Cardiovasc Prev Pharmacother. 2019 Oct;1(2):57-62
https://doi.org/10.36011/cpp.2019.1.e6
eISSN 2671-700X

Cardiovascular Prevention and Pharmacotherapy

Special Article

Check for updates

Improving Causal Inference in Observational Studies: Propensity Score Matching

Min Heui Yu, MS[1], Dae Ryong Kang, PhD[1,2]

[1]Department of Biostatistics, Yonsei University Wonju College of Medicine, Wonju, Korea
[2]Department of Precision Medicine, Yonsei University Wonju College of Medicine, Wonju, Korea

OPEN ACCESS

Received: Oct 9, 2019
Accepted: Oct 20, 2019

Correspondence to
Dae Ryong Kang, PhD
Department of Biostatistics, Yonsei University Wonju College of Medicine, 20 Ilsan-ro, Wonju 26426, Korea.
E-mail: dr.kang@yonsei.ac.kr
kdr.bmc@gmail.com

ORCID iDs
Min Heui Yu
https://orcid.org/0000-0003-3787-795X
Dae Ryong Kang
https://orcid.org/0000-0002-8792-9730

Conflict of Interest
The authors have no financial conflicts of interest.

Author Contribution
Writing - original draft: Yu M; Writing - review & editing: Kang DR.

ABSTRACT

Propensity score matching (PSM) is a useful statistical methods to improve causal inference in observational studies. It guarantees comparability between 2 comparison groups are required. PSM is based on a "counterfactual" framework, where a causal effect on study participants (factual) and assumed participants (counterfactual) are compared. All participants are divided into 2 groups with the same covariates matched as much as possible. Propensity score is used for matching, and it reflects the conditional probabilities that individuals will be included in the experimental group when covariates are controlled for all subjects. The counterfactuals for the experimental group are matched between groups with characteristics as similar as possible. In this article, we introduce the concept of PSM, PSM methods, limitations, and statistical tools.

Keywords: Causality; Epidemiologic studies; Logic; Observational study; Propensity score

INTRODUCTION

In clinical studies comparing 2 groups, researchers must show that the effect of an explanatory variable on a response variable is due only to the studied intervention. To do so, all subjects in the study should have the same characteristics. In a randomized control study, random treatment allocation ensures that the treatment status is not confounded with either measured or unmeasured baseline characteristics.[1] However, in observational studies, limited control of confounding variables makes it difficult to explain causality and uncover clear medical evidence. Improving causal inference in observational studies requires statistical methods that are able to guarantee comparability between 2 comparison groups; propensity score matching (PSM) is one of these methods. For example, in their study of aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery diseases, Gum et al.[2] used PSM to match patients between the aspirin group and no aspirin group. Seung et al.[3] applied PSM in their study of stents versus coronary-artery bypass grafting (CABG) for left main coronary artery disease, matching patients between the stent group and CABG group. Yao et al.[4] used PSM in their study comparing surgical left atrial appendage occlusion (LAAO) with no surgical LAAO.

PSM is based on a "counterfactual" framework derived from Rubin's causal model.[5] The counterfactual condition assumes that an individual in an experiment group did not participate in the experiment. Therein, the causal effect of interest to researchers would be the difference between the effect on a subject from participating in a clinical study (factual) and the effect on the subject if he/she did not participate in the study (counterfactual). For example, in a study of the effects of a new drug, the effects derived from taking the drug are factual, and those derived from not taking the drug are counterfactual. In general experimental studies, it is difficult to compare treatment effects in the same person. Since the experimental group cannot observe the situation where they do not participate in the study, a control group is established as a counterfactual to estimate the potential effect of the intervention.[6]

PSM divides all subjects into 2 groups with the matching covariates as much as possible. This matching is based on propensity scores (PSs), which reflect the conditional probabilities that individuals will be included in the experimental group when covariates are controlled for all subjects. That is, the counterfactuals for the experimental group are matched between groups with close to homogeneous characteristics.

In this article, we introduce the concept of PSM, PSM methods, limitations, and statistical tools for conducting PSM.

