BMJ Open

Evaluation of propensity score used in cardiovascular research: a cross-sectional survey and guidance document

Michelle Samuel [1,2] Brice Batomen [2] Julie Rouette [2] Joanne Kim [2] Robert W Platt [2] James M Brophy [1,2] Jay S Kaufman [2]

To cite: Samuel M, Batomen B, Rouette J, *et al*. Evaluation of propensity score used in cardiovascular research: a cross-sectional survey and guidance document. *BMJ Open* 2020;**10**:e036961. doi:10.1136/bmjopen-2020-036961

► Prepublication history and additional material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2020-036961).

Received 06 March 2020
Revised 11 July 2020
Accepted 13 July 2020

[1]Center for Health Outcomes Research and Evaluation, Research Institute of the McGill University Health Centre, Montreal, Quebec, Canada
[2]Epidemiology, Biostatistics, and Occupational Health, McGill University, Montreal, Quebec, Canada

Correspondence to
Professor Jay S Kaufman; jay.kaufman@mcgill.ca

ABSTRACT

Background Propensity score (PS) methods are frequently used in cardiovascular clinical research. Previous evaluations revealed poor reporting of PS methods, however a comprehensive and current evaluation of PS use and reporting is lacking. The objectives of the present survey were to (1) evaluate the quality of PS methods in cardiovascular publications, (2) summarise PS methods and (3) propose key reporting elements for PS publications.

Methods A PubMed search for cardiovascular PS articles published between 2010 and 2017 in high-impact general medical (top five by impact factor) and cardiovascular (top three by impact factor) journals was performed. Articles were evaluated for the reporting of PS techniques and methods. Data extraction elements were identified from the PS literature and extraction forms were pilot tested.

Results Of the 306 PS articles identified, most were published in *Journal of the American College of Cardiology* (29%; n=88), and *Circulation* (27%, n=81), followed by *European Heart Journal* (15%; n=47). PS matching was performed most often, followed by direct adjustment, inverse probability of treatment weighting and stratification. Most studies (77%; n=193) selected variables to include in the PS model a priori. A total of 38% (n=116) of studies did not report standardised mean differences, but instead relied on hypothesis testing. For matching, 92% (n=193) of articles presented the balance of covariates. Overall, interpretations of the effect estimates corresponded to the PS method conducted or described in 49% (n=150) of the reviewed articles.

Discussion Although PS methods are frequently used in high-impact medical journals, reporting of methodological details has been inconsistent. Improved reporting of PS results is warranted and these proposals should aid both researchers and consumers in the presentation and interpretation of PS methods.

Strengths and limitations of this study

- To our knowledge, this is the most recent and largest comprehensive systematic review of propensity score methods used in cardiovascular research to date.
- Although each article was reviewed by two independent reviewers, some differences in interpretation may remain.
- The current manuscript discusses mainstream propensity score methods, however, many more approaches exist, and details are provided elsewhere.

INTRODUCTION

Since its introduction in 1983, the use of propensity score (PS) methods has steadily increased in observational studies. By attempting to reduce confounding, the goal of using the PS is to provide better estimates of the causal effect of treatments on outcomes.[1] In large randomised controlled trials (RCTs), the distribution of risk factors is balanced between treatment groups through randomisation; thus, confounding is absent in expectation.[2–4] In observational studies, treatment may be assigned based on systematic differences that influence outcomes, thus potentially reducing the required comparability between exposure groups to make causal inferences.[2–5]

The PS is an estimate of the probability of receiving treatment conditional on observed baseline covariates.[2–4] By conditioning on the PS, the distribution of measured, but not unmeasured, covariates becomes balanced between treatment groups.[2–5] PS methods include matching, inverse probability of treatment weighting (IPTW), stratification and direct adjustment.

Previous PS evaluations found insufficient, inappropriate and inaccurate reporting of methods and accompanying statistics.[6–8] An earlier review of the cardiology literature (2004–2006) showed that, of 44 papers using PS matching, 45% did not report how matching was performed, 68% did not assess its success and 75% used inappropriate statistical testing.[6] These findings were confirmed in another review.[9] Prior reviews, however, are now outdated, included a limited number of articles, were not comprehensive and did not assess the causal interpretations of the results.

Due to an ever-increasing number of PS articles, an updated and systematic assessment of these methods in recently published cardiovascular literature is warranted. The objectives of our cross-sectional survey were to: (1) comprehensively evaluate PS methods, reporting and interpretations in cardiovascular literature published between 2010 and 2017 in high-impact journals, (2) summarise PS methods and techniques and (3) propose guidelines outlining key elements to report in PS publications.

