Review

Ann Coloproctol 2022;38(6):398-402
https://doi.org/10.3393/ac.2022.01060.0151

Annals of
Coloproctology

pISSN 2287-9714 eISSN 2287-9722
www.coloproctol.org

Applications of propensity score matching: a case series of articles published in *Annals of Coloproctology*

Hwa Jung Kim

Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Korea

Propensity score matching (PSM) is an increasingly applied method of ensuring comparability between groups of interest. However, PSM is often applied unconditionally, without precise considerations. The purpose of this study is to provide a nonmathematical guide for clinicians at the stage of designing a PSM-based study. We provide a seed of thought for considering whether applying PSM would be appropriate and, if so, the scope of the list of variables. Although PSM may be simple, its results could vary substantially according to how the propensity score is constructed. Misleading results can be avoided through a critical review of the process of PSM.

Keywords: *Propensity score; Propensity score matching; Observational study; Selection bias; Indication bias*

INTRODUCTION

Although well-designed and well-constructed randomized controlled trials (RCTs) occupy a higher evidence level than other observational studies, such as cohort or case-control studies, it is also well known that RCTs can be difficult to conduct on certain topics. In particular, for ethically fraught issues, RCTs may not even be possible to consider. Moreover, to investigate safety issues or real-world efficacy data, additional observational studies should be conducted even after an RCT.

The ability to reach an unbiased conclusion from an observational study is premised on comparability between groups. However, there could be differences in underlying factors related to the selection of treatment, and these differences (e.g., in the severity or duration of disease) could confound the association with the outcome. Therefore, propensity score matching (PSM) is widely adopted as a method to compare outcomes between groups that receive different treatments.

For example, assume that a colorectal cancer surgeon would like to investigate the recurrence rate of patients treated with various modalities (e.g., surgery, chemotherapy, radiation therapy, or targeted drug therapy). Designing an arm-to-arm comparative observational study would yield a biased result due to factors (such as the stage) that influence the selection of treatment. Likewise, the results of observational studies comparing different groups are most clearly interpreted if the patients in each group have the same baseline characteristics [1, 2]. This is why RCTs are at the top of the evidence pyramid; randomization secures a balance of covariates between groups, including both observed and unobserved factors that may affect the results [3]. Therefore, RCTs are the best method to prove causality, compared to other observational studies, which are limited to associations. Additionally, the comparability between groups should be ensured (i.e., similarity in various underlying characteristics), because if not, the results could reflect differences in population characteristics rather than any true difference in treatment efficacy. In such cases, PSM could be a good choice to select comparable groups of patients.

PSM could be a good alternative to random allocation in a retrospective observational study setting; this method selects similar patients with a propensity score (i.e., the estimated probability for each individual in the study to be assigned to the group) from each comparison group by calculating the probability of allocation with various potential confounders. Subjects with similar propensity scores can be considered to have a similar distribution of all confounding variables used to construct the propensity

Received: Dec 1, 2022 • Accepted: Dec 2, 2022

Correspondence to: Hwa Jung Kim, M.D., Ph.D.
Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-ro 43-gil, Songpa-gu, Seoul 05505, Korea
Tel: +82-2-3010-5636, Fax: +82-2-3010-7304
E-mail: rsvp@amc.seoul.kr
ORCID: https://orcid.org/0000-0003-1916-7014

score [2, 4, 5]. Therefore, subjects with similar propensity scores are comparable, since the confounding variables are balanced, and an unbiased result could be derived from a comparison of groups consisting of the selected study participants with similar propensity scores.

Still, PSM may leave the wrong impression unless each methodological step is correctly conducted. It sounds obvious that the propensity score should be calculated correctly in order to reach a solid conclusion. As PSM has been widely adopted in the field of coloproctology, we should critically review the methods before believing the results. This article was conducted to review the process of PSM conducted in original articles published in *Annals of Coloproctology*. There were 5 articles using PSM, as described in Table 1 [6-10], and each was reviewed for issues related to comparability between the comparison groups and the appropriateness of applying PSM.

COMPARABILITY BETWEEN COMPARISON GROUPS

The premise of PSM is the "exchangeability" between comparison groups; as mentioned before, RCTs produce the most robust evidence due to randomization, and properly constructed PSM could be a good alternative to random allocation in a retrospective observational study.