DEFINITION OF PS

PS is defined as the "conditional probability" that a patient *i* will be in the treatment group given his/her measured covariates.[7] In other words, the PS is a balancing score that summarizes the covariates. If X is a dummy variable that takes on value 1 for treatment and 0 for control, and assuming that the covariate to be matched is c, the PS can be expressed as Equation 1.

$$PS=e(x_i)=p(X_i=1|C_i=c_i) \text{ (Equation 1)}$$

Matching the same covariates, as much as possible, across two groups using the PS values from Equation 1 enables the treatment and control groups to be reconstructed with reduced selection bias. The PS corresponds to the predicted probability of the logistic regression model. In this case, the predictive probability should be obtained by using a full non-parsimonious model that includes all possible covariates investigated. Moreover, some interaction terms or squared terms should also be included. In addition, the c-statistic and suitability (Hosmer–Lemeshow statistic) of the predictive model should be examined.[8] Once estimated, PS can be used to reduce bias through matching, stratification (sub-classification), regression adjustment, or some combination of all 3.[8]

Three assumptions are needed to apply PSM (**Figure 1**).[9]

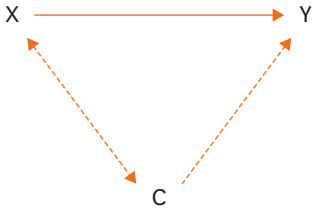
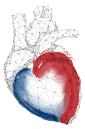


Figure 1. Relationships among the treatment, covariates, and outcomes.[9]

https://doi.org/10.36011/cpp.2019.1.e6



When X represents a treatment, C denotes covariates, and Y is the outcome, the following 3 assumptions can be drawn.

First, the treatment is related to the covariates:

$$X \Leftrightarrow C$$

Second, with the given covariates, the treatment is independent of the outcome:

$$(X \perp Y) \mid C$$

Third, with the given covariates, the treatment is still related to the outcome because of hidden selection bias:

$$X \Leftrightarrow Y \mid C$$

PSM methods

The following methods are utilized for PSM: local optimal (greedy) algorithms, global optimal algorithms, and Mahalanobis metric matching.[7] Of these, greedy algorithms are used most often and include the nearest neighbor, caliper, and radius methods. The nearest neighbor method is the most basic method of matching cases in the control group with the smallest PS difference from the treatment cases. Next, the caliper method matches the case closest to a control's PS within the tolerance level of the treatment's PS (caliper±propensity). Caliper refers to the maximum allowable distance between 2 objects. When a caliper is specified, it will not match cases if the Euclidean distance is higher than the specified value. A caliper of 1/4 the standard deviation of the logit transformation of the PS can also work well to reduce bias.[10] Lastly, the radius method uses all comparison units within a pre-defined PS.[11] One benefit of this matching approach is that it uses only as many comparison units as are available within the calipers, allowing for the use of extra (fewer) units when good matches are (not) available.[11]

Global optimal algorithms provide a method for determining the “differences” in PSs between all treatment and control cases, and for sorting these differences in order to select the smallest value as the matching value. Mahalanobis metric matching uses PSs and different covariate values.

BALANCE CHECK AFTER PSM

After PSM, a check is needed to verify that it has been performed properly. One way to check whether the PSM is appropriate is to calculate standardized mean differences. Recently, L1 statistics[12] and D2 statistics[13] have been proposed as alternatives. Based on these statistics, a researcher can plan PSM again.

The standardized mean difference reflects the difference in covariates between two groups using mean, variance, and proportion values, and can be assessed using the equations below. A standardized mean difference greater than 10 percent represents meaningful imbalance.[14] Equations 2 and 3 reflect the case of continuous variables and categorical variables, respectively.

https://doi.org/10.36011/cpp.2019.1.e6

$$d = \frac{100 \cdot (\bar{x}_{treatment} - \bar{x}_{control})}{\sqrt{(s^2_{treatment} + s^2_{control}) / 2}} \quad \text{(Equation 2)}$$

$\bar{x}$: mean of covariates in the treatment group and in the control group
s^2: variance of covariates in the treatment group and in the control group

$$d = \frac{100 \cdot (p_{treatment} - p_{control})}{\sqrt{[p_{treatment}(1 - p_{treatment}) + p_{control}(1 - p_{control})] / 2}} \quad \text{(Equation 3)}$$

p: proportion of covariates in the treatment group and in the control group

L1[12] is a multivariate imbalance measure for matching balance. Values of L1 are easily interpretable. If the distributions of two groups are completely different, then L1=1, indicating that the PSM results are imbalanced. Conversely, if the 2 distributions overlap exactly, then L1=0, indicating good balance.

D2[13] is a test statistic for determining the overall imbalance between treatment and control groups, that is, to evaluate whether the linear combination of covariates and covariates are unbalanced after matching. If the null hypothesis is not rejected, the structure of the 2 groups is similar, and hence, the matching is considered good. As this statistic is sensitive to the number of subjects (N), researchers should apply it with caution if N is too small.