METHODS

Identification and selection of PS publications

Cardiovascular articles using PS published between January 1, 2010 and December 31, 2017 in the five highest impact general medical journals (*New England Journal of Medicine (NEJM), Lancet, Annals of Internal Medicine, Journal of the American Medical Association (JAMA), British Medical Journal (BMJ)*) and three highest impact cardiovascular journals (*Journal of the American College of Cardiology (JACC), European Heart Journal (EHJ) and Circulation*) were considered eligible for review. A PubMed search strategy, similar to prior systematic reviews, was used to identify studies with the keyword *propensity* in targeted journals (further described in the online supplementary appendix).[6 7] In addition, we searched for the terms: *inverse probability weighting, inverse probability of treatment weighting, marginal structural models, targeted maximum likelihood estimation* and *doubly robust* as these are PS-based methods. Titles and abstracts were examined by two reviewers (MS, BK) to determine inclusion. Studies included in the cross-sectional survey were (1) published in one of the target journals, (2) used a PS-based method and (3) focused on cardiovascular diseases, outcomes, interventions or techniques. Cardiovascular disease categories were identified from the 10th revision of the International Classification of Diseases codes (listed in online supplementary appendix).[10]

A total of 315 articles were identified from title and abstract review and 306 articles remained in the final sample after full-text review. Excluded articles were meta-analyses (4), commentaries (2) and articles using prognostic scores (1) or non-PS matching (2). The main manuscript and all online supplemental materials were evaluated in the full-text review.

Criteria for data extraction

Data extraction elements were identified from a literature review of methodological articles on PS use, methods and interpretations.[1–6 11–17] Data collection forms were created and reviewed by all authors before a pilot test of 16 articles (two articles per journal) was conducted by two reviewers (MS, BK). Review criteria were further modified based on pilot results and input from all authors. Information was extracted on: (1) bibliographic information, (2) PS assumptions, (3) model selection and assessment of model success, (4) type of study and data source, (5) incidence of the outcome, (6) type of PS methods (matching, IPTW, stratification and/or direct adjustment) and specifics to include with each method, and (7) type of causal interpretation based on the parameter estimated (average treatment effect in the treated (ATT), average treatment effect in the untreated (ATU) and average treatment effect (ATE)[18]; also defined in table 1) and its consistency with the written interpretation of the effect estimates. The final extraction form is in the online supplementary appendix, with interpretations and example quotes from selected reviewed articles (online supplementary table A1).

Each article was reviewed by two reviewers in two teams (MS, BB, JR, JK; 153 articles per team). All variables were binary or categorical and reported as percentages. Descriptive statistical analyses were conducted using SAS V.9.4 (SAS Institute).

Table 1 Key terms

Terms	Description
Average treatment effect	Average treatment effect of moving the entire population from untreated to treated, regardless of the treatment received.
Average treatment effect in the treated	Average effect among treated subjects. Treated sample becomes the reference group to which the treated and untreated subjects are being standardised.
Average treatment effect in the untreated	Average effect in subjects who were untreated. Untreated samples become the reference group to which the untreated and treated subjects are being standardised.
Conditional effects	Treatment effect at the individual-level and consists of moving individual subjects with the same covariate pattern from untreated to treated.
Marginal effects	Average treatment effect at the population level.
Variance ratio	Analytic toll to assess balance by comparing the variances of baseline characteristics between treatment groups.
Positivity assumption	Probability of subjects being assigned treatment, non-treatment or varying levels of treatment is greater than 0%.

Patient and public involvement

No patients involved.

RESULTS

Of the 306 cardiovascular articles using PS published between January 2010 and December 2017, most articles were published in *JACC* (88 articles; 29%) and *Circulation* (81 articles; 27%), followed by *EHJ* (47 articles; 15%), *JAMA* (36 articles; 12%), *BMJ* (31 articles; 10%), *NEJM* (10 articles; 3%), *Annals of Internal Medicine* (10 articles; 3%) and *Lancet* (3 articles; <1%).

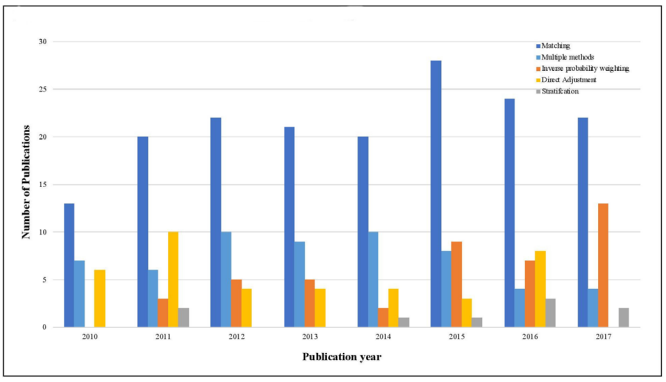

Figure 1 Number of articles by propensity score method over time.

Overall study characteristics and PS model selection (all articles)

In 36% of publications, a rare (<5%) primary outcome was investigated (pre-matching). A majority (81%) of studies were multicentre, however only 31% accounted for possible heterogeneity due to centre differences in the PS or statistical analyses with regression, matching or clustering. PS methods were used as sensitivity analyses in 24% of studies. Heterogeneity of effect was assessed in 59% of articles.

PS matching was performed most often (52%) followed by combination of methods (19%), direct adjustment (13%), IPTW (12%) and stratification (3%). Overall, the number of articles using IPTW increased over time, while the use of direct adjustment appeared to decrease (figure 1). Based on the methods used and described, ATT was the most common (55%) intended effect estimate, followed by ATE (19%) and conditional effects (13%) (figure 2).

