ELSEVIER

Journal of Clinical Epidemiology ■ (2014) ■

Journal of Clinical Epidemiology

REVIEW ARTICLE

Reporting of covariate selection and balance assessment in propensity score analysis is suboptimal: a systematic review

M. Sanni Ali[a], Rolf H.H. Groenwold[a], Svetlana V. Belitser[a], Wiebe R. Pestman[a], Arno W. Hoes[a], Kit C.B. Roes[a], Anthonius de Boer[a], Olaf H. Klungel[a,b,c,*]

[a]*Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Universiteitsweg 99, Utrecht, The Netherlands*

[b]*Division of Pharmacoepidemiology & Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, Utrecht, The Netherlands*

[c]*Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands*

Accepted 1 August 2014; Published online xxxx

Abstract

Objectives: To assess the current practice of propensity score (PS) analysis in the medical literature, particularly the assessment and reporting of balance on confounders.

Study Design and Setting: A PubMed search identified studies using PS methods from December 2011 through May 2012. For each article included in the review, information was extracted on important aspects of the PS such as the type of PS method used, variable selection for PS model, and assessment of balance.

Results: Among 296 articles that were included in the review, variable selection for PS model was explicitly reported in 102 studies (34.4%). Covariate balance was checked and reported in 177 studies (59.8%). *P*-values were the most commonly used statistical tools to report balance (125 of 177, 70.6%). The standardized difference and graphical displays were reported in 45 (25.4%) and 11 (6.2%) articles, respectively. Matching on the PS was the most commonly used approach to control for confounding (68.9%), followed by PS adjustment (20.9%), PS stratification (13.9%), and inverse probability of treatment weighting (IPTW, 7.1%). Balance was more often checked in articles using PS matching and IPTW, 70.6% and 71.4%, respectively.

Conclusion: The execution and reporting of covariate selection and assessment of balance is far from optimal. Recommendations on reporting of PS analysis are provided to allow better appraisal of the validity of PS-based studies.

Keywords: Balance; Confounding; Pharmacoepidemiology; Propensity score; Reporting; Variable selection

1. Introduction

In observational studies, treated and control subjects often differ systematically on prognostic factors leading to treatment-selection bias or confounding in estimating the (adverse) effect of treatment on an outcome. Analytical tools such as the propensity score (PS) methods are applied to correct for such confounding bias. In their seminal article, Rosenbaum and Rubin described the PS as a balancing score: treated and untreated subjects with the same PS tend to have similar distributions of measured confounders given the PS [1]. In other words, assuming no unmeasured confounding and having adequately measured confounders, conditioning on the PS allows one to obtain an unbiased estimate of the average treatment effect at that value of the PS.

PS analysis involves two key steps: deriving the PS from the data and estimating the treatment effect by using the PS to control for confounding. The first step involves an iterative process of fitting a PS model (eg, using logistic regression) on selected covariates until an optimal balance on those covariates is achieved [2]. Despite the growing popularity of PS methods in epidemiology, criteria for selecting variables for a PS model are not well developed compared with variable selection for conventional outcome models [3,4]. Once the PSs

Funding: The PROTECT project is supported by Innovative Medicines Initiative (IMI) Joint Undertaking (www.imi.europa.eu) under Grant Agreement no. 115004, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies in kind contribution. In the context of the IMI Joint Undertaking, the Department of Pharmacoepidemiology, Utrecht University, also received a direct financial contribution from Pfizer. The views expressed are those of the authors only and not of their respective institution or company.

Conflict of interest: O.H.K. received unrestricted funding for pharmacoepidemiological research from the Dutch private-public funded top institute Pharma.

* Corresponding author. Tel.: +31 6 288 31 313; fax: +31 30 253 91 66.
E-mail address: O.H.Klungel@uu.nl (O.H. Klungel).

http://dx.doi.org/10.1016/j.jclinepi.2014.08.011
0895-4356/

What is new?

Key findings

- Balance of confounders between treatment groups is not properly checked and reported in propensity score (PS) analysis.
- *P*-values from significance tests are the most commonly used statistical tools for checking balance.

What this adds to what was known?

- Reporting of PS analysis including assessment of balance of confounders is far from optimal in the medical literature.
- Balance is more often checked in articles using PS matching and inverse probability of treatment weighing.

What is the implication and what should change now?

- The reporting of aspects of PS analysis such as covariate selection and balance assessment should be improved.

are derived, an intermediate step is using one of the four possible methods: matching, stratification or subclassification, covariate adjustment, and inverse probability of treatment weighting (IPTW) using the PS and checking the balance of covariate distribution between treatment groups using appropriate metric [2]. The choice of the PS method depends on the specific research question, the target population, and inferential goals of the study [5–7], and it affects the way balance on covariates is assessed. Finally, the effect of treatment on the outcome is estimated using one of the PS methods chosen in the previous step.

Although the use of PS methods has shown a dramatic increase in the medical literature [8], previous literature reviews indicated that most authors do not adequately report information on the PS model development [9,10], the balance of covariates between the treatment groups in PS analysis [8,9,11,12], and those who report, often use inappropriate diagnostics [8,9,11]. In addition, researchers often ignore explicit discussion of the PS estimate (estimand) and its relationship with their research question [5]. However, the reviews were limited to PS matching [8,11], and detailed information on the current practice is very limited.

