Annals of Clinical Epidemiology 2022;4(4):101–109

Annals of Clinical Epidemiology

SEMINAR

Theory and practice of propensity score analysis

Yohei Hashimoto[1,2], Hideo Yasunaga[1]

[1] *Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo*

[2] *Department of Ophthalmology, Graduate School of Medicine, The University of Tokyo*

ABSTRACT

Propensity score analysis has been widely used in observational studies to make a causal inference. This study introduces three assumptions for causal inferences—conditional exchangeability, positivity, and consistency—and five steps for propensity score (PS) analysis—1) construct appropriate PS models, 2) check overlap in PS, 3) apply appropriate weighting (inverse probability of treatment weighting, standardized mortality ratio weighting, matching weights, and overlap weights) or matching methods according to the target of inference, 4) check the balance of covariates, and 5) estimate the effect of exposure appropriately. Finally, the advantages of PS analyses over conventional multivariable regression are discussed.

KEY WORDS

Propensity score, Cohort studies, Multivariable analysis

1. INTRODUCTION

Propensity score (PS) analysis has become a reliable approach in medical research to make a causal inference on the association between exposure and outcome using observational data [1, 2]. PS analysis is based on several important assumptions for causal inference and there exists a diversity of PS methods that vary by the target of inference. This study introduces the steps to conduct PS analysis, mainly focusing on appropriately choosing the most suitable PS method. We will also review the advantages of PS analysis compared with conventional multivariable regression.

2. THREE ASSUMPTIONS FOR CAUSAL INFERENCE

Generally, three assumptions are needed to identify the causal effects of exposure (A) on an outcome (Y) from observational studies: 1) conditional exchangeability, 2) positivity, and 3) consistency [3].

Conditional exchangeability means that, within levels of confounders (L), all other covariates are equally distributed between the exposed (treatment) and unexposed (control) groups [3]. Suppose that A is a binary variable of a new surgery (1, with surgery; 0, without surgery), Y is mortality (1, dead; 0, alive), and L is sex (1, male; 0, female). If a male patient is more likely to receive this surgery and there are no other confounders, in the subset $L = 1$, the exposed and unexposed are exchangeable. If the exposed individuals remained unexposed, they would experience the same average outcome as the unexposed individuals did and vice versa [4]. However, L is a set of *measured* confounders and does not include *unmeasured* confounders. Suppose that an individual without a history of stroke (U) is more likely to undergo this surgery, but the information on history of stroke is unmeasured. When the history of stroke is distributed differently between the exposed and unexposed, conditional exchangeability given L does not hold. Therefore, the assumption of conditional exchangeability has the same meaning as no existence of unmeasured confounders.

The second assumption is positivity. Suppose that exposure A is a new surgery and the outcome Y is mortality, as described above. If the doctors assign all patients to receive the exposure level $A = 1$, it is impossible to calculate the average causal effect because

there is nobody in $A = 0$. The probability of being assigned to each of the exposure levels should be over 0 (positive). This assumption should be satisfied for all the strata made by confounders L that are required for exchangeability. Suppose that L includes age L_1 (1, ≥65 years old; 0, <65 years old) and sex L_2 (1, male; 0, female). In the strata meeting the conditions of aged ≥65 years ($L_1 = 1$) and male ($L_2 = 1$), one or more individuals must exist both in the exposed group ($A = 0$) and the unexposed group ($A = 1$). This should be true of the other strata.

The third assumption is consistency, which means that researchers should pay attention to multiple versions of exposures [3]. Suppose the above example again: the exposure A is a surgery and the outcome Y is mortality. Now suppose that the operative time (short, normal, or long) and the number of nurses engaged in post-operation procedures (1 nurse per ≤7 patients or per >7 patients) had multiple versions. In this context, the prognosis of a patient undergoing surgery with a very long operative time would be different from that with a short operative time. Similarly, more nurses per patient after an operation would have a better effect on the patient's prognosis. Thus, the exposure should be sufficiently well-defined by us [3]. In the above example, we have to define the inclusion criteria for the operative time and the number of nurses. Specifically, we use the term "sufficiently" because we do not have to define irrelevant things. For example, whether the surgeon is right-handed or left-handed would be unrelated to the outcome; thus, there is no need to define the dominant hand of the surgeon. How strictly researchers should define the exposure and to what extent they allow the treatment-variation irrelevance depend on the researchers' expert knowledge. Making clinical questions without vagueness is important for accurate causal inference. Generally, biological (e.g., blood pressure) and social (e.g., socioeconomic status) exposures tend to leave high vagueness, whereas interventional exposure (e.g., surgery or medical treatments) do not [3].

