European Journal of Cardio-Thoracic Surgery 53 (2018) 1112–1117
doi:10.1093/ejcts/ezy167 Advance Access publication 19 April 2018

REVIEW

Cite this article as: Benedetto U, Head SJ, Angelini GD, Blackstone EH. Statistical primer: propensity score matching and its alternatives. Eur J Cardiothorac Surg 2018;53:1112–7.

Statistical primer: propensity score matching and its alternatives†

Umberto Benedetto[a,*], Stuart J. Head[b], Gianni D. Angelini[a] and Eugene H. Blackstone[c]

[a] Bristol Heart Institute, University of Bristol, School of Clinical Sciences, Bristol, UK
[b] Department of Cardiothoracic Surgery, Erasmus Medical Center, Rotterdam, Netherlands
[c] Department of Thoracic and Cardiovascular Surgery & Clinical Investigations, Cleveland Clinic Foundation, Cleveland, OH, USA

* Corresponding author. Bristol Heart Institute, University of Bristol, Upper Maudlin St., Bristol BS2 8HW, UK. Tel: +44-117-9230000; e-mail: umberto.benedetto@bristol.ac.uk (U. Benedetto).

Received 8 November 2017; received in revised form 16 March 2018; accepted 25 March 2018

Summary

Propensity score (PS) methods offer certain advantages over more traditional regression methods to control for confounding by indication in observational studies. Although multivariable regression models adjust for confounders by modelling the relationship between covariates and outcome, the PS methods estimate the treatment effect by modelling the relationship between confounders and treatment assignment. Therefore, methods based on the PS are not limited by the number of events, and their use may be warranted when the number of confounders is large, or the number of outcomes is small. The PS is the probability for a subject to receive a treatment conditional on a set of baseline characteristics (confounders). The PS is commonly estimated using logistic regression, and it is used to match patients with similar distribution of confounders so that difference in outcomes gives unbiased estimate of treatment effect. This review summarizes basic concepts of the PS matching and provides guidance in implementing matching and other methods based on the PS, such as stratification, weighting and covariate adjustment.

Keywords: Statistics • Propensity score • Matching • Weighting • Stratification

INTRODUCTION

Randomized trials are deemed to be the most scientifically rigorous study design to investigate the efficacy of treatment while minimizing systematic bias. In fact, subjects are randomly assigned to the treatment or control group, thus allowing an equal distribution between the 2 groups of measured and unmeasured confounders (variables that influences both the dependent variable and independent variable causing a spurious association, referred to as covariates in regression context) [1]. However, randomized trials can be difficult to conduct, and observation studies can provide important evidence. In observational studies, subjects in treatment and control groups likely differ for confounders and differences in outcomes can reflect differences in baseline conditions rather than a real treatment effect. Matching each subject in the treatment group with subjects in the control group with comparable baseline confounders is an intuitive way to minimize confounding in observational studies. However, matching simultaneously on few confounders is a very complex process and often results in a very limited number of similar matches. An alternative method is matching based on the propensity score (PS) [2]. The PS is the probability of a subject to receive a treatment T conditional on the set of confounders (X), and it is commonly estimated via logistic regression. The purpose of estimating the PS is to simplify the matching process by collapsing all confounders into a single value. Matching patients with a similar estimated PS creates approximate balance for all the confounders, and difference in outcomes within groups with a similar PS gives unbiased estimate of treatment effect [3, 4]. PS matching can circumvent few limitations of standard multivariable regression modelling [5] (Table 1), and it has increasingly appeared in cardiovascular researches [6]. This article provides a guidance in implementing PS-based methods to foster transparency and consistency and to facilitate interpretation on study findings.

PROPENSITY SCORE METHODS

Four different PS-based methods exist: (i) matching: matches 1 or more control cases with a PS that is (nearly) equal to the PS for each treatment case, (ii) stratification (subclassification): divides sample into strata based on rank-ordered PSs and comparisons between groups are performed within each stratum, (iii) weighting: weights cases by the inverse of the PS. Similar to the use of survey sampling, weights are used to ensure that samples are representative of specific populations and (iv) regression adjustment: this includes PSs as a covariate in a regression model used to estimate the treatment effect.

PS method should be primarily chosen based on the estimand of interest, which depends on the research question and the target population. The most common estimands are the 'average

†Presented at the 31st Annual Meeting of the European Association for Cardio-Thoracic Surgery, Vienna, Austria, 7–11 October 2017.

