Korean Journal of Anesthesiology

Beyond balance: propensity scores and causal reasoning

Eunjin Ahn

Department of Anesthesiology and Pain Medicine, Chung-Ang University Gwangmyeong Hospital, Chung-Ang University College of Medicine, Gwangmyeong, Korea

Editorial

Korean J Anesthesiol [Epub ahead of print]
https://doi.org/10.4097/kja.261051
pISSN 2005–6419 • eISSN 2005–7563

Received: August 30, 2026
Accepted: September 9, 2026

Corresponding author:
Eunjin Ahn, M.D., Ph.D.
Department of Anesthesiology and Pain Medicine, Chung-Ang University Gwangmyeong Hospital, Chung-Ang University College of Medicine, 110 Deogan-ro, Gwangmyeong 14353, Korea
Tel: +82-2-2222-1557
Fax: +82-2-2610-9210
Email: aej0226@cauhs.or.kr
ORCID: https://orcid.org/0000-0001-6321-5285

© The Korean Society of Anesthesiologists, 2026
This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Real-world evidence (RWE) has become an indispensable complement to randomized controlled trials (RCTs) in anesthesiology. This is particularly important in our field because clinically relevant exposures, including anesthetic techniques, intraoperative management, and perioperative treatment strategies, are often difficult to study through randomization in routine clinical settings. Large hospital databases, multicenter registries, and clinical data warehouses allow investigators to study rare outcomes, long-term effects, and treatment patterns that RCTs are often poorly suited to capture. Propensity score (PS) methods have become standard tools for emulating the balance achieved through randomization. The Statistical Round article by Ahn and Kwak [1] reminds us of the uncomfortable truth that a well-executed PS analysis is not the same as a valid causal conclusion. The gap between the two is filled by careful causal reasoning and interpretation rather than by more sophisticated algorithms.

Propensity score methods and causal interpretation

PS matching (PSM), inverse probability of treatment weighting (IPTW), and directed acyclic graphs (DAGs) are well-established [2–5]. PSM and IPTW have been widely used in anesthesiology research to address confounders. Therefore, the important issue is not whether these methods are available, but how they are used, and ultimately, how their results are interpreted. The article by Ahn and Kwak [1] is thus more than a niche statistical concern; it is increasingly relevant to clinical decision-making. The following four questions are particularly important: What is the estimand? Were covariates selected based on the underlying causal structure? Are adequate positivity and overlap present? How sensitive are the findings to unmeasured confounders?

The two case studies drawn from recent issues of this journal make these concepts concrete [6,7]. In the first case [6], a conventional PSM analysis of preoperative hyperglycemia and cardiac events is typically interpreted as targeting an average treatment effect on the treated (ATT) (i.e., the treatment effect among treated patients represented in the matched sample), yet the finding can easily be read as applying to "all surgical patients." This distinction is important because matching restricts the analysis to patients for whom comparable counterparts are available, potentially affecting the target population.

In the second case [7], an IPTW analysis of tricuspid regurgitation after liver transplantation targeted a population-level average treatment effect (ATE) (i.e., the ATE across the population represented by the study sample). However, the credibility of this estimate depends on adequate positivity, overlap, and weight stability. When the overlap is poor, extreme weights can produce an apparently population-level estimate that is disproportionately dependent on a small number of observations.

Importantly, the lesson is not that either study used PS methods incorrectly. Rather, the

two examples reflect the same underlying gap between methodological rigor and interpretation. In the PSM analysis, this separation takes the form of overgeneralization; whereas, in the IPTW analysis, it takes the form of a causal interpretation that may exceed what the underlying assumptions support. In both cases, the problem emerges not in the statistics themselves, but in the sentence that follows them: the leap from "balanced covariates" to "causal effect."

Why, then, is covariate balance insufficient for causal interpretation? PS methods address only one part of the causal problem (i.e., measured confounders) using measured baseline characteristics. They cannot be used to aid a study with poorly defined treatment initiation, inappropriate time zero, post-treatment selection, informative censoring, or misclassified exposure and outcomes. A PS model also cannot adjust for confounders that were never measured.

This limitation is why the balance should not be equated with exchangeability. Covariate balance is evidence that measured variables have become more comparable after adjustment; however, it does not prove that all relevant sources of confounders have been eliminated. Balance is a diagnostic of what was measured and does not provide evidence that the causal assumptions have necessarily been satisfied.

