Letter to Editor

Propensity Score Matching in Non-Interventional Studies: A Step-by-Step Guide for Clinicians and Researchers

Dear Editor

Selection bias occurs when a study's sample is not representative of the entire population, which can lead to incorrect conclusions.[1] Propensity score matching (PSM) is a method used to address this issue, particularly in non-randomized studies where randomization is impractical or unethical.[2] PSM matches participants based on their probability of receiving a treatment, estimated from observable variables. Rather than comparing all treated and untreated units directly, the method compares units with similar propensity scores, making the groups more comparable and improving the validity of the research.

The propensity score is most often estimated using logistic regression, with the treatment indicator as the dependent variable (e.g., treated=1, untreated=0).[3] The independent variables are the covariates believed to influence treatment assignment. To illustrate this for those unfamiliar with statistics, consider the following example.

Example: Suppose that a study analyzes the causal effect of smoking on lung cancer by setting comparable groups of smokers and non-smokers based on propensity scores. Data for five patients with their scores are presented in table 1.

A logistic regression model is used to estimate the likelihood of smoking (1=yes, 0=no), conditional on observed covariates such as age (x_1) and health score (x_2):

A logistic regression model is used to estimate the probability of smoking (1=yes, 0=no), conditional on observed covariates such as age (x_1) and health score (x_2):

$$log_e \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \Rightarrow \frac{p}{1-p} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2} \Rightarrow p = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

Assuming that fitting logistic regression to this data provides estimates for β_0, β_1, and β_2 as follows:

$$\beta_0 = -2, \qquad \beta_{1(age)} = 0.03, \qquad \beta_{2\,(health\ score)} = 0.50$$

Propensity score for patient 1:

$$p1 = \frac{e^{(-2+0.03\times 45+0.5\times 7)}}{1+e^{(-2+0.03\times 45+0.5\times 7)}} = 0.945$$

We repeat this calculation for all observations to obtain the propensity scores. Smokers and non-smokers with similar scores are then matched to ensure the groups are comparable. Following matching, researchers compare the rate of lung cancer between these matched groups to measure the causal effect of smoking, having reduced the impact of confounding factors.

Once propensity scores are produced, there are several methods to use them for identifying comparable groups and investigating causal associations. The following are commonly used approaches for this purpose:[4, 5]

Table 1: Characteristics of five patients, such as age, health score, and smoking status

ID	Age	Health score	Smoking (1=yes, 0=no)	$P_{i=propensity\ score}$
1	45	7	1	0.945
2	60	4	0	…
3	50	6	1	…
4	70	3	0	…
5	55	8	1	…

1. Matching

Matching involves the comparison of treated (e.g., smoker) and untreated (e.g., non-smoker) units with similar propensity scores to create comparable groups. Common matching designs include:[6]

- ***Nearest Neighbor Matching:*** A treated unit is matched to an untreated unit with the closest propensity score.
- ***Caliper Matching:*** A match is only permitted if the propensity scores of treated and untreated units fall within a specified range or caliper.
- ***Kernel Matching:*** This method uses a weighted average of all, or many, untreated units.

Strength

It produces balanced treatment and control groups.

Results are relatively simple to interpret.

It can reduce bias related to confounding variables.

Weaknesses

It can lead to data loss if an excessive number of unmatched units are picked, with no covariate distance.

The choice of matching criteria and caliper can significantly influence the results.

Residual confounding may persist.

2. Stratification (Sub-classification)

The sample is stratified (or binned) into groups, usually in the form of quintiles or deciles, based on the propensity scores. The goal is to make comparisons among groups that are similar (figure 1).[6]

Strengths

Simple to implement and interpret.

Creates balance within all strata.

Can perform very well with large groups.

Weaknesses

The choice of strata size is critical.

Some strata may be too small for reliable estimation.

It fails to fully control for confounding by continuous variables.