In a study comparing 2 different diagnostic tests, the design would take intrapatient comparability for granted (i.e., conducting both tests in every study participant so that a paired test could be performed). However, in studies comparing treatments, each patient only receives 1 treatment, meaning that each group involves different subject groups. Patients are typically not allocated to treatment randomly in clinical practice; instead, treatment is assigned based on the clinician's subjective judgment or the patient's choice. Consequently, the study participants included in each comparison group would reflect selection bias or indication bias, and retrospective analyses would lead to substantially biased results for treatment (e.g., measures of death or recurrence). Therefore, a mere comparison of "face values" without accounting for all potential confounders may lead to a false conclusion. This ultimately translates to incorrect medical practice, once described as a "scandal of poor medical research" [11].

Most researchers believe that PSM enhances the quality of observational studies by increasing the comparability between each group through a reduction in the extent of the unequal distribution of various clinical factors due to selection bias or indication bias. Nonetheless, unreasonable comparisons should be avoided, and an "exchangeable" comparison group should be selected. In other words "comparing apples and oranges" should be avoided.

In the study by Kataoka et al. [6], the participants were divided into 2 according to the C-reactive protein to albumin ratio (CAR) in a preoperative examination and at postoperative days 1 and 7 using cut-off values of CAR determined by analyzing the receiver operating characteristic curves for the survival rate. However, the CAR has been reported to be associated with colorectal cancer prognosis [12, 13]. As the classification rationale is directly associated with the results, indication bias would occur. Indication bias,

Table 1. The list of articles applying PSM published in *Annals of Coloproctology*

Study	PICO	Description (No. of patients before/after PSM)
Kataoka et al. [6]	Patients	Colorectal cancer
	Intervention vs. comparison	Patients with a high C-reactive protein to albumin ratio (CAR, n = 75/n = 72) vs. patients with low CAR (n = 245/n = 72)
	Outcome	Relapse-free survival, overall survival
Hyun et al. [7]	Patients	Clinical T1 rectal cancer
	Intervention vs. comparison	Local excision (n = 106/n = 91) vs. total mesorectal excision (n = 965/n = 91)
	Outcome	Overall survival
Park et al. [8]	Patients	Colorectal cancer with synchronous resectable liver metastasis
	Intervention vs. comparison	Neoadjuvant chemotherapy followed by operation (n = 66/n = 60) vs. surgical resection only (n = 363/n = 60)
	Outcome	Disease-free survival, overall survival
Nasir et al. [9]	Patients	Locally advanced rectal tumors were identified as T3B or T4
	Intervention vs. comparison	Locally advanced rectal cancer (LARC, ND/n = 109) vs. non-LARC (NLARC, ND/n = 109)
	Outcome	Short-term outcomes related to the operation (e.g., conversion to open surgery, clinical anastomotic leak, readmission, reoperation, 30-day mortality)
Yun et al. [10]	Patients	Colorectal cancer
	Intervention vs. comparison	Signet-ring-cell carcinomas (n = 71/n = 71) vs. adenocarcinomas (n = 12,570/n = 142)
	Outcome	Clinical outcome, overall survival, disease-free survival

PSM, propensity score matching; ND, not described.

also referred to as confounding by indication, is a specific type of selection bias that is primarily responsible for the incomparability between groups in retrospective analyses of clinical data [1, 14]. This bias occurs when a patient's condition that determines the selection of any particular treatment is also associated with the outcome of the treatment. Therefore the previously mentioned "exchangeability" is not established between patients with high and low CARs. This violates the "counterfactual assumption" of PSM, making the results unreliable.

In the study by Nasir et al. [9], locally advanced rectal cancer (LARC) and non-LARC (NLARC) patients were compared using PSM. However, the criterion for defining each comparison group was invasion depth, which is not a parameter that we can modify at will. As in the abovementioned study by Kataoka et al. [6], this would involve "comparing apples and oranges." It would rather be better to compare outcomes according to whether patients underwent laparoscopy in each stratum (LARC or NLARC), and the efficacy of laparoscopy in patients with NLARC should be determined according to differences in the quality of surgical specimens, morbidity, and mortality.

Likewise, in the study by Yun et al. [10], histology—as a non-"exchangeable" parameter—was the criterion used for the classification of comparison groups. As signet-ring-cell carcinoma is a rare subtype compared to adenocarcinoma, it would instead be preferable to select a comparable set of adenocarcinoma by directly matching a set of variables [15], including various underlying characteristics that could affect the prognosis of patients (e.g., age, sex, preoperative carcinoembryonic antigen [CEA] levels, location, operation method, stage, lymphatic invasion, vascular invasion, perineural invasion, and adjuvant treatment). A propensity score is the conditional probability of receiving a specific treatment exposure given a set of covariates, and the key property of a propensity score is exchangeability between comparison groups [16]. If the exchangeability assumption is unclear or inapplicable, it would be helpful to design an RCT with each comparison group.

