Proceedings of CONF-CIAP 2026 Symposium: Applied Mathematics and Statistics
DOI: 10.54254/2753-8818/2026.CH30759

Causal Inference with Observational Data via Propensity Score Matching: A Simple, Hands-On Guide

Jingyu Xiao

Department of Statistics, Dalhousie University, Halifax, Canada
jn445431@dal.ca

Abstract. Propensity score matching (PSM) was presented as a practical method for making fair comparisons in studies without random assignment. This paper explained PSM in accessible terms and provided a step-by-step workflow that could be followed by readers. The research was conducted to address the issue of differences between treated and untreated individuals. The key ideas behind PSM, the assumptions required for it to work, and approaches to verifying these assumptions with simple plots and summary statistics were described. Furthermore, the main steps were outlined: defining the target question, selecting baseline covariates, estimating the propensity score, performing matching or reweighting, checking balance, estimating the treatment effect, and conducting basic sensitivity analyses. To maintain a hands-on perspective, a small numerical example and a concise checklist were included to support replication in other studies. Besides, common pitfalls were discussed, such as weak overlap, overly strict calipers, and post-matching analyses that ignored the matched structure. This paper aimed to help students and applied researchers develop PSM analyses that were transparent, reproducible, and easy to communicate to non-specialists.

Keywords: propensity score matching, causal inference, observational studies, balance diagnostics, sensitivity analysis

1. Introduction

1.1. Background

Many real-world questions cannot be answered with randomized experiments. Hospitals, schools, and public programs often adopt a new approach first for certain people and not for others. Simply comparing outcomes between the two groups may be misleading, because the groups were different to begin with. Propensity score matching (PSM) is a design-stage tool that helps us form more comparable groups before estimating the effect [1,2]. Propensity score matching (PSM) works by estimating each individual's probability of receiving a treatment given their observed characteristics and then pairing treated and untreated individuals with similar scores [1,2]. This approach helps mimic the balance of a randomized experiment and reduces selection bias in observational studies [3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.CH30759

This paper explains PSM in plain language. It focus on the specific operational steps, the reasons for its importance, and how to determine if the design is adequate enough. While there are many technical variations, most applied projects share a small number of recurring decisions. This study summarizes these decisions as a simple workflow and gives tips that reduce avoidable mistakes [1-3].

1.2. What PSM tries to solve

Suppose a hospital offers an intensive program to some patients. Patients who join may be younger, more motivated, or have better family support. Even if the program has no real effect, these differences could make outcomes look better for the treated group. PSM tries to balance observable baseline factors so that treated and control groups look similar at the start [3]. When the groups are similar on things we measured, the remaining differences in outcomes are more likely due to the program itself.

PSM does not guarantee fairness on unmeasured factors. However, by matching on a rich set of baseline variables, significant bias can be reduced. In short, PSM helps us compare like with like.

1.3. Key assumptions (what needs to be true)

Every method has conditions. For PSM, three ideas matter most:

First, consistency (sometimes bundled into SUTVA) [4]. The treatment must be clearly defined, and one person's outcome should not be affected by another person's treatment. If the treatment is fuzzy or people affect each other in strong ways, matching will not fix the problem.

Second, conditional exchangeability [4]. Once we control for measured baseline variables, treatment assignment should be as if random. This is a strong idea: if we miss an important confounder, PSM may still be biased. The best defense is careful design—talk to subject-matter experts, use pre-treatment variables only, and record how decisions were made.

Third, positivity (overlap) [4]. For each person, there should be a non-zero chance of being treated and not treated. If entire subgroups always get the treatment (or never do), we cannot make a fair comparison for them. In practice, we check the overlap of the propensity scores and may trim or focus on the region where groups overlap well.

2. Methodology

2.1. A Step-by-step workflow

PSM projects work best when we follow a clear plan. Here is a practical sequence you can reuse:

1) Define the target question. Be specific about the population, the treatment, the time the treatment starts (time-zero), and the outcome window. Write a one-sentence version that a non-expert can understand.

2) Pick baseline covariates. Include variables that predict both treatment and outcome, measured before time-zero. Avoid variables created after treatment starts, and avoid adjusting for strong mediators. If you are unsure, record your reasoning and plan sensitivity checks.

