal relationship between insulin therapy and CVD is not easy because of many covariate and confounders such as blood pressure, BMI (Body Mass Index), lipid profile and etc. Moreover, the statistical methods for adjusting numerous covariates (for example regression models) need a large sample size and include complex interpretations. In RCTs, random treatment assignment allows one to establish causation (the intervention causes improvement in outcome) and to obtain an unbiased assessment of the treatment effect (2). Therefore, we need a method to obtain causal relationships in observational studies (relation between insulin therapy and CVD in our example). Rosenbaum and Rubin described a score for observational study in which the probability of a subject’s treatment (exposure) group is determined as a function of the measured covariates for that subject (3). This score was named “propensity score”, which is expressed as:

\[ p_i = \Pr(Z_i = 1 | X_i) \]  

Assuming that Z is the treatment (exposure) variable, and X is the background variables. Conditioning on this probability can produce an unbiased estimation of the average treatment effect (4). Bias due to unmeasured covariates may still exist (3). It should be noted that the propensity score as defined by Rosenbaum and Rubin implies a
A brief guide to propensity score analysis

Fig. 1. Identification of studies with “propensity score” term in title/abstract from 1987 until 2016 in PubMed

Treatment with two levels, for example, treatment versus control, or new therapy versus standard therapy. In our example, “z” is binary variable (insulin therapy or oral drug), “X” is a vector of covariates such as blood pressure, BMI, lipid profile and etc. and “Y” is the incidence of CVD (yes or no).

The systematic review that was published in 2006 illustrated an increase in the use of propensity scores term within the past several years (5). Searching this term in PubMed, we noticed this growing trend in literature as well (Fig. 1). Moreover, medical researchers used the Propensity Score (PS) in important topics in recent years (6-10). Examples of applying this method in medical literature are: to show an association between depression and subsequent substance use for men and women; to assess the effect of teenage alcohol use on education attainment; and to compare the results of regression and PS methods for right heart catheterization (7, 11, 12).

However, some clinical researchers are not familiar with the applications of PS and its assumptions. The aim of this paper is to provide useful information for clinicians and researchers on how to apply propensity score analysis as a tool for analyzing observational data. Moreover, another goal of this study is to guide researchers when and how to use this method.

Method of estimation of the propensity score

The propensity score is often estimated using a logistic regression model. In this model, treatment (exposure) status is regressed on observed characteristics (covariates). In the assumed example, insulin variable is regressed on blood pressure, BMI, lipid profile and etc. The estimated propensity score is the predicted probability of the fitted regression model (3). The PS is able to incorporate a larger number of background covariates because it uses the covariates to estimate a single number (8). After estimating the propensity score, there are four methods of using this score to control covariates: matching, stratification, inverse probability of treatment weighting, and covariate adjustment.

Methods of using the PS

Propensity score matching

In PS matching, a subject in the treatment group (exposure group) is selected randomly and matched with an untreated subject base on their propensity score (3). The common implementation of propensity score matching is one-to-one matching, in which pairs of treated and untreated subjects have similar values of the propensity score (13). Matching can be done with or without replacement, but matching with replacement can decrease bias and is helpful where the numbers of controls are limited (14).

The final consideration for matching between subjects is what “close” means in terms of distance between propensity scores. There are some methods which are used to define this. Rosenbaum & Rubin suggested using a caliper of 0.25 of the propensity score, which has been shown to remove 98% of the bias due to measured covariates (15).

Stratification on the propensity score

Stratification (sub-classification), divides subjects into separate subsets based on their propensity scores. The literature showed that five strata are adequate to reduce at least 90% of the bias associated with a confounding variable (16). With a large sample size, we can use between 10 or 20 strata (14).

Inverse probability of treatment weighting (IPTW) using the propensity score

Inverse probability of treatment weighting (IPTW) uses the propensity score as a weight. Assume $Z_i$ be an indicator variable denoting whether or not the $i^{th}$ subject was treated (or exposed); and let $e_i$ as the propensity score. The weights for subject $i$ is defined as (17):

$$w_i = \frac{1}{\hat{p}_i^{1-Z_i}}$$

This weight is equal to the inverse of the probability of receiving the treatment (or exposure) that the subject actually received.

