Educ Psychol Rev (2011) 23:273–278
DOI 10.1007/s10648-011-9164-9

REVIEW

Using Propensity Score Analysis for Making Causal Claims in Research Articles

Haiyan Bai

Published online: 15 May 2011

Abstract The central role of the propensity score analysis (PSA) in observational studies is for causal inference; as such, PSA is often used for making causal claims in research articles. However, there are still some issues for researchers to consider when making claims of causality using PSA results. This summary first briefly reviews PSA, followed by discussions of its effectiveness and limitations. Finally, a guideline of how to address these concerns is also provided for researchers to make appropriate causal claims using PSA results in their research articles.

Keywords Propensity score · Observational studies · Causal effects · Causal inference

Propensity score analysis (PSA) developed by Rosenbaum and Rubin (1983) has become a popular approach in estimating treatment effects for observational studies. Rosenbaum and Rubin (1983) highlighted that the central role of PSA in observational studies is for causal inference regarding treatment effects. PSA aims at increasing the validity of causal inference from observational studies through balancing the distributions of the observed covariates between the treatment and comparison groups (Rubin 1997). As such, PSA has been widely applied in the fields of education and other social and behavioral sciences. A Web of Science (Thomson Corporation 2009) search using "propensity score" as a topic keyword found that the frequency of using PSA in published studies has exponentially increased from January 1983 to July 2009 (see Fig. 1). Among these publications, many empirical studies were concluded with claims of causality for treatment effects; however, the perspective statements with causal claims can be fraught with pitfalls because the research findings from PSA results may rely on unwarranted effects of bias correction (Rubin 1997). With this concern, this summary will discuss some issues for researchers to

H. Bai (✉)
Department of Educational and Human Sciences, University of Central Florida, P.O. Box 161250, Orlando, FL 32816-1250, USA
e-mail: Haiyan.Bai@UCF.edu

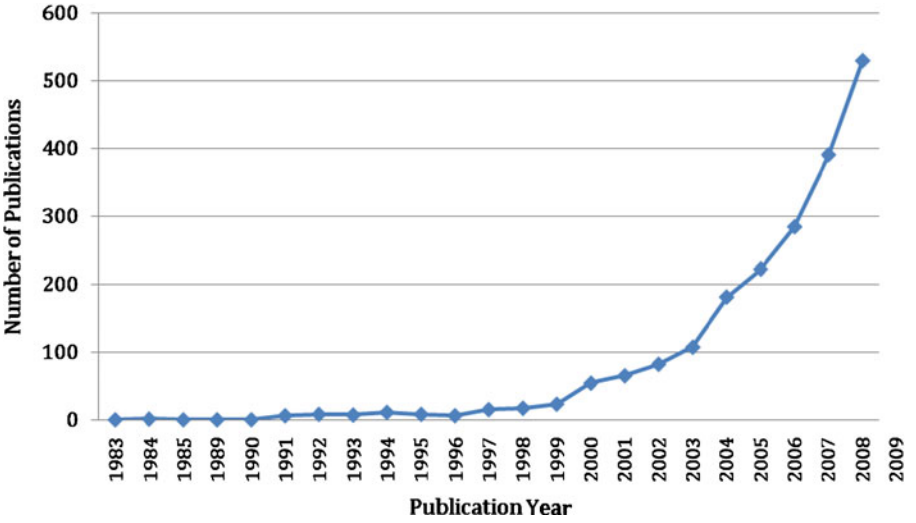

Fig. 1 Number of articles in a Web of Science search for articles published from January 1983 to July 2009 by publication year using "propensity score" as a topic keyword

consider when they tend to make claims of causality using PSA results in their perspective statements. In the following sections, specifically, the concept of PSA will be briefly reviewed, followed by discussions of the effectiveness and limitations of PSA and suggestions for researchers to make appropriate causal claims using PSA in their research articles.