3) Estimate the propensity score (the chance of treatment given the covariates). A simple logistic model is often enough. If needed, add interactions and non-linear terms. Machine-learning models can help when there are many covariates, but they do not replace good design.

4) Match or reweight. Nearest-neighbor matching with or without replacement is easy to explain. A caliper (for example, 0.2 times the standard deviation of the logit of the propensity score [2])

DOI: 10.54254/2753-8818/2026.CH30759

reduces poor matches. Alternative designs like optimal matching, kernel matching, or overlap weighting can perform well in tough overlap settings [5,6].

5) Check balance [3]. Look at standardized mean differences (SMD) and Love plots. A common rule of thumb is $|SMD|<0.1$ for each covariate and a small average $|SMD|$ overall. If the balance is not good, adjust your model, caliper, or method and try again [2,3].

6) Estimate the effect with the matched/weighted sample. Use standard errors that respect the matched sets or the weights. Report the estimand (for example, effect on the treated) so readers know who the result applies to.

7) Run a few sensitivity checks. For example, vary the caliper, use a different matching ratio, or apply Rosenbaum bounds to ask how strong an unmeasured bias would need to be to overturn the finding [2,7].

8) Report transparently. Include a flow diagram, a balance table, and enough detail for someone else to reproduce your steps [3].

2.2. How to choose a matching or weighting method

There is no single best method for all problems. Here is a quick guide:

Nearest-neighbor with a caliper: simple and transparent; good when overlap is okay [2]. Watch for sample loss if the caliper is too small [2].

Optimal matching: finds the best overall pairing; good when you want to minimize total distance; may take longer to compute.

Kernel matching: uses weighted averages of many controls; retains more data; depends on bandwidth choice.

Cardinality matching: enforces balance constraints first, then maximizes the sample size; reliable but can be complex to set up [3,8].

Overlap weighting: not matching, but often solves weak overlap by focusing on the region where groups are most similar; changes the target population to the area of overlap [5,6].

2.3. Balance diagnostics in plain language

After matching or weighting, we want the treated and control groups to look alike on baseline variables. We quantify this using standardized mean differences (SMD). Think of SMD as a unit-free gap; smaller is better. Plots that show SMD before and after matching help readers see improvement.

Do not rely only on p-values for balance: with large samples, tiny imbalances look significant; with small samples, big imbalances may not. SMD and simple variance ratios are more stable summaries of design quality [3].

2.4. Effect estimation and uncertainty

Once the balance is acceptable, estimate the treatment effect using the matched or weighted data [3]. Respect the design: if you matched pairs, cluster, or use robust standard errors that account for pairs. If you created matched sets or weights, use methods that recognize them. Clearly state the effect type (difference in means, risk difference, odds ratio, hazard ratio) and the target population (for example, average treatment effect on the treated).

DOI: 10.54254/2753-8818/2026.CH30759

2.5. Sensitivity analysis (what if we missed something?)

No observational study can rule out all hidden bias. Sensitivity tools ask how strong an unmeasured confounder would need to be to change the conclusion [9,10]. Rosenbaum bounds report a number, often called Γ (Gamma). If the finding remains significant up to Γ=1.4, then an unmeasured factor would need to increase the odds of treatment by 40% to explain it away. Other quick checks include trying different calipers, ratios, or an alternative design such as overlap weighting [6]. If possible, negative controls or falsification tests should be used to probe design choices.

2.6. A mini walk-through with toy numbers

Assume the study included 1,000 patients; 400 joined a program and 600 did not. We collected age, sex, baseline risk score, and prior hospitalizations. A logistic model gived each person a propensity score. We performed nearest-neighbor matching with replacement and a caliper of 0.2 on the logit scale. After matching, 380 treated patients and their best matches were retained.

Before matching, several covariates had |SMD| around 0.20–0.35. After matching, all |SMD| values are under 0.08, and the average |SMD| is 0.04. The outcome is readmission within 30 days. The matched analysis shows a risk difference of −5.2 percentage points (95% CI: −8.1 to −2.3), clustered by matched pairs. A quick Rosenbaum bound suggests Γ≈1.5 to overturn the result. We also try overlap weighting and obtain a similar effect of −4.9 points, which increases our confidence in the direction and size.

