

Putting Propensity Score Matching to Good Use in Medical Education Research

Dmitry Tumin, PhD

AUTHOR AFFILIATION:

Department of Academic Affairs,
Brody School of Medicine at East
Carolina University, Greenville, NC

CORRESPONDING AUTHOR:

Dmitry Tumin, Department of
Academic Affairs, Brody School of
Medicine at East Carolina University,
Greenville, NC,
tumind18@ecu.edu

HOW TO CITE: Tumin D. Putting
Propensity Score Matching to Good
Use in Medical Education Research.
Fam Med. 2026;X(X):1–2.
doi: [10.22454/FamMed.2026.583931](https://doi.org/10.22454/FamMed.2026.583931)

FIRST PUBLISHED: January 22, 2026

© Society of Teachers of Family
Medicine

TO THE EDITOR:

I read with interest Shimkin and colleagues' evaluation of the Under-served Pathway (UP) at the University of Washington School of Medicine, which used propensity score matching (PSM) to examine the association of UP participation with career outcomes.¹

PSM is an elegant approach to handling self-selection into voluntary programs in statistical analysis comparing outcomes between program participants and non-participants. By limiting the sample to “cases” (UP participants) and matched “controls” (non-UP participants who had a similar “propensity” to have participated in the program, based on other characteristics), PSM can reduce confounding bias in a different manner than single-equation techniques, such as multivariable regression of career outcomes on UP participation.² However, after PSM is used to create the matched sample, analyses are still subject to confounding by unobserved variables, special methods are needed to analyze the matched sample correctly, and special limitations apply to the conclusions from this analysis.^{2,3}

Fundamentally, throwing away “unmatched” controls trades off a smaller sample size (384, relative to the initial 2,027 students) for arguably lower bias in the estimate of the UP effect.⁴ The latter point is critical. A propensity-matched sample should be used only for estimating the effect of the primary exposure (UP participation) on one or more outcomes. Therefore, among the odds ratios (ORs) shown in Tables 2-5, only the OR for UP participation can be considered generalizable. The other ORs, for example, OR of “race” in Table 3, are not generalizable to the school's graduates (let alone to other institutions), because they are calculated from a sample purposefully selected to be balanced

on the propensity of participating in UP. Conclusions about other drivers of specialty choice and practice location, apart from UP participation, are likely to be different if those exposures were studied in the entire original sample, or if each exposure were studied separately with its own PSM analysis.

The implementation of PSM in this study would have benefitted from two other technical modifications, as described in more detail in prior literature on this method.^{2,3,5} First, the balance of covariates in the matched sample (Table 1), should have been analyzed using standardized differences (which are not sensitive to sample size), rather than null hypothesis testing methods, such as analysis of variance (where an educationally significant difference could appear to not be “statistically significant” if the sample size were small enough). Second, multivariable analyses of each outcome should have accounted for the matched, or clustered, nature of the sample. A conventional logistic regression, for example, assumes that all observations were sampled independently of one another. In this case, conditional (or “fixed-effects”) logistic regression should have been used to reflect the fact that each control was selected to match a specific case.

PSM is a powerful tool for reducing the confounding bias that vexes many medical education projects. Yet, results from PSM must be interpreted with caution, and its singular purpose can make it unsuitable for broader research questions, such as asking which among many exposures can predict a certain outcome.

CONFLICT DISCLOSURE

D.T. discloses salary support from William T. Grant Foundation, Kate B. Reynolds Charitable Trust, and Lilly Grant Office for research and quality improvement projects unrelated to this work.

REFERENCES

1. Shimkin G, Kardonsky K, Cassell A, et al. Using propensity-score matched cohorts to evaluate career outcomes for medical students completing the underserved pathway. *Fam Med*. 2025;57(9):627–633. doi:10.22454/FamMed.2025.305728
2. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399–424. doi:10.1080/00273171.2011.568786
3. Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. *Eur J Cardiothorac Surg*. 2018;53(6):1112–1117. doi:10.1093/ejcts/ezy167
4. Urschel S. Apples, oranges, and statistical magic: Limitations of registry studies and need for collaborative studies. *J Heart Lung Transplant*. 2015;34(9):1136–1138. doi:10.1016/j.healun.2015.05.013
5. Staffa SJ, Zurakowski D. Five Steps to Successfully Implement and Evaluate Propensity Score Matching in Clinical Research Studies. *Anesth Analg*. 2018;127(4):1066–1073. doi:10.1213/ANE.0000000000002787