To summarize, researchers should always account for (1) conditional exchangeability (no unmeasured confounders), (2) positivity (the existence of one or more individuals in all strata), and (3) consistency (sufficiently well-defined exposure) to make a firm causal inference.

3. TARGET OF INFERENCE (ESTIMAND)

There are four methods for PS: matching, weighting, adjustment as a covariate, and stratification. In this study, we will focus on the two most widely used methods: matching and weighting.

Before expounding on the details of the two methods, we have to consider the target of inference (estimand). If researchers assume that the exposure of interest could be applied to all the individuals in the study, the target of inference would be the average treatment effect (ATE) [5]. An example may be in a study comparing metformin versus sodium-glucose cotransporter-2 inhibitors for the prevention of acute myocardial infarction in patients with diabetes. These two drugs are both indicated for the treatment of relatively young patients with diabetes unless they have contraindications; thus, all the patients in the study can be candidates for receiving both drugs.

However, if researchers assume that the individuals who were actually exposed in the study are a unique population that has certain characteristics and that the individuals who were unexposed in the study do not have a possibility of being exposed, the target of inference would be the average treatment on the treated (exposed) (ATT) [5]. An example may be a study investigating whether extracorporeal membrane oxygenation (ECMO) can decrease the risk of death for patients with COVID-19. As the conditions of patients undergoing ECMO are naturally more severe than those not undergoing ECMO, not all the individuals in the unexposed group can be candidates for receiving ECMO.

Finally, if researchers are interested in comparing the subsets of exposed and unexposed groups that have similar characteristics and are assumed to be able to be equally assigned to both exposure and non-exposure, the target of inference would be ATE in a subset with clinical equipoise. An example may be a study comparing 30-day mortality between open surgery and laparoscopic surgery groups for colorectal cancer. Various factors such as sex, age, comorbidities, and social economic status may affect the assignment of the open surgery versus laparoscopic surgery, but patients with similar characteristics (i.e., close PS values) would exist in both two groups. Such patients are the target in this case.

4. PROCESS OF PS ANALYSIS

In this section, we examine the process of PS analysis in five steps. **Fig. 1** shows the flowchart of these steps.

Step 1. Calculating PS with Correctly Specified Models

Researchers should avoid misspecification of the PS model; that is, they should pay attention to the variable selection. Direct acyclic graph (DAG), which shows

