OPEN

PRS GLOBAL OPEN

SPECIAL TOPIC
Research

Propensity Scoring in Plastic Surgery Research: An Analysis and Best Practice Guide

Jacqueline J. Chu, BA*
Meghana G. Shamsunder, MPH*
Shen Yin, PhD*
Robyn R. Rubenstein, MD*
Hanna Slutsky, BS*
John P. Fischer, MD, MPH†
Jonas A. Nelson, MD, MPH*

Summary: Randomized controlled trials, though considered the gold standard in clinical research, are often not feasible in plastic surgery research. Instead, researchers rely heavily on observational studies, leading to potential issues with confounding and selection bias. Propensity scoring—a statistical technique that estimates a patient's likelihood of having received the exposure of interest—can improve the comparability of study groups by either guiding the selection of study participants or generating a covariate that can be adjusted for in multivariate analyses. In this study, we conducted a comprehensive review of research articles published in three major plastic surgery journals (*Plastic and Reconstructive Surgery*, *Journal of Plastic, Reconstructive, & Aesthetic Surgery*, and *Annals of Plastic Surgery*) to determine the utilization of propensity scoring methods in plastic surgery research from August 2018 to August 2020. We found that propensity scoring was used in only eight (0.8%) of 971 research articles, none of which fully reported all components of their propensity scoring methodology. We provide a brief overview of propensity score techniques and recommend guidelines for accurate reporting of propensity scoring methods for plastic surgery research. Improved understanding of propensity scoring may encourage plastic surgery researchers to incorporate the method in their own work and improve plastic surgeons' ability to understand and analyze future research studies that utilize propensity score methods. *(Plast Reconstr Surg Glob Open 2022;10:e4003; doi: 10.1097/GOX.0000000000004003; Published online 9 February 2022.)*

INTRODUCTION

The field of plastic surgery has increasingly adopted the tenets of evidenced-based medicine, which has led to better quality research over time. The average level of evidence of plastic surgery research articles has improved,[1] and a recent systematic review found that almost 40% of articles published between 2008 and 2017 were cohort studies or randomized controlled trials (RCTs).[2] RCTs are considered the gold standard of research because well-executed RCTs have the lowest risk of study-design bias.[3] Two-arm RCTs are specifically designed to balance measured and unmeasured confounding factors between study groups, leading to groups that should be exchangeable on everything other than the exposure of interest. However, in plastic surgery research, RCTs may not be appropriate, feasible, or ethical; so, many researchers depend on observational studies. Indeed, while 35% of published plastic surgery articles in the aforementioned systematic review were cohort studies, only 3% were RCTs.[2]

Unlike their counterparts in RCTs, patients in observational studies are assigned to their exposure. This may mean that unadjusted or direct comparisons of outcomes between study groups are at risk for bias, since selection into study groups was not an unbiased (ie, not a randomized) process.[4] This typically results in nonexchangeable treated and control groups,[4] and requires the use of analytic methods to handle issues of confounding and selection bias that threaten the validity of causal inferences drawn from the study findings. Propensity score methods aim to estimate the probability that individuals with particular baseline characteristics (confounders) received the treatment of interest, and to use these probabilities (ie, propensity scores) to match, stratify, or weight participants such that both the exposed and the unexposed groups have a similar distribution of scores or similar likelihood of having received treatment.[5] This method seeks to make

*From *Plastic and Reconstructive Surgery Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, N.Y.; and †Division of Plastic Surgery, University of Pennsylvania, Philadelphia, Pa.*

Received for publication August 30, 2021; accepted October 28, 2021.

DOI: 10.1097/GOX.0000000000004003

Disclosure: *The authors have no financial interest to declare in relation to the content of this article. This study was funded in part through the NIH/NCI Cancer Center Support Grant P30CA008748.*

observational groups more exchangeable while balancing confounders related to treatment self-selection. Despite the benefit of propensity-score methodologies, recent systematic reviews have revealed that, even in high-impact journals, propensity score methods very often are used improperly or described inadequately, potentially leading to an accumulation of biased results in the literature.[6,7] For plastic surgery research, specifically, it is unknown how frequently or appropriately propensity score methods are used.

In this study, we aim to determine the current utilization of propensity score methods in plastic surgery research. We then provide a primer on how such methods can be employed to improve causal inference within plastic surgery research, and present best practice recommendations for presentation of propensity score methodology and results.

Takeaways

Question: How is propensity scoring being used in plastic surgery research?

Findings: This comprehensive review found that only eight of 971 clinical research articles published in three major plastic surgery journals used propensity scoring. However, each study missed at least one important component when reporting their propensity scoring methodology.

Meaning: Propensity scoring can be a powerful tool for observational studies but is underutilized in plastic surgery research. So we provide a best practice guide to help plastic surgeons understand and use propensity scoring methods.

METHODS

Study Selection

We performed a comprehensive review of the study design characteristics of every research article published in *Plastic and Reconstructive Surgery*, *Annals of Plastic Surgery*, and *Journal of Plastic, Reconstructive, & Aesthetic Surgery* from August 2018 to August 2020. Research articles were included if they were RCTs, cohort studies, case-control studies, cross-sectional studies, quasi-experimental studies, or case series with at least 10 patients. Systematic reviews, meta-analyses, and nonclinical research in which patients were not the study population and/or outcomes were not clinical (eg, basic science or translational research, resident education research, survey research on plastic surgeon perspectives) were excluded.

