European Heart Journal (2011) **32**, 1704–1708
doi:10.1093/eurheartj/ehr031

STATISTICAL REVIEW

An overview of the objectives of and the approaches to propensity score analyses

Georg Heinze[1]* and Peter Jüni[2,3]

[1]Section for Clinical Biometrics, Center for Medical Statistics, Informatics and Intelligent Systems, Medical University of Vienna, Spitalgasse 23, A-1090 Vienna, Austria; [2]Division of Clinical Epidemiology and Biostatistics, Institute of Social and Preventive Medicine, University of Bern, Finkenhubelweg 11, CH-3012 Bern, Switzerland; and [3]CTU Bern, Bern University Hospital, CH-3010 Bern, Switzerland

Received 26 January 2010; revised 17 December 2010; accepted 27 January 2011; online publish-ahead-of-print 28 February 2011

The assessment of treatment effects from observational studies may be biased with patients not randomly allocated to the experimental or control group. One way to overcome this conceptual shortcoming in the design of such studies is the use of propensity scores to adjust for differences of the characteristics between patients treated with experimental and control interventions. The propensity score is defined as the probability that a patient received the experimental intervention conditional on pre-treatment characteristics at baseline. Here, we review how propensity scores are estimated and how they can help in adjusting the treatment effect for baseline imbalances. We further discuss how to evaluate adequate overlap of baseline characteristics between patient groups, provide guidelines for variable selection and model building in modelling the propensity score, and review different methods of propensity score adjustments. We conclude that propensity analyses may help in evaluating the comparability of patients in observational studies, and may account for more potential confounding factors than conventional covariate adjustment approaches. However, bias due to unmeasured confounding cannot be corrected for.

Keywords Bias • Causality • Confounding by indication • Non-randomized studies • Observational studies

Introduction

In the *European Heart Journal*'s recent issues, many studies applied propensity score techniques to investigate effects of interventions in observational studies. The increasing popularity of these methods in cardiology warrants a critical appraisal from a statistical point of view.

In a randomized controlled trial (RCT), random allocation of patients to either an experimental or a control arm guarantees that treatment allocation is unrelated to measured and unmeasured patient characteristics between groups. It enables researchers to draw unbiased conclusions about a treatment effect, provided that the number of randomized patients is large enough to minimize random variation. However, viability of RCTs may be limited by ethical or economical constraints. Moreover, the patient population is often highly selected in an RCT due to restrictive inclusion criteria, and diagnostic and therapeutic interventions may be atypically intensive when compared with the interventions used for target patients in routine clinical settings. Therefore, treatment effects observed in RCTs are often explored in observational studies, with investigators wishing to infer causal effects of a treatment. Since treatment allocation is not randomized in such studies, treated and non-treated groups may differ considerably in their pre-treatment characteristics, which may seriously hamper the validity of conclusions. Even though part of this imbalance may be due to randomness, as in an RCT, the treatment decision is likely to depend on a patient's pre-treatment characteristics. If these are imbalanced and associated with the study outcome, the assessment of the treatment effect from an observational study suffers from bias caused by *confounding by indication*.[1] Two statistical methods, among others, which may reduce this bias are, first, post-hoc matching of all experimental patients to control patients with similar pre-treatment characteristics and, second, evaluating treatment effects in a multivariable model including those characteristics as covariates that are considered to potentially confound the treatment effect.

Problems may arise in both approaches if the number of covariates considered for matching or adjustment is high. A high number of covariates precludes finding a control patient for each patient undergoing experimental treatment. For multivariable models, it was recommended that there should be at least 10 outcome events per covariate considered (EPC).[2] Thus, these approaches cannot adequately account for confounding when outcome events are rare or when the number of confounding variables is high.

* Corresponding author. Tel: +43 1404006684, Fax: +43 1404006687, Email: georg.heinze@meduniwien.ac.at

Both problems can often be solved by applying propensity score analyses, which typically consist of two stages: first, propensity scores are estimated using a multivariable logistic regression model, the *propensity model*. Second, the estimated propensity scores are used in the *treatment effect model* to adjust the effect of treatment on the outcome under study. The type of the treatment effect model depends on the nature of the outcome variable; usually logistic regression[3] will be used to model a binary outcome, while Cox regression[4] is the standard method for time-to-event outcomes.