A Primer of Propensity Score Analysis

The concept and effectiveness of PSA are discussed at length in the literature (e.g., Abadie and Imbens 2006; Dehejia and Wahba 2002; Gu and Rosenbaum 1993; Heckman *et al.* 1998; Hill and Reiter 2006; Hirano *et al.* 2003; McCandless *et al.* 2008; Rosenbaum 1987; Rubin and Thomas 1996). Here, only a primer of PSA is introduced. Rosenbaum and Rubin (1983) described that PSA is a method to use balancing scores, namely the propensity score, to compare groups so that direct comparisons of the observational data are more meaningful with the groups balanced on the covariates (Gu and Rosenbaum 1993; Rosenbaum and Rubin 1983; Weitzen *et al.* 2004). A propensity score is the conditional probability $e(\mathbf{x})=Pr(Z=1 \mid \mathbf{X}=\mathbf{x})$ of receiving the treatment given covariates **x**. In other words, a propensity score used to reduce the selection bias through balancing groups based on the observed covariates is the probability of a unit (e.g., student, classroom, and school) being assigned to a particular condition in a study given a set of observed covariates (e.g., age, gender, ethnicity, socioeconomic status, or prior performance scores). Propensity score methods allow adjustment for multiple covariates simultaneously in order to balance comparison groups for estimating treatment effects (Gu and Rosenbaum 1993; Rubin 1997). PSA commonly has four basic steps: (a) identifying and measuring as many covariates as possible based on the theory and prior research, (b) estimating propensity scores using a logistic regression or linear discriminant analysis as suggested by Rosenbaum and Rubin (1984, 1985), (c) matching each of the cases in the treatment group with one or more in the comparison group based on the propensity score or

stratifying the study sample using the propensity score, and (d) conducting the intended analysis on the matched sample or with propensity score adjustment. Researchers can use introductory articles (e.g., Caliendo and Kopeinig 2008; Hahs-Vaughn and Onwuegbuzie 2006; Guo *et al*. 2006; Rosenbaum and Rubin 1983) for further details of the above procedures. The following section will focus on the discussion of the Effectiveness and Limitations of PSA in practice.

Effectiveness and Limitations of PSA

After Rosenbaum and Rubin's (1983) seminal work on PSA, many studies about the rationale and effectiveness of PSA flooded the literature. Glynn *et al.* (2006) concluded that researchers have established reasons in favor of using PSA. In terms of the effectiveness of PSA, Gu and Rosenbaum (1993) found that propensity score matching usually outperformed other matching techniques because it balanced on many covariates simultaneously, which potentially approximated the balance achieved through randomization. Therefore, PSA is regarded as an effective strategy from alternative designs (Glynn *et al*. 2006) and an optimal tool for causal questions with large datasets as well (Rubin 1997).

The effectiveness of PSA is also echoed by Heckman *et al*. (1998), who indicated that propensity scores are robust to choice-based sampling. Furthermore, PSA also can be used to identify interactions between propensity of treatment and treatment effects on outcomes (Kurth *et al*. 2006). In sum, the effectiveness of PSA can be summarized as "if the conditional independence is assumed between the treatment assignment and potential outcomes given the observed covariates (strongly ignorable treatment assignment), it is possible to obtain unbiased estimates of treatment effects for causal inferences" (Imai and Van Dyk 2004, p. 854).

While embracing PSA as an effective methodological remedy to the unavoidable flaws from observational studies, methodologists, statisticians, and researchers also discerned the limitations of the popular method for improving the validity of claims on causality from observational data. Guo *et al.* (2006) stated that PSA could not provide ultimate answers to causal questions or treatment effects. Based on the literature (Guo *et al.* 2006; Rubin 1997; Rosenbaum and Rubin 1983), major limitations of using PSA are threefold. First, propensity scores cannot adjust for the "hidden bias" from *unobserved* covariates (Joffe and Rosenbaum 1999; Rosenbaum and Rubin 1983; Rubin 1997). Propensity scores can only be obtained from the observed covariates, but there could be other unknown confounders to influence the treatment effect. Therefore, the accuracy and precision of estimates from logistic models or discriminate analysis used for either adjusting estimates of the effects or predicting outcomes could be seriously affected by missing predictors or confounders (Greenland 1989; Hosmer and Lemeshow 2000; Rothman and Greenland 1998; Weitzen *et al*. 2004). Second, propensity scores may only work well with large samples (Rubin 1997) because there could be very little overlap between the treatment and comparison groups with respect to the distribution of the propensity scores in a small sample. This may result in a large proportion of lost cases due to the lack of matching, which consequently only leaves a few cases for analysis (Weitzen *et al*. 2004). Third, propensity scores consider only the covariates that are related to the treatment assignment but not the outcome. This is in contrast to the analysis of a covariate, which is associated with both the outcome and the treatment assignment (Rubin 1997; Rubin and Thomas 1996). In this case, the inclusion of irrelevant covariates usually reduces the efficiency of balancing on the relevant covariates (Rubin 1997).