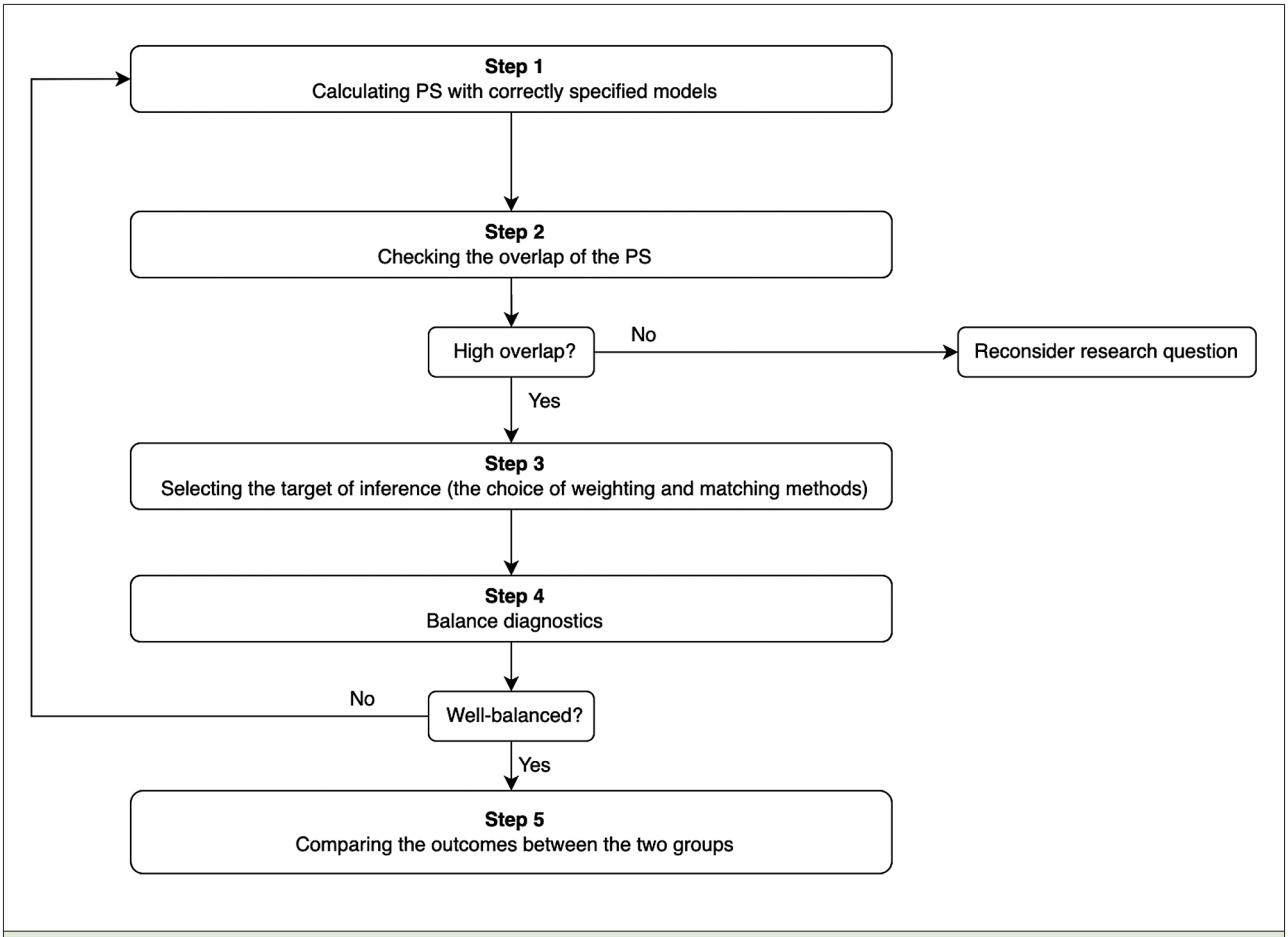

Fig. 1 **Flowchart showing the five steps of propensity score analysis**

causal diagrams, will be useful for this purpose (**Fig. 2**). DAG can clearly show the relationships between the exposure, outcome, and covariates based on subject-matter knowledge rather than statistical associations [3]. The covariates that should be included in the PS model are confounders (L_1 in **Fig. 2**) and those (L_2) affecting the outcome but not affecting the exposure. Conversely, the covariates that should not be included are mediators (M) and those (L_3) affecting the exposure but not affecting the outcome. This is because adding the variable L_2 to the PS model decreases the bias of the estimator [6], whereas adding the variable L_3 increases the variance [6] and adding the mediator M increases the bias (called overadjustment for mediators) [3].

Logistic regression is usually used to estimate the PS. Other machine-learning algorithms such as decision trees, support vector machines, and neural networks can also be used for the estimation of PS [7]. These machine-learning methods have the advantage of automatically taking account of interaction terms, splines, and highly order polynomials, which may get closer to the correctly specified PS model, but extensive simulation studies are needed to use these algorithms in practice [7].

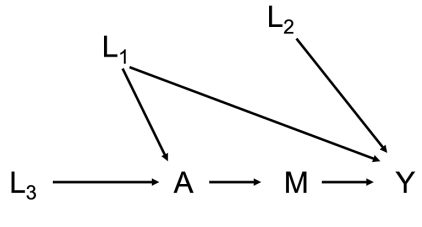

Fig. 2 **Direct acyclic graph**

A is exposure and Y is outcome. L_1 is a confounder because L_1 is a common cause of the exposure A and outcome Y. M is a mediator because M is affected by the exposure A and affects the outcome Y. L_2 is a cause of Y, but not a cause of A. L_3 is a cause of A, but not a cause of Y. Thus, L_2 and L_3 are not confounders. L_1 and L_2 should be included in the propensity score model, but M and L_3 should not be included in the propensity score model.

Step 2. Checking the Overlap of the PS

The next step is checking the propensity score overlap between the exposed and unexposed groups. High overlap in the PS indicates a clinical equipoise between the two groups and we can reasonably compare them (**Fig. 3**). However, low overlap indicates that the two groups are not comparable (**Fig. 4**), as exposed individuals with a certain PS have to be compared with unexposed individuals with a similar PS (relating to the positivity assumption [8]). At a PS of 0.80 in **Fig. 4**, for example, exposed individuals do exist, but unexposed do not. In this case, the weights for the exposed individuals become extremely large, resulting in biased estimates. The trimming and truncation methods are often used to deal with the problem; the details are described below. However, because low overlap in PS indicates the two groups are not comparable, researchers should reconsider the research question.