In 92% of articles, the variables included in the PS model were potential confounders and temporality between the confounders, treatment and outcome was clearly established. PS model variables were predefined in 77% of articles, selected with statistical testing in 17% of articles or both in 5% of articles (no details for 1% of articles).

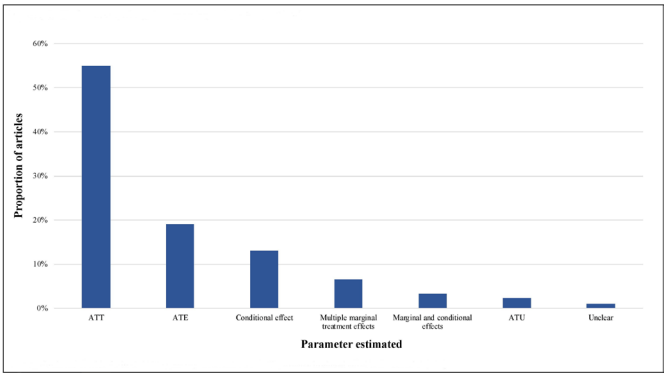

Figure 2 Estimated effects inferred based on propensity score method. ATE, average treatment effect; ATT, average treatment effect in the treated; ATU, average treatment effect in the untreated.

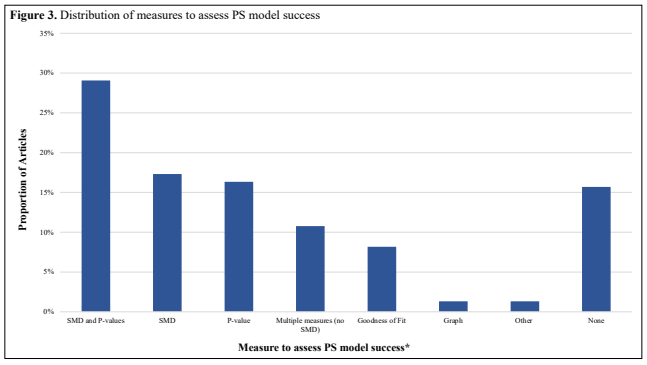

Figure 3 Distribution of measures to assess propensity score (PS) model success. *Articles using multiple measures to assess PS model success were included in multiple categories. SMD, standardised mean difference.

The degree of covariate balance achieved by the PS analysis was formally assessed in 29% of articles with both standardised mean differences (SMDs) and hypothesis testing, however, a measure of balance was not reported for 16% of articles (figure 3). Only 5% of articles reported absolute SMDs of the PS and <1% presented a variance ratio (defined in table 1).

Matching

Matching was performed in 160 (52%) articles and in combination with another PS method in 50 (16%) articles. Most publications (92%) presented the pre-match distribution of baseline characteristics and 89% of articles compared the post-match balance of covariates. After matching, 26% of studies had ≤10% of unmatched treated subjects, while 18% had >50% of unmatched treated subjects (figure 4).

The reported use of specific matching techniques in reviewed articles is presented in table 2. Most studies conducted a 1:1 match and nearest neighbour matching with callipers was the most common method to find matches (57%); however, 20% of studies did not report type of matching (figure 5).

Post-match balance of covariates was often assessed by SMDs and hypothesis testing (table 2) and only 14% of articles compared PS graphically between

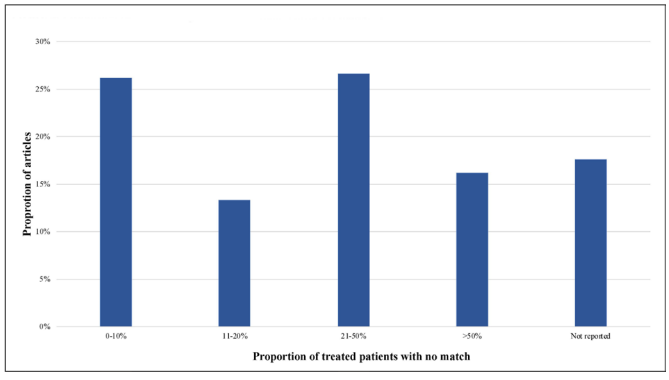

Figure 4 Proportion of treated subjects with no match after propensity score matching.