Given the aforementioned limitations, there have been some doubts of the interpretations on the results from PSA concerning its scientific validity in causal claims. The often serious concerns of the potential selection bias in observational studies, however, suggest that it is still desirable to use PSA if a reduction in bias can be achieved (Stürmer *et al*. 2006). As such, researchers should be responsible to provide sufficient empirical evidence for controlling the limitations, addressing the unsolved issues, and interpreting their research findings from PSA appropriately. The next section will discuss how researchers can address the above concerns when making causal claims using PSA.

Strategies for Making Appropriate Causal Claims with Results of PSA

As discussed, the use of propensity scores is not guaranteed to reduce bias due to evident limitations. However, the use of PSA is recommended in observational studies for treatment effects conditionally because it effectively adjusts the existing bias if the concerns are well addressed, which therefore, improves the validity of estimations (Glynn *et al.* 2006). The related literature suggests various strategies for the correct use of PSA in empirical investigations for the study of causal effects.

To deal with "hidden bias", Rubin (1997) recommended performing sensitivity analysis (Caliendo and Kopeinig 2008; Lechner 2001; Rosenbaum and Rubin 1983) and testing different sets of conditioning variables to address the limitations of the lack of balancing unobserved variables (Guo *et al.* 2006; Michalopoulos *et al*. 2004). Researchers should be aware that it is essential to test the robustness of results to departures from the identifying assumption; and if the sensitivity of estimated effects was found with respect to a failure of the unconfoundedness assumption, researchers should either consider adopting alternative identifying assumptions or combine propensity score matching procedures with other evaluation approaches (Caliendo and Kopeinig 2008). However, if the quality of matching is reasonable, researchers can use PSA in estimating the treatment effects (Caliendo and Kopeinig 2008). In this case, it can be appropriate for researchers to make causal claims for the treatment effects.

With regards to the third issue, the efficiency of estimating or calculating propensity scores with inclusion of covariates associated with treatment assignment but not outcomes, Rubin and Thomas (1996) again suggested that it may not be a substantial issue in practice if modest or large datasets are used in the analysis. However, researchers are responsible in making legitimate judgments (Rubin and Thomas 1996; Weitzen *et al*. 2004) in selecting covariates so that the use of PSA is more efficient and the analysis results are more accurate for causal claims.

There are a few other caveats that also need to be considered when making causal claims using PSA. First, to appropriately claim the treatment effect, the researcher should be responsible to include information on variable selection. Otherwise, it is difficult to assess whether all potential and available confounders are adjusted for the appropriateness of using PSA (Weitzen *et al*. 2004). Variable selection is an important procedure which is often ignored by researchers (Brookhart *et al*. 2006; Greenland 2007; Weitzen *et al*. 2004). The appropriate way to select covariates to be included in PSA is to identify the related confounding variables from theory to existing literature (Brookhart *et al*. 2006; Greenland 2007). PSA performs better when the researcher includes as many theoretically important covariates as possible in the analysis, and therefore, if any theoretically identified covariate cannot be included in PSA because it is immeasurable, researchers should address the

concern for any causal claims related to the treatment effects (Dehejia and Wahba 2002; Hirano *et al*. 2003).

