EuroIntervention 2014;10:761-767

Tools & Techniques - Statistics: Propensity score techniques

Bruno R. da Costa[1,2,3], PhD; Brigitta Gahl[4], MSc; Peter Jüni[1,2]*, MD, FESC

1. Institute of Social and Preventive Medicine, University of Bern, Bern, Switzerland; 2. CTU Bern, Department of Clinical Research, University of Bern, Bern, Switzerland; 3. Department of Physical Therapy, Nicole Wertheim College of Nursing & Health Science, Florida International University, Miami, FL, USA; 4. Department of Cardiovascular Surgery, Bern University Hospital (Inselspital) and University of Bern, Bern, Switzerland

Abstract

Propensity score (PS) techniques are useful if the number of potential confounding pretreatment variables is large and the number of analysed outcome events is rather small so that conventional multivariable adjustment is hardly feasible. Only pretreatment characteristics should be chosen to derive PS, and only when they are probably associated with outcome. A careful visual inspection of PS will help to identify areas of no or minimal overlap, which suggests residual confounding, and trimming of the data according to the distribution of PS will help to minimise residual confounding. Standardised differences in pretreatment characteristics provide a useful check of the success of the PS technique employed. As with conventional multivariable adjustment, PS techniques cannot account for confounding variables that are not or are only imperfectly measured, and no PS technique is a substitute for an adequately designed randomised trial.

Introduction

Well-designed large-scale randomised clinical trials (RCTs) are the most reliable approach for investigating the causal effect between an intervention and clinical outcomes. Adequately concealed randomisation[1] of several hundred or even several thousand patients ensures that patient groups are almost identical regarding any confounding factors, measured or unmeasured, known or unknown at the time of randomisation. Confounding factors are variables that could distort our interpretation of the presence, direction and magnitude of a true causal effect between interventions and clinical outcomes[2]. The feasibility of RCTs may be called into question by ethical or economic constraints. The generalisability of results from RCTs may also be limited since the patient population is often highly selected due to restrictive inclusion criteria. In addition, interventions may be more carefully implemented in RCTs, with a high compliance of carers and patients to a standardised protocol, and may not be representative of what is actually implemented in routine clinical settings.

Observational studies are therefore often conducted, either to explore further how results of RCTs translate to routine clinical settings or to obtain information about the effect of an intervention in the absence of RCTs. In these studies, the allocation of patients to treatment groups is based on referral patterns and the clinical reasoning of the therapist. Because the choice of an intervention is probably related to the risk factor of patients, results of observational studies may be confounded. This type of confounding is known as "confounding by indication"[3].

**Corresponding author: Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, 3012 Bern, Switzerland. E-mail: juni@ispm.unibe.ch*

DOI: 10.4244/EIJV10I6A130

Different statistical methods can be used to minimise the influence of confounding by indication. Arguably the most commonly used method, a multivariable regression model, adjusts for patient characteristics (i.e., covariates) that are likely to confound the estimation of the treatment effect. When too many covariates are used for adjustment or the number of outcome events is too small, the validity of multivariable models may be compromised. A rule of thumb suggests that there should be at least 10 outcome events for every covariate included in a multivariable model to prevent overfitting of the model to the data[4,5], which results in spurious associations and/or instability of the model.

Propensity score modelling

Propensity score (PS) models may be used to control for confounding by indication even when the number of covariates is large or outcome events are rare. Unlike conventional multivariable adjustments, which model outcomes (i.e., they are based on estimating the association of baseline variables with clinical outcome), PS are based on modelling treatment selection (i.e., they are based on estimating the causal association of pre-treatment variables with interventions). In the typical case there will be considerably more patients who received a given intervention than patients who experienced a given outcome; therefore, the risk of overfitting is smaller for this approach than for conventional multivariable adjustment, which is one of the three major advantages of PS techniques.

A PS quantifies the probability of a patient from zero to one to receive the experimental intervention on which the PS is modelled, given all the pretreatment characteristics included in the model that were used for selecting the treatment a patient received. For instance, in an observational study comparing transcatheter aortic valve implantation (TAVI) with surgical aortic valve replacement (SAVR) in patients with severe aortic stenosis published in 2009 in EuroIntervention, elderly patients and more severely diseased patients were more likely to receive the experimental TAVI than the control SAVR **(Table 1)**[6]. The PS model estimates the associations with all relevant variables in a multivariable manner.