Table 2 Characteristics of studies using PS matching (N=210)

	Yes (n (%))	Not reported (n (%))
Matches:		8 (3.8)
1:1	178 (84.7)	
1:2	8 (3.8)	
1:3	5 (2.4)	
1:4	5 (2.4)	
Other	4 (1.9)	
Multiple matching ratios	2 (1)	
Calliper scale (nearest neighbour matching with callipers):		2 (1.7)
PS	52 (43.3)	
SD of Logit PS	39 (32.4)	
SD of PS	21 (17.5)	
Logit PS	3 (3.3)	
Mahalanobis distance	2 (1.7)	
Matches found without replacement	79 (37.6)	120 (57.1)
Matching algorithm:		120 (57.1)
Greedy	80 (38.1)	
Optimal	10 (4.8)	
Assessment of balance		
Balance of covariates assessed by:		17 (8.1)
Hypothesis testing	68 (32.4)	
SMDs	57 (27.1)	
Both	60 (28.6)	
Stated balance was assessed	8 (3.8)	
Threshold for successful balance explicitly stated (eg, 10% SMDs, p>0.05)	142 (67.6)	---
SMDs threshold:		39 (33.3)
<10%	71 (60.7)	
<20%	3 (2.6)	
Other	4 (3.4)	

PS, propensity score; SMD, standardised mean difference.

treatment groups. The post-match balance of covariates was successful in 67% articles; however, balance diagnostics were not presented in 15% of articles. Of the 18% which did not achieve balance, 79% did not account for the difference and 13% added the unbalanced covariates in the outcome regression model.

Most articles (87%) specifically described the statistical methods used to compare matched groups, however only 30% accounted for the matched pairs including Cox proportional hazard models stratified on matched pairs, McNemar's test, regression with generalised estimating equation methods, signed rank test and methods with bootstrapping.

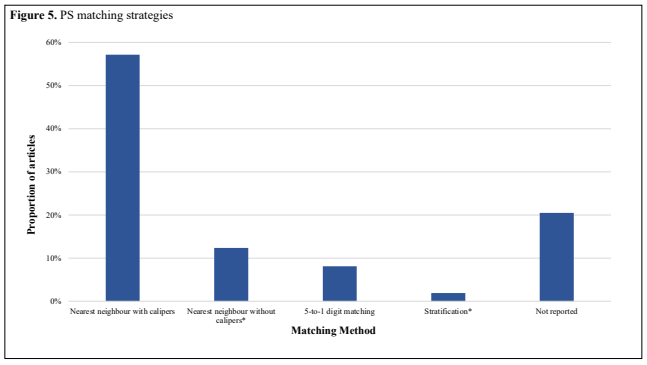

Figure 5 Propensity score matching strategies. *One article used both nearest neighbour matching without callipers and stratification.

Inverse probability of treatment weighting

IPTW was used in 63 (21%) articles, of which 40 used it as the only PS method conducted. A majority (92%) of studies applied weights throughout the study population, 3% applied subgroup-specific weights and 5% did not report the application of weights. Balance was assessed only in 27% of articles. Approximately 19% of studies reported that weights were stabilised and 13% of studies performed trimming. None of the articles truncated extreme weights.

Stratification

Twenty-two studies stratified on the PS. A majority (86%) of these used equal-sized strata and 36.4% reported the balance of covariates within strata. Most studies (86%) created five or more strata of PS. Trimming was performed in only 18% of articles.

Causal interpretations of treatment effect (all articles)

Although 93% of articles clearly stated the population to which the results applied, only half (51%) of all articles interpreted the treatment effect consistently with the primary PS method used and described. Of the 168 articles that estimated an ATT effect, only 20% correctly interpreted the treatment effect as ATT. In contrast, ATE was correctly interpreted in 73% of the 52 studies estimating an ATE. ATU was estimated in only seven articles, of which only 14% correctly interpreted the treatment effect. Excluding studies where PS was a sensitivity analyses led to similar results (not presented).

DISCUSSION

To our knowledge, this is the most recent and largest comprehensive survey of PS methods used in cardiovascular research to date. We found that PS methods were often used in high-impact journals; however, the reporting of details was often inconsistent. Detailed reporting of PS methods is important to: (1) increase transparency, (2) evaluate the appropriateness of the specific PS method

Table 3 Summary of recommendations for reporting propensity score (PS) methods

Elements to be reported	Methodological recommendations
Variable selection strategy for PS model	► Potential confounders ► Select variables a priori ► Optional: strong predictors of the outcome
Balance diagnostics	► Standardised mean differences (threshold <10%) ► Graphical representation of PS distribution* ► Optional: variance ratio
Matching 1. Ratio for matches 2. Matching strategy 3. Number of subjects and balance diagnostics pre-match and post- match 4. Variance estimation	► 1:1 or 1:2 matching is sufficient ► Nearest-neighbour with callipers strongly preferred ► 0.2 SD of the logit of the PS ► Untreated subjects chosen with or without replacement ► Without replacement—untreated matches chosen with greedy or optimal matching ► Account for matched pairs in outcome model with clustering, stratification or regression ► Account for matching with replacement
Inverse probability weighting 1. Application of weights 2. Extraneous values 3. Variance estimation	► Throughout; otherwise, if heterogeneity in treatment effect expected, apply subgroup-specific weights ► Use stabilisation, trimming and truncation, if appropriate ► Non-parametric bootstrap method preferred
Stratification 1. Number of strata 2. Size of strata 3. Combine estimates	► Five strata ► Equal-sized or unequal-sized strata ► Pool stratum-specific estimates using the proportion of subjects in each stratum
Direct adjustment Balance diagnostics	► Conditional standardised difference or quantile regression
Causal interpretations 1. Inclusion criteria 2. Target population a. ATE b. ATT c. ATU	► Methods consistent with target population ► Describe the inclusion criteria and: ► Treatment effect in treated and untreated groups ► Treatment effect in the treated subgroup only ► Treatment effect in the untreated subgroup only

*Kernel density plots, histograms, cumulative distribution functions, quantile–quantile plots, side-by-side box plots, etc.
ATE, Average treatment effect; ATT, average treatment effect in the treated; ATU, average treatment effect in the untreated.

applied, (3) determine the precise population to which the results apply and (4) interpret the effect estimates. To highlight areas for improvements, we make several recommendations of key elements that should be reported in PS articles (see table 3).