Step 3. Selecting the Target of Inference (the Choice of Weighting and Matching Methods)

After confirming a high overlap in PS, we can proceed to the selection of the target of inference. There are three options: ATE, ATT, and ATE in a subset with clinical equipoise.

ATE and ATT

If researchers are interested in ATE, they should choose the inverse probability of treatment weighting (IPTW). This method involves weighting each individual by the inverse probability of receiving the exposure that they actually received [9]; that is, the weights for the exposed and unexposed individuals are calculated as 1/PS and 1/(1 − PS), respectively. If researchers are interested in ATT, standardized mortality ratio weighting (SMRW) should be selected, and weights for the exposed and unexposed individuals are calculated as 1 and PS/(1 − PS), respectively. For the IPTW and SMRW, weights can become extremely large and lead to increased variance and bias of the effect estimates, thus requiring trimming or truncation. Trimming removes the individuals with extreme weights. There are several ways of trimming: 1) the common range method (lower cutpoint = lowest PS in the exposed; upper cutpoint = highest PS in the unexposed) [10], 2) the Stürmer method (lower cutpoint, 5th PS percentile in the exposed; upper cutpoint, 95th PS percentile in the unexposed) [11], 3) the Walker method

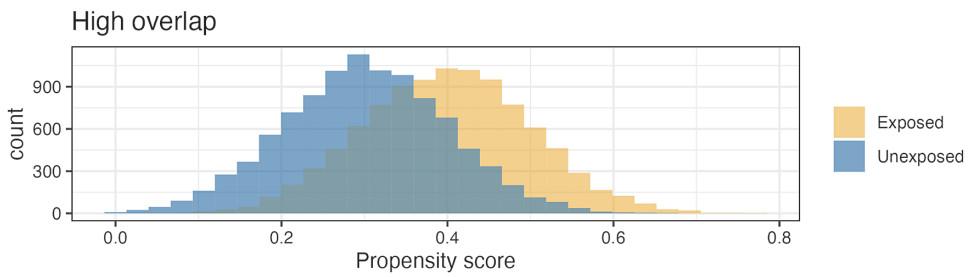

Fig. 3 **High overlap in the propensity score**

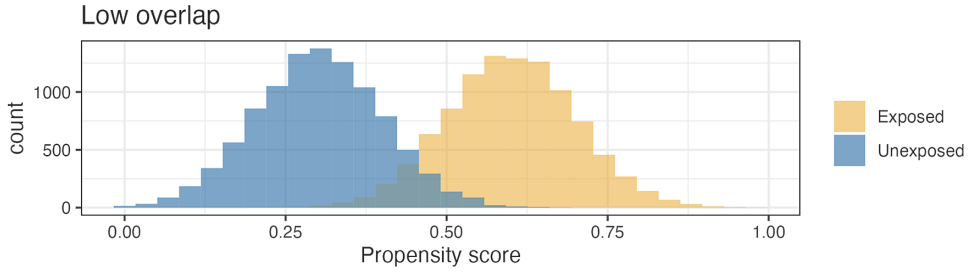

Fig. 4 **Low overlap in the propensity score**

(lower cutpoint, preference score ≤0.3; upper cutpoint, preference score ≥0.7) [12], and 4) the Crump method (lower cutpoint, PS ≤0.1; upper cutpoint, PS ≥0.9) [13]. A simulation study reported that the Stürmer and Walker methods consistently reduced bias when unmeasured confounding was concentrated in the tails of the PS distribution [10]. Truncation is to replace the value of weights larger than percentile p with the value of percentiles p. The 1st and 99th percentiles are typically used because these cutoff values are superior to others in terms of bias-variance trade-off [8].

ATE in a subset with clinical equipoise (overlap weights and matching weights)

If researchers are interested in ATE in a subset with clinical equipoise, they can select weighting with matching weights [14] or overlap weights [15]. Matching weights and overlap weights were firstly published in 2013 and 2018, respectively. Overlap weights have been especially popular in recent medical research (**Fig. 5**).