Propensity score modelling

The propensity score of a patient is defined as the probability to receive the experimental treatment conditional on pre-treatment covariates. These covariates summarize all of what is known about that patient prior to treatment assignment, and are used as independent variables in the propensity model. Two conditions are necessary to obtain unbiased treatment effects by use of propensity scores. These conditions have been summarized by Rosenbaum and Rubin[5] as the assumption of a 'strongly ignorable treatment assignment'. The first condition is that once covariates have been controlled for, *to be assigned to* experimental or control treatment is not by itself a prognostic factor. This also means that all confounding variables must be known. Second, it is required that for each patient a real choice existed between experimental and control treatment at the time point of treatment selection. In other words, the patient had a non-null probability to receive either of the two, experimental or control treatment. Given that the above conditions hold, patients undergoing experimental and control treatments with equal propensity scores will on average have equal pre-treatment covariate values, and thus propensity scores equalize the covariate distributions in the experimental and control patients. Thus, adjusting for the propensity score will allow for unbiased estimation of the treatment effect.

Ideally, all variables potentially confounding the treatment effect (*confounders*) should be included in the propensity model. A confounder is defined by three conditions: first, it is a covariate available prior to the treatment assignment; second, it may influence the treatment decision; and third, it may influence the outcome of a patient.[1] This definition excludes any post-treatment measurements, because any imbalance in post-treatment covariates may already be a reflection of a treatment effect. Such variables would rather be called *mediators* of the treatment effect.

In the propensity model, significance is not a necessary condition for inclusion of covariates in the model. Since the model uses treatment status rather than clinical outcome status as the dependent variable, the number of potential confounders to be used is not limited by the number of clinical outcome events. Therefore, more covariates can be used than in conventional regression approaches. Nevertheless, propensity models need to be developed with the same care as other multivariable models. This includes checking the propensity model for interactions or non-linear effects of continuous confounders. One should be aware that including interactions will not necessarily yield better balance in individual covariates between experimental and control groups, but rather balance certain *combinations* of covariate values between the groups. Therefore, the appropriateness of including interactions into the propensity score model cannot be determined by investigating the balance of the covariates after propensity score adjustment.[6]

The ultimate goal of a propensity model is not to maximize the prediction of treatment status,[7] but to reduce the bias in the estimated treatment effect. In this light, recent investigations revealed that the inclusion of covariates that correlate with the treatment decision but not with the outcome will not improve the results from a propensity analysis; such covariates rather increase the imprecision of the treatment effect estimate.[8] In contrast, it may be more important to include covariates that are strongly related to outcome, but have only minor relationship with the treatment decision.[7] As a simple rule, one could determine whether relevant changes are observed in the estimated treatment effect when potential confounders are deleted from the propensity model. Variables identified as an unlikely confounder by this approach may be dropped from the propensity model provided their deletion leads to a gain in precision (i.e. narrower confidence interval).[1]

There is some debate as to how adequacy of propensity models can be assessed. Many use the c-index (area under the receiver operating characteristic curve), a measure of discrimination defined as the estimated probability that a randomly selected experimental patient has a higher propensity score than a randomly selected control patient. However, the c-index is not suitable to detect confounders omitted from the model.[9] Furthermore, a high c-index indicates non-overlap of propensity scores between experimental and control patients, and may render any evaluation of a treatment effect questionable because of the lack of comparability of the characteristics of experimental and control patients.[7] Piazza *et al.*[10] report a c-index for a propensity model amounting to 0.92, and compare graphically the distribution of propensity scores among aortic stenosis patients either undergoing transcatheter aortic valve implantation or surgical aortic valve replacement. Their *Figure 1* reveals that sufficient overlap of propensity scores is only provided in the subset of patients with propensity scores >0.675, which the authors said 'could be eligible for a randomized trial'. As also pointed out by Glynn *et al.*[7], such a graphical comparison of propensity scores can identify areas of non-overlap that are otherwise difficult to describe in conventional regression analysis with many covariates influencing treatment decisions. Often a c-index of 0.8 is considered as confirming adequate model fit. If such a value cannot be reached by a propensity model, it should not automatically be concluded that propensity analysis is inappropriate. Provided that the propensity model was built with care, a c-index close to 0.5 may even indicate that the amount of randomness in the treatment assignment, given the known covariates, is similar to that of a randomized trial. Under this circumstance, one cannot expect any benefit from incorporating propensity scores in the estimation of the treatment effect. However, a low c-index could also mean that important covariates are missing in the propensity model, because they are not known to the investigator. The problem of detecting such *unmeasured confounding* is impossible to address by statistical techniques, and its presence means that obtaining an unbiased estimate of the treatment effect is not possible. Some proposals have been made to assess the sensitivity of results to the potential presence of unmeasured confounding.[11]

Five important points to remember when modelling propensity scores are lined out in *Table 1*.