The PS methodology has evolved over the last few years, during which researchers have proposed recommendations on variable selection for PS model [4,13–16] and statistical tools for checking balance and/or selecting the optimal PS models [17–21] and advised against the use of some statistics such as significance testing or prematching C statistics for evaluating balance and appropriateness of a PS model [17,22–24]. However, the current practice on selecting variables for PS model, choosing a specific PS method as well as measuring and reporting of covariate balance, is not well documented. Therefore, the objective of this study was twofold (1) to systematically review the practice of variable selection and PS model building with emphasis on assessment and reporting of balance when using PS analysis in the medical literature and (2) to provide practical recommendations on the reporting of PS analysis.

2. Methods

We performed a PubMed search to identify studies that used different PS methods. The search was conducted on June 2, 2012, using keywords: ‘‘propensity score(s)’’ or ‘‘propensity matching’’ in all fields (title, abstract, body, or references) identifying 2,317 unduplicated references. To assess the current practice, we limited our search to 6 months (December 2011–May 2012). Articles were excluded if they addressed only methodological or statistical aspects of the PS, if they are unrelated to medical research or published in languages other than English, or if they were reviews, editorials, or letters.

All authors discussed on identifying aspects of the PS analysis on which data had to be collected, but the extraction was by one of the investigators (M.S.A.). From each article included for the review, we extracted information on the type of PS method used, the methods used to estimate the PS, how variables were selected for inclusion in the PS model, whether balance on confounder was checked, methods used for checking balance, and the ‘‘appropriateness’’ of PS model. When PS matching was used, we recorded information on whether the articles mentioned the matching algorithm applied, the treated/untreated matching ratios used, size of the matched pairs and the starting population, and whether matching was taken into account in the analysis. When stratification on the PS was applied, we extracted information on the quantile of the PS used (deciles, quintiles, quartiles, or tertiles). In addition, information on the impact factor (IF) of the journals [25] and the SCImago Journal Rank (SJR) indicator from Scopus [25–27], a measure of quality of the journals, extracted for articles included in the review to allow direct comparison of sources in different subject fields. Chi-squared tests were used to compare the frequency of reporting balance assessment and the use of different balance metrics among quintiles of the IF and the SJR of the journals in which the reviewed articles were published.

3. Results

The PubMed search identified 388 articles, of which 92 were excluded: methodological or statistical articles ($n = 20$), articles unrelated to medical research ($n = 63$),

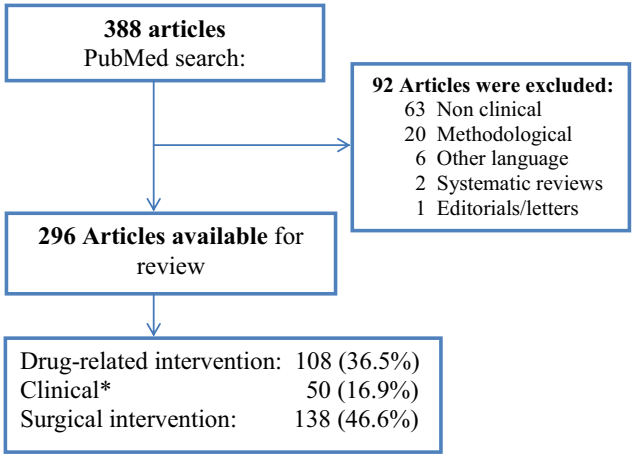

Fig. 1. Flow chart of abstract or article extraction for the systematic review. *Studies which did not involve drug-related or surgical interventions were classified as clinical.

articles published in languages other than English ($n = 6$), reviews ($n = 2$), and editorials or letters ($n = 1$; Fig. 1). This resulted in 296 articles published in the medical literature during December 2011—May 2012 that used PS methods in empirical data (the articles can be found in the Appendix at www.jclinepi.com).

The articles included for analysis were related to cardiovascular research (148, 50.0%), cancer research (41, 13.9%), renal research (18, 6.1%), neurological research (16, 5.4%), respiratory research (15, 5.1%), and other fields of medical research (57, 19.2%). Surgical interventions and drug-related evaluation studies constituted most of the articles included in the review, 138 (46.6%) and 108 (36.5%), respectively (Table 1).

3.1. Variable selection and PS estimation

Most articles (194, 65.5%) did not explicitly mention how variables were selected for the PS model. Variables' association with treatment, outcome, and both treatment and outcome was considered and reported in 38 (12.8%), 39 (13.2%), and 30 (10.1%) of the studies, respectively. Background knowledge was specifically mentioned in 14 studies (4.7%); only four of these articles explicitly reported that they took into account variables' association with the outcome and/or treatment. Inclusion of interaction or higher order terms in the PS model was reported in 17 articles (5.7%), but none of these articles mentioned any motivation for the inclusion of interaction and higher order terms. Only seven articles (2.4%) reported the PS model itself and how the variables were modeled. Other methods considered include stepwise variable selection methods, C statistics ($n = 41$, 13.9%), Hosmer–Lemeshow goodness-of-fit tests ($n = 25$, 8.4%), and balance measures ($n = 48$, 16.2%). Almost all studies (283, 95.6%) reported the variables included in the PS model.