LIMITATIONS OF PSM

While PSM can minimize selection bias, its limitations in practical applications should also be considered. PSM analysis requires large cohorts; an insufficient number, depending on the number of observed covariates, can make it difficult to use. It is also inappropriate to use PSs to estimate covariates that lack observations. Next, PSM relies on unverifiable assumptions (e.g., a strongly ignorable treatment assignment given the observed covariates or unobserved confounders), and therefore necessitates a sensitivity analysis. In addition, PSM requires substantial overlap between the treatment and control groups. If there are only a few overlapping regions, it can be difficult to select subjects with similar characteristics, thereby potentially leading to a significant loss of data. Moreover, in longitudinal studies, PS application requires the construction of time-dependent PSs (sequential matching) and inverse probability of treatment weighted estimator methods.[15]

STATISTICAL TOOLS FOR PSM

PSM can be implemented using statistical programs in SAS (SAS Institute, Cary, NC, USA), R (R Foundation, Vienna, Austria), SPSS (IBM Corp., Armonk, NY, USA), and STATA (StataCorp, College Station, TX, USA). After PS values are obtained using the PROC LOGISTIC function, matching is possible through the %PSmatching macro in SAS,[16] MatchIt package in R,[17] PSMATCH2 algorithm in STATA,[18] or by integrating the R package in SPSS.[19] If researchers find using SAS and R challenging, SPSS is recommended. To do so, R should be

https://doi.org/10.36011/cpp.2019.1.e6

installed according to the SPSS version being used; this is possible after executing a program (SPSS® Statistics-Essentials for R) that links SPSS and R.

CONCLUSION

PSM is a useful approach for researchers to improve causal inference in observational studies. However, there are some limitations and precautions that warrant consideration.[20] Accordingly, researchers ought to proceed in a manner that is appropriate for their given data.

REFERENCES

1. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399-424.
PUBMED | CROSSREF
2. Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. JAMA 2001;286:1187-94.
PUBMED | CROSSREF
3. Seung KB, Park DW, Kim YH, Lee SW, Lee CW, Hong MK, Park SW, Yun SC, Gwon HC, Jeong MH, Jang Y, Kim HS, Kim PJ, Seong IW, Park HS, Ahn T, Chae IH, Tahk SJ, Chung WS, Park SJ. Stents versus coronary-artery bypass grafting for left main coronary artery disease. N Engl J Med 2008;358:1781-92.
PUBMED | CROSSREF
4. Yao X, Gersh BJ, Holmes DR Jr, Melduni RM, Johnsrud DO, Sangaralingham LR, Shah ND, Noseworthy PA. Association of surgical left atrial appendage occlusion with subsequent stroke and mortality among patients undergoing cardiac surgery. JAMA 2018;319:2116-26.
PUBMED | CROSSREF
5. Rubin DB. Bayesian inference for causal effects: the role of randomization. Ann Stat 1978;6:34-58.
CROSSREF
6. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;66:688-701.
CROSSREF
7. Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. J Econ Surv 2008;22:31-72.
CROSSREF
8. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265-81.
PUBMED | CROSSREF
9. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.
CROSSREF
10. Stuart EA, Rubin DB. Best practices in quasi-experimental designs: matching methods for causal inference. In: Osborne J, editor. Best Practices in Quantitative Methods. Thousand Oaks, CA: Sage Publications; 2008. pp. 155-76.
11. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. Rev Econ Stat 2002;84:151-61.
CROSSREF
12. Iacus SM, King G, Porro G. Causal inference without balance checking: Coarsened exact matching. Polit Anal 2012;20:1-24.
CROSSREF
13. Hansen BB, Bowers J. Covariate balance in simple, stratified and clustered comparative studies. Stat Sci 2008;23:219-36.
CROSSREF
14. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med 2009;28:3083-107.
PUBMED | CROSSREF

https://doi.org/10.36011/cpp.2019.1.e6

15. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550-60.
PUBMED | CROSSREF
16. Coca-Perraillon M, editors. Local and global optimal propensity score matching. SAS Global Forum 2007; 2007 Apr 16–19; Orlando, FL. Cary: SAS Institute; 2007 Apr. 9 p.
17. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. J Stat Softw 2011;42:1-28.
18. Leuven E, Sianesi B. PSMATCH2: stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. [place unknown]: RePEc; [cited 2018 February]. Available from http://EconPapers.repec.org/RePEc:boc:bocode:s432001.
19. Thoemmes F. Propensity score matching in SPSS [Internet]. Ithaca, NY: arXiv, Cornell University Library; [cited 2012 January]. Available from https://arxiv.org/abs/1201.6385.
20. King G, Nielsen R. Why propensity scores should not be used for matching. Polit Anal 2016:1-20.