© The Author(s) 2018. Published by Oxford University Press on behalf of the European Association for Cardio-Thoracic Surgery. All rights reserved.

Table 1: Advantages of the PS matching over standard MV regression

Problem with MV regression	Comments	Advantages of the PS matching
Restricted number of confounders in the model	In the MV model, the number of confounders is limited by the number of events. A common rule-of-thumb is 1 covariate for every 8–10 events. This limits the application of the MV model, particularly in case of large number of confounders and relatively low number of events	For the calculation of the PS, the number of confounders used in the PS model is not limited by the number of outcome events. The collapsing of covariates into 1 score allows the investigator to include all potential confounders that otherwise may not have been possible to include and may improve statistical efficiency. Therefore, the use of the PS may be warranted when the number of confounders is large or the number of outcomes is small
Invalidity of the study due to confounding by indication	Patients with contraindications to the experimental treatment (or those with absolute indications) may have no comparable exposed subjects (or unexposed subjects) for valid estimation of relative or absolute differences in outcomes. These subjects are not usually recognized with conventional response modelling and might be influenced due to effect measure modification or model misspecification	Matching on the PS focuses directly on the indications for the experimental treatment. Graphical comparison of the PSs in exposed versus unexposed subjects can identify these areas of non-overlap that are otherwise difficult to describe in a multivariate setting with many factors influencing treatment decisions
Modelling assumption	The MV regression model relies on the modelling assumptions of linearity between covariates and the natural logarithm of the odds of the outcome	Matching by the PS eliminates the linearity assumption between the PS and outcomes
Model design not separated from outcome analysis	MV regression models adjust for confounders by modelling relationship between covariates and outcome and, therefore, model specification can be influenced by researcher expectation to prove the original hypothesis	PS matching estimates the treatment effect by modelling covariates and treatment assignment. PS matching mirrors a randomized experiment because the study design (PS model and matching) is separated from the outcome analysis. This protects against actual or suspected bias on the part of the researcher

MV: multivariable; PS: propensity score.

effect of the treatment on the treated' (ATT), which is the effect for those in the treatment group, and the 'average treatment effect' (ATE), which is the effect on all individuals (treatment and control). The ATE is of more interest if every treatment potentially might be offered to every subject, whereas the ATT is preferable when patient's characteristics are more likely to determine the treatment received. Matching can estimate only the ATT, weighting can estimate either effects based on how weights are defined, stratification can estimate either effects based on how strata are weighted and finally, covariate adjustment can estimate only marginal effect but neither the ATT nor the ATE. When estimating treatment effect on binary outcomes (odds ratio), matching results in estimates with less bias than stratification or covariate adjustment. Inverse probability of treatment weighting (IPTW) should be used for estimating risk differences particularly when the interest is in estimating the ATE [7]. When estimating treatment effect on time-to-event outcomes, matching and IPTW result in less biased estimates than stratification or covariate adjustment (Fig. 1) [8]. Based on the above considerations, we propose an algorithm for selecting the PS methods (Fig. 1). Currently, several programmes can perform the PS analysis but they are primarily written in R or consist of special macros in Stata or SPSS (Supplementary Material, Table S1). We recommend the use of the following R packages: MatchIt or non-random matching, non-random for stratification and twang for weighting.

Steps in propensity score-based analysis

The following are the basic steps for removing the effects of confounding from the treatment effect: (i) decide on confounders for which balance must be achieved, (ii) estimate the distance measure (e.g. the PS), (iii) condition on the distance measure (e.g. using matching, weighting or subclassification), (iv) assess balance on the covariates of interest, the PS-based analysis is an iterative process and alternative PS-based methods should be attempted until a well-balanced sample is attained and (v) estimate the treatment effect in the conditioned sample.

Selection of confounders

Confounders (X) used for the PS model must not be influenced by the treatment (T), and they should be measured (observed) before T is given. Possible explanation for the treatment assignment should be provided including physician preference, local policies or temporal change in practice. Pre-existing conditions in control units for whom a given treatment is not applicable are removed from the study population. A non-parsimonious selection of confounders is recommended to reduce residual bias [3, 4]. However, the inclusion of many confounders can reduce the number of good matches and, therefore, decreased the precision. A reasonable approach is to include both confounders related to outcome and treatment assignment if the sample size is large and to concentrate on variables believed to be strongly related to the outcome if the sample is small.