3. Common pitfalls and solutions

Mixing pre-treatment and post-treatment variables can distort causal inference because post-treatment information is partly determined by the treatment itself [9,10]. To avoid this, only covariates measured before treatment initiation should be included, excluding any variable that could have been affected by treatment.

Poor overlap of propensity scores occurs when most treated or untreated individuals have probabilities close to 0 or 1, making it impossible to find comparable matches. In such cases, trimming the sample to the region of common support or applying overlap weighting can help, while clearly stating that the target population has been redefined [5,6].

Over-tuning the model to achieve “perfect” balance can make the analysis opaque and unstable. Instead, researchers should aim for a reasonable balance using a transparent, parsimonious model and document any small residual imbalances when they are minor [3].

Ignoring the matched or weighted design when estimating uncertainty can lead to underestimated standard errors and inflated confidence. To correct this, variance estimation methods consistent with the design should be used, such as clustered or robust standard errors for matched data, or sandwich estimators and bootstrap methods for weighting analyses.

Reporting results without showing a balance prevents readers from assessing whether the design succeeded [3]. Therefore, before-and-after balance summaries—both tables and visual plots—should always be included to demonstrate that treated and control groups became comparable [3].

4. A short reporting checklist

Readers should be able to reproduce your steps. Include at least:

1) Clear question, population, treatment, time-zero, and outcome window;
2) List of baseline covariates and why you chose them;

DOI: 10.54254/2753-8818/2026.CH30759

3) How you estimated the propensity score;
4) Matching/weighting details (caliper, ratio, replacement, method);
5) Balance results (tables or plots) and any iterations;
6) Effect estimate with appropriate standard errors and the target population;
7) Sensitivity analysis and any alternative designs tried;
8) Limits of the study and implications for practice.

5. Limitations and external validity

PSM improves internal validity when overlap is decent and key confounders are measured, but it does not fix unmeasured bias or design flaws. Results often apply to the overlap population, which may differ from the original group you started with. Be explicit about who the findings are for and who they may not represent well [11,12] .

6. Conclusion

Propensity score matching (PSM) provides a practical and transparent framework for drawing causal conclusions from observational data when randomized trials cannot be conducted. The method helps researchers reduce selection bias and form more comparable groups by aligning treated and untreated individuals with similar background characteristics. This study summarized the fundamental workflow for PSM, including defining the research question, selecting relevant covariates, estimating propensity scores, applying matching or weighting, checking covariate balance, estimating the treatment effect, and performing sensitivity analyses. Each step was explained in plain language so that readers can reproduce the process in their own projects with confidence and clarity.

The discussion also highlighted common challenges—such as poor overlap, excessive model tuning, and neglecting matched structures—and provided practical strategies to address them. By following these guidelines, applied researchers and students can improve both the credibility and the communicability of their causal analyses. Ultimately, the use of PSM strengthens the bridge between statistical design and real-world decision-making, promoting evidence-based practices in fields ranging from health care and social policy to economics and education. With clear documentation, balanced reporting, and modest sensitivity checks, PSM can make non-randomized research more rigorous, reproducible, and easier to trust.

Acknowledgements

I would like to thank my instructors and classmates for helpful discussions on study design and matching methods.

References

[1] Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Statistics in Medicine, 28(25), 3083–3107. https: //doi.org/10.1002/sim.3697
[2] Austin, P. C. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions. Pharmaceutical Statistics, 10(2), 150–161. https: //doi.org/10.1002/pst.433
[3] Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424. https: //doi.org/10.1080/00273171.2011.568786