Most articles reported that binary logistic regression was used for estimating the PS (260, 87.5%). Four articles reported that they used multinomial logistic regression where the exposure was categorical with more than two levels and two other articles reported the use of recursive partitioning for estimating PS in binary exposures. Other methods reported include the probit model ($n = 1$, 0.3%) and high-dimensional PS ($n = 3$, 1%).

3.2. Balance assessment and PS methods

Balance of confounders between treatment groups was checked and reported in 177 (59.8%) of the articles, and the most commonly used statistical tools to report balance were the *P*-value (125 of 177, 70.6%) from hypothesis testing (eg, chi-square test or t-test). Standardized difference (SDif) was used in 45 (25.4%) of the studies where balance was reported and 11 (6.2%) used graphical displays such as SDif plots, PS boxplots, kernel plots, and histograms to assess covariate balance (Table 2). Hosmer–Lemeshow goodness-of-fit test and the C statistic of the PS model were reported in 26 (8.8%) and 39 (13.2%) of the reviewed studies, respectively. Frequency of balance assessment did not seem to differ among studies involving surgical interventions (61.6%), drug-related intervention (57.4%), and clinical studies (58%; $P > 0.05$).

The PS can be used in different ways to control for confounding: matching, stratification, PS adjustment, or weighting. Matching on the PS was the most commonly used approach (204 of 296, 68.9%), followed by PS adjustment (62 of 296, 20.9%), and stratification using the PS (41

Table 1. Classification of based articles included in the review by body system and type of exposure (number and percentage)

Type of research	Drug related ($n = 108$)	Surgical ($n = 138$)	Clinical[a] ($n = 50$)	Total (296)
Cardiovascular	46 (42.6)	85 (61.6)	17 (34.0)	148 (50.0)
Cancer	11 (10.2)	27 (19.6)	3 (6.0)	41 (13.9)
Renal	4 (3.7)	9 (6.5)	5 (10.0)	18 (6.1)
Respiratory	7 (6.5)	4 (2.9)	4 (8.0)	15 (5.1)
Neurological	10 (9.3)	2 (1.4)	4 (8.0)	16 (5.4)
Infection	9 (8.3)	1 (0.7)	5 (10.0)	15 (5.1)
General/nonspecific	11 (10.2)	4 (2.9)	9 (18.0)	24 (8.1)
Others[b]	10 (9.3)	6 (4.3)	3 (6.0)	18 (6.1)

Abbreviation: PS, propensity score.

[a] Studies that did not involve drug-related or surgical interventions were classified as clinical.

[b] Others include digestive, lymphatic, musculoskeletal, and ophthalmic studies.

Table 2. The frequency of different methods used for checking balance of confounders between treatment groups among the different methods

Methods	Number of articles (n)[a]	Matching (n = 204)	Covariate adjustment (n = 62)	Stratification (n = 41)	IPTW (n = 21)
SDif	45 (25.4)	42 (20.6)	6 (9.7)	1 (2.4)	3 (14.3)
P-values[b]	125 (70.6)	105 (51.5)	15 (24.2)	13 (31.7)	10 (47.6)
Graphical displays	11 (6.20)	6 (2.9)	3 (4.8)	2 (4.9)	—
Eye balling	4 (2.3)	3 (1.5)	1 (1.6)	2 (4.9)	—
Others[c]	13 (7.3)	10 (4.9)	5 (8.0)	1 (2.4)	1 (4.7)

Abbreviations: PS, propensity score; IPTW, inverse probability of treatment weighting; SDif, standardized difference.

[a] Number of studies include those in which balance was checked and reported (n = 177); the total does not add up to 177 because some of the articles may have used more than one type of PS methods (matching, covariate adjustment, stratification, or IPTW).

[b] *P*-values from hypothesis testing (chi-square test or Fisher's exact test, t-test, Wilcoxon's signed-rank test, McNemar's test, Mann–Whitney U-test, the Kruskal–Wallis test, and logistic regression).

[c] Others include Kolmogorov–Simonov test, Lévy distance, overlapping coefficient, multivariate models, and percent reduction in bias.

of 296, 13.9%). Weighting was applied in 21 articles (7.1%). Three studies did not mention the PS methods used, and 26 articles (8.8%) used combinations of two or more of the PS methods. Ten studies reported that they sensitivity analysis, and three studies specifically indicated that the PS was used as a sensitivity analysis.

Among the studies that used PS matching, one-to-one matching was the most frequently reported approach (118 of 204, 57.8%); the matching ratio was neither explicitly reported nor clear from the data in 73 (35.8%) of the studies using PS matching (Table 3). The matching algorithm used to form the matched pairs was reported only in 67 studies (32.8%) that applied PS matching, and greedy matching with the nearest neighbor matching was more often reported (n = 42). Other matching approaches reported include "5-to-1-digit" greedy matching [6] (n = 10), Greedy matching on Mahalanobis distance (n = 6), "8-to-1-digit" greedy matching [6] (n = 3), stratified matching (n = 3), optimal matching (n = 1), and exact matching (n = 2). The reporting of caliper width was poor and inconsistent (ranging 0.001–0.06 standard deviations on the logit of PS), the frequently used being 0.02 standard deviations on the logit of the PS (n = 20). Unmatched subjects were reported excluded from the analysis in four studies and retained in one study for efficiency reason.