Data Collection and Analysis

We recorded data on the included studies' methodology, including their study design, sample size, number of study groups, method of confounder adjustment, and use of propensity score. These characteristics were summarized as proportions or medians with interquartile ranges (IQRs) using GraphPad Prism (version 8.4.2, GraphPad Software, San Diego, Calif.). While we included all clinical observational studies, other than case reports, in our initial review, the denominator for these characteristics were based on the number of cohort and cross-sectional studies, as propensity scoring is only applicable to these study types.

RESULTS

Overall, 971 studies were included in the analysis. Of these studies, 463 (48%) were cohort studies, 286 (29%) were case series, 133 (14%) were cross-sectional studies, 47 (4.8%) were case-control studies, 41 (4.2%) were RCTs, and one (0.1%) was a quasi-experimental study (Fig. 1).

In the 596 cohort and cross-sectional studies, the number of participants ranged from 10 to 499,766, with a median of 106 (IQR: 51–333) (Fig. 2). Of these studies, 344 (58%) were comparative studies, meaning they examined differences in study outcomes between at least two study groups (Table 1). More than half of studies (n = 348, 58%) did not adjust for confounders. Of the methods used to adjust or control for confounders, multivariate regression was used in 162 studies (28%) and propensity score analysis in eight studies (1.3%).

Among the propensity scoring studies, six (75%) used propensity score matching and two (25%) used propensity score weighting (Table 2). The analysis of methodological reporting quality revealed that all of propensity scoring studies failed to describe one or more important components of generating or utilizing propensity scores (Table 3). Only four of the eight articles (50%) justified the covariates used to generate the propensity score, and only two of the six articles that used propensity score matching (33%) adequately described their matching methodology.

DISCUSSION

Plastic surgery researchers often rely on observational studies to evaluate surgical interventions. However, if their results are to be applied to clinical practice, these studies must account for inherent issues of selection bias and confounding. Our review of the plastic surgery literature highlights a need for plastic surgery researchers to consider methodologies to address bias and improve the quality and applicability of clinical research available for evidence-based medicine. We found that propensity score methodologies are rarely utilized (they accounted for just 1% of studies in our sample), despite their ability to control or adjust for confounding. This lack of utilization may be due to plastic surgery researchers' being unfamiliar with propensity score methods and uncertain about how to apply them to their research. Here, we provide an overview of four propensity score methods, with a special focus on propensity score matching, a commonly used technique in clinical research.

Generating a Propensity Score

The propensity score is the probability of having received treatment and is generated from baseline predictors selected a priori. The first step for generating a

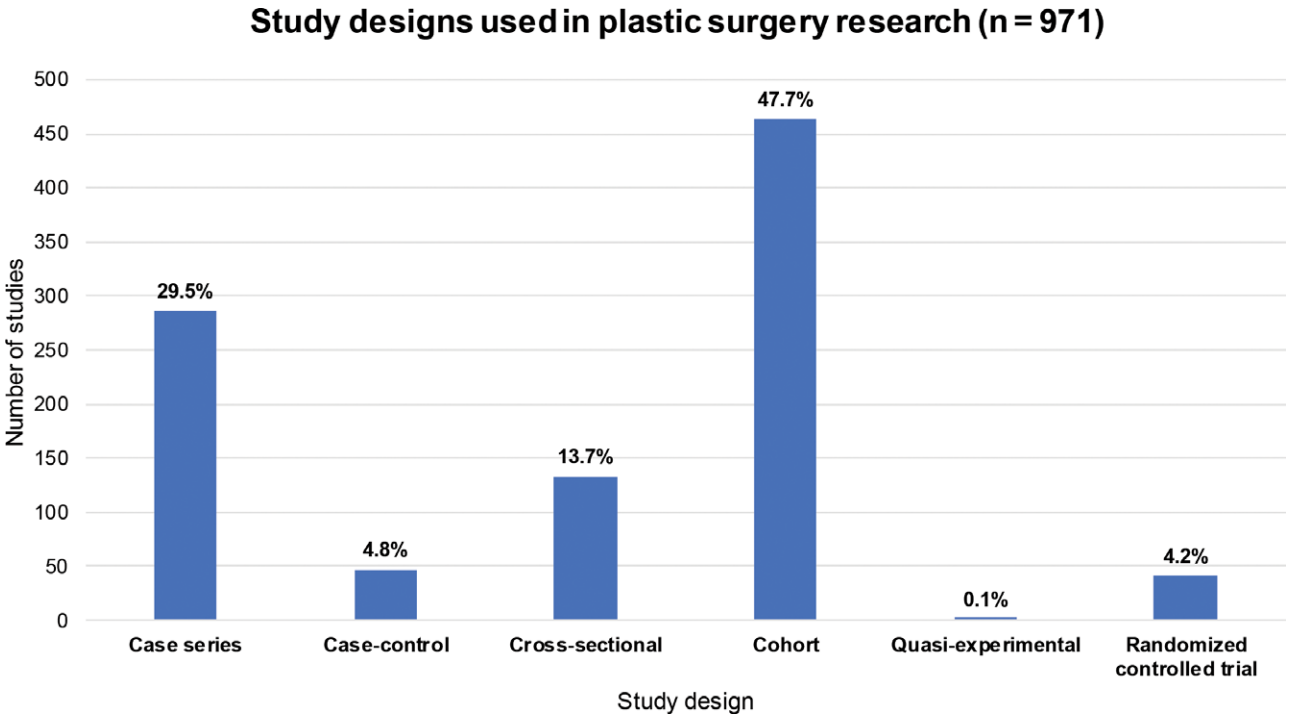

Fig. 1. Study designs used in plastic surgery research (N = 971).