Propensity score calculation

The great majority of applications of the PS have used logistic regression to estimate the score [3, 4] with treatment assignment used as dependent variable and all selected confounders forced as covariates. Other approaches such as classification, regression trees and neural networks can also be considered. In building the

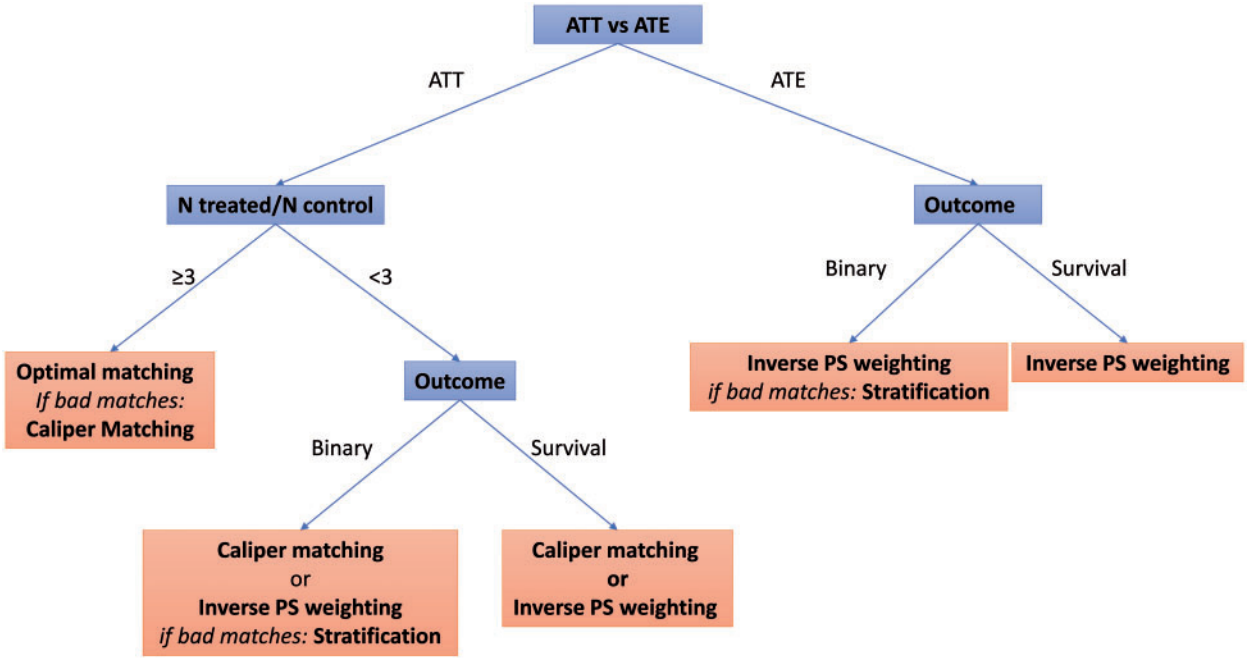

Figure 1: An algorithm to select the most appropriate PS method. ATE: average treatment effect; ATT: average effect of the treatment on the treated; PS: propensity score.

PS, consideration of interaction terms is recommended [9]. The predictive ability of model included should not represent a limitation in building the PS model. In fact, the PS model is not employed for inferential purposes but simply for creating a balancing score, and, therefore, the common practice of reporting the C-statistic as a measure of the adequacy of a PS is questionable [10]. A very high C-statistic can indicate non-overlap in the distribution of the PS between treated and untreated subjects and suggests an inability to make comparisons between treated and untreated subjects. Additionally, a high C-statistic cannot be taken as evidence that the PS included every important confounder.

Matching

Two commonly selected matching methods are the nearest neighbour matching and optimal matching [3, 4]. Nearest neighbour relies on a greedy algorithm which selects a treated participant at random and sequentially moves through the list of participants and matches the treated unit with the closest match from the comparison group. The optimal matching algorithm minimizes the overall distance across matched groups.

