Debate Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2018(7 Dec);32.122. https://doi.org/10.14196/mjiri.32.122

A brief guide to propensity score analysis

Ameneh Ebrahim Valojerdi[1], 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. Appling 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

Conflicts of Interest: *None declared*
Funding: *None*

Cite this article as: Ebrahim Valojerdi A, Janani L. A brief guide to propensity score analysis. *Med J Islam Repub Iran. 2018* (7 Dec);32:122. https://doi.org/10.14196/mjiri.32.122

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 larg e 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:

$$e_i = Pr(Z_i = 1 | X_i) \quad (1)$$

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

Corresponding author: *Dr Leila Janani, Janani.l@iums.ac.ir*

1. Endocrine Research Center, Institute of Endocrinology and Metabolism, Iran University of Medical Sciences, Tehran, Iran
2. Department of Biostatistics, School of Public Health, Iran University of Medical Sciences, Tehran, Iran, & Preventive Medicine and Public Health Research Center, Iran University of medical sciences, Tehran, Iran

↑What is "already known" in this topic:
Application of propensity score as a method for analyzing observational study is very useful.

→What this article adds:
This article explains how and when we can use the propensity score.

[DOI: 10.14196/mjiri.32.122]

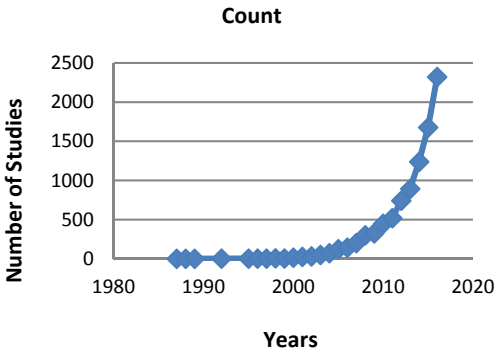

Fig. 1. Identification of studies with "propensity score" term in title/abstract from 1987 until 2016 in PubMed

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 ith subject was treated (or exposed); and let e_i as the propensity score. The weights for subject i is defined as (17):

$$\frac{Z_i}{e_i} + \frac{1-Z_i}{1-e_i} \quad (2)$$

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:

$$y = \alpha_0 + \alpha_1 * z_i + \alpha_2 * e_i \quad (3)$$

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-

[DOI: 10.14196/mjiri.32.122]

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

1. LeLorier J, Gregoire G, Benhaddad A, Lapierre J, Derderian F. Discrepancies between meta-analyses and subsequent large randomized, controlled trials. New England J Med. 1997;337(8):536-42.
2. Freidlin B, Korn E. Assessing causal relationships between treatments and clinical outcomes: always read the fine print. Bone Marrow transplant. 2012;47(5):626-32.
3. Rosenbaum PR, Rubin DB. The central role of the propensity score in

[DOI: 10.14196/mjiri.32.122]

observational studies for causal effects. Biometrika. 1983;70(1):41-55.
4. d'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998;17(19):2265-81.
5. Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol. 2006;59(5):437. e1-. e24.
6. Fukuta H, Goto T, Wakami K, Ohte N. Effect of renin-angiotensin system inhibitors on mortality in heart failure with preserved ejection fraction: a meta-analysis of observational cohort and randomized controlled studies. Heart Fail Rev. 2017:1-8.
7. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. Health Serv Outcom Res Methodol. 2001;2(3):259-78.
8. D'Agostino RB. Propensity scores in cardiovascular research. Circulation. 2007;115(17):2340-3.
9. Kazmi WH, Obrador GT, Khan SS, Pereira BJ, Kausz AT. Late nephrology referral and mortality among patients with end-stage renal disease: a propensity score analysis. Nephrol Dialys Transplant. 2004;19(7):1808-14.
10. Karkouti K, Beattie WS, Dattilo KM, McCluskey SA, Ghannam M, Hamdy A, et al. A propensity score case-control comparison of aprotinin and tranexamic acid in high-transfusion-risk cardiac surgery. Transfusion. 2006;46(3):327-38.
11. Green KM, Stuart EA. Examining moderation analyses in propensity score methods: Application to depression and substance use. J Consul Clin Psychol. 2014;82(5):773.
12. Staff J, Patrick ME, Loken E, Maggs JL. Teenage alcohol use and educational attainment. J Stud Alcohol Drugs. 2008;69(6):848-58.
13. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariat Behav Res. 2011;46(3):399-424.
14. Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical science: a review. J Institute Math Stat. 2010;25(1):1.
15. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39(1):33-8.
16. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics. 1968:295-313.
17. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv Outcom Res Methodol. 2001;2(3-4):169-88.
18. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med. 2008;27(12):2037-49.
19. Rubin DB. Using multivariate matched sampling and regression adjustment to control bias in observational studies. J Am Stat Assoc. 1979;74(366a):318-28.
20. Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. Stat Med. 2004;23(19):2937-60.
21. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med. 2013;32(19):3388-414.
22. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc. 1984;79(387):516-24.
23. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med. 2015;34(28):3661-79.
24. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol. 2006;163(12):1149-56.
25. Emsley R, Lunt M, Pickles A, Dunn G. Implementing double-robust estimators of causal effects. Stat J. 2008;8(3):334-53.
26. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods. 2010;15(3):234.
27. Millimet DL, Tchernis R. On the specification of propensity scores, with applications to the analysis of trade policies. J Bus Econ Stat. 2009;27(3):397-415.
28. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stati Med. 2007;26(4):734-53.
29. Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol. 2008;37(5):1142-7.
30. Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol. 2003;158(3):280-7.
31. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996;49(12):1373-9.
32. Funk MJ, Westreich D, Wiesen C, Stürmer T, Brookhart MA, Davidian M. Doubly robust estimation of causal effects. Am J Epidemiol. 2011;173(7):761-7.
33. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. Stat Med. 2009;28(1):94-112.

[DOI: 10.14196/mjiri.32.122]