Another caveat is that model selections regarding the methods used for obtaining and matching propensity scores are essential in PSA. Researchers should fully understand the role of the recommended criteria for logistic regression in the estimation of useful propensity scores (Cepeda *et al*. 2003) and the effectiveness of various matching methods in order to select the most efficient strategy for bias reduction.

Finally, PSA like any other statistical methods should not be mechanically regarded as a preferable and sole method to control for confounding variables in observational studies but rather as a promising addition (Stürmer *et al*. 2006). Therefore, we should not expect PSA to provide definitive answers to causal questions of treatment effects. As Rubin (1997) commented, "In observational studies, confidence in causal conclusion must be built by seeing how consistent the obtained answers are with other evidence…" (p. 762). Guo *et al*. (2006) also emphasized that researchers using PSA with observational data "should be cautious about these limitations and make efforts to warrant that interpretation of study results does not go beyond the limits of data and their analytical methods" (p. 380).

In conclusion, for observational studies using PSA to improve the validity of causal inference, researchers need to address the above issues and justify their evidence before reaching any conclusive causal claims or prescriptive statements of their research articles.

References

Abadie, A., & Imbens, G. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica, 74*(1), 235–267.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149–1156.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31–72.

Cepeda, M. S., Boston, R., Farrar, J. T., & Strom, B. L. (2003). Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American Journal of Epidemiology, 158*(3), 280–287.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *The Review of Economics and Statistics, 84*(1), 151–161.

Glynn, R. G., Schneeweiss, S., & Stürmer, T. (2006). Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clinicla Pharmacology Toxicology, 98*(3), 253–259.

Greenland, S. (1989). Model and variable selection in epidemiologic analysis. *American Journal of Public Health, 79*(3), 340–349.

Greenland, S. (2007). Invited commentary: Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology, 167*(5), 523–529.

Gu, X. S., & Rosenbaum, P. R. (1993). Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics, 2*(4), 405–420.

Guo, S., Barth, R., & Gibbons, C. (2006). Propensity score matching strategies for evaluating substance abuse services for child welfare clients. *Children and Youth Services Review, 28*, 357–383.

Hahs-Vaughn, D. L., & Onwuegbuzie, A. J. (2006). Estimating and using propensity score analysis with complex samples. *The Journal of Experimental Education, 75*(1), 31–65.

Heckman, J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing selection bias using experimental data. *Econometrica, 66*(5), 1017–1098.

Hill, J., & Reiter, J. P. (2006). Interval estimation for treatment effects using propensity score matching. *Statistics in Medicine, 25*, 2230–2256.

Hirano, K., Imbens, G., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161–1189.

Hosmer, D. L., & Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: Wiley.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association, 99*, 854–866.

Joffe, M. M., & Rosenbaum, P. R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology, 150*, 327–333.

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., et al. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology, 163*, 262–270.

Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In M. Lechner & F. Pfeifer (Eds.), *Econometric evaluation of labour market policies* (pp. 1–18). Heidelberg: Physica.

McCandless, L. C., Gustafson, P., & Austin, P. C. (2008). Bayesian propensity score analysis for observational data. *Accepted in Statistics in Medicine, 28*, 94–112.

Michalopoulos, C., Bloom, H. S., & Hill, C. J. (2004). Can propensity-score methods match the findings from a random assignment evaluation of mandatory welfare-to-work programs? *The Review of Economics and Statistics, 86*, 156–179.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *Journal of the American Statistical Association, 82*, 387–394.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41–50.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*, 516–524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*(1), 33–38.

Rothman, K. J., & Greenland, S. (1998). *Modern epidemiology* (2nd ed.). Philadelphia: Lippincott-Raven.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine, 127*(8), 757–763.

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*, 249–264.

Stürmer, T., Joshi, M., Glynn, R. J., Avorn, J., Rothman, K. J., & Schneeweiss, S. (2006). 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. *Journal of Clinical Epidemiology, 59*(5), 437–461.

Thomson Corporation (2009*). Web of science*. Retrieved from http://scientific.thomson.com/products/wos/

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: A systematic literature review. *Pharmacoepidemiology and Drug Safety, 13*, 841–853.