We feel the need to emphasise that the model should include all pretreatment variables that are likely to be associated with the clinical outcome of interest, irrespective of whether they are associated with treatment selection. So, in the example of **Table 1**, all variables should be used for the PS model, including diabetes mellitus, for example, which does not appear to be associated with the type of intervention (p=0.76), but is certainly associated with the clinical outcome of interest, in this instance overall mortality. Conversely, a variable that is associated with treatment selection, but not with outcome (i.e., an instrumental variable), should not be

Table 1. Crude and adjusted between-group comparisons of pretreatment characteristics.

		Aortic valve replacement						Cardioplegic solution					
		Experimental	Control	Crude		IPTW		Experimental	Control	Crude		IPTW	
				Diff	*p*-value	Diff	*p*-value			Diff	*p*-value	Diff	*p*-value
Age, yrs (SD)		82.8 (5.5)	69.9 (11.4)	1.44	<0.001	–0.43	0.19	64.6 (13.6)	67.3 (13.0)	–0.21	0.35	–0.18	0.13
Female, n (%)		64 (56.1)	408 (41.5)	0.32	0.001	–0.22	0.35	253 (35.0)	21 (25.3)	0.21	0.087	0.05	0.71
Logistic EuroSCORE, % (SD)		20.1 (13.4)	9.1 (10.2)	0.93	<0.001	0.00	0.98	8.9 (10.8)	10.2 (11.9)	–0.11	0.007	0.11	0.27
NYHA Class, n (%)	I to II	15 (13.2)	550 (55.6)	–0.97	<0.001	–0.38	0.089	499 (69.1)	62 (74.7)	–0.12	0.32	–0.07	0.61
	III	78 (68.4)	356 (35.3)	0.70	<0.001	0.07	0.79	191 (26.5)	17 (20.5)	0.13	0.26	0.03	0.80
	IV	21 (18.4)	102 (10.1)	0.24	0.008	0.33	0.21	32 (4.4)	4 (4.8)	–0.02	0.89	0.08	0.50
Diabetes mellitus, n (%)		26 (22.8)	243 (24.1)	–0.03	0.76	–0.22	0.13	102 (14.1)	11 (13.3)	0.02	1.00	0.02	0.87
Hypertension, n (%)		72 (63.1)	631 (62.6)	0.01	0.91	–0.32	0.22	462 (64.0)	50 (60.2)	0.08	0.47	0.14	0.29
Coronary artery disease, n (%)		64 (56.1)	512 (50.8)	0.11	0.28	–0.15	0.53	123 (17.0)	23 (27.7)	–0.26	0.023	–0.03	0.82
Previous coronary bypass surgery, n (%)		28 (24.6)	42 (4.2)	0.61	<0.001	–0.06	0.51	26 (3.6)	5 (6.0)	–0.11	0.36	0.02	0.81
Left ventricular ejection fraction, n (%)	>50%	67 (58.8)	834 (82.7)	–0.55	<0.001	0.29	0.083	617 (85.5)	66 (79.5)	0.16	0.16	–0.02	0.84
	30–50%	40 (35.1)	135 (13.4)	0.52	<0.001	–0.22	0.17	88 (12.2)	14 (16.9)	–0.13	0.23	–0.02	0.88
	<30%	7 (6.1)	39 (4.9)	0.10	0.25	–0.18	0.18	17 (2.4)	3 (3.6)	–0.07	0.50	0.11	0.42
Atrial fibrillation, n (%)		22 (19.3)	90 (8.9)	0.30	<0.001	–0.20	0.11	60 (8.3)	10 (12.0)	–0.12	0.30	–0.05	0.71
Cerebrovascular disease, n (%)		20 (17.5)	50 (5.0)	0.41	<0.001	–0.17	0.060	47 (6.5)	4 (4.8)	0.08	0.81	0.03	0.86
Peripheral vascular disease, n (%)		21 (18.4)	47 (4.7)	0.44	<0.001	0.09	0.74	30 (4.2)	3 (3.6)	0.03	1.00	0.17	0.13
COPD, n (%)		24 (21.1)	134 (13.3)	0.21	0.024	0.05	0.85	84 (11.6)	10 (12.0)	–0.01	0.86	0.06	0.63
Pulmonary hypertension, n (%)		34 (29.8)	86 (8.5)	0.56	<0.001	0.09	0.74	35 (4.8)	3 (3.6)	0.06	0.79	0.09	0.40
Creatinine above 200 µmol/L, n (%)		7 (6.1)	31 (3.1)	0.15	0.086	–0.02	0.90	11 (1.5)	3 (3.6)	–0.13	0.17	–0.02	0.85
MI within 90 days of procedure, n (%)		4 (3.5)	34 (3.4)	0.01	0.94	–0.21	0.005	6 (0.8)	2 (2.4)	–0.12	0.20	–0.08	0.49