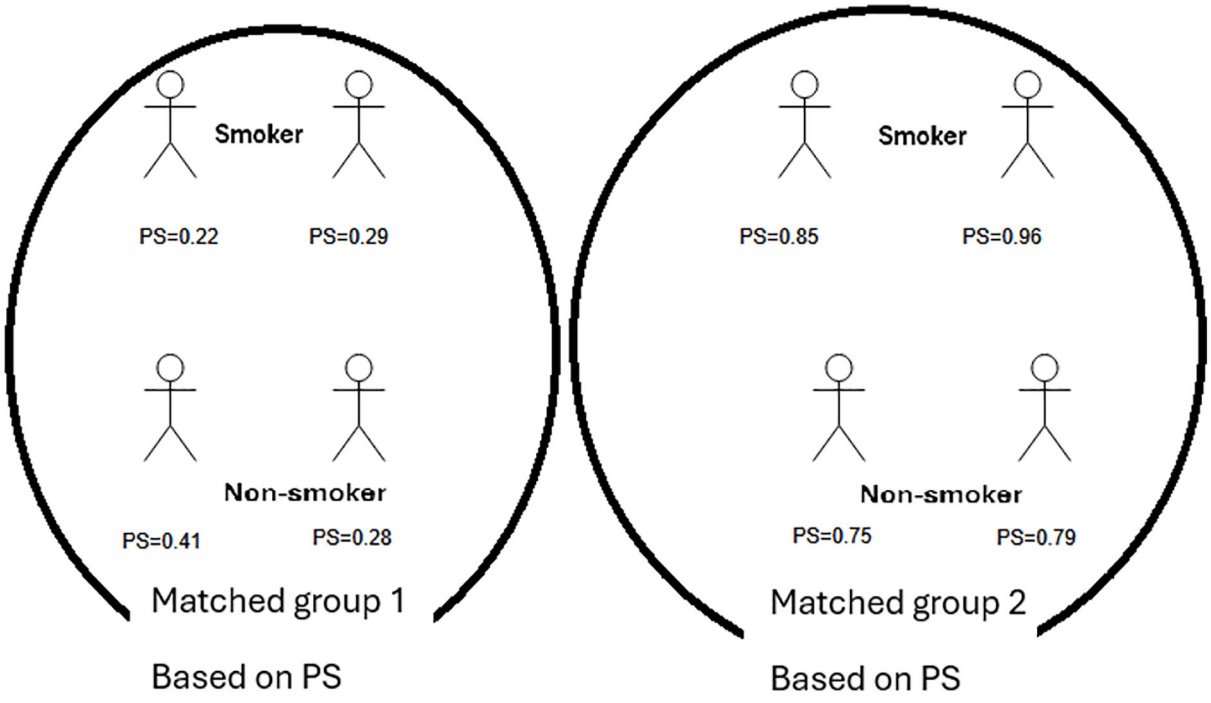

Figure 1: The patients are stratified based on their propensity score (PS) values, using a cutoff equals 0.6.

3. Inverse Probability Weighting (IPW)

Inverse probability of treatment weighting (IPW) employs the estimated propensity scores as weights. For each subject, the weight is calculated as the inverse of the probability of receiving the treatment they actually received. The weight formulas are:[5, 6]

For treated individual:

$$W_i = \frac{1}{\text{propensity score i}}$$

For an untreated individual:

$$W_i = \frac{1}{1 - \text{ propensity score i}}$$

Strengths

Uses all available data without ignoring/refusing observations.

Can eliminate possible bias when the model is correctly specified.

Provides flexibility to adjust for covariates.

Weaknesses

It requires a correctly specified propensity score model to avoid bias.

Large weights can lead to unstable estimates.

Results are sensitive to model misspecification.

4. Covariate Adjustment with Propensity Scores

Instead of matching or weighting, the propensity score itself can be included as an adjustment variable in a regression model to control for selection bias.[6]

Strengths

Easy to implement in regression models.

Can control for multiple covariates in a single analysis.

Can be a useful alternative when matching or weighting is not feasible.

Weaknesses

It can introduce collinearity between the propensity scores and other covariates in the model.

Generally, less effective than matching or weighting for achieving balance across all covariates.

The validity of the estimate depends on the correct specification of the regression model.

5. Doubly Robust Estimation

Doubly robust estimation is an approach that combines both regression modeling with weighting techniques based on propensity scores.[6] The resulting estimator is unbiased if either the propensity score model or the outcome regression model is correctly specified. This technique helps minimize bias in the comparison between treated and untreated groups to better estimate causal effects.[5]

Strengths

Less dependent on the correct specification of both the propensity score and outcome models.

Provides more robust causal estimates under a wider set of assumptions.

Reduces bias in the estimation of the causal effect.

Weaknesses

More complex to implement than standard methods such as regression adjustment.

Involves specifying two models, doubling the potential for misspecification.