Table 3. The frequency of the different methods and balance assessment

Method	Number of articles (n)	Balance checked
PS matching[a]	204 (68.9)	144 (70.6)
1:1 matching	118	92
1:2 matching	3	2
1:3 matching	4	3
1:4 matching	5	3
Covariate adjustment using PS	62 (20.9)	25 (40.3)
Stratification using PS[a]	41 (13.9)	17 (41.5)
Quintiles of PS	21	10
Deciles of PS	8	1
Quartiles of PS	3	2
Tertiles of PS	5	3
IPTW	21 (7.1)	15 (71.4)
Mixed[b]	26 (8.8)	18 (69.2)

Abbreviations: PS, propensity score; IPTW, inverse probability of treatment weighting.

[a] Some articles did not mention ratio of treated to untreated patients used in matching and quantiles used in stratifictaion using PS.

[b] Studies used a combination of two or more of the different PS methods (matching, covariate adjustment, stratification, or IPTW).

Covariate balance was more often checked and reported in studies using PS matching (144 of 204, 70.6%) and IPTW (15 of 21, 71.4%; Table 4). Covariate balance before and after matching or stratification or weighting was compared only in 110 articles (37.2%).

Among the studies that applied stratification on the PS, stratification based on quintiles of the PS was the most common application (21 of 41, 51.2%). Four studies (9.8%) did not mention the PS quantiles used for stratification. Four studies reported that observations in the first and fifth quintiles of PS were excluded due to lack of overlap ("nonpositivity").

3.3. Use and reporting of different diagnostic methods in different journals

The IF of the journals from which articles were included for the review ranged from 0.1 (the Korean Journal of Thoracic Cardiovascular Surgery) to 53.3 (The New England Journal of Medicine, NEJM); there was no association between journal's IF and the frequency of reporting PS analysis ($P > 0.05$). The SJR indicator of the journals varied between 0.11 (Managed Care) and 10.16 (NEJM). The frequency of balance checking was similar among studies within quintiles of the SJR indicator (64% and 57% in the first and fifth quintile of SJR indicator, respectively). However, the use and reporting of *P*-values from hypothesis testing, C statistic, and goodness-of-fit tests of the PS model was less common in studies published in journals from the fifth quintiles of SJR ranking compared with those published in journal from the first quintiles (37.3% vs. 49.2%, 10.2% vs. 16.9%, and 6.8% vs. 13.8% for *P*-values, C statistic, and goodness-of-fit tests, respectively, $P > 0.05$). In addition, the use of absolute SDif for measuring and reporting balance was higher, although not significant, in studies published in journals with in the higher quintiles of SJR indicator (18.6% and 19.6% in the fourth and fifth quintiles vs. 9.7% and 12.1% in the first and second quintiles of SJR indicator, $P > 0.05$). Similarly,

Table 4. Comparison of different statistical methods for assessing and reporting model fit and/or covariate balance

Balance diagnostic	Short description	Strength	Limitation
Test of significance	• Assess evidence in favor of some claim about the population from which the sample has been drawn • Frequently used to compare the distribution of measured baseline covariates between treated and control subjects in the PS analysis [8]	• Easy to use • Easy to interpret • Can be derived from nonparametric tests • Scale invariant	• Influenced by sample size • It is not a characteristic of a sample (relates to a hypothetical population) [28] • Arbitrary cutoff (threshold) • Gives little or no information whether the PS model has been correctly specified (hence, bias)
C statistic	• It refers to the ability of the PS model to accurately distinguish treated subjects from untreated ones [29–31] • For binary exposure, it is identical to the area under the receiver operating characteristic curve [31] • The value ranges between 0.5 (classification no better than a pure chance) to 1.0 (perfect classification)	• Easy to use • Easy to interpret • Gives information on the full model • Scale invariant • Nonparametric	• Gives no signal whether the PS model has been correctly specified or key confounders have been omitted from the PS model (hence, no indication of bias) [8,22,32] • Higher C statistics may not necessarily indicate optimal balance [8,22,24] • Arbitrary cutoff (threshold) • Influenced by sample size.
Goodness-of-fit test	• A measure the compatibility of the observed values from the data with the predicted values from the model in question, that is, they show how well the selected model describes the data [22,31]	• Easy to use • Easy to interpret • Semi and nonparametric alternatives • Gives information on the full model • Scale invariant	• Influenced by sample size • Does not indicate bias [22] • Indicate only model fit not balance of covariates • Arbitrary cutoff (threshold)
Overlapping Coefficient (OVL)	• The amount of overlap in the density of covariate distributions for treated and untreated subjects • For continuous covariates, nonparametric kernel density can be used to estimate the OVL [19,33]. For dichotomous covariates, it is the proportion of overlap [19] • Range between zero (no overlap, that is, perfect imbalance) and one (complete overlap, that is, ''perfect'' balance)	• Characteristic of a sample • Graphical presentation of PS distribution • Scale invariant • Nonparametric • Good indicator of bias in large sample size [19,21]	• Influenced by sample size[a] [19,21] • Estimation is complex • Relies on densities [33] • Arbitrary cutoff (threshold)
Kolmogorov–Smirnov distance (D)	• The maximum vertical distance between two cumulative distribution functions of a certain covariate for treated and untreated subjects expressed as relative frequencies [19,34,35] • Range between zero (''perfect'' balance) and one (complete imbalance)	• Characteristic of a sample • Nonparametric • Does not need densities • Scale invariant [34] • Clear interpretation	• Influenced by sample size[a] [19,21] • Estimation is complex • Fails to capture convergence of distribution • Arbitrary cutoff (threshold)
Lévy distance (L)	• The side length of the largest square that can be inscribed between the cumulative distribution functions of a certain covariate for treated and untreated subjects with sides parallel to the coordinate axes [19,35] • This distance can range between zero (''perfect'' balance) and one (complete imbalance)	• Nonparametric • Characteristic of a sample • Does not need densities • Captures convergence of distribution	• Not scale invariant • Estimation is complex • Influenced by sample size[a] [19,21] • Interpretation is complex • Arbitrary cutoff (threshold)
Absolute standardized difference	• The absolute difference in means between the two treatment groups divided by an estimate of the common standard deviation of that variable in the two treatment groups, that is, the pooled standard deviation [17,23,36,37] • Describes the observed bias in the means (or proportions) of covariates across treatment groups, expressed as a percentage of the pooled standardized deviation [38]	• Easy to calculate • Nonparametric • Clear interpretation • Scale invariant • Not influenced by sample size [19,21] • Characteristic of a sample	• Arbitrary cutoff (threshold)