COPD: chronic obstructive pulmonary disease; Diff: standardised differences between patients in the experimental and control groups; IPTW: inverse probability treatment weighting after trimming of 2.5% of tails as described in the text; MI: myocardial infarction; NYHA: New York Heart Association

included in the model[7,8]. This may seem counterintuitive at first, but makes sense when the aim of PS models is considered: to minimise the effect of confounding on estimating the causal relationship between an intervention and an outcome. If the choice of TAVI, for some esoteric reason, were associated with the signs of the zodiac, this variable should therefore not be included in the model since we do not really have a scientific rationale to suggest that signs of the zodiac are causally related to overall mortality. If it is unclear for a variable that is associated with treatment selection whether it is also associated with outcome or whether it is merely an instrumental variable, we will include it in the PS model[7]. Technically, a multivariable logistic or probit regression[9] with type of treatment as dependent variable and all the pretreatment variables discussed above as independent variables will be used to generate PS. Variables collected at or after treatment initiation absolutely must not be used[7,10]. These variables need to be considered a result of the treatment decision and can by definition not be causally related to the treatment decision.

Propensity score models eliminate confounding when two conditions are met[11]. The first condition is that all relevant confounding variables are measured and used to calculate the patients' PS to receive the experimental treatment. Once all confounding variables are controlled for, the treatment decision itself (before the actual treatment takes place) should no longer predict the outcome. The second condition is that all patients included in the study have a non-null probability to receive either the experimental or control treatment. This means that, at the time of inclusion in the study, all patients had to be potential candidates to receive either of the two compared interventions. This is more or less analogous to patients included in an RCT: only patients who qualify for either of the two compared interventions will be included in the trial, but not those who only qualify for one of the two interventions, but not the other. Only when both conditions are fully met will PS models provide unbiased estimates of the treatment effect.

Visual inspection

Once PS are generated, their distribution can and should be visually inspected. This is the second major advantage of PS techniques. We use smoothed Kernel probability density estimates[12,13] of the PS of the two groups to do so, but simple histograms will also do the job. **Figure 1A** shows the distribution of PS in the above-mentioned comparison of TAVI with SAVR in patients with aortic stenosis[6]. The distribution of PS in experimental patients receiving TAVI is almost uniformly distributed between zero and one. Conversely, in control patients receiving SAVR, the distribution is completely skewed with a thin long tail towards the right and the Kernel probability density falling to values near null at PS greater than 0.2. So, even if there are few patients with PS between 0.2 and 0.85, which corresponds to probabilities of 20% and 85% of SAVR patients to receive TAVI given their pretreatment characteristics, the thin long tail spread along the null between 0.2 and 0.85 indicates that we cannot be certain that the actual probability of these patients to receive TAVI is anything else but zero.