Comparison to prior PS surveys on cardiovascular publications

Compared with prior evaluations of PS methods in published research, the present study demonstrated comparable reporting. Only one previous evaluation of randomly sampled coronary artery disease publications (N=48) evaluated the use of all PS methods.[9] It found that matching was the most frequently used PS method (56.3%), with a rate consistent with the present study (52%).[9] Two additional evaluations by Austin were limited to the evaluation of articles using PS matching in cardiology[6] and cardiac surgery[8] literature between 2004 and 2006. These studies found that post-match balance was not assessed 18%–48% of reviewed articles, of which our studied found a slightly reduced rate (11%).[6 8 9] Matching 1:1 was also the most common matching ratio (treated to untreated) and callipers were used in approximately 50%–70% of reviewed articles (present study reported 60%).[6 8 9] These evaluations, however, were limited in the (1) PS characteristics extracted (including interpretations), (2) type of PS method used and (3) cardiovascular topics and years of included publications.

DESCRIPTION OF PS METHODS AND KEY ELEMENTS

Variable selection for PS model

A clearly defined selection strategy for variables to include in the PS model is a critical first step to successfully control for confounding. The inclusion of variables that only influence treatment and are not related to the outcome could decrease the precision of the effect estimate,[16 19] while variables only related to the outcome reduce the variance of the estimate.[1 2 16 20] Although only observed in 8% of articles, inclusion of variables that only influence treatment and are not related to the outcome

could decrease the precision of the effect estimate.[16 19] Therefore, potential confounders are the most appropriate variables for the PS model as they effectively reduce confounding bias.[1 2 16 20]

Whereas an a priori variable selection strategy is preferred, 17% of publications used statistical testing to identify PS model variables. The use of statistical testing is problematic considering the influence of sample size on p values. Also, consideration of only the exposure-covariate association (overlooking strong covariate-outcome associations) could lead to residual confounding.[1 16 20]

Diagnostics

Once the PS model is specified, researchers should evaluate and report on the success of the model to remove systematic differences between treatment groups. SMDs are preferred to compare proportions (or means) of individual characteristics between treatment groups conditional on the PS because such measures are not influenced by measurement scale or sample size.[11 14] Typically, a value of less than 0.1 indicates sufficient balance.

Variance ratios and graphical representations of the PS distribution can be used to further assess balance and verify the positivity assumption[18] by comparing the distribution of covariates or PS between treatment groups.[21] Variance ratios compare variances of baseline characteristics between treatment groups and help determine whether the PS model is correctly specified. In addition, graphical methods such as kernel density plots, histograms, cumulative distribution functions, quantile–quantile plots and side-by-side box plots[11 14] provide information on the overall distribution of the PS and the exact population to which the results apply (the region of PS overlap), extraneous values, proportions of excluded subjects and heterogeneity.[22]

The C-statistic and other goodness-of-fit scores (eg, Hosmer-Lemeshow) should not generally be used to judge the success of the PS prediction model, because variables that improve the prediction of the treatment do not necessarily remove bias from causal estimates, and in fact can reduce precision.[23]

Matching

PS matching was the most commonly used method among reviewed articles. Treated and untreated subjects with the similar PS scores are matched, which makes this PS method relatively simple to understand.[1 2] Further, PS matching is more effective in reducing bias compared with stratification and direct adjustment, and less sensitive to slight misspecification of PS model than IPTW.[1]

In the present review, 85% of PS matching articles matched treated to untreated patients in a 1:1 ratio. Simulations have shown that 1:1 matching sufficiently reduces the mean square error,[12] and thus is appropriate for use. Applying higher fixed matching ratios can induce substantial bias due to the exclusion of treated subjects without sufficient matches while only minimally increasing precision.[24] It is recommended that the proportion of unmatched treated and untreated patients be reported, as a significant number of unmatched treated (or untreated) subjects can induce bias and limit the generalisability of the target parameter.[1 25]

An exact PS match between a treated and untreated subject cannot always be made. Its closest match (nearest neighbour) should then be used, either within a predefined PS distance (calliper) or without. In contrast to other matching strategies, the use of callipers ensures better comparability between treatment groups and reduces confounding bias. When using logistical regression to derive the PS, a calliper width of 0.2 SD of the logit of the PS has been recommended, as it has been shown to eliminate 99% of the bias due to measured confounding.[13]