propensity score is identifying predictors that indicate potential sources of selection bias in the observational data. Investigators should have a strong clinical rationale for including any particular predictor in propensity scoring and avoid including potential mediator variables. Binary regression models (eg, logistic regressions) are commonly used in estimating propensity scores. Alternatives, such as nonparametric models, can be employed as well. It is worth noting that model selection and overfitting has little effect on estimated propensity scores; rather, it is the choice of predictors that matters.[16] As an illustration, consider a recent study performed at our institution comparing patient-reported outcomes among patients who had undergone implant-based reconstruction following nipple-sparing or skin-sparing mastectomies.[17] The primary aim was to compare these surgical methods and their impact on patient-reported outcomes by studying patients with an equal likelihood of receiving either nipple-sparing or skin-sparing mastectomy (mimicking the equal likelihood of treatment allocation

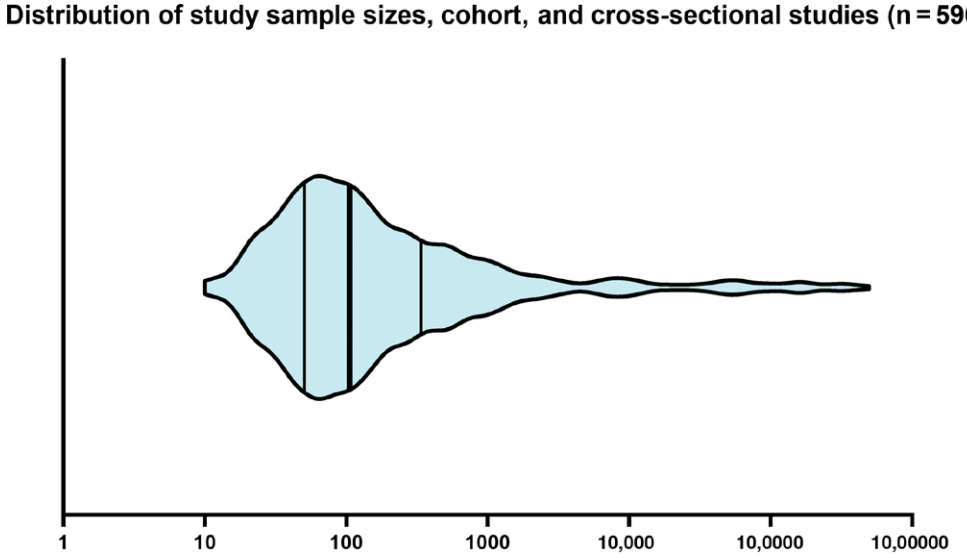

Fig. 2. Distribution of sample sizes in cohort and cross-sectional studies (N = 596). Solid lines represent first quartile, median, and third quartile.

Table 1. Characteristics of Cohort and Cross-sectional Studies (n = 596)

Study Characteristic	
Sample size	
Median sample size (IQR)	106 (51–333)
Range	10–499766
Study groups	
Noncomparative study	252 (42.3%)
Comparative study	344 (57.7%)
2 study groups	*268 (77.9%)*
>2 study groups	*76 (22.1%)*
Method of confounder adjustment	
Propensity scoring	8 (1.3%)
ANCOVA	8 (1.3%)
Multivariate matching	1 (0.2%)
Matching by common variable	14 (2.3%)
Stratification	48 (8.1%)
Multivariate logistic or linear regression	162 (27.2%)
No confounder adjustment	348 (58.4%)
Cannot determine the method used for confounder adjustment	7 (1.2%)

in RCTs). Because we had only observational data from electronic medical records, patient selection needed to account for differences between the two groups in age, body mass index, race, smoking history, use of neoadjuvant chemotherapy, bra size, and psychiatric health, all of which affect mastectomy selection and patient-reported outcomes. Indeed, unadjusted comparisons of patient characteristics revealed that individuals who had chosen nipple-sparing surgery were younger, had a lower mean body mass index, had a smaller mean bra size, and were more likely to be White than those who had chosen skin-sparing mastectomy. Therefore, we used this collection of predictors to generate a propensity score for each patient. As another example, Retrouvey et al, in their retrospective study on postoperative anticoagulation in digit replantation and revascularization outcomes,[14] recognized that certain variables may have influenced both the use of anticoagulation and the success of replantation and revascularization. Thus, they generated propensity scores that reflected patients' probability of having received anticoagulation; predictors used to generate the scores included age, smoking status, digit injury mechanism, number of injured digits, procedure type, and use of a vein graft.

After propensity scores have been generated, a variety of methods—such as matching, stratification, inverse probability weighting, and adjustment (ie, using propensity score as an additional covariate in multivariate regression)—can then be used to balance study groups for statistical comparison. Selection of the appropriate propensity score method can be challenging, and we recommend discussing the best option for any particular study with a data analyst or a statistician who has experience using propensity scoring in clinical research.

Propensity Score Matching

Propensity score matching selects and matches treatment and control participants on the basis of their estimated propensity scores (likelihood/probability of being in a study group). The purpose of this method is to create study groups with similar propensity score distributions and, thereby, balance measured and unmeasured confounding[5] (Fig. 3). Components of this technique include (1) identifying the ratio of control to treated participants, (2) matching with or without replacement, (3) choosing a matching algorithm, and (4) using a caliper to minimize differences between treated and control patients.