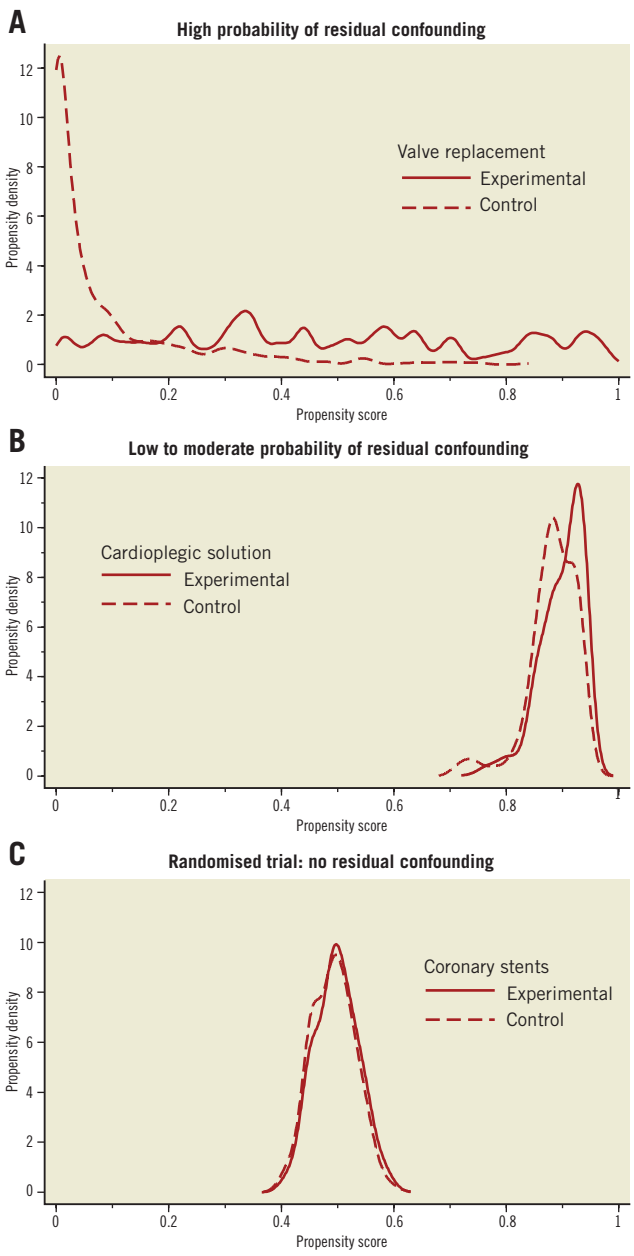

Figure 1. *Crude distribution of propensity scores for three clinical examples. A) High probability of residual confounding. B) Low to moderate probability of residual confounding. C) Randomised trial: no residual confounding.*

An 81-year-old SAVR patient with a PS of 0.5, for example, may look on paper as if she has had a 50% chance to receive TAVI, but at clinical inspection one would immediately grasp that she appears biologically much younger, has only minor severity of her comorbid conditions, is participating fully in life and shows no signs of frailty, which made her an extremely likely candidate to receive SAVR despite her age and rendered her practically ineligible for receiving TAVI[6]. In our group, we consider this issue when the Kernel probability density estimate permanently drops below 0.5, but we are not aware of a generally agreed cut-off. In any case, long thin tails of PS that spread near the null point indicate a high

probability of residual confounding (here biological rather than chronological age, frailty, severity of comorbid conditions), which may severely distort results, as indeed was the case here[6]: conventional analyses with multivariable adjustment indicated a threefold increase in the odds of death with TAVI as compared with SAVR (OR 3.05, 95% CI: 1.09 to 8.51) at 30 days[6], when the only randomised trial available at the time indicated a trend towards a lower overall mortality at 30 days with TAVI as compared with SAVR (OR 0.53, 95% CI: 0.24 to 1.14)[14].

Figure 1B shows the distribution of PS in an observational comparison of two cardioplegic solutions used in patients undergoing SAVR at our centre (unpublished). Compared with **Figure 1A**, there is a much higher overlap of the distributions of PS between the experimental and control groups, tails are short, Kernel density estimates hardly fall below 0.5, and there is considerably less asymmetry of the two distributions than observed in **Figure 1A** for SAVR patients. While we acknowledge that the data are observational in nature and therefore distortion of results through residual confounding is always possible, we deem residual confounding considerably less likely here than in the first example. For further comparison, we show the optimal distribution of PS, as derived from a randomised trial **(Figure 1C)**. Here, we calculated PS for being treated with an experimental coronary stent as compared to a control stent in patients with acute myocardial infarction in a trial with a 1:1 randomisation[15]. PS of both groups scatter around 0.5, corresponding to a 50% chance to receive an experimental stent, with a narrow, symmetrical distribution and near superimposition of scores of the two groups.