Several options exist to increase the quality of matches: matching with replacement and matching with caliper adjustment. In matching with replacement, a control participant could be paired multiple times if that person's PS provides the closest match to multiple intervention participants. Matching with replacement requires the standard errors to be estimated using more complex methods, e.g. sandwich estimators as data are no longer independent and results in precision loss. Caliper matching uses a prespecified distance within which matches are considered acceptable. If the best match is outside of the caliper distance, the matches are not included in the final set. The designated distance is usually a fraction of a standard deviation of the logit of the PS (e.g. 0.20 SD) [9]. Monte Carlo simulations have shown that when compared with other methods, caliper matching results in estimates with less bias when compared with optimal and nearest neighbour matching and shows the best performance when assessed using mean squared error; matching with replacement does not have superior performance when compared with caliper matching without replacement [11]. Finally, the number of comparison units selected for each treated unit should be >1 when there are few good matches in the control groups for each treated unit as a higher ration increases precision. However, if there is a limited number of comparison units, a ratio >1 can be selected as badly matched, thus leading to bias.

Stratification

Stratification subclassifies the individuals based on quantiles of the PSs [12]. The outcomes of the individuals are then compared within each of the strata, and a common estimator of the treatment effect is derived by combining the results over the 5 strata. A common practice is to divide the PS into 5 strata; this has been shown to eliminate 90% of the bias from measured confounders. Stratification approximates matching without running the risk of losing unmatched patients. Another advantage of the stratification technique is that it allows the calculation of both the ATE and ATT. Stratum-specific estimates of effect are weighted by the proportion of subjects who lie within that stratum. Thus, when the sample is stratified into n equal-size strata, stratum-specific weights of $1/n$ are commonly used when pooling the stratum-specific treatment effects, allowing one to estimate the ATE. The use of stratum-specific weights that are equal to that proportion of treated subjects that lie within each stratum allow one to estimate the ATT. A disadvantage of stratification is that it reduces biases less than other methods in particular to survival analysis [8]. Another disadvantage is the complexity of pooling the strata effects (e.g. the use of the Cochran–Mantel–Haenszel method).

IPTW

The PS can also be used as inverse weights in estimates of the ATE, known as IPTW [13]. The weight of each participant is calculated using 2 variables: T (indicator of the participant's treatment status being 0 if in the control arm and 1 if in the treatment arm) and PS of each participant. The weight (w) of the participant [wATE = T/PS + (1 - T)/(1 - PS)] is equal to the inverse of the probability of receiving the treatment the participant received. In this approach, the contributions of the study subjects are weighted by 1/PS for experimental patients and by 1/(1 - PS) for control patients. However, a different set of weights permit estimation of the ATE in the treated (ATT): wATT: T + PS(1-T)/(1-PS). Treated subjects receive a weight of 1. Thus, the treated sample is being used as the reference population to which the treated and control samples are being standardized. Moreover, for settings with more than 2 treatments, inverse propensity score weighting (IPSW) with the PS estimated via generalized boosted models can be implemented using those scores to estimate weights and causal effects. The advantages of using IPSW are that it retains all the patient data and reduces bias more than stratification and covariate adjustment [8].

Covariate adjustment

The PS can be used as a covariate in adjusting the treatment effect for baseline differences. Its advantage is that the PS itself can include many covariates along with interactions; this allows for the subsequent covariate regression model to be more parsimonious, including only the relevant covariates along with the propensity score variable. However, a formal assessment of balance between treatment groups is not possible. Moreover, it produces more biased estimates [8] and wrong assumptions about the functional relationship of the PS and outcome (linearity and proportional hazards) may then directly lead to biased estimates.

Balance check

The quality of the matches is based on the comparison of confounders distribution in the matched sample [14]. The use of hypothesis tests and P-values to compare balance is not appropriate because there are no inferences being made in relation to a population. They also conflate changes in balance with changes in statistical power. Standardized biases (also known as standardized mean difference) are recommended to assess the balance of covariates between the 2 groups. The standardized mean difference compares the difference in means in units of the pooled standard deviation [14]. A value higher than 0.10 (10% in case it is reported as percentage) is commonly considered index of residual imbalance. The lack of balance can indicate the need to add higher order or non-linear terms. Interaction terms should also be considered, in particular between the most unbalanced covariates. The analysis can also be restricted only to those subjects with the PS that overlap with other group (common support) [15]. Graphical diagnostics can be helpful for getting a quick assessment of the covariate balance in the presence of many covariates. The first step is to examine the distribution of the PS in the original and matched groups and the PS overlapping using mirrored histogram (Fig. 2A), while a Love plot of the standardized differences of means (Fig. 2B) gives us a quick overview of whether balance is adequate.