Ratio of Control to Treated Participants

Control and treated participants can be matched on either a one-to-one basis (one control to one treated) or a many-to-one basis (multiple controls to one treated, also known as k:1). The selection of a match ratio is based on several factors, including the statistical power and sample size needed for the study, the number of participants available for matching, and the ability to obtain optimal and similar distributions of propensity scores in each study group. Both types of matching are acceptable, though one-to-one matching (used in three of the six propensity score matching studies identified in this review)[8,9,11] can increase ease of statistical analysis and interpretability. However, 2:1 (or 3:1, etc.) matching may be advantageous when there are many more controls than treatment participants, as it allows larger sample sizes and greater power. For example, we recently used 2:1 matching for an analysis of pain severity scores after preoperative paravertebral blocks in patients who had undergone implant-based breast reconstruction. At our institution, most patients receive paravertebral blocks, and a considerably smaller proportion do not receive any form of nerve block. By matching two paravertebral block patients to each no-block patient, we were able to increase our sample size by 50%. Similarly, in their study comparing breast satisfaction and well-being in breast cancer patients to that in the general population, Mundy et al took advantage of the disproportionate sizes of their study groups and matched one normative volunteer to up to five breast cancer volunteers.[13]

It is important to know that when implementing many-to-one matching this schema requires appropriately weighted analyses (eg, weighted means, weighted Student *t* test). In addition, matches beyond the first one (ie, beyond 1:1) may have increasingly dissimilar propensity scores (especially if there is no utilization of a caliper), defeating the purpose of propensity score matching.[18] Alternatives to one-to-one and many-to-one matching include matching with a variable number of untreated subjects[19] and full matching.[20] Full matching uses all available participants to match individuals into "sets" that contain at least one treated subject and at least one control subject. There is no limit to how many similar subjects can be in the same set, thus making it more flexible than many-to-one matching.[21]

With or without Replacement

Matching without replacement means that each control can be matched to at most one treated subject (Fig. 4). Conversely, matching with replacement allows a control subject to form pairs with more than one treated subject. Matching with replacement can be useful if only a small sample of controls is available. For example, we could utilize propensity score matching with replacement when

Table 2. Studies Utilizing Propensity Scoring (n = 8)

Study	Study Population	Independent Variable	Outcome of Interest	Use of Propensity Score	Starting Sample Size	Sample Size Used for Analysis
Calotta et al[8]	Breast reduction mammaplasty patients	Surgical setting	ER visits and readmissions	Matching	Not reported	2474 patients • Outpatient: 1237 • 23-hour observation: 1237 • Inpatient: 1237
Fu et al[9]	Plastic and general surgery patients	Smoking	Postoperative complications	Matching	294,903 patients • Plastic surgery smokers: 3889 • Plastic surgery nonsmokers: 32,565 • General surgery smokers: 49,719 • General surgery nonsmokers: 208,730	103,196 patients • Plastic surgery smokers: 3787 • Plastic surgery nonsmokers: 3787 • General surgery smokers: 47,811 • General surgery nonsmokers: 47,811
Kaltenborn et al[10]	Carpal tunnel release patients	Discontinuation of platelet inhibitors during surgery	Postoperative bleeding complications	Adjustment	635 wrists • Platelet inhibitors: 90 • No platelet inhibitors: 545	635 wrists • Platelet inhibitors: 90 • No platelet inhibitors: 545
Kouwenberg et al[11]	Mastectomy patients	Type of breast reconstruction	Score on EQ-5D-5L questionnaire	Matching	463 patients: • Autologous: 202 • Implant: 103 • No reconstruction: 158	268 patients • Autologous: 67 • Implant: 67 • No reconstruction: 134
Kouwenberg et al[12]	Breast cancer patients	Type of breast cancer surgery	Scores on EQ-5D-5L questionnaire	Adjustment	1871 patients • Breast-conserving surgery: 615 • Mastectomy: 507 • Autologous reconstruction: 330 • Implant-based reconstruction: 419	1294.4 patients • Breast-conserving surgery: 434.0 • Mastectomy: 386.3 • Autologous reconstruction: 178.6 • Implant-based reconstruction: 295.5
Mundy et al[13]	Army of Women participants	History of breast cancer	BREAST-Q scores	Matching	8040 women • Breast cancer: 6840 • No breast cancer: 1200	5265 women • Breast cancer: 4343 • No breast cancer: 922
Retrouvey et al[14]	Digit replantation and revascularization patients	Postoperative anticoagulation	Digit failure	Matching	282 patients • Anticoagulation: 69 • No anticoagulation: 213	199 patients • Anticoagulation: 68 • No anticoagulation: 131
Sheckter et al[15]	Burn patients	Burn-related operation	Scores on Short Form-12/Veterans RAND 12 health survey	Matching	1359 patients • Burn-related operation: 372 • No burn-related operation: 987	Not reported

Table 3. Methodology Reporting for Articles Utilizing Propensity Scoring

Propensity Scoring Component	No. Articles Reporting (%)
All articles (n = 8)	
Type of regression model used to generate the propensity score	6 (75.0%)
Covariates used to generate the propensity score	7 (87.5%)
Justification for the covariates used	4 (50.0%)
Predictive ability of the propensity score	3 (37.5%)
Sensitivity analysis	0 (0.0%)
Articles using propensity score matching (n = 6)	
Unmatched cohort characteristics	5 (83.3%)
Matched sample size	5 (83.3%)
Matching algorithm	5 (83.3%)
Matching with or without replacement	2 (33.3%)
Covariate balance assessment	3 (50.0%)

comparing patient-reported outcomes between autologous and implant-breast reconstruction patients. As demonstrated in a long-term analysis, the majority of patients at our institution who have completed PROMs underwent implant reconstruction.[22] As such, a 1:1 matching schema could result in a smaller sample size with insufficient power from which to draw conclusions. A 2:1 schema would improve the sample size but could lead to matched pairs with dissimilar propensity scores (ie, pairs having differing probabilities of receiving autologous versus implant-based breast reconstruction based on body mass index, history of radiation therapy, etc.). By utilizing matching with replacement of subjects in the implant breast reconstruction group, in addition to 2:1 matching, we could increase both the sample size and the similarity of matched pairs.