THE APPROPRIATENESS OF APPLYING PROPENSITY SCORE MATCHING

The process of PSM could be divided into 5 steps, including (1) estimation of the propensity score, (2) evaluation of the propensity score distribution, (3) PSM, (4) assessment of the balance in covariates between groups after PSM, and (5) the main analysis of between-group differences after PSM. Thus, PSM is of no use if the first step is not performed properly [2].

Multivariable logistic regression modeling is commonly used for estimating the propensity score, in which potential confounders for group allocation to be adjusted are included as independent variables ("x" variables), and the group assignment is included as the dependent variable ("y" variable). Unlike randomization, which ensures a random allocation of unmeasured confounders, the propensity score method cannot overcome biases caused by variables that are not included in the model [2, 5, 17].

Thus, it is generally better to include as many potential confounders as possible in the propensity score model as independent variables. Although debate continues regarding variable selection during propensity score model construction, variables that are related to both group assignment and the outcome can satisfy the ignorable group assignment and minimize study bias [2, 5, 17-19].

Propensity scores were estimated by Kataoka et al. [6] based on a logistic regression model addressing the following 11 variables: age as a continuous variable, sex, American Society of Anesthesiologists physical status classification, serum CEA level, serum carbohydrate antigen 19-9 (CA19-9) level, tumor location, tumor size, tumor histotype, lymphatic invasion, venous invasion, and TNM stage as categorical variables. While the variable selection for propensity score estimation seems to be appropriate, the propensity score could be different if the serum CEA level or CA19-9 level is included as a continuous variable in the model. Likewise, age could be included as a categorical variable in a clinically meaningful manner (such as age of <65 or ≥ 65 years, as used by Yun et al. [10]).

In contrast, Hyun et al. [7] constructed a propensity score with only 5 variables: sex, age, tumor location, tumor size, and T classification. Even Nasir et al. [9] used only tumor height (low vs. middle vs. high) to estimate the propensity score. While PSM is expected to select an experimental study-like dataset mimicking randomization by removing sources of incomparability between groups, a proper selection of independent variables during the propensity score estimation is absolutely crucial for the validity of the propensity score method [2, 20].

Therefore, it is necessary to think about which variables should be selected among various candidates with clinical importance. Moreover, how the selected variables are included could affect the estimation of the propensity score. If categories are used for continuous covariates, clinically meaningful thresholds (e.g., normal vs. abnormal serum CEA levels) are recommended over data-driven classifications (e.g., quartiles). Moreover, when continuous variables are believed to have non-linear associations, adding quadratic or even cubic terms (e.g., $age^2 + age^3$) or a transformation (e.g., logarithm) could allow more flexible fitting of the data. Interaction terms could also be considered [21]. For this reason, the propensity score could function like a black box, and the result could be different according to how the propensity score is estimated. Therefore, a sensitivity analysis should be conducted to reveal whether the result is robust regardless of the propensity score model. While the c-index (the area under the receiver operating characteristic curve of the logistic regression model) is often used to assess the adequacy of a propensity score model, empirical evidence may be used to augment preexisting knowledge [2, 18, 21, 22].

Additionally, any possible changes in the extent of the imbalance of covariates used to construct the propensity score after PSM should be checked thoroughly, and a P-value is not sufficient because it depends on the sample size. Intergroup differences are usually measured with the standardized mean difference (SMD) [23] and an improved balance after PSM compared to that before PSM could be described in terms of the change in the SMD for each variable used to estimate the PS. However, all 5 articles analyzed herein presented P-values to represent the balance in covariates between groups after PSM. Still, as mentioned above, only the observed and included variables for propensity score estimation could be balanced. Therefore, residual confounding is possible, so a multivariable model to compare the selected comparison groups after PSM could be considered.

CONCLUSION

While PSM is a good alternative to randomization for retrospective observational studies, the covariates for propensity score estimation should be selected carefully among those with clinical importance. However, this process is often ignored and applied unconditionally. Therefore, it is necessary to review issues in the PSM process, including whether the comparison group shows comparability or exchangeability, as well as the appropriateness of applying PSM. A step-by-step checklist for each process of PSM could be used for objective and transparent reporting [21, 24, 25], and sensitivity analyses with various propensity score models should be conducted actively to reveal whether the results are robust. Well-conducted PSM using a well-estimated propensity score can be a superb surrogate for RCTs using real-world data.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

FUNDING

None.