Abbreviation: PS, propensity score.

[a] The correlation with bias is influenced by sample size only when covariates are continuous or mixed binary and continuous [19,21].

PS weighting for estimating treatment effect was often used in studies published in the highest quintile of the SJR indicator (16.9% in the fifth quintile vs. 3.4% in first quintile, $P < 0.05$).

4. Discussion

The PS method has become a commonly used method for controlling confounding in observational studies. This systematic review reveals that the process of variable selection, assessment, and/or reporting covariate balance as well as PS model fit is inconsistent. Moreover, a limited number of studies reported critical aspects of the PS model development or the use of appropriate statistical methods for checking balance. In general, other observational studies, the conduct and the reporting of the PS methods are poor in the medical literature despite the tremendous methodological discussions the topic in the last few years [3−5,8,9,11,16,17,19,22,39−41]. In the Appendix at www.jclinepi.com to this manuscript, major methodological contributions in the PS methods are summarized.

Our study is consistent with previous systematic reviews with respect to low quality of reporting and/or conduct of PS analysis [8,9,11,12]. However, our review included a large number of recently published studies to evaluate the current status in conducting and reporting of PS analysis. To our knowledge, this review is the first to specifically address the current practice of variable selection based on clinical knowledge as well treatment and/or outcome association apart from algorithmic methods, discrimination, and goodness-of-fit tests of the model. In addition, much effort was put in extracting detailed information on other important aspects of the PS that could help investigators and readers appraise the validity of a given PS analysis. This review is representative of the medical literature because the studies included were restricted neither to epidemiological nor to high impact journals. It could be possible that authors performed detailed analysis but only reported limited information due to journal revision and editorial restrictions [8,9]; however, this does not seem to be a strong justification for poor reporting of the results for two main reasons. First, those studies that reported aspects of the PS analysis used inappropriate statistical methods such as significance testing or C statistics. Second, inconsistency of reporting was observed irrespective of the IF of the journal in which the article was published and even with in a specific journal, which is also in line with a previous systematic review [12].

Lack of well-established standards for conducting and reporting of PS analysis may contribute to inconsistent and poor execution and reporting of the PS analysis despite substantial advances in the PS methodology in the last few years [3,17,18,22,24,40,42]. We, therefore, propose that critical items in relation to PS analysis should be incorporated in guidelines on the reporting of observational studies, such as the STROBE statement [43] and the ENCePP guide on methodological standards in pharmacoepidemiology to improve the quality of reporting [44]. Before we come to recommendations for the reporting of studies that use PS methods, we summarize important issues in the conduct of PS analysis.

First, the variable selection for PS model should be performed with a great care because the choice of variables has tremendous effect on both the bias and the precision of the treatment effect estimate [2−4,16,41,45]. The choice of variables, interaction, and/or higher order terms should be primarily based on prior clinical knowledge [15,41]. Obviously, confounders should always be included in the PS model. Variable selection based on their association with the outcome irrespective of the exposure can improve the precision of an estimated treatment effect without increasing bias [3,15]. In contrast, variables that are strongly related to the treatment but not to the outcome (instrumental variables) or weakly related to the outcome should not be included in the PS model because such variables could amplify bias in the presence of unmeasured confounding particularly in nonlinear models and decrease the precision of the treatment effect estimate [14−16,21,41].