Matching weights for the exposed and unexposed individuals are calculated as $\{min(PS, 1 - PS)\}/PS$ and $\{min(PS, 1 - PS)\}/(1 - PS)$, respectively [14]. Overlap weights for the exposed and unexposed individuals are calculated as $1 - PS$ and PS, respectively [15]. Basically, these two methods are analogues to PS matching [14, 15].

As stated above, IPTW (ATE) and SMRW (ATT) need cutoff values for trimming or truncation, which can be arbitrarily chosen and which cause the variance of the resulting subpopulation from study to study. Especially when trimming is used, many individuals are discarded and the target of inference is modified (this problem is true of PS matching). These problems might be mitigated by overlap weights and matching weights. The two weights do not need researchers to select the cutoff values for trimming or truncation. Furthermore, these weights are bound between 0 and 1 by design (down-weighing) and become smaller for extreme PS values; thus they can avoid extraordinarily amplifying the outlier individuals who are nearly always exposed (PS approximately 1) or who are most unlikely to receive exposure (PS approximately 0) [16, 17]. These outliers contribute little to the weighted samples, whereas the individuals who have similar characteristics greatly contribute to them. Thus, the resulting subpopulation can emphasize the individuals at clinical equipoise without excluding the outliers [15, 17]. The matching weights and overlap weights are useful especially when the baseline covariates are greatly different in the exposed and unexposed groups (low PS overlap) [16, 17]. Another advantage of matching weights and overlap weights is that they can be easily extended to the comparison of three or more groups (generalized matching weights and generalized overlap weights) unlike PS matching [17, 18].

ATE in a subset with clinical equipoise (PS matching)

Another method to investigate ATE in a subset with clinical equipoise is PS matching. The most commonly implemented way would be one-to-one matching, in which one exposed individual with a certain PS is

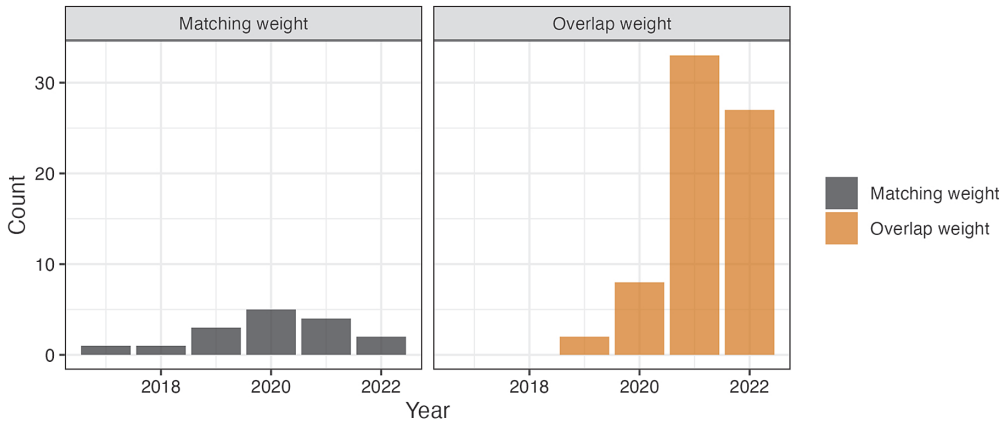

Fig. 5 **The number of papers using matching weight and overlap weight found in Pubmed**

Papers using matching weight were found with the following query: ("matching weight" [Title/Abstract] OR "matching weighting" [Title/Abstract] OR "matching-weighted" [Title/Abstract] OR "matching weighted" [Title/Abstract]) AND propensity.
Papers using overlap weight were found with the following query: ("overlap weight" [Title/Abstract] OR "overlap weighting" [Title/Abstract] OR "overlap-weighted" [Title/Abstract] OR "overlap weighted" [Title/Abstract]) AND propensity. These searches were performed on August 20, 2022.