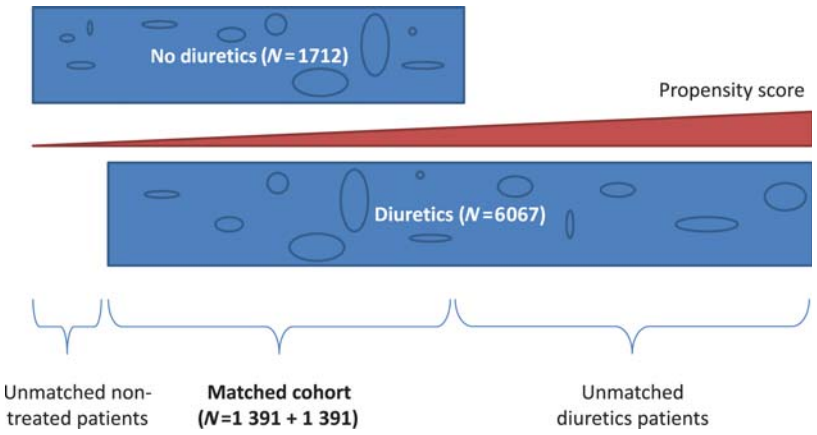

Figure 1 Graphical representation of matching by propensity score in the study of Ahmed *et al.*[17] Patients not treated by diuretics are matched to corresponding diuretics-treated patients. Patients with similar baseline characteristics are symbolized by ellipses of similar shape. First, propensity scores are estimated from 19 baseline covariates. These propensity scores represent the probability of being treated by diuretics given the covariates. Second, propensity scores are used to find matches of non-treated patients in the diuretics-treated group. Only 1391 of the non-treated patients could be matched; the remaining 321 patients had no counterparts in the treated group with similar characteristics. Analogously, 4676 of the diuretics patients were left unmatched.

Table 1 Five points to remember for propensity score modelling

Item	Description
1	Propensity analysis can be used to estimate treatment effects in observational studies, when confounding is present.
2	Propensity scores are estimated by logistic regression of the treatment assignment on pre-treatment covariates.
3	The advantage of propensity score techniques is that usually the number of covariates adjusted for can be higher than in conventional multivariable models.
4	The c-index should not be used for judging the adequacy of a propensity model.
5	Unlike randomization, propensity score techniques cannot account for unknown or unmeasured potential confounding factors.

Estimation of the treatment effect

After propensity scores have been estimated, they can be incorporated in the treatment effect model in the following four different ways:

Matching

Propensity scores can be used to enable numerical matching algorithms to find best matches between experimental and control patients. This method is fairly robust against misspecification of the propensity model. A recent study suggested that matching on the logit of the propensity score with a calliper width of 0.2 standard deviations of the logit of the propensity score may be superior to other methods used in the medical literature.[12] A drawback of matching is an often substantially reduced sample size since 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. Matching by propensity score should be followed by an analysis of the treatment effect that accounts for the matched pairs,[12] such as a stratified log-rank test or a Cox regression analysis stratified by matched pairs for time-to-event outcomes, or conditional logistic regression for binary outcomes.

Stratification

A stratified analysis sub-classifies the individuals based on quintiles of the propensity scores, computed over the combined treatment groups. 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 five strata. The quasi-standard of using five subclasses originates from Cochran[13] who showed that this may already reduce 90% of the imbalance of a confounder between the groups. Cochran[13] also showed that using more than five strata may reduce imbalance even better and, thus, more than five strata should be used in large data sets. Stratification does not impose any strong assumptions about functional form or time dependency of the effect of propensity scores on survival. It approximates matching without running the risk of losing unmatched patients, since balance in the proportions of experimental and control patients within each stratum is not required.

Covariate adjustment

The propensity score could be included in the treatment effect model as a covariate adjusting for baseline differences. However, wrong assumptions about the functional relationship of propensity scores and outcome (linearity, proportional hazards etc.) may then directly lead to biased estimates.[5]