Second, fitting the PS model and extracting the PS values using methods such as ordinary logistic regression or recursive partitioning. Machine learning techniques such as neural networks and classification and regression trees [46] have shown superior performance in terms of bias reduction and more consistent 95% confidence interval coverage than logistic regression approach, particularly under conditions of both nonadditivity and nonlinearity [47,48].

Third, using appropriate PS methods: matching or stratification or inverse probability weighting. The choice of the PS method affects the way balance on covariates should be assessed (fourth step), for example, in PS matching and stratification on the PS, balance can easily be assessed by looking at the distribution of covariates between matched groups or within strata of the PS, respectively. It also dictates the treatment effect estimation (fifth step) and its interpretation (sixth step); hence, it should be based on the inferential goal of the research [5,7,49].

Fourth, assessing balance on measured baseline characteristics between treated and untreated patients in the (matched or stratified or weighted) sample using appropriate balance diagnostics that are specific to a sample and not influenced by sample size [28,50]. Accordingly, balance should be assessed on a selected set of covariates that are confounders and/or independent predictors of outcome (first step). The use of prematching C statistic and goodness-of-fit tests for covariate selection and assessment of PS model fit should be avoided because such methods neither provide sufficient information on whether the PS model is correctly specified nor detect confounders in the PS model [22,24]. In our previous studies [19−21],

we recommended the absolute SDif as a balance measure of choice because it has shown superior performance over other balance measures such as the overlapping coefficient, both in simulation and empirical studies. In addition, it is a very familiar, easy-to-calculate, and reasonably well-understood tool for epidemiologists compared with other balance measures. Once the SDif is calculated per covariate, covariate squares, and important pairwise interactions, covariate-specific SDif can be pooled into a single measure using empirically derived weights based on the strength of association between covariates (or covariate terms) and outcome as suggested by Belitser et al. [19]. Recently, C statistic of the PS model [51] has been suggested as an overall measure of the balance across covariates. It may be simpler to evaluate and has shown comparable performance with SDif in terms of indicating bias; however, unlike the SDif, an assessment of covariate's potential for confounding (by checking balance on the covariate's scale) and identification of whether the imbalances are due to a set of related covariates are difficult [21,38,51]. An iterative process of fitting the PS model, checking balance on covariates, and respecifying the PS model has been suggested by Rosenbaum and Rubin [2]. A comparison of different statistical methods for assessing and reporting of balance and PS model fit is summarized in Table 4. R codes for computing different balance metrics are provided in the Appendix at www.jclinepi.com.

Fifth, estimating treatment effect using appropriate statistical methods depending on the type of the outcome (for example, Cox proportional hazard model for time-to-event data or logistic regression for binary data). When the PS matching is used, the matched nature of the data or lack of independence between observations should be taken into account in the analysis particularly when matching