Estimation of treatment effect

Matched data should be analysed using procedures for matched analyses, such as paired t tests for continuous variable, while McNemar's test, conditional logit or mixed effect (matched pairs as

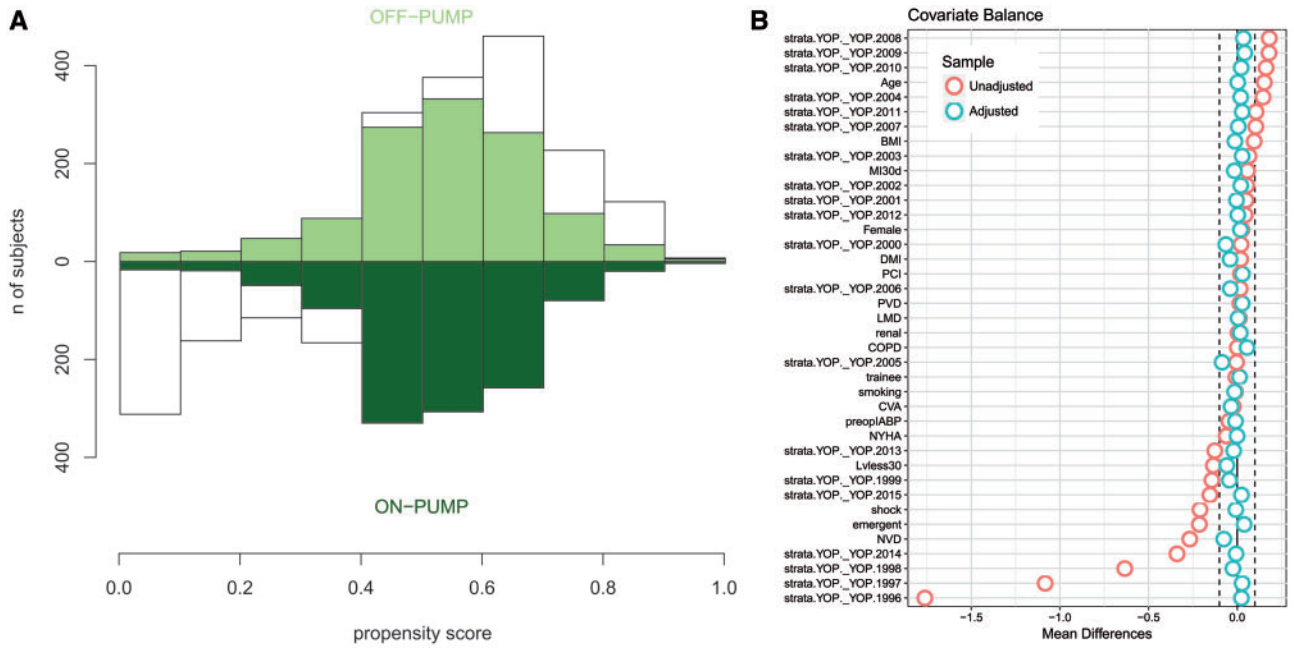

Figure 2: (**A**) Mirrored histogram showing the propensity score distribution and overlapping in unmatched (white) and matched (green) samples in the treatment (off-pump, top) and in the control groups (on-pump, bottom). (**B**) The Love plot showing changes in standardized mean difference before (red) and after (blue) matching. BMI: body mass index; COPD: chronic obstructive pulmonary disease; CVA: cerebrovascular accident; DMI: diabetes mellitus on insulin; IABP: intra-aortic balloon pump; LMD: left main disease; LV: left ventricular; MI: myocardial infarction; NVD: number of vessel diseased; NYHA: New York Heart Association functional class; PCI: percutaneous coronary intervention; PVD: peripheral vascular disease; YOP: year of procedure.

random effect) logistic regression can be used for binary outcomes. For time-to-event outcomes (survival), stratified log-rank test, stratified Cox model or mixed effect Cox model are required [3]. Data can also be analysed using a standard regression in the matched sample that includes a treatment indicator and the variables used in the PS model (double robust), where the regression adjustment is used to 'clean up' small residual covariate imbalance between the groups. Methods for paired samples provide better estimates [16], and non-paired methods can be presented as sensitivity analysis.

Propensity score and missing data

In the presence of missing data, multiple imputation can be used to create completed datasets from which the PS can be estimated. There are 2 proposed methods [17]: (i) averaging of the PSs after multiple imputation, followed by causal inference, or (ii) causal inference using each set of the PSs from the multiple imputations followed by averaging of the causal estimates. It is advisable to include the outcome in the imputation model.