Implications for authors

For authors conducting PS-based studies, the first step is to define the target estimand before selecting a PS method. Second, the covariates should be selected using clinical and causal reasoning rather than statistical significance or data availability. Third, the overlap and weight stability should be examined in addition to conventional balance diagnostics. Finally, investigators should explicitly consider the possibility of unmeasured confounders and interpret their findings based on these assumptions.

These considerations are particularly important when reporting PS-based results. A Love plot assesses whether the measured covariates are balanced, whereas a PS distribution provides information on common support. Weight diagnostics can reveal whether a small number of observations disproportionately influenced an IPTW estimate. These are related but not interchangeable questions. Therefore, authors should report diagnostics that correspond to each assumption, rather than treating a single balance statistic as evidence that the entire causal analysis is sound.

Sensitivity analysis is another important layer. The E-value can quantify how strong an unmeasured confounder would need to be to explain away an observed association [8]. However, this should not be a ritualistic number appended to every PS analysis. An E-value is most useful when its assumptions and effect scale are appropriate. In other settings, alternative quantitative or qualitative sensitivity analyses may be more informative. The broader principle is not "always report an E-value," but rather "do not treat measured covariate balance as the end of the confounder discussion."

Implications for reviewers

For reviewers, these considerations provide a practical framework for evaluating whether the statistical analysis supports the causal claims made in the manuscript. Reviewers should first ask whether the target estimand is clearly defined, and whether the chosen PS method is appropriate for that estimand. They should then assess whether the covariate selection is supported by clinical and causal reasoning, whether adequate overlap and weight stability have been demonstrated, and whether the authors have appropriately considered unmeasured confounding factors. Specifically, reviewers should distinguish between evidence of measured covariate balance and evidence supporting the broader assumptions required for causal interpretation.

Reviewers may also consider the target trial perspective when evaluating the overall coherence of an observational analysis. Hernán and Robins [9] emphasized that observational studies can be strengthened by explicitly specifying the hypothetical trial they are intended to emulate. This perspective encourages investigators to define not only with whom they wish to compare but also when a treatment begins, when follow-up starts, which outcomes are relevant, and what causal contrast is actually being estimated. A PS can help address confounders within such a design, but it cannot be used to define the design for the investigator.

Conclusions

This is not an argument against PS methods. In contrast, PS methods remain valuable tools for reducing confounding factors in observational studies. The argument is simply that they should be used honestly and interpreted within the assumptions that make them informative. The broader lesson from Ahn and Kwak's article [1] is that causal inference is not achieved by selecting the correct algorithm alone. This requires an alignment between the clinical question, target estimand, causal structure, observed data, statistical methods, and final interpretation. Their framework provides clinicians and reviewers with a practical method to make the alignment more explicit.

Perhaps the most important message is also the simplest: a PS can balance the measured characteristics, but cannot determine

https://doi.org/10.4097/kja.261051

what effect was estimated, whether the comparison was credible, or how far the conclusion should travel. These remain matters of causal reasoning, and ultimately, of human judgment.

Funding

None.

Conflicts of Interest

EunJin Ahn has been a member of the Statistical Round of the *Korean Journal of Anesthesiology*. There were no other potential conflicts of interest relevant to this article.

References

1. Ahn E, Kwak SG. From association to causation: interpreting propensity score-based analyses in real-world evidence. Korean J Anesthesiol 2026. Advance Access published on Jun 29, 2026. doi: 10.4097/kja.26332.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983; 70: 41-55.
3. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011; 46: 399-424.
4. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Stat Med 2015; 34: 3661-79.
5. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology 1999; 10: 37-48.
6. Choi B, Oh AR, Park J, Yang K, Lee DY, Park B, et al. Association between preoperative hyperglycemia and adverse cardiac events after non-cardiac surgery: a multicenter cohort study. Korean J Anesthesiol 2025; 78: 535-46.
7. Kim KS, Ha SY, Yang SM, Kwon HM, Kim SH, Jun IG, et al. Liver transplantation outcomes in patients with primary tricuspid regurgitation with coaptation defects: a retrospective analysis in a high-volume transplant center. Korean J Anesthesiol 2025; 78: 261-71.
8. VanderWeele TJ, Ding P. Sensitivity analysis in observational research: introducing the E-value. Ann Intern Med 2017; 167: 268-74.
9. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol 2016; 183: 758-64.