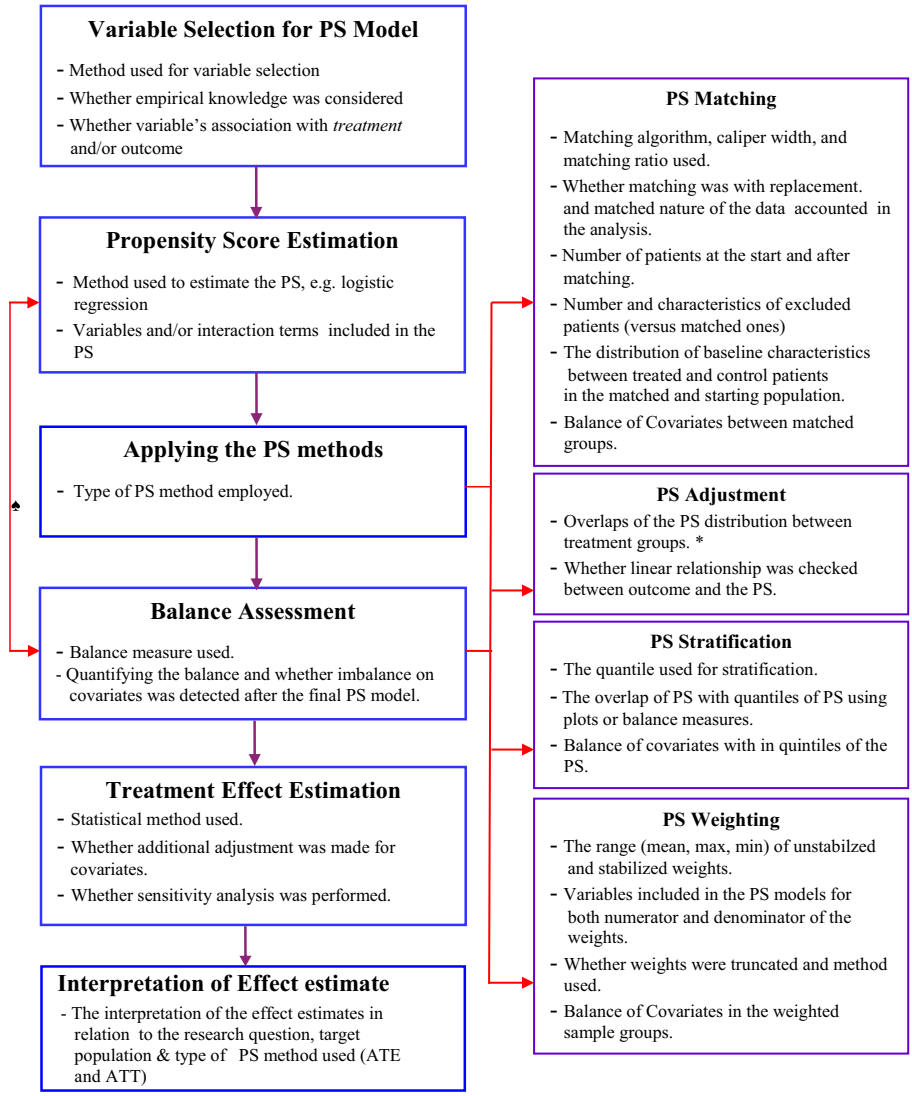

Fig. 2. Flow chart summarizing relevant information to be reported when conducting PS analysis. The PS estimation can be iterative until an ''optimal'' balance on covariates is reached (♠). It is not relevant to report goodness-of-fit tests, prematching C statistic of the PS model, the actual PS values, the PS model itself, *P*-values, and model coefficients from the PS model. *Overlap does not quantify balance and should be complemented with balance measures on individual covariates.

is done with replacement [5,8,11]. In IPTW, the use of stabilizing weights could help "normalize" the range of the inverse probabilities and increase efficiency of the analysis [49,52–55]. Because there is no strong theory regarding when balance is close enough, examining the sensitivity of results to a range of PS specifications is recommended [5,24].

Last but not least, careful interpretation of the treatment effect estimate (estimand) and explanation on the relationships among this estimand, their research question, and the target population in mind [5]. For example, MSM using IPTW estimates a marginal treatment effect [52,53], which on average is similar to the treatment effect in randomized studies; thus, the estimand can be directly interpreted as the average causal treatment effect between treated and patients. However, covariate adjustment and stratification using PS give conditional treatment effect estimates, and their interpretation is not straightforward [39]. This is particularly the case when noncollapsible effect measures such as odds ratio and hazard ratio are used where the conditional and marginal effect estimates differ in the presence of a nonnull treatment effect [7,49,56,57]. On the other hand, PS matching typically focuses on the effect of the treatment either in the treated or in the untreated, not on the average treatment effect on the whole population [5,6]. It is important to note that exclusion of unmatched patients from the analysis not only affect the precision of the effect estimate but also have consequences on the generalizability of results [5,6]. More sophisticated methods such as full matching or one-to-many matching can make use of all available data and may improve the performance of PS matching in terms of reducing bias [58]. In addition, the choice of appropriate caliper and matching algorithm in PS matching deserves great attention for achieving good balance thereby reducing bias. We refer to the literature for detailed aspects of PS matching [6,42,58–60].

Our systematic review of current reporting of PS methods in the medical literature shows that the quality of reporting variable selection, assessing covariate balance, and other important aspects of the PS analysis is far from optimal. The conduct of studies that use PS methods could be split into six essential steps, each of which should be clearly reported. Recommendations for the reporting of PS methods are summarized in Fig. 2. These recommendations may improve the quality of reporting methods, which allows for a better appraisal of the PS-based studies.

Acknowledgments

The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European ConsorTium, www.imi-protect.eu), which is a public–private partnership coordinated by the European Medicines Agency.

Supplementary data

Supplementary data related to this chapter can be found at http://dx.doi.org/10.1016/j.jclinepi.2014.08.011.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.

[2] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. J Am Stat Assoc 1984;79:516–24.

[3] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149.

[4] Patrick AR, Schneeweiss S, Brookhart MA, Glynn RJ, Rothman KJ, Avorn J, et al. The implications of propensity score variable selection strategies in pharmacoepidemiology: an empirical illustration. Pharmacoepidemiol Drug Saf 2011;20:551–9.

[5] Hill J. Discussion of research using propensity-score matching: comments on 'a critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2055–61.

[6] Lunt M. Selecting an appropriate caliper can be essential for achieving good balance with propensity score matching. Am J Epidemiol 2014;179:226–35.

[7] Ali MS, Groenwold RH, Klungel OH. Propensity score methods and unobserved covariate imbalance: comments on "squeezing the balloon". Health Serv Res 2014;49:1074–82.

[8] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27:2037–49.

[9] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841–53.

[10] Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005;58:550–9.

[11] 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;134:1128–35.

[12] D'ascenzo F, Cavallero E, Biondi-Zoccai G, Moretti C, Omedè P, Bollati M, et al. Use and misuse of multivariable approaches in interventional cardiology studies on drug-eluting stents: a systematic review. J Interv Cardiol 2012;25:611–21.

[13] Brookhart MA, Stürmer T, Glynn RJ, Rassen J, Schneeweiss S. Confounding control in healthcare database research: challenges and potential approaches. Med Care 2010;48:S114.

[14] Pearl J. On a class of bias-amplifying variables that endanger effect estimates. In: Grünwald P, Spirtes P, Eds. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence (UAI 2010). 2010; Corvallis, OR: Association for Uncertainty in Artificial Intelligence; 201: 425–432.