Propensity score trimming

A logical consequence from the reasoning above is to drop patients with PS in areas suspected of residual confounding by trimming the tails at both ends of the distribution, which is the third major advantage of PS models. **Figure 2** shows this for the observational comparison of cardioplegic solutions. Here, the larger value of the 2.5th percentile of the PS was found at 0.797 in the experimental group, while the smaller value of the 97.5th percentile was found at 0.937 in the control group. Observations in both groups that have PS beyond these two cut-offs are dropped[16]. We do not advocate trimming based on the common support assumption, even though we previously used it[6], since the cut-off will depend heavily on the parameters used for the Kernel probability density function[12,13]. After trimming, the study population becomes more selected and, again analogous to an RCT, results after trimming may not be generalisable to all patients seen in routine clinical practice.

Use of propensity scores for adjustment

After PS are estimated, they can be incorporated in regression models used to estimate the treatment effect as specified in **Table 2**. We mostly use matching on PS or inverse probability of treatment weighting (IPTW). For matching, we advocate 1:1 or 1:n nearest neighbour matching within a calliper of a fifth of the pooled standard deviation of the PS observed in the two groups after trimming.

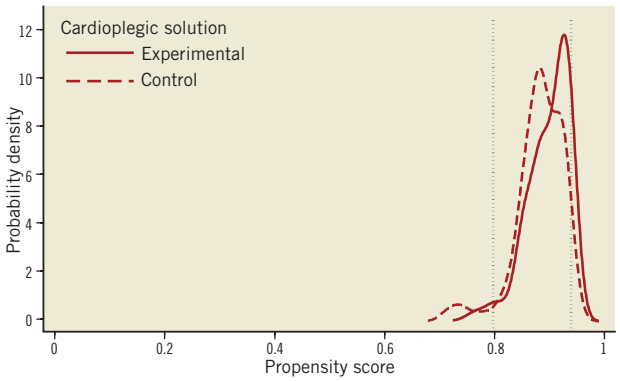

Figure 2. *Trimming of propensity score distributions in both groups with cut-offs below the 2.5 percentile in the experimental group and above 97.5 percentile in the control group for the comparison of cardioplegic solutions.*

If the pooled standard deviation of the PS is 0.15, for example, we use a calliper width for matching of 0.03 on the PS scale, which ranges from zero to one. Using this approach for the logit of the PS rather than the PS may even be preferable[17]. Callipers should not be chosen too narrow, since this will lead to an unnecessary loss of observations. Matching should be accounted for in the analysis, such as the use of a Cox proportional hazards model for time to event data stratified by matched pairs[18], or conditional logistic regression for binary data.

Figure 3 presents the distribution of PS after 1:1 matching. **Figure 3A** shows a multimodal distribution for the TAVI versus SAVR comparison, which suggests instability of the approach even though matching worked perfectly and the distribution of PS scores

Table 2. How propensity scores can be implemented to minimise confounding by indication.

Use of propensity scores (PS) as continuous covariate	Two explanatory variables are entered into the model: type of treatment and PS as a linear term.
"Doubly robust" multivariable analysis	Multiple explanatory variables are entered in the model: type of treatment, PS as a linear term, and all pretreatment characteristics.
Univariable analysis with matching on PS	Use of PS for matching patients in the experimental group to patients in the control group who have similar pretreatment characteristics.
Univariable analysis stratified by PS	Patients are stratified according to their PS. At least five strata should be created. Treatment effects are estimated within stratum and then combined across strata.
Univariable analysis weighted with inverse of the probability of treatment weights (IPTW)	Use of PS to generate IPTW, which are then implemented in the analysis.
"Doubly robust" multivariable analysis weighted with IPTW	IPTW weighted analysis adjusted for all pretreatment characteristics.