ORCID

Hwa Jung Kim, https://orcid.org/0000-0003-1916-7014

REFERENCES

1. Psaty BM, Siscovick DS. Minimizing bias due to confounding by indication in comparative effectiveness research: the importance of restriction. JAMA 2010;304:897-8.
2. Baek S, Park SH, Won E, Park YR, Kim HJ. Propensity score matching: a conceptual review for radiology researchers. Korean J Radiol 2015;16:286-96.
3. Rosenberger WF, Lachin JM. Randomization and the clinical trial. In: Rosenberger WF, Lachin JM, eds. Randomization in clinical trials: theory and practice. New York: Wiley Interscience; 2002. p. 1-14.
4. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41-55.
5. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249-64.
6. Kataoka M, Gomi K, Ichioka K, Iguchi T, Shirota T, et al. Clinical impact of C-reactive protein to albumin ratio of the 7th postoperative day on prognosis after laparoscopic colorectal cancer surgery. Ann Coloproctol 2022 Jun 13 [Epub]. hppts://doi.org/10.3393/ac.2022.00234.0033.
7. Hyun JH, Alhanafy MK, Park HC, Park SM, Park SC, Sohn DK, et al. Initial local excision for clinical T1 rectal cancer showed comparable overall survival despite high local recurrence rate: a propensity-matched analysis. Ann Coloproctol 2022;38:166-75.
8. Park SH, Shin JK, Lee WY, Yun SH, Cho YB, Huh JW, et al. Clinical outcomes of neoadjuvant chemotherapy in colorectal cancer patients with synchronous resectable liver metastasis: a propensity score matching analysis. Ann Coloproctol 2021;37:244-52.
9. Nasir IU, Shah MF, Panteleimonitis S, Figueiredo N, Parvaiz A. Spotlight on laparoscopy in the surgical resection of locally advanced rectal cancer: multicenter propensity score match study. Ann Coloproctol 2022;38:307-13.
10. Yun SO, Cho YB, Lee WY, Kim HC, Yun SH, Park YA, et al. Clinical significance of signet-ring-cell colorectal cancer as a prognostic factor. Ann Coloproctol 2017;33:232-8.
11. Altman DG. The scandal of poor medical research. BMJ 1994;308:283-4.
12. Shibutani M, Maeda K, Nagahara H, Iseki Y, Hirakawa K, Ohira M. The significance of the C-reactive protein to albumin ratio as a marker for predicting survival and monitoring chemotherapeutic effectiveness in patients with unresectable metastatic colorectal cancer. Springerplus 2016;5:1798.
13. Shibutani M, Maeda K, Nagahara H, Iseki Y, Ikeya T, Hirakawa K. Prognostic significance of the preoperative ratio of C-reactive protein to albumin in patients with colorectal cancer. Anticancer Res 2016;36:995-1001.
14. Salas M, Hofman A, Stricker BH. Confounding by indication: an example of variation in the use of epidemiologic terminology. Am J Epidemiol 1999;149:981-3.
15. Guy D, Karp I, Wilk P, Chin J, Rodrigues G. Propensity score matching versus coarsened exact matching in observational comparative effectiveness research. J Comp Eff Res 2021;10:939-51.
16. Shiba K, Kawahara T. Using propensity scores for causal inference: pitfalls and tips. J Epidemiol 2021;31:457-63.
17. Rubin DB. On principles for modeling propensity scores in medical research. Pharmacoepidemiol Drug Saf 2004;13:855-7.
18. 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.

19. Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiol Drug Saf 2011;20:551-9.
20. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. Am J Epidemiol 1993;138:923-36.
21. Yang JY, Webster-Clark M, Lund JL, Sandler RS, Dellon ES, Stürmer T. Propensity score methods to control for confounding in observational cohort studies: a statistical primer and application to endoscopy research. Gastrointest Endosc 2019;90:360-9.
22. Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317-20.
23. Linden A, Samuels SJ. Using balance statistics to determine the optimal number of controls in matching studies. J Eval Clin Pract 2013;19:968-75.
24. Staffa SJ, Zurakowski D. Five steps to successfully implement and evaluate propensity score matching in clinical research studies. Anesth Analg 2018;127:1066-73.
25. Eikenboom AM, Le Cessie S, Waernbaum I, Groenwold RH, de Boer MG. Quality of conduct and reporting of propensity score methods in studies investigating the effectiveness of antimicrobial therapy. Open Forum Infect Dis 2022;9:ofac110.