Depending on the availability of close matches, a treated subject can be matched to one (without replacement) or more (with replacement) untreated subjects, and replacement should be accounted for in variance estimation.[26] If untreated subjects are used without replacement, those must be matched in a 'greedy' or 'optimal' process. In greedy matching, treated subjects are selected in a random order and paired with their closest untreated match, regardless of whether that untreated subject would be more suitably matched to another treated subject.[1 2 27] Optimal matching forms pairs that minimise the global within-pair difference in PS (eg, Mahalanobis distance) to ensure efficient matching overall.[1 2] Optimal matching marginally improves the balance in matched samples compared with greedy matching.[28]

Approximately half of the PS matching articles did not account for the lack of independence between matched pairs in the statistical analyses. Matched subjects are more likely to have similar outcomes than randomly selected subjects,[2 17] therefore variance estimators that account for matching (eg,. paired t-tests, McNemars test, Cox models stratified on matched pairs, generalised estimating equations accounting for matched pairs) should be used.[2 11 29]

Finding an untreated match for each treated subject is the most common strategy, and for simplicity is the only matching strategy described. However, it is also possible to find a treated match for each untreated subject or randomly finding a match for a subject in the sample. This method estimates the ATU (described in the section Interpretation of treatment effect).

Inverse probability of treatment weighting

In IPTW, subjects are weighted by the inverse of the probability of receiving the treatment that the subject received (1/PS for treated and 1/(1–PS) for untreated subjects),[1 2] creating a pseudo-population in which measured baseline characteristics are independent of treatment status. Compared with other PS methods, IPTW allows for the adjustment of time-dependent covariates, and unlike matching will not lose power from the reduced sample size that results from unmatched observations.[1 2 15 25] IPTW estimates, however, may be more sensitive to misspecification of the PS model and extreme PS values.[1]

Having treated subjects with low probability of treatment or untreated subjects with high probability of treatment result in large weights, increasing the variance of the effect estimate. Mitigation strategies include stabilisation, trimming and truncation. Stabilisation multiplies the weight by a constant; truncation sets any values exceeding a set threshold to that threshold (often based on the quantile distribution of weights, for example, 1st and 99th percentiles) and trimming removes subjects with weights beyond a set threshold (weight quantiles or non-overlap region).[1 15 25] Specifying the use of these techniques and the proportion of subjects exceeding the thresholds provides insight into the precision and generalisability of the effect estimates.

With IPTW, correct estimation of standard errors is limited to the use of robust, sandwich-type estimators or non-parametric bootstrap methods. While the former adjusts for the lack of independence in the weighted sample,[15 25] bootstrapping accounts for PS sampling variability, resulting in more accurate variance estimation, and therefore is recommended for use with IPTW.[15 25]

Stratification

Stratification divides the entire sample into mutually exclusive subgroups based on the PS and estimates treatment effects within each stratum. Stratum-specific estimates are then pooled or averaged to estimate the overall effect.[1 2] Stratification can be less sensitive to slight misspecification of the PS model than IPTW or direct adjustment.[1] It can, however, result in more biased treatment effect estimates than IPTW or matching, especially in survival analyses.[29] Stratification of PS is often used in complex survey designs.[30]

The majority of stratification articles created five or more strata. Stratification based on quantiles of PS eliminates 90% of bias from measured confounders, which is only minimally reduced with each additional stratum.[31 32] When strata sizes are unequal, combining stratum-specific estimates weighted by the proportion of subjects in each stratum, rather than the inverse variance, performs better in the presence of heterogeneity.[33] In addition, trimming can be used for extreme PS values and should be reported accordingly.

Direct adjustment

In direct adjustment, the outcome is regressed on the PS,[1 34] which can include large number of covariates and interaction terms to create a more parsimonious model.[1 2] Conditioning on the PS thus occurs in the analysis phase of the study, whereas for all other PS methods, it occurs in the design phase without regard to the outcome, which is one key advantage of the other PS methods.[1 4 15 16] Consequently, direct adjustment can lead to more biased effect estimates than other PS methods if not correctly specified (eg, using a spline).[1 2 34 35] In contrast to the other PS methods, direct adjustment typically produces conditional, instead of marginal, effect estimates. Conditional effects are interpreted at the individual level and consist of moving individual subjects with the same covariate pattern from untreated to treated.[2]

Standard methods for assessing balance between treatment groups cannot be used in direct adjustment because the PS model is incorporated into the outcome model.[14] Instead, alternative diagnostics methods including weighted conditional standardised difference and quantile regression comparing the distribution of baseline covariates should be performed.[1 14 34] The former integrates the standardised difference over the distribution of the PS in the study sample and compares the means of the baseline covariates.[14] Quantile regression compares the conditional distribution of baseline covariates between treatment groups[14] to show whether the treatment effect is constant or heterogeneous across PS for each covariate.[36]

Heterogeneity

Although more than half of reviewed articles assessed heterogeneity, it was conducted inconsistently. Heterogeneity results when effect estimates differ by magnitude and/or direction between subgroups of a population. Articles either presented effect estimates by strata of potential risk factors/effect modifiers, by strata of PS, or only stated heterogeneity was assessed. Overall, it is important to present details of the heterogeneity assessment for transparency and accurate interpretation of results.