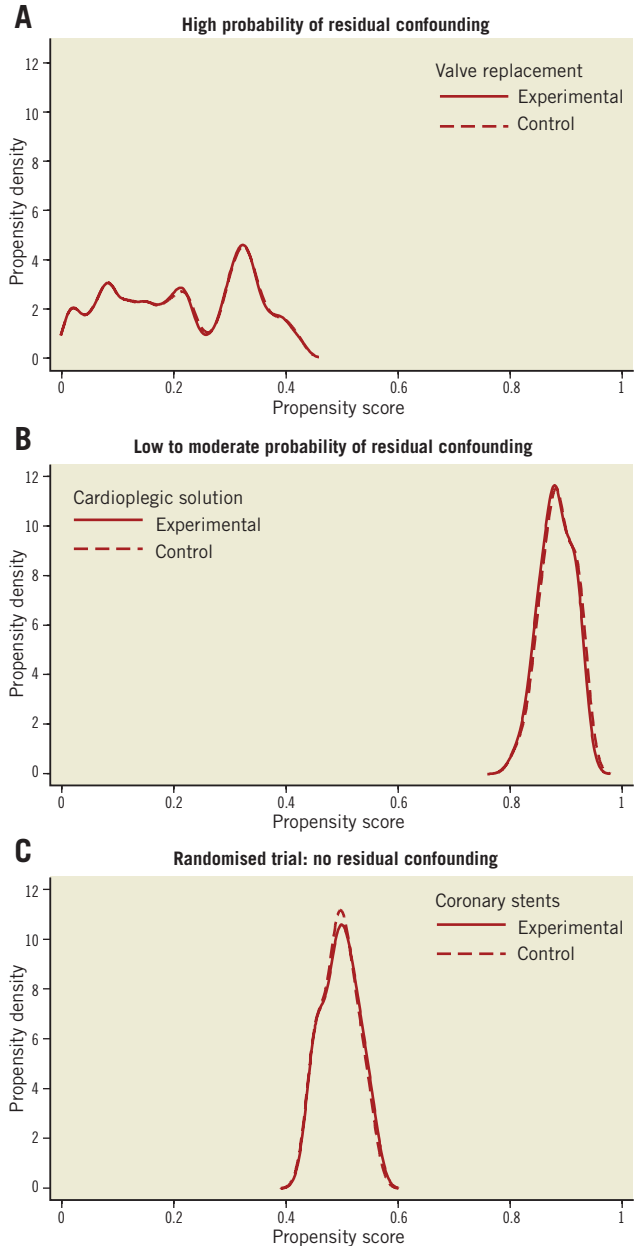

Figure 3. *Distribution of propensity scores after trimming and 1:1 matching. A) High probability of residual confounding. B) Low to moderate probability of residual confounding. C) Randomised trial: no residual confounding.*

from experimental and control groups are nearly superimposed. This suspicion is confirmed when using 1:1 matching for the analysis of 30-day mortality, since we still find completely biased results: an OR of 5.00 (95% CI: 0.58 to 42.8) when the best estimate from an RCT is 0.53[14]. Conversely, the distribution of the PS of the comparison of cardioplegic solutions after 1:1 PS matching **(Figure 3B)** looks much like the distribution found in the randomised trial **(Figure 3C)**, which suggests that we can be rather confident when interpreting results of the 1:1 matched comparison of cardioplegic solutions.

A disadvantage of PS matching, particularly when using a ratio of 1:1, is the loss of observations. Therefore, we perform weighted univariable analyses using inverse probability of treatment weights (IPTW) more and more often[19]. IPTW analyses use the inverse of the propensity score as weights in patients who received the experimental intervention and the inverse of one minus the propensity score in patients who received the control intervention (assuming that PS is modelled on the experimental intervention). Patients who had a low probability of receiving the experimental intervention but received it anyway will be upweighted in the analysis, whereas patients with a high probability of receiving the experimental intervention who actually received it will be downweighted in the analysis. An alternative to IPTW, which also avoids the loss of observations, is stratification of the analysis by PS. For this purpose, PS should be stratified at least in quintiles[20]. If the number of observations is large, even smaller strata can be used. For an extended description of the different methods, see Heinze and Jüni[21]. Doubly robust methods essentially combine multivariable adjustment with PS methods. Two simple approaches are also specified in **Table 2**. Our group has only little experience with these approaches[6], and we are unaware of clear evidence to suggest that these methods are superior to PS matching and IPTW.