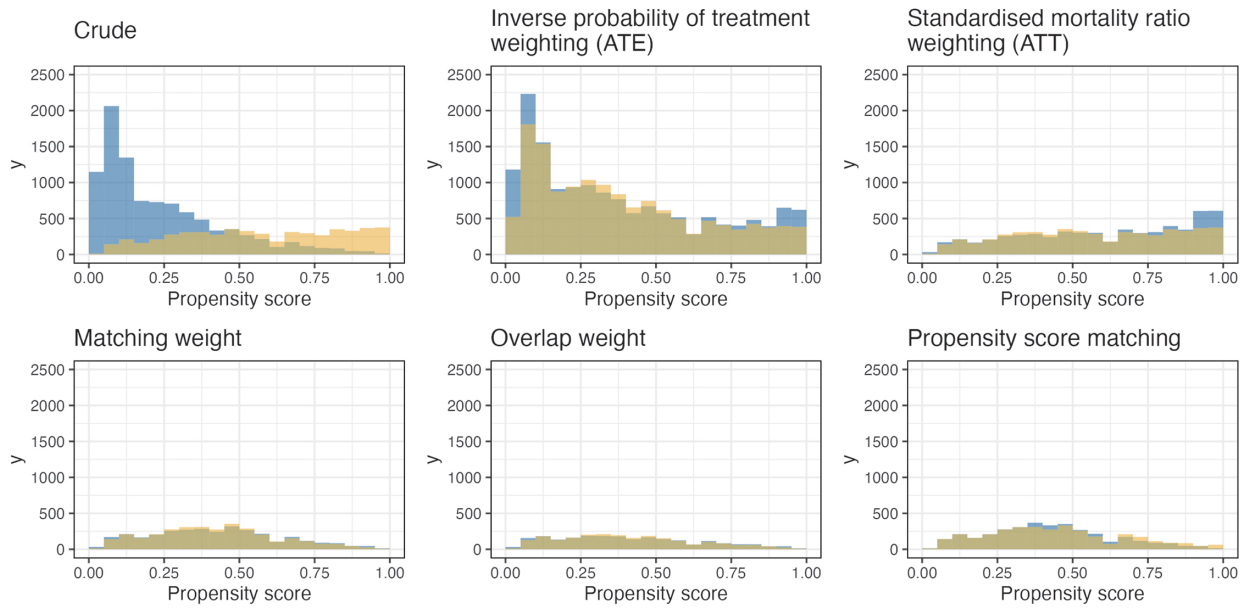

Fig. 6 **Overlap in propensity score before and after each weighting and matching method.**

Orange bars represent exposed individuals and blue bars represent unexposed individuals. ATE = average treatment effect, ATT = average treatment effect in the treated.

matched to one unexposed individual who has a similar PS. We consider one-to-one matching here. Researchers have to make several decisions to perform PS matching [19]. First, they must choose between matching with replacement and without replacement [20]. For matching without replacement, an unexposed individual is matched to an exposed individual only once; the unexposed individual is no longer a candidate for subsequent matching. By contrast, for matching with replacement, an unexposed individual can be matched to multiple exposed individuals. Matching with replacement can minimize the PS distance between the matched exposed and unexposed individuals, which leads to a greater reduction in bias compared with matching without replacement, especially when there are few unexposed individuals similar to the exposed individual [21]. This is an advantage of matching with replacement, but what is troublesome is that researchers should use statistical methods accounting for the nature of repeated occurrence [22].

Second, researchers have to decide between greedy nearest neighbor matching or optimal matching [19]. Greedy nearest neighbor matching selects an exposed individual and then selects an unexposed individual whose PS is closest to that of the exposed individual [23]. Optical matching is a method to minimize the average within-pair difference in PS [23]. Greedy nearest neighbor matching is frequently used in medical literature and shows better performance than optical matching [23].

Third, a caliper distance should be specified for greedy nearest neighbor matching; that is, how distant PS values are allowed for researchers to create pairs of exposed and unexposed individuals [19]. Although there is no consensus about this threshold value, a simulation study recommends that researchers use a caliper of width equal to 0.2 of the standard deviation of the logit of the PS [24].

Fig. 6 shows the overlap in propensity score before (crude) and after the five methods (IPTW, SMRW, matching weight, overlap weight, and PS matching) using a sample dataset. The figure demonstrates that individuals with extreme weight are up-weighted greatly for the IPTW and SMRW, whereas they are down-weighted for matching weight, overlap weight, and PS matching. Furthermore, the PS distribution after weighting with matching weight and overlap weight is almost the same as that after PS matching.