Assessment of heterogeneity applies to all study designs. In the context of PS, commonly used methods include: (1) presenting results by strata of the PS,[22] (2) calculating PS within strata of strong risk factors or (3) for IPTW, incorporating strong risk factor in the numerator of the weight calculations and as an interaction term in the outcome model.[18] Other methods are still in development. Overall, it is important to present the details of the heterogeneity assessment for transparency and accurate interpretation of results.

Interpretation of treatment effect

After careful application of PS methods, a fundamental aspect of using PS methods is an accurate interpretation of effects, specifically, to clarify which target population the results apply to. Similar to RCTs, PS methods allow the researcher to estimate marginal treatment effects.[2–4] A marginal treatment effect is the average treatment effect at the population level and includes ATE, ATT and ATU.[2] Conceptually, the ATE consists of moving the entire population from untreated to treated, regardless of the treatment actually received.[2] The treated sample becomes the reference group to which the treated and untreated subjects are being standardised for the ATT and the untreated sample become the reference group to which subjects are being standardised for the ATU.[2]

As different treatment effects refer to different target populations, careful and precise interpretation of results that specifically identifies the correct population is necessary. For ATE, the interpretation should include

the inclusion criteria and the effect for the entire population (treated and untreated). For example, a correct interpretation of an ATE would be 'our findings show that in an unselected heart failure population (entire population studied), candesartan (treatment) was associated with lower all-cause mortality (outcome) compared with losartan (standard of care treatment) (overall effect specified and not according to treatment group)'.[37] For studies estimating ATT, conclusions are limited to inclusion criteria and the treatment effect in the treated subgroup only. For example, a correct ATT interpretation would be 'among heart failure patients discharged from the emergency department (entire population studied), the risk of death and morbidity of recurrent hospital visits (outcome) was reduced in those who received care within 30 days after discharge that was shared by a primary care physician and a cardiac specialist (treatment effect for treated patients only described)'.[38] For ATU studies, the following interpretation would be correct: 'the risk for intraoperative/perioperative use of blood products such as fresh frozen plasma and cryoprecipitate was lower in the aprotinin era patients (untreated), as was the overall risk for the use of rFVlla'; where post-aprotinin era is the treatment group.[39] Additional examples of interpretations of individual treatment effect are provided in the online supplementary appendix.

Unless otherwise stated, matching typically seeks to estimate the ATT. IPTW and stratification generally estimate the ATE, and direct adjustment estimates conditional effects.[17 19] Any PS method can estimate either ATE, ATT or ATU after specific analytical steps, which should be reported to ensure correct effect estimate interpretation.[1 2 15] Further traditional regression adjustment (non-PS) following PS methods will make the interpretation of effect estimates more difficult due to mixing conditional and marginal effects, however, it can lead to multiple robustness which is a desirable property.[2] Thus, all statistical methods used in the study should be carefully reported for correct effect estimate interpretation.

Limitations

First, this survey captures PS methodological details as reported in published articles. As recommended in the present review, these key elements are important for transparency and interpretation of results and should be included, at a minimum, in appendices. The present study, however, could not investigate differences in validity of studies based on differences in PS techniques used. Second, although each article was reviewed by two independent reviewers, some differences in interpretation may remain. Third, the current manuscript discusses mainstream PS methods, however, many more approaches exist, and details are provided elsewhere.[2 29 40–44] Also, although general medical journals may publish more articles using PS methods, only articles focused on cardiovascular topics were included in the present study. Finally, as only the number of articles using PS methods was reported, the proportion of total articles published in each journal using PS was not measured.

CONCLUSION

Although PS methods are frequently used in high-impact cardiovascular and medical journals, reporting of methodological details has been inconsistent. We have proposed a guidance document outlining the necessary elements to report when using PS.

Twitter James M Brophy @brophyj

Contributors All authors (MS, BB, JR, JK, RWP, JB, JSK) contributed to the development of the research proposal, development of the data extraction form and revised the manuscript. MS, BB, JR and JK completed the data extraction and MS wrote the manuscript.

Funding MS, BB and JK are supported by doctoral funding grants from Fonds de Research Santé Quebec (FRQS) and JR is supported by a doctoral funding grant from the Canadian Institutes of Health Research (CIHR).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. Article is a systematic review and all data is presented in the manuscript or supplement.

ORCID iDs
Michelle Samuel http://orcid.org/0000-0002-9674-6565
Brice Batomen http://orcid.org/0000-0002-5772-120X
Julie Rouette http://orcid.org/0000-0003-3882-0998
Joanne Kim http://orcid.org/0000-0001-7458-2128
Robert W Platt http://orcid.org/0000-0002-5981-8443
James M Brophy http://orcid.org/0000-0001-8049-6875
Jay S Kaufman http://orcid.org/0000-0003-1606-401X

REFERENCES

1 Deb S, Austin PC, Tu JV, *et al*. A review of propensity-score methods and their use in cardiovascular research. *Can J Cardiol* 2016;32:259–65.
2 Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399–424.
3 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;70:41–55.
4 Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med* 2007;26:20–36.
5 Abdia Y, Kulasekera KB, Datta S, *et al*. Propensity scores based methods for estimating average treatment effect and average treatment effect among treated: a comparative study. *Biom J* 2017;59:967–85.
6 Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. *Circ Cardiovasc Qual Outcomes* 2008;1:62–7.
7 Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med* 2008;27:2037–49.
8 Austin PC. Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *J Thorac Cardiovasc Surg* 2007;134:e3:1128–35.