REPORTING

Although the use of this analytical approach has increased significantly in clinical research, current reporting is often inadequate and ambiguous, and this results in problems with study reproducibility and interpretation. To improve consistency and reproducibility, a set of items to be reported has been recommended [18] (Supplementary Material, Table S2). These items should be then integrated with the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) categories for reporting observational studies [19].

EXAMPLE

In available randomized trials, off-pump failed to improve hospital mortality when compared with on-pump in coronary artery bypass grafting (CABG). However, patients enrolled in randomized studies are highly selected and at low risk. Whether off-pump can provide survival benefit in high-risk patients in the real-world practice remains unclear. Herein, we implemented the PS matching to compare hospital mortality after off-pump versus on-pump in high-risk patients (EuroSCORE ≥6) undergoing isolated first-time CABG using the Bristol Heart Institute (UK) database.

Study population

A total of 3424 patients with preoperative EuroSCORE ≥6 underwent isolated first-time CABG from 1996 to 2015 at a single

Table 2: Preoperative confounders distribution in the original and matched samples

	Off-pump (original)	On-pump (original)	SMD before matching	Off-pump (matched)	On-pump (matched)	SMD after matching
Set of confounders	1670	1754		1199	1199	
Age (years), mean (SD)	74 (7)	73 (7)	0.158	74 (7)	74 (7)	0.004
Female, *n* (%)	486 (29.1)	489 (27.9)	0.027	332 (27.7)	336 (28.0)	0.019
NYHA III/IV, *n* (%)	680 (40.7)	767 (43.7)	0.061	497 (41.5)	506 (42.2)	<0.001
MI within 30 days, *n* (%)	666 (39.9)	649 (37.0)	0.059	463 (38.6)	474 (39.5)	0.016
Prior PCI, *n* (%)	91 (5.4)	88 (5.0)	0.019	69 (5.8)	65 (5.4)	0.030
IDDM, *n* (%)	163 (9.8)	161 (9.2)	0.020	113 (9.4)	121 (10.1)	0.039
Smoking, *n* (%)	164 (9.8)	177 (10.1)	0.009	115 (9.6)	120 (10.0)	0.014
Creatine >200 mmol/l, *n* (%)	99 (5.9)	102 (5.8)	0.005	70 (5.8)	65 (5.4)	0.018
COPD, *n* (%)	254 (15.2)	266 (15.2)	0.001	195 (16.3)	182 (15.2)	0.056
CVA, *n* (%)	111 (6.6)	126 (7.2)	0.021	77 (6.4)	90 (7.5)	0.033
PVD, *n* (%)	440 (26.3)	450 (25.7)	0.016	334 (27.9)	311 (25.9)	0.029
NVD, *n* (%)			0.299			0.092
1	101 (6.0)	44 (2.5)		51 (4.3)	30 (2.5)	
2	416 (24.9)	282 (16.1)		261 (21.8)	197 (16.4)	
3	1153 (69.0)	1428 (81.4)		887 (74.0)	972 (81.1)	
LMD, *n* (%)	529 (31.7)	545 (31.1)	0.013	392 (32.7)	399 (33.3)	0.005
LVEF <30%, *n* (%)	237 (14.2)	331 (18.9)	0.126	179 (14.9)	203 (16.9)	0.055
Cardiogenic shock, *n* (%)	23 (1.4)	67 (3.8)	0.154	20 (1.7)	28 (2.3)	0.006
Preoperative IABP, *n* (%)	59 (3.5)	77 (4.4)	0.044	49 (4.1)	54 (4.5)	0.009
Emergency, *n* (%)	64 (3.8)	139 (7.9)	0.175	53 (4.4)	59 (4.9)	0.037
BMI, mean (SD)	27 (4)	27 (5)	0.092	27 (4)	27 (5)	0.012
YOP, mean (SD)	2006 (4)	2005 (6)	0.328	2006 (5)	2006 (5)	0.051
Performed by trainee, *n* (%)	411 (24.6)	437 (24.9)	0.007	294 (24.5)	293 (24.4)	0.014
Estimation of treatment effect			*P*-value			*P*-value
In-hospital mortality, *n* (%)	54 (3.2)	75 (4.3)	0.11[a]	36 (3.0)	55 (4.6)	0.06[a]
Fully adjusted logistic, OR (95% CI)	0.72 (0.48–1.07)		0.10			
Conditional logit, OR (95% CI)				0.66 (0.43–0.99)		0.04
Doubly robust logistic, OR (95% CI)				0.61 (0.39–0.95)		0.03
Mixed effect logistic, OR (95% CI)				0.64 (0.42–0.99)		0.04

[a]The χ^2 test.