DOI: 10.54254/2753-8818/2026.CH30759

[4] Greifer, N. (2025). Using WeightIt to estimate balancing weights (vignette). CRAN. https: //cran.r-project.org/web/packages/WeightIt/vignettes/WeightIt.html
[5] Andrew, B. Y., Brookhart, M. A., Pearse, R., & Karthik, N. (2023). Propensity score methods in observational research: A brief review and guidance for authors. British Journal of Anaesthesia, 131(5), 805–809. https: //doi.org/10.1016/j.bja.2023.06.054
[6] Chen, J. W., et al. (2022). Best practice guidelines for propensity score methods in medical research: Consideration on theory, implementation, and reporting. Arthroscopy, 38(2), 541–555. https: //doi.org/10.1016/j.arthro.2021.08.056
[7] Cheng, C., Li, F., Thomas, L. E., & Li, F. (2021). Addressing extreme propensity scores in survival analysis via overlap weights. arXiv. https: //arxiv.org/abs/2108.04394
[8] Dahabreh, I. J., Hernán, M. A., & Wang, W. (2024). Causal inference about the effects of interventions from observational studies. JAMA. https: //doi.org/10.1001/jama.2024.6512
[9] Greifer, N. (2025). Covariate balance tables and plots: A guide to the cobalt package (vignette). CRAN. https: //cran.r-project.org/web/packages/cobalt/vignettes/cobalt.html
[10] Hernán, M. A., & Robins, J. M. (2024). Causal Inference: What If (updated ed.). Harvard T.H. Chan School of Public Health. https: //content.sph.harvard.edu/wwwhsph/sites/1268/2024/01/hernanrobins_WhatIf_2jan24.pdf
[11] Langworthy, B. W., et al. (2023). An overview of propensity score matching methods for clustered data. Statistical Methods in Medical Research, 32(4), 592–612. https: //doi.org/10.1177/09622802221133556
[12] Li, F., Morgan, K. L., & Zaslavsky, A. M. (2018). Balancing covariates via propensity score weighting. Journal of the American Statistical Association, 113(521), 390–400. https: //doi.org/10.1080/01621459.2016.1260466

Supplementary

A. Frequently Asked Questions (FAQ)

Q: Do I always need machine learning to estimate the propensity score?

A: No. Start simple. A logistic model with sensible interactions often balances just as well. Use ML when you have many variables or complex shapes, but still judge success by balance, not by predictive accuracy.

Q: How many controls per treated unit should I match?

A: One-to-one is easiest to explain. One-to-many (for example, 1:2 or 1:3) can improve precision when you have many controls with similar scores. If the pool is thin, forcing more matches may hurt balance.

Q: Should I include the outcome in the propensity model?

A: No. Use only pre-treatment covariates. Including outcome information breaks the design logic and can hide bias.

Q: What if I lose a lot of treated units because of the caliper?

A: That signals weak overlap. Consider overlap weighting or redefining the target population so that overlap is acceptable. Be clear in the write-up about the new population.

Q: Are p-values useful for checking balance?

A: They are sensitive to sample size and do not tell you how large an imbalance is. Report SMDs and variance ratios instead, with a simple visual.

B. Practical do's and don'ts

Do: Write down time-zero and keep all covariates measured before it. Do: Show a flow diagram so readers see what was excluded and why. Do: Keep a short design log so you remember what you tried and what finally worked.

Don't: Match on variables that happen after treatment starts. Don't: Chase perfect balance with dozens of tweaks—document and move on if residual gaps are small. Don't: Forget to use matched-set or clustered standard errors when estimating uncertainty.

DOI: 10.54254/2753-8818/2026.CH30759

C. Glossary (plain language)

Propensity score: The chance of receiving the treatment based on baseline facts. Two people with similar scores had a similar chance of treatment, even if one actually received it and the other did not.

Caliper: A rule that blocks poor matches by requiring scores to be close. Smaller calipers mean stricter matching, but can drop more data.

Standardized mean difference (SMD): A size-free measure of how far apart groups are on one variable. Values near zero mean the groups look alike on that variable.

Overlap: The region where treated and control people look similar enough to be compared fairly. Outside this region, comparisons are guessy and should be avoided.

Appendix: a reusable reporting skeleton

Use this as a fill-in-the-blanks template:

• Question and estimand. “We estimated the effect of [treatment] on [outcome] over [time window] for [population], targeting [ATE/ATT/ATO].”

• Data and timing. “We defined time-zero as [definition]. Covariates were measured from [window] and included [list].”

• Propensity model. “We estimated the propensity score using [logistic/ML] with [key interactions / non-linear terms].”

• Design. “We used [matching/overlap weighting], with [ratio, replacement, caliper/bandwidth].”

• Balance. “All covariate SMDs were below [threshold]; the average |SMD| was [value].”

• Effect estimation. “We estimated [effect] with [method], using [clustered/robust] standard errors.”

• Sensitivity. “Results were stable to [caliper/ratio/bounds with Γ=…].”

• Limitations. “Main threats include [unmeasured confounding, weak overlap, measurement error].”