Step 4. Balance Diagnostics

After selecting the target of inference and calculating the weights for individuals, researchers must check the balance of baseline covariates. This process corresponds to assessing whether the propensity score model has been

correctly specified [19]. After weighting or matching with PS was performed, the covariates used for the construction of the PS model are expected to be well-balanced between the exposed and unexposed groups; however, there may exist an imbalance of covariates. To check the balance, standardized differences are usually used [19]. The standardized difference is the difference in the average of a variable between two groups divided by the pooled standard deviation (weighted average of standard deviations for the two groups) [19, 25]. Its values are within the range of 0 to 1 and smaller values mean better balance. For continuous variables, the standardized difference d is denoted as [19]

$$d = \frac{\left(\overline{x}_{exposed} - \overline{x}_{unexposed}\right)}{\sqrt{\frac{s^2_{exposed} + s^2_{unexposed}}{2}}},$$

where $\overline{x}_{exposed}$ and $\overline{x}_{unexposed}$ represent the sample average for the covariate in exposed and unexposed individuals, respectively, and $s^2_{exposed}$ and $s^2_{unexposed}$ represent the sample variance of the covariate in exposed and unexposed individuals, respectively. For dichotomous variables, the standardized differences d is denoted as [19]

$$d = \frac{\left(\widehat{p}_{exposed} - \widehat{\overline{p}}_{unexposed}\right)}{\sqrt{\frac{\widehat{p}_{exposed}\left(1 - \widehat{p}_{exposed}\right) + \widehat{p}_{unexposed}\left(1 - \widehat{p}_{unexposed}\right)}{2}}},$$

where $\widehat{p}_{exposed}$ and $\widehat{p}_{unexposed}$ represent the prevalence or average of the dichotomous variable in exposed and unexposed individuals, respectively. The standardized difference is not affected by sample size and the units of covariates, whereas significance testing is dependent on sample size (e.g., p-value is likely to be below 0.05 just because of the large sample size) [25]. Thus, standardized differences are preferable to significance testing. Although there is no consensus on the threshold of the standardized difference that indicates good balance, the value of 0.1 is often used [25]; that is, standardized difference <0.1 means that the covariate is well-balanced. If some covariates show a standardized difference ≥0.1 even after weighting or matching, researchers should return to the step of constructing PS models (step 1). The c-statistic of the PS model is sometimes used for balance diagnostics [26]; however, it only indicates how well the PS model has discriminated the exposed and unexposed individuals [19]. Previous studies show that the c-statistic does not provide any information on the covariate balance [27, 28]. Thus, researchers do not need to report the c-statistics.

Step 5. Comparing the Outcomes between the Two Groups

After confirming the balance of covariates between the weighted or matched groups, researchers are finally able to compare the outcomes between the two groups. When weighting (IPTW, SMRW, matching weight, overlap weight) has been chosen, researchers will use a generalized linear model with identity link to compare continuous variables (e.g., length of stay) and a generalized linear model with logit link to compare dichotomous variables (e.g., death). What is important here is that they must use the robust variance to calculate the confidence interval because variance estimation must account for the weighted nature of the sample [29]. When matching has been chosen, researchers will use t-tests to compare continuous variables and chi-squared tests to compare dichotomous variables. Like the weighting method, they must take account of the matched nature of the sample; thus, paired t-tests and McNemer's tests should be used [19]. A previous study using Monte Carlo simulations demonstrated that statistical methods accounting for the relationship within pairs resulted in estimated standard errors that more closely reflected the sampling variability of the estimated treatment effect [30], compared with the methods that do not consider the relationship within pairs.

5. ADVANTAGES OF PS ANALYSIS VS. CONVENTIONAL REGRESSION

The flow of PS analyses has been shown so far. In this section, we will outline the advantages of PS analyses compared with conventional multivariable regression models. First, PS methods work against p-hacking, as PS models are decided *before* looking at outcome data [31]. In other words, researchers can go back to the construction of PS models if the covariates are not well-balanced after weighting or matching, but this process does not include the step involving the comparison of the outcome. However, when researchers construct and run a conventional regression model on statistics software, the resulting estimate and p-value will be produced instantly. Looking at these results, they may reconstruct the model (e.g., adding interaction terms). This process allows them to change the model many times until the results become convenient for them. Therefore, conventional multivariable regression is more subject to p-hacking.

Second, researchers can consider the important assumption of causal inference "positivity" through the step of checking the PS overlap between the two groups.

However, conventional regression does not include the step of checking the positivity assumption [32], which may lead to inappropriate results unless researchers consider the positivity assumption explicitly.

Third, when the outcome is rare, the performance of PS analyses is better than conventional regression [33]. This is because PS converts high-dimensional covariates into a single variable.

Fourth, PS analyses are robust to model misspecification [34, 35]. Researchers cannot know the true PS model, but they can bring the model closer to the true one by constructing the PS model until a good covariate balance is achieved.