BMI: body mass index; CI: confidence interval; COPD: chronic obstructive pulmonary disease; CVA: cerebrovascular accident; IABP intra-aortic balloon pump; IDDM: insulin dependent diabetes mellitus; LMD: left main disease; LVEF: left ventricular ejection fraction; MI: myocardial infarction; NYHA: New York Heart Association; NVD: number of vessels diseased; OR: odds ratio; PCI: percutaneous coronary intervention; PVD: peripheral vascular disease; SMD: standardized mean difference; YOP: year of procedure.

institution (Bristol). Off-pump was used in 1670 patients, and on-pump was used in 1754 patients.

Propensity score calculation

We selected 20 confounders for hospital mortality to compare between the 2 groups (Table 2). We estimated the PS for off-pump by logistic regression using linear terms, no interaction and stratification by year of surgery.

Matching process

To create matched pairs, we used caliper matching (0.20 SD of logit of the PS) without replacement; matching identified 1199 pairs. Table 2 shows baseline confounders distribution in the original and matched samples.

Balance check

After matching, the 2 groups were comparable for all confounders (standardized mean difference <0.10). Visual inspection of mirrored histogram showing an adequate PS overlapping (Fig. 2A), and the Love plot confirmed no significant imbalance for any of covariates included (Fig. 2B).

Estimation of treatment effect

In the original sample, off-pump was not significantly associated with reduced mortality, also in a fully adjusted multivariable logistic regression. However, in the matched sample conditional logit, at doubly robust and mixed effect logistic regression, off-pump was found to be significantly associated with reduced in-hospital mortality (Table 2) (R codes are provided in Supplementary Material).

LIMITATIONS

The main limitation of the PS methods is their inability to control for unmeasured confounding. A drawback of matching is an often substantially reduced sample size because for some patients, matches may not be found. This may significantly affect the study's final conclusions, which then apply only to the selected subset of patients that could be matched. The PS tends to work better in larger samples. Significant imbalances of certain covariates may be unavoidable despite a well-constructed PS secondary to a small number of observations. As randomized studies, the PS methods generate an average effect and therefore, they do not address what treatment may be right for a given patient.

CONCLUSIONS

PS methods reduce a set of confounders into a single, intuitive variable which optimizes matching and makes it possible to statistically adjust when the ratio of events to confounders is low. They also may reveal cases in which the patient populations are too divergent to make meaningful comparisons. Proposed reporting guidelines should be followed to foster transparency and consistency and to facilitate interpretation on study findings.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *EJCTS* online.

Conflict of interest: none declared.

REFERENCES

[1] Pocock SJ, Elbourne DR. Randomized trials or observational tribulations? N Engl J Med 2000;342:1907–9.
[2] Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52:249–64.
[3] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav Res 2011;46:399–424.
[4] Stuart EA. Matching methods for causal inference: a review and a look forward. Stat Sci 2010;25:1–21.
[5] 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.
[6] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Stat Med 2008;27:2037–49.
[7] Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. Stat Med 2007;26:3078–94.
[8] Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. Stat Med 2013;32:2837–49.
[9] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163:1149–56.
[10] Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. Pharmacoepidemiol Drug Saf 2011;20:317–20.
[11] Austin PC. A comparison of 12 algorithms for matching on the propensity score. Stat Med 2014;33:1057–69.
[12] Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 1968;24:295–313.
[13] McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. Stat Med 2013;32:3388–414.
[14] 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.
[15] Blackstone EH. Comparing apples and oranges. J Thorac Cardiovasc Surg 2002;123:8–15.
[16] Austin PC. Comparing paired vs non-paired statistical methods of analyses when making inferences about absolute risk reductions in propensity-score matched samples. Stat Med 2011;30:1292–301.
[17] Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. Stat Med 2009;28:1402–14.
[18] Yao X, Wang X, Speicher PJ, Hwang ES, Cheng P, Harpole DH *et al.* Reporting and guidelines in propensity score analysis: a systematic review of cancer and cancer surgical studies. J Natl Cancer Inst 2017;109.
[19] von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE Initiative. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ 2007;335:806–8.