When matching with replacement, researchers must adjust for correlations among the matched pairs.[23] In addition, they must account for the repeated use of controls in matched pairs—usually by conducting weighted analyses. Given these methodological complexities, matching without replacement is more commonly used, such as in studies by Sheckter et al and Calotta et al.[8,15,24]

Matching Algorithm and Caliper Use

There are two common matching algorithms that researchers can employ when using propensity score matching: greedy (or "nearest neighbor") matching and optimal matching[20] (Fig. 5). In greedy matching, treated subjects are first selected in a random order, and the control subject whose propensity score is nearest to the selected treated subject's is chosen to form a pair. This process is called greedy because the closest match is made at each step regardless of whether the control subject would have been a better match for a subsequent treated subject. In contrast, optimal matching seeks to minimize the total within-pair distance of the propensity score. Greedy matching performs comparably to optimal matching in balancing the matched samples,[20] and because of its simplicity, it is more commonly employed.[24]

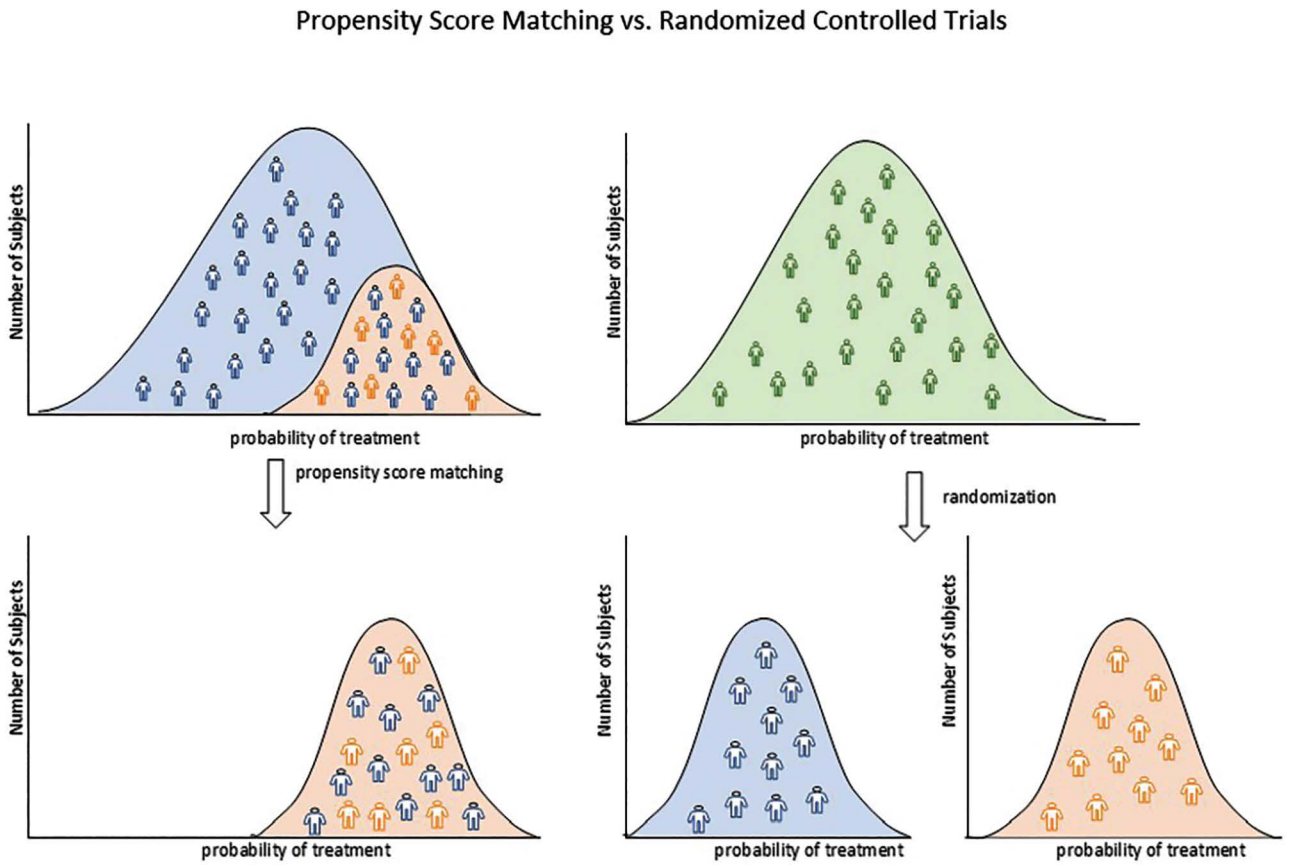

Fig. 3. RCTs use randomization to ensure comparability of study groups, whereas observational studies can use propensity scoring to account for selection bias and confounding.

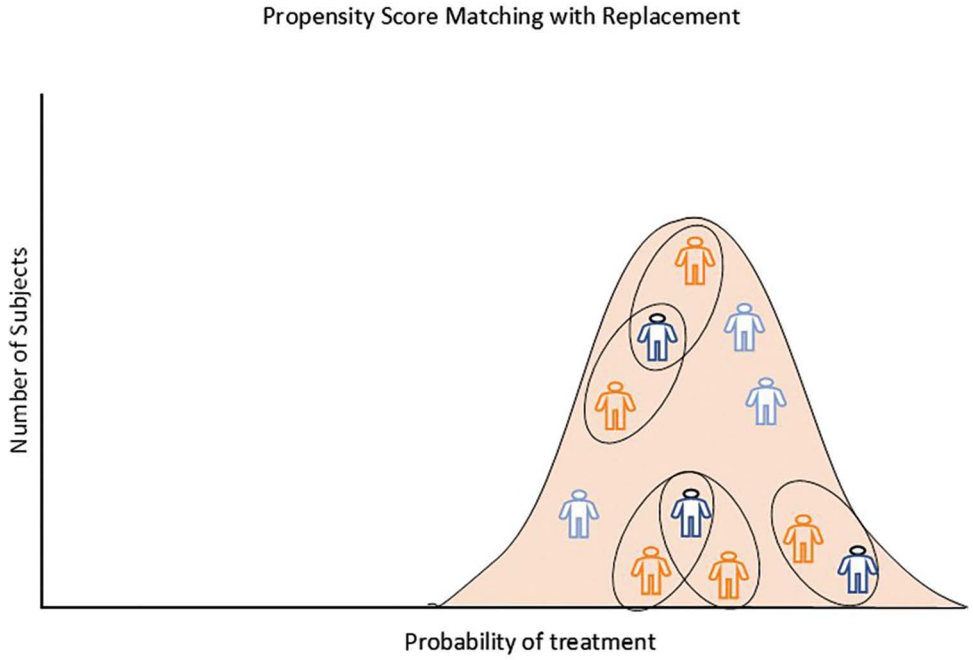

Fig. 4. Matching with replacement allows subjects in one group to be matched multiple times.

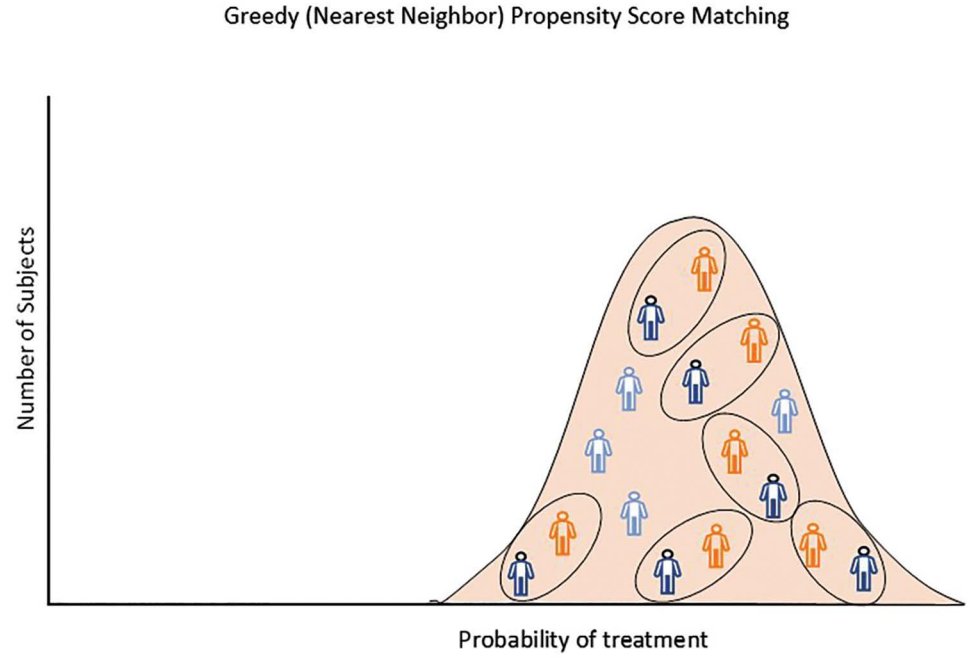

Fig. 5. In greedy (or nearest neighbor) matching, subjects in the control and treatment groups are paired to yield the smallest difference in propensity scores.

Matching can also be accomplished by specifying a "caliper" (ie, a maximum absolute difference in the estimated propensity scores of each pair). Recent studies suggest using a caliper difference (or "distance") of 0.2 times the SD of the logit transformed propensity score,[24,25] although a variety of caliper distances—ranging from 0.01 to 0.2—were used in the studies we reviewed.[9,13–15]

Assessing the Results of Propensity Score Matching

After implementing the above strategies, one must ensure that propensity score distributions are similar between study groups. Visually, propensity scores can be assessed using histograms or jitter plots to demonstrate how well the matching algorithm worked. Figure 6 demonstrates two different methods (jitter plot and histogram) for visualizing the distribution of propensity scores before and after matching.

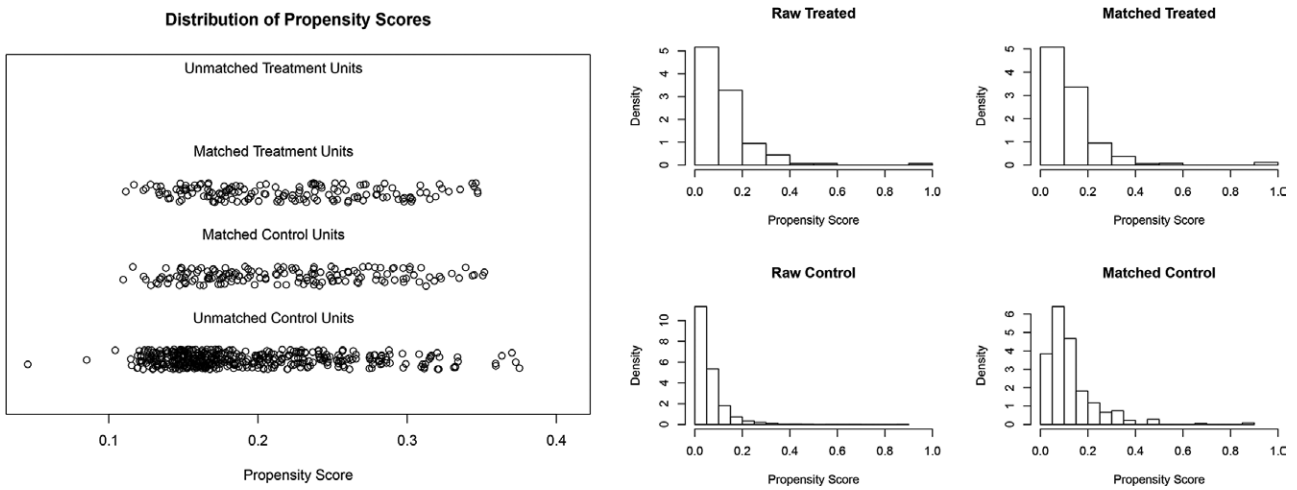

Fig. 6. Methods for assessing the results of propensity score matching. Jitter plot (A) and histogram (B) comparing similarity of cohorts before and after propensity score matching.

Adequate matching results in overlapping propensity score distributions; small visual deviances may be acceptable depending on other diagnostic criteria, such as assessment of standardized differences. Standardized differences are independent of sample size and reflect the matched sample's characteristics. A difference no greater than 0.1 is thought to be an ideal (albeit arbitrary) threshold.[6]

Additional Methods Utilizing Propensity Scores

Beyond propensity-score matching, other techniques, including covariate adjustment, inverse probability weighting, and stratification, can be used to balance study groups on the basis of propensity scores. Covariate adjustment was utilized in two of the eight studies we reviewed.[10,12] In this technique, differences between study groups are analyzed via traditional regression techniques, but propensity scores are included as a covariate for adjustment. The propensity score itself is estimated using a separate model. Covariate adjustment aims to control for confounding with one variable and may help create a more parsimonious model. Because the approach assumes that the model relating the propensity score and the outcome have been correctly specified (this may be difficult to assess), model selection and appropriate model fitting are critical for estimating treatment effect and its standard errors.

Like covariate adjustment, inverse probability weighting utilizes the entire patient population. In this technique, logistic regression is used to estimate the probability of exposure, given a set of predictors. These probabilities are then used for inverse probability weighted statistical analyses. The final technique (stratification) produces a set of quasi-RCTs in which the treatment effect can be estimated by comparing outcomes directly between treated and control subjects within strata.[25] Strata are formed by categorizing patients (eg, into quintiles or deciles) according to their estimated propensity scores. Although increasing the number of strata could potentially eliminate more bias attributed to measured confounders, it may result in noninformative strata (ie, strata that contain subjects from only the control group or only the treatment group). Studies have shown that stratification into five levels based on the estimated propensity scores can remove as much as 90% of the bias.[26,27]

Reporting of Propensity Score Methods

Propensity score methods are becoming more commonplace in clinical research, but reviews of the medical literature have consistently noted the inadequacy and inconsistency of reporting for propensity score methods in clinical research studies.[6,7] These reviews found that authors failed to report important components of their methodology in generating or utilizing the propensity score; indeed, every study failed to report at least one important component of the methodology. Our study demonstrates that the plastic surgery literature suffers from the same shortcoming. We therefore recommend that propensity score analyses follow the reporting guidelines outlined in Table 4. Following these recommended reporting standards will help ensure transparency of research and facilitate reproducibility of results. Other

Table 4. Standardized Reporting Guidelines for Propensity Score Analyses

Components to Report
Study design
• Study question and aims
• A priori hypothesis
• Clear treatment and control groups
Generating propensity scores
• Method of estimating propensity scores
• Predictors selected for propensity score estimation
• Rationale for choice of predictors
Analysis
• How propensity score is used to balance study groups (ie, matching, covariate adjustment, inverse probability weighting, stratification)
• Display and/or discussion of propensity score diagnostics
Propensity score matching
• Matching ratio
• Sample size of control and treatment groups before and after matching
• Matching algorithm (greedy, optimal)
• Caliper size
• Specification of with or without replacement

more comprehensive guidelines for reporting propensity score analyses, such as Yao et al,[28] are also available and can be reviewed before study design.

Limitations of Propensity Scoring

Propensity score methods, although robust, have their limitations. Statistical methods are only as good as the data collected; if the observational data are being collected from a poorly designed study, this method may be insufficient to address systemic bias. Additionally, although groups may be similar with respect to matched variables, other unknown or unassessed variables were not accounted for; so one should not assume that propensity scoring will produce exchangeable groups as a randomized controlled trial would. To address this concern, some have advocated for the use of sensitivity analysis that can assess for unaccounted for selection bias.[29] Ultimately, propensity scoring is a pseudo-randomization method and likely cannot overcome the weaknesses of observational studies in comparison with RCTs.

The choice of which propensity scoring method to use depends on several factors, in particular, the sample size and original research question. For example, matching may lead to more comparable study groups but may reduce sample size and statistical power. Therefore, propensity scoring may not be appropriate for smaller sample sizes because it may weaken the ability of investigators to draw more definitive, appropriately powered conclusions. The reduction in sample sizes is also a concern when performing balance assessments or diagnostics after utilization of propensity scoring; hypothesis testing should not be used to assess balance as it is dependent on sample size, meaning that nonsignificant differences may be attributed to sample size, rather than actual balance between study groups.[6]

Finally, each strategy for utilizing propensity scores has its strengths and weaknesses. Implementing the appropriate strategy, selecting the right variables, and designing a high-quality study should ideally be discussed with a statistician or data analyst before study initiation.

CONCLUSIONS

Propensity score methods are underutilized in plastic surgery research. In this review, we provide a framework for understanding and utilizing propensity score methods and guidelines for reporting of important methodology components. This may empower plastic surgery researchers to consider propensity score methods for their own studies and may improve their ability to understand and analyze future research studies that utilize propensity score methods.

Jonas A. Nelson, MD, MPH
Plastic and Reconstructive Surgery Service
Memorial Sloan Kettering Cancer Center
321 E 61st St, New York, NY
E-mail: nelsonj1@mskcc.org

ACKNOWLEDGEMENTS

We thank Tajah Bell and Craig Davis for their graphical design assistance. We also thank Peter Doskoch and Dagmar Schnau for their editorial assistance.

REFERENCES

1. Loiselle F, Mahabir RC, Harrop AR. Levels of evidence in plastic surgery research over 20 years. *Plast Reconstr Surg.* 2008;121:207e–211e.
2. Rifkin WJ, Yang JH, DeMitchell-Rodriguez E, et al. Levels of evidence in plastic surgery research: a 10-year bibliometric analysis of 18,889 publications from 4 major journals. *Aesthet Surg J.* 2020;40:220–227.
3. Burns PB, Rohrich RJ, Chung KC. The levels of evidence and their role in evidence-based medicine. *Plast Reconstr Surg.* 2011;128:305–310.
4. Fisher RA. *The Design of Experiments.* Edinburgh, UK: Oliver & Boyd; 1951.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41–55.
6. Granger E, Watkins T, Sergeant JC, et al. A review of the use of propensity score diagnostics in papers published in high-ranking medical journals. *BMC Med Res Methodol.* 2020;20:132.
7. Grose E, Wilson S, Barkun J, et al. Use of propensity score methodology in contemporary high-impact surgical literature. *J Am Coll Surg.* 2020;230:101–112.e2.
8. Calotta NA, Merola D, Slezak S, et al. Outpatient reduction mammaplasty offers significantly lower costs with comparable outcomes: a propensity score-matched analysis of 18,780 cases. *Plast Reconstr Surg.* 2020;145:499e–506e.
9. Fu RH, Toyoda Y, Li L, et al. Smoking and postoperative complications in plastic and general surgical procedures: a propensity score-matched analysis of 294,903 patients from the National Surgical Quality Improvement Program Database from 2005 to 2014. *Plast Reconstr Surg.* 2018;142:1633–1643.
10. Kaltenborn A, Frey-Wille S, Hoffmann S, et al. The risk of complications after carpal tunnel release in patients taking acetylsalicylic acid as platelet inhibition: a multicenter propensity score-matched study. *Plast Reconstr Surg.* 2020;145:360e–367e.
11. Kouwenberg CAE, Kranenburg LW, Visser MS, et al. The validity of the EQ-5D-5L in measuring quality of life benefits of breast reconstruction. *J Plast Reconstr Aesthet Surg.* 2019;72:52–61.
12. Kouwenberg CAE, de Ligt KM, Kranenburg LW, et al. Long-term health-related quality of life after four common surgical treatment options for breast cancer and the effect of complications: a retrospective patient-reported survey among 1871 patients. *Plast Reconstr Surg.* 2020;146:1–13.
13. Mundy LR, Rosenberger LH, Rushing CN, et al. The evolution of breast satisfaction and well-being after breast cancer: a propensity-matched comparison to the norm. *Plast Reconstr Surg.* 2020;145:595–604.
14. Retrouvey H, Solaja O, Baltzer HL. Role of postoperative anticoagulation in predicting digit replantation and revascularization failure: A propensity-matched cohort study. *Ann Plast Surg.* 2019;83:542–547.
15. Sheckter CC, Carrougher GJ, McMullen K, et al. Evaluation of patient-reported outcomes in burn survivors undergoing reconstructive surgery in the rehabilitative period. *Plast Reconstr Surg.* 2020;146:171–182.
16. Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable selection for propensity score models. *Am J Epidemiol.* 2006;163:1149–1156.
17. Kokosis G, Stern CS, West S, et al. Nipple-sparing mastectomy and immediate implant-based breast reconstruction: a propensity score matched analysis of clinical outcomes and health-related quality of life. *Plast Reconstr Surg.* Forthcoming 2021.
18. Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci.* 2010;25:1–21.

19. Ming K, Rosenbaum PR. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*. 2000;56:118–124.
20. Gu XS, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Comput Graph Stat*. 1993;2:405–420.
21. Stuart EA, Green KM. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Dev Psychol*. 2008;44:395–406.
22. Nelson JA, Allen RJ Jr, Polanco T, et al. Long-term patient-reported outcomes following postmastectomy breast reconstruction: an 8-year examination of 3268 patients. *Ann Surg*. 2019;270:473–483.
23. Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med*. 2006;25:2230–2256.
24. Allan V, Ramagopalan SV, Mardekian J, et al. Propensity score matching and inverse probability of treatment weighting to address confounding by indication in comparative effectiveness research of oral anticoagulants. *J Comp Eff Res*. 2020;9:603–614.
25. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399–424.
26. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24:295–313.
27. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*. 1984;79:516–524.
28. Yao XI, Wang X, Speicher PJ, et al. Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. *J Natl Cancer Inst*. 2017;109:djw323.
29. Liu W, Kuramoto SJ, Stuart EA. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev Sci*. 2013;14:570–580.