Comparison of pretreatment characteristics

Once the PS model is implemented, its success in minimising imbalances in pretreatment characteristics needs to be checked. For this purpose, standardised differences will be calculated, dividing the difference in arithmetic means or the difference in proportions by the respective pooled standard deviation. Differences are expressed as standard deviation units and interpreted as effect sizes, as classically described by Cohen[22]: ±0.20 standard deviations difference represent a small biological difference, ±0.50 a moderate and ±0.80 a large difference. If a PS technique is successful, one would expect standardised differences of ±0.10 standard deviation units or less, which is considered an irrelevant difference.

Table 1 presents a comparison of pretreatment characteristics of patients included in the comparison of TAVI versus SAVR[6], which was characterised by the grotesquely skewed distribution of PS in SAVR patients and minimal or no overlap of PS with the majority of TAVI patients **(Figure 1)**. The crude analysis of pretreatment characteristics of all patients shows dramatic differences, with four standardised differences beyond ±0.80, five beyond ±0.50, and another five beyond ±0.20. Unsurprisingly, 14 differences are statistically significant at $p<0.05$. The IPTW model used after trimming for calculating adjusted standardised differences improves the situation somewhat, but nine standardised differences are still beyond ±0.20, even though only one is statistically significant, which suggests residual baseline imbalances. Of these, six change sign, which additionally suggests instability of the approach and untrustworthy results. This time, the OR of 30-day mortality is 1.87 (95% CI: 0.50 to 6.99), which appears less biased than results from the PS matched analysis above, but still at odds with the randomised trial[14]. Even though the previously used approach towards

trimming did not differ much from the approach discussed here, results of the corresponding IPTW analysis were completely different (OR 0.35, 95% CI: 0.04 to 2.72)[6].

Table 1 presents a comparison of pretreatment characteristics of patients included in the comparison of cardioplegic solution. Here, crude standardised differences are considerably less pronounced than in the previous example, and IPTW analysis after trimming yields minimal standardised differences, with none beyond ±0.20 and only one beyond ±0.10. Together with **Figure 3B**, which shows a distribution of PS much like that from an RCT, this suggests that we can be reasonably confident of the results yielded by PS techniques in this example.

Conclusion

In conclusion, PS techniques are useful if the number of potential confounding pretreatment variables is large and the number of analysed outcome events is rather small so that conventional multivariable adjustment is hardly feasible. The steps to take when performing a PS analysis are specified in **Table 3**. Only pretreatment characteristics should be chosen for the PS model, and only when they are probably associated with outcome. A careful visual inspection of PS will help to identify areas of no or minimal overlap, which suggests residual confounding, and trimming of the data according to the distribution of PS will help to minimise residual confounding. Standardised differences in pretreatment characteristics provide a useful check of the success of the PS technique. As with conventional multivariable adjustment, PS techniques cannot account for confounding variables that are not or are only imperfectly measured, and no PS technique is a substitute for an adequately designed randomised trial.

Table 3. Steps to take when conducting a propensity score analysis.

1. Select variables that will be used to calculate PS. It is paramount that variables are pretreatment characteristics only and probably associated with the outcome.
2. Calculate PS for all patients in the data set.
3. Plot Kernel probability density estimates of PS per group and inspect.
4. Trim the tails at both ends of the distribution of PS to drop patients in areas suspected of residual confounding.
5. Calculate standardised differences expressed in standard deviation units of patients' pretreatment characteristics with the same approach as was used to incorporate PS in the outcome analysis.
6. Use PS in outcome model to estimate adjusted treatment effect estimates, preferably with inverse probability of treatment weighting or PS matching.