Covariate adjustment using the propensity score

In regression adjustment, PS is employed as a covariate in the regression model. Consider this model:

$$y = \alpha_0 + \alpha_1 z_1 + \alpha_2 e_i + \epsilon_i$$

Let $Z$ is the treatment indicator and $e_i$ is the estimated propensity score. Regression adjustment is attractive because it can allow for incorporation of many covariates (4). One systematic review have shown that regression adjustment is the most commonly used propensity score method (18). However, researchers have advised that this technique should be used with caution (4), because Rubin (19) showed that bias may increase when the variance in the treated and untreated groups are very different (actually, the untreated group variance is much larger than the treated groups variance).

Some important issues

Assumptions of PS Analysis

Application of PS has several assumptions. One of these assumptions is that all covariates that are related to both the outcome and the treatment (exposure) are measured and included in the propensity score model. Many authors (7, 13, 20) highlighted a fact that, this is a strong assump-
tion, and it is untestable, because it is an assumption about unmeasured variables (21). Another major assumption of PS is the Stable Unit Treatment Value Assumption (SUTVA). This assumption says that the treatment effect for one individual is not affected by the treatment status of another. Other assumptions are the logistic regression's assumptions.

Check balance with propensity score
The final goal of PS is balancing the distribution of covariates between treatment (exposure) groups. Rosenbaum and Rubin (1984), used simple bar charts to compare proportions of particular covariates within subclasses, or strata, defined on the propensity score quintiles (22). It should be noted that the covariates for treatment and control groups after balancing on the propensity score should be balanced on their entire distributions, not solely their means or medians (13), so bar charts may not be sufficiently informative. It seems that boxplots are the most graphical approaches employed for assessing the balance (23).

Variable selection
Many authors (13, 22, 24-26), have explored the question of which covariates are important to include in a logistic regression model for estimating the propensity scores. There is some controversy in the literature (27). A few authors say that including all measured covariates in the propensity score model is the simplest approach and enhances the precision of the estimates (25). Other authors have performed simulations to illustrate that covariates related to the outcome is required for obtaining the least biased estimates of treatment effect (24).

Simulations shows including variables that are related to the exposure but not to the outcome will increase the variance of the estimated exposure effect without decreasing bias (24). Moreover, in a Monte Carlo simulation study, four propensity score models were compared; the model that included only true confounders; the model that included all variables associated with the outcome; the model that included all measured variables; and the model that included all variables associated with treatment selection; for the first two PS models, reduction in bias was greater when stratification on the quintiles of the propensity score model was employed (28).

Comparing between PS and regression
Stürmer et al. in their review published in 2006, compared the results of propensity score methods to the usual regression model for the control of confoundings. In this review, in only 13% of studies, effect size using propensity scores changed by more than 20% in comparison of conventional models (5).

On the other hand, Martens et al. showed in a simulated population that estimation of the PS methods for a general treatment effect is closer to the true marginal treatment effect than a logistic regression model (29).

However, some authors reported that in studies with small number of events relative to the number of confounders (fewer than eight events per confounder), analysis based on propensity scores yielded estimates with less biased, more robust, and more precise than a regression model (30, 31).

Alternative methods
The mentioned classic methods have some limitations; therefore, two newer methods were introduced recently:

Doubly robust propensity score
Both outcome regression and propensity score methods are unbiased only if the statistical model is correctly specified. Doubly robust method estimates the causal effect of an exposure on an outcome by combining a form of outcome regression with a model for the exposure (i.e., the propensity score). This method needs only 1 of the 2 models to be correctly specified to obtain an unbiased effect estimator.

Doubly robust estimator is a relatively new method. Although this approach has been described in the statistical literature, it is not yet well known among the researchers (32).

Bayesian propensity score
Despite their popularity, conventional propensity score estimation methods do not take into account uncertainties in propensity scores. McCandless et al. in 2009 introduced Bayesian propensity score estimators to model the joint likelihood of both propensity score and outcome in one step, which naturally incorporates such uncertainties into causal inference. They modeled the joint distribution of the data with the propensity score as a latent variable and suggested Markov chain Monte Carlo (MCMC) method to simulate from the posterior distribution for estimating model parameters (33).

Conclusion
Application of propensity score as a method for analyzing observational study is very useful, but we should know when and how to use this method. New methods of propensity score analysis such as Bayesian and doubly robust approaches were established in recent years, and these methods could be more useful for researchers in estimating causal effect from observational studies. Doubly robust estimator is unbiased when there is a misspecification in the outcome or propensity score model and Bayesian approach can take into account uncertainties in estimations.

Conflict of Interests
The authors declare that they have no competing interests.

References
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in

http://mjiri.iums.ac.ir
Med J Islam Repub Iran. 2018 (7 Dec); 32.122.
A brief guide to propensity score analysis