6. CONCLUSION

We provided an overview of PS analysis with five steps: 1) construct appropriate PS models; 2) check the overlap in PS; 3) apply appropriate weighting or matching methods according to the target of inference; 4) check the balance of covariates; and 5) estimate the effect of exposure appropriately. Researchers should follow these steps to perform PS analysis. Furthermore, they must understand the advantages of PS analyses over conventional multivariable regression.

CONFLICTS OF INTEREST

No conflicting relationship exists for any author.

ACKNOWLEDGMENTS

None.

REFERENCES

1. Granger E, Watkins T, Sergeant JC, Lunt M. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol* 2020;20:132.
2. Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2008;27:2037–49.
3. Hernán MA, Robins JM. *Causal Inference: What If* 2020.
4. Hernan MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health* 2004;58:265–71.
5. Desai RJ, Franklin JM. Alternative approaches for confounding adjustment in observational studies using weighting based on the propensity score: a primer for practitioners. *BMJ* Published online 2019:15657.
6. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
7. Westreich D, Lessler J, Funk MJ. Propensity score estimation: machine learning and classification methods as alternatives to logistic regression. *J Clin Epidemiol* 2010;63:826–33.
8. Cole SR, Hernán MA. Adjusted survival curves with inverse probability weights. *Comput Method Program Biomed* 2004;75:45–9.
9. Chesnaye NC, Stel VS, Tripepi G, Dekker FW, Fu EL, Zoccali C, Jager KJ. An introduction to inverse probability of treatment weighting in observational research. *Clin Kidney J* 2022;15:14–20.
10. Stürmer T, Webster-Clark M, Lund JL, Wyss R, Ellis AR, Lunt M, et al. Propensity score weighting and trimming strategies for reducing variance and bias of treatment effect estimates: a simulation study. *Am J Epidemiol* 2021;190:1659–70.
11. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution—a simulation study. *Am J Epidemiol* 2010;172:843–54.
12. Walker A, Patrick A, Lauer M, Hornbrook M, Marin M, Platt R, et al. A tool for assessing the feasibility of comparative effectiveness research. *CER* Published online January 2013:11.
13. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* 2009;96:187–99.
14. Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat* 2013;9:215–34.
15. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc* 2018;113:390–400.
16. Thomas LE, Li F, Pencina MJ. Overlap weighting: a propensity score method that mimics attributes of a randomized clinical trial. *JAMA* 2020;323:2417.
17. Yoshida K, Hernández-Díaz S, Solomon DH, Jackson JW, Gagne JJ, Glynn RJ, Franklin JM. Matching weights to simultaneously compare three treatment groups: comparison to three-way matching. *Epidemiology* 2017;28:387–95.
18. Li F. Propensity score weighting for causal inference with multiple treatments. *arXiv:180805339 [stat]*. Published online June 28, 2019. Accessed July 20, 2021. http://arxiv.org/abs/1808.05339
19. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivar Behav Res* 2011;46:399–424.
20. Rosenbaum PR. *Observational Studies*. Springer New York; 2002. Accessed August 17, 2022. http://link.springer.com/10.1007/978-1-4757-3692-2
21. Dehejia RH, Wahba S. Propensity Score-Matching Methods for Nonexperimental Causal Studies. *Rev Econ Stat* 2002;84:151–61.
22. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med* 2006;25:2230–56.
23. Austin PC. A comparison of 12 algorithms for matching on the propensity score. *Stat Med* 2014;33:1057–69.
24. Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 2011;10:150–61.
25. Austin PC. Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Commun Stat Simul Comput* 2009;38:1228–34.
26. 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;59:437–47.
27. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf* 2005;14:227–38.
28. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making* 2009;29:661–77.
29. Joffe MM, Ten Have TR, Feldman HI,

Kimmel SE. Model Selection, Confounder Control, and Marginal Structural Models: Review and New Applications. *Am Stat* 2004;58:272–9.

30. Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. *Stat Med* 2011;30:1292–301.

31. Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. *Serv Outcomes Res Methodol* 2001;2:169–88.

32. Shiba K, Kawahara T. Using Propensity Scores for Causal Inference: Pitfalls and Tips. *J Epidemiol* 2021;31:457–63.

33. Cepeda MS. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol* 2003;158: 280–7.

34. Rubin DB. The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies. *Biom* 1973;29:185–203.

35. Drake C. Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect. *Biom* 1993;49:1231–6.