Conflict of interest statement

CTU Bern, which is part of the University of Bern, has a staff policy of not accepting honoraria or consultancy fees. However, CTU Bern is involved in design, conduct, or analysis of clinical studies funded by Abbott Vascular, Ablynx, Amgen, AstraZeneca, Biosensors, Biotronik, Boehringer Ingelheim, Eisai, Eli Lilly, Exelixis, Geron, Gilead Sciences, Nestlé, Novartis, Novo Nordisc, Padma, Roche, Schering-Plough, St. Jude Medical, and Swiss Cardio Technologies. P. Jüni is an unpaid steering committee or statistical executive committee member of trials funded by Abbott Vascular, Biosensors, Medtronic and Johnson & Johnson. The other authors have no conflicts of interest to declare.

References

1. Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ.* 2001;323:42-6.
2. Smith GD, Ebrahim S. Data dredging, bias, or confounding. *BMJ.* 2002;325:1437-8.
3. Walker AM. Confounding by indication. *Epidemiology.* 1996;7:335-6.
4. Harrel F. Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis. New York, NY, USA: Springer Inc; 2001.
5. van Domburg R, Hoeks S, Kardys I, Lenzen M, Boersma E. Tools and techniques--statistics: how many variables are allowed in the logistic and Cox regression models? *EuroIntervention.* 2014;9:1472-3.
6. Piazza N, van Gameren M, Jüni P, Wenaweser P, Carrel T, Onuma Y, Gahl B, Hellige G, Otten A, Kappetein AP, Takkenberg JJ, van Domburg R, de Jaegere P, Serruys PW, Windecker S. A comparison of patient characteristics and 30-day mortality outcomes after transcatheter aortic valve implantation and surgical aortic valve replacement for the treatment of aortic stenosis: a two-centre study. *EuroIntervention.* 2009;5:580-8.
7. Stürmer T, Wyss R, Glynn RJ, Brookhart MA. Propensity scores for confounder adjustment when assessing the effects of medical interventions using nonexperimental study designs. *J Intern Med.* 2014;275:570-80.
8. Myers JA, Rassen JA, Gagne JJ, Huybrechts KF, Schneeweiss S, Rothman KJ, Joffe MM, Glynn RJ. Effects of adjusting for instrumental variables on bias and precision of effect estimates. *Am J Epidemiol.* 2011;174:1213-22.
9. Bliss CI. The method of probits. *Science.* 1934;79:38-9.
10. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health.* 2000;21:121-45.
11. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika.* 1983;70:41-55.
12. Cao R, Cuevas A, Gonzalez-Manteiga W. A comparative study of several smoothing methods in density estimation. *Computational Statistics and Data Analysis.* 1994;17:153-76.
13. Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*. 1996;91:401-7.
14. Smith CR, Leon MB, Mack MJ, Miller DC, Moses JW, Svensson LG, Tuzcu EM, Webb JG, Fontana GP, Makkar RR,

Williams M, Dewey T, Kapadia S, Babaliaros V, Thourani VH, Corso P, Pichard AD, Bavaria JE, Herrmann HC, Akin JJ, Anderson WN, Wang D, Pocock SJ; PARTNER Trial Investigators. Transcatheter versus surgical aortic-valve replacement in high-risk patients. *N Engl J Med.* 2011;364:2187-98.

15. Räber L, Kelbæk H, Ostojic M, Baumbach A, Heg D, Tüller D, von Birgelen C, Roffi M, Moschovitis A, Khattab AA, Wenaweser P, Bonvini R, Pedrazzini G, Kornowski R, Weber K, Trelle S, Lüscher TF, Taniwaki M, Matter CM, Meier B, Jüni P, Windecker S; COMFORTABLE AMI Trial Investigators. Effect of biolimus-eluting stents with biodegradable polymer vs bare-metal stents on cardiovascular events among patients with acute myocardial infarction: the COMFORTABLE AMI randomized trial. *JAMA.* 2012;308:777-87.

16. Stürmer T, Rothman KJ, Avorn J, Glynn RJ. Treatment effects in the presence of unmeasured confounding: dealing with observations in the tails of the propensity score distribution--a simulation study. *Am J Epidemiol.* 2010;172:843-54.

17. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res.* 2011;46:399-424.

18. 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.

19. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11:550-60.

20. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics.* 1968;24:295-313.

21. Heinze G, Jüni P. An overview of the objectives of and the approaches to propensity score analyses. *Eur Heart J.* 2011;32:1704-8.

22. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates; 1988.