Computationally intensive, especially with large sample sizes (n>1000)

Recommendation

After calculating the propensity score, the best method depends on the specific context, data quality, and research questions. Currently, IPW and doubly robust estimation are often recommended because

they effectively reduce bias and improve balance.[6] If the data allows for good matching, nearest neighbor matching with a caliper can also produce good results, particularly in smaller datasets. Ultimately, the choice of method depends on considerations of computation time, implications for subsequent analysis, and the degree of bias reduction required.

Despite its advantages, PSM faces challenges. Researchers using PSM in observational studies must carefully attend to the selection of the matching algorithm, the choice of covariates, the assessment of balance, and the credibility of the outcome results.[6] Furthermore, researchers transparently report all process choices and assumptions. The matching process, selected covariates, and key decisions must be documented to ensure the study can be replicated. A major concern, as previously noted, is overfitting. Researchers must avoid including too many covariates in the propensity score model as this can lead to spurious findings and undermine the validity of the study's results.[6]

In brief, similar to many statistical methods, PSM has both strengths and weaknesses. Its strengths include:[5]

- Reducing selection bias based on observable covariates.
- Not requiring exact matching on all relevant variables.
- Being relatively straightforward to implement and interpret.

The weaknesses of PSM are as follows:[5]

- Unobserved confounding: PSM cannot adjust for bias from unmeasured variables.
- Sample size reduction: The matching process may discard many unmatched units.
- Model dependence: The results depend on the specification of the propensity score model and the variables selected for it.

In conclusion, PSM is a powerful tool for enhancing the rigor of observational studies in medical research. When applied thoughtfully and transparently, it can significantly improve the validity and reliability of findings, ultimately benefiting patients and healthcare providers. Despite the challenges, PSM remains a valuable tool for researchers in the field of medical research. By carefully considering its strengths and limitations, researchers can leverage PSM to produce high-quality evidence and contribute to the advancement of medical knowledge.

Authors' Contribution

S.P.: Conceptualization and drafting, F.M.: Conceptualization and revising; All authors have read and approved the final manuscript and agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Conflict of Interest: None declared.

Keywords ● Propensity scores ● Naturalistic observation study ● Selection biases ● Sampling bias ● Sampling errors

Saeedeh Pourahmad[1], PhD; Farzan Madadizadeh[2], PhD

[1]Department of Biostatistics, School of Medicine, Shiraz University of Medical Sciences, Shiraz, Iran;
[2]Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, Yazd, Iran

Correspondence:
Farzan Madadizadeh, PhD;
Center for Healthcare Data Modeling, Departments of Biostatistics and Epidemiology, School of Public Health, Shahid Sadoughi University of Medical Sciences, Postal Code: 89169-78477, Bahonar Sq., Yazd, Iran
Email: madadizadehfarzan@gmail.com
Received: 26 January 2025
Revised: 16 April 2025
Accepted: 31 May 2025

Please cite this article as: Pourahmad S, Madadizadeh F. Propensity Score Matching in Non-Interventional Studies: A Step-by-Step Guide for Clinicians and Researchers. Iran J Med Sci. doi: 10.30476/ijms.2025.105595.3947.

References

1 Rojas-Saunero LP, Glymour MM, Mayeda ER. Selection Bias in Health Research: Quantifying, Eliminating, or Exacerbating Health Disparities? Curr Epidemiol Rep. 2024;11:63-72. doi: 10.1007/s40471-023-00325-z. PubMed PMID: 38912229; PubMed Central PMCID: PMCPMC11192540.

2 Caliendo M, Kopeinig S. Some practical guidance for the implementation of propensity score matching. Journal of economic surveys. 2008;22:31-72. doi: 10.1111/j.1467-6419.2007.00527.x.

3 Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg. 2018;53:1112-7. doi: 10.1093/ejcts/ezy167. PubMed PMID: 29684154.

4 Li M. Using the propensity score method to estimate causal effects: A review and practical guide. Organizational Research Methods. 2013;16:188-226. doi: 10.1177/1094428112447816.

5 Duhamel A, Labreuche J, Gronnier C, Mariette C. Statistical Tools for Propensity Score Matching. Ann Surg. 2017;265:E79-E80. doi: 10.1097/SLA.0000000000001312. PubMed PMID: 28486297.

6 Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med. 2014;33:1057-69. doi: 10.1002/sim.6004. PubMed PMID: 24123228; PubMed Central PMCID: PMCPMC4285163.