[15] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Effects of adjusting for instrumental variables on bias and precision of effect estimates. Am J Epidemiol 2011;174:1213–22.

[16] Pearl J. Invited commentary: understanding bias amplification. Am J Epidemiol 2011;174:1223–7.

[17] 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.

[18] Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. Pharmacoepidemiol Drug Saf 2008;17:1218−25.

[19] Belitser SV, Martens EP, Pestman WR, Groenwold RHH, Boer A, Klungel OH. Measuring balance and model selection in propensity score methods. Pharmacoepidemiol Drug Saf 2011;20:1115−29.

[20] Groenwold RHH, Vries F, Boer A, Pestman WR, Rutten FH, Hoes AW, et al. Balance measures for propensity score methods: a clinical example on beta-agonist use and the risk of myocardial infarction. Pharmacoepidemiol Drug Saf 2011;20:1130−7.

[21] Ali MS, Groenwold RH, Pestman WR, Belitser SV, Roes KC, Hoes AW, et al. Propensity score balance measures in pharmacoepidemiology: a simulation study. Pharmacoepidemiol Drug Saf 2014;23:802−11.

[22] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf 2005;14:227−38.

[23] 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.

[24] Westreich D, Cole SR, Funk MJ, Brookhart MA, Stürmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317−20.

[25] Falagas ME, Kouranos VD, Arencibia-Jorge R, Karageorgopoulos DE. Comparison of SCImago journal rank indicator with journal impact factor. FASEB J 2008;22:2623−8.

[26] Gonzalez-Pereira B, Guerrero-Bote VP, Moya-Anegón F. A new approach to the metric of journals' scientific prestige: the SJR indicator. J Informetr 2010;4:379−91.

[27] Bornmann L, Marx W, Gasparyan AY, Kitas GD. Diversity, value and limitations of the journal impact factor and alternative metrics. Rheumatol Int 2012;32:1861−7.

[28] Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal 2007;15:199−236.

[29] Ash A, Shwartz M. R2: a useful measure of model performance when predicting a dichotomous outcome. Stat Med 1999;18:375−84.

[30] Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 1983;148:839−43.

[31] Hosmer DW Jr, Lemeshow S. Applied logistic regression. 2nd ed. New York, NY: John Wiley & Sons; 2004.

[32] Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics 1993;49:1231−6.

[33] Silverman BW. Density estimation for statistics and data analysis. London, UK: Chapman & Hall/CRC; 1986.

[34] Stephens MA. Use of the Kolmogorov-Smirnov, Cramér-Von Mises and related statistics without extensive tables. J R Stat Soc Ser B Meth 1970;32:115−22.

[35] Pestman WR. Mathematical statistics: an introduction. 2nd ed. Berlin, Germany: Walter De Gruyter Inc; 1998.

[36] Fleiss JL, Levin B, Paik MC. Statistical methods for rates and proportions. 2nd ed. Hoboken, NJ: John Wiley & Sons; 2013.

[37] Hartung J, Knapp G. Statistical inference in adaptive group sequential trials with the standardized mean difference as effect size. Sequential Anal 2011;30:94−113.

[38] Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates Publishers; 1988.

[39] Martens EP, Pestman WR, De Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008;37:1142−7.

[40] Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol 2006;98:253−9.

[41] Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, et al. Myers et al. Respond to "understanding bias amplification". Am J Epidemiol 2011;174:1228−9.

[42] Stuart EA. Developing practical recommendations for the use of propensity scores: discussion of 'a critical appraisal of propensity score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine. Stat Med 2008;27:2062−5.

[43] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 2007;45:247−51.

[44] ENCePP Guide on Methodological Standards in Pharmacoepidemiology. EMA/95098/2010. Available at www.encepp.eu/standards_and_guidances. Accessed June 22, 2013

[45] Mortimer KM, Neugebauer R, Van Der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. Am J Epidemiol 2005;162:382−8.

[46] Westreich D, Lessler J, Funk MJ. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. J Clin Epidemiol 2010;63:826−33.

[47] Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS One 2011;6:e18174.

[48] Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. Pharmacoepidemiol Drug Saf 2008;17:546−55.

[49] Ali MS, Groenwold RH, Pestman WR, Belitser SV, Hoes AW, de Boer A, et al. Time-dependent propensity score and collider-stratification bias: an example of beta2-agonist use and the risk of coronary heart disease. Eur J Epidemiol 2013;28:291−9.

[50] 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.

[51] Franklin JM, Rassen JA, Ackermann D, Bartels DB, Schneeweiss S. Metrics for covariate balance in cohort studies of causal effects. Stat Med 2014;33:1685−99.

[52] Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561−70.

[53] Robins JM, Hernán MÁ, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550−60.

[54] Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology 2004;15:615−25.

[55] Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. Am J Epidemiol 2006;163:262−70.

[56] Greenland S, Pearl J. Adjustments and their consequences−collapsibility analysis using graphical models. Int Stat Rev 2010;79:401−26.

[57] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734−53.

[58] Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. Pharmacoepidemiol Drug Saf 2012;21:69−80.

[59] 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.

[60] Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2013;33:1057−69.