Inverse probability of received treatment weighting

In this approach, the contributions of the study subjects are weighted by 1/propensity score for experimental patients and by 1/(1−propensity score) for control patients.[14] These weights assure that for each combination of baseline characteristics (leading to a particular value of the propensity score), the sum of contributions of all experimental and control patients are equal. As an example, consider a combination of covariate values (e. g. male, smoking, non-hypertensive) present in 20 patients. Assume further that a propensity score of 0.3 was estimated for these patients, when out of the 20 patients, 6 (30%) received experimental treatment and 14 (70%) the control intervention. Inverse probability of received treatment weighting (IPTW) assigns a weight of $1/0.3 = 3.33$ to each of the 6 experimental patients, and a weight of $1/0.7 = 1.43$ to each of the 14 control patients. The sum of weights of the experimental patients, $6 \times 3.33 = 20$, is equal to the sum of weights of the control patients, $14 \times 1.43 = 20$. Thus, IPTW generates a pseudo-population in which each covariate combination is perfectly balanced between treatment groups. In IPTW analyses, experimental patients with low propensity scores, and control patients with high propensity scores obtain disproportional high weight, compared with experimental patients with high propensity of receiving experimental treatment and control patients with low propensity scores. In some analyses, the weights assigned to some patients can become disproportionally high. Then the IPTW estimate of the treatment effect will be imprecise, which is reflected by a wide confidence interval. This unfavourable property is related to problems with finding matches for these patients.[7] In such cases, likely to be encountered in small-sized studies, it may be indicated to restrict analysis and conclusions to those patients where weights are homogenous, i.e. where the propensity scores are neither close to zero nor to one. Clearly, IPTW crucially depends on a correct specification of the propensity model. Inverse probability of received treatment weighting allows for a population-based interpretation of results, as if the study population would have undergone a randomized trial in which, counter to fact, both treatments were applied to each subject.[14]

Recent simulation studies[15,16] have demonstrated that matching on propensity score and IPTW tend to eliminate systematic differences between experimental and control subjects to a greater degree than stratification or covariate adjustment. While matching on the propensity score may result in the exclusion of some subjects, in particular those who cannot be matched, the other approaches always use the entire sample. Therefore, the choice of approach depends on the particular data situation and on how certain the investigator is about the optimal set of potential confounders to be included in the propensity model. Thus, a general recommendation towards a particular approach cannot easily be derived.

In practice, the success of propensity score modelling is often judged by whether balance on covariates is achieved between experimental and control group after its use. Balance on covariates is assessed by computing the standardized differences d for each covariate, which is defined as

$$d = \frac{100 \times |\bar{x}_E - \bar{x}_C|}{\sqrt{(s_E^2 + s_C^2)/2}} \quad (1)$$

for continuous variables, and as

$$d = \frac{100 \times |\hat{p}_E - \hat{p}_C|}{\sqrt{(\hat{p}_E(1-\hat{p}_C) + \hat{p}_C(1-\hat{p}_E))/2}} \quad (2)$$

for binary variables.[16] Here, $\bar{x}_E$, s_E, $\bar{x}_C$, and s_c denote the mean and standard deviation of a covariate in the experimental and control groups, respectively. Likewise, $\hat{p}_E$ and $\hat{p}_C$ denote the prevalence of one of the categories of a binary variable (e.g. female sex) in the experimental and control groups, respectively. These formulas directly apply if matching or stratification has been used. For IPTW, standardized differences can be computed with the quantities in Equations 1 and 2 replaced by their weighted equivalents.[16] Austin[16] further describes adjustments of these basic formulas to be used with covariate adjustment. The standardized difference d should be preferred to significance testing of covariates between experimental and control groups, as it is not confounded by sample size or the statistical power of the test employed, which can be substantially lower for binary than for continuous variables.

If for some covariates balance has not been achieved, one may include these covariates in the treatment effect model, provided that the EPC rule is respected. However, the estimated effects of those covariates cannot be meaningfully interpreted, because some of their association with the outcome variable will have been captured through their inclusion in the propensity score model.

Example: diuretics and mortality in heart failure patients

Ahmed *et al.*[17] used propensity scores to adjust their comparison of mortality between heart failure patients who received non-potassium sparing diuretics or no diuretics in an observational study. A total of 6067 patients with diuretics and 1712 patients without were included in the analysis. A propensity model based on 19 variables, such as New York Heart Association class, cardiothoracic ratio, use of potassium sparing diuretics, and clinically meaningful interactions was used to estimate the propensity of receiving diuretics. The propensity score was used to match patients with and without diuretics. Eventually, 1391 pairs of patients could be identified, which corresponds to 81% of the no-diuretic patients being matched to diuretic patients (*Figure 1*). The success of propensity matching was demonstrated by comparing the standardized differences, estimated by formulas similar to Equations 1 and 2, before and after matching. In the unmatched group, the absolute standardized difference was >40% in each of the above-mentioned variables, while it was reduced to <