9 Ellis AG, Trikalinos TA, Wessler BS, *et al*. Propensity score-based methods in comparative effectiveness research on coronary artery disease. *Am J Epidemiol* 2018;187:1064–78.
10 Sundbøll J, Adelborg K, Munch T, *et al*. Positive predictive value of cardiovascular diagnoses in the Danish national patient registry: a validation study. *BMJ Open* 2016;6:e012832.
11 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;28:3083–107.
12 Austin PC. Statistical criteria for selecting the optimal number of untreated subjects matched to each treated subject when using many-to-one matching on the propensity score. *Am J Epidemiol* 2010;172:1092–7.
13 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.
14 Austin PC. Goodness-Of-Fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiol Drug Saf* 2008;17:1202–17.
15 Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med* 2016;35:5642–55.
16 Brookhart MA, Schneeweiss S, Rothman KJ, *et al*. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
17 Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat* 2004;86:4–29.
18 Hernan MA, Robins JM. *Causal inference*. FL: CRC Boca Raton, 2010.
19 Heckman J, Ichimura H, Smith J, *et al*. Characterizing selection bias using experimental data. *Econometrica* 1998;66:1017–98.
20 Weitzen S, Lapane KL, Toledano AY, *et al*. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2004;13:841–53.
21 Imai K, King G, Stuart EA. Misunderstandings between experimentalists and observationalists about causal inference. *J R Stat Soc Ser A Stat Soc* 2008;171:481–502.
22 Shrier I, Pang M, Platt RW. Graphic report of the results from propensity score method analyses. *J Clin Epidemiol* 2017;88:154–9.
23 Westreich D, Cole SR, Funk MJ, *et al*. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol Drug Saf* 2011;20:317–20.
24 Rosenbaum PR, Rubin DB. The bias due to incomplete matching. *Biometrics* 1985;41:103–16.
25 Curtis LH, Hammill BG, Eisenstein EL, *et al*. Using inverse probability-weighted estimators in comparative effectiveness analyses with observational databases. *Med Care* 2007;45:S103–7.
26 Hill J, Reiter JP. Interval estimation for treatment effects using propensity score matching. *Stat Med* 2006;25:2230–56.
27 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010;25:1–21.
28 XS G, Rosenbaum PR. Comparison of multivariate matching methods: structures, distances, and algorithms. *J Computational Graphical Stat* 1993;2:405–20.
29 Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med* 2014;33:1242–58.
30 Zanutto EL. A comparison of propensity score and linear regression analysis of complex survey data. *J Data Sci* 2006;4:67–91.
31 Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;24:295–313.
32 Hullsiek KH, Louis TA. Propensity score modeling strategies for the causal analysis of observational data. *Biostatistics* 2002;3:179–93.
33 Rudolph KE, Colson KE, Stuart EA, *et al*. Optimally combining propensity score subclasses. *Stat Med* 2016;35:4937–47.
34 Zou B, Zou F, Shuster JJ, *et al*. On variance estimate for covariate adjustment by propensity score analysis. *Stat Med* 2016;35:3537–48.
35 Alam S, Moodie EEM, Stephens DA. Should a propensity score model be super? the utility of ensemble procedures for causal adjustment. *Stat Med* 2019;38:1690–702.
36 Stürmer T, Joshi M, Glynn RJ, *et al*. 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.e1–437.e24.
37 Eklind-Cervenka M, Benson L, Dahlström U, *et al*. Association of candesartan vs losartan with all-cause mortality in patients with heart failure. *JAMA* 2011;305:175–82.
38 Lee DS, Stukel TA, Austin PC, *et al*. Improved outcomes with early collaborative care of ambulatory heart failure patients discharged from the emergency department. *Circulation* 2010;122:1806–14.
39 DeSantis SM, Toole JM, Kratz JM, *et al*. Early postoperative outcomes and blood product utilization in adult cardiac surgery: the post-aprotinin era. *Circulation* 2011;124:S62–9.
40 Austin PC, Schuster T. The performance of different propensity score methods for estimating absolute effects of treatments on survival outcomes: a simulation study. *Stat Methods Med Res* 2016;25:2214–37.
41 de Los Angeles Resa M, Zubizarreta JR. Evaluation of subset matching methods and forms of covariate balance. *Stat Med* 2016;35:4961–79.
42 DE H, Imai K, King G, *et al*. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *J Political analysis* 2007;15:199–236.
43 Li L, Greene T. A weighting analogue to pair matching in propensity score analysis. *Int J Biostat* 2013;9:215–34.
44 Austin PC. Advances in propensity score analysis. *Stat Methods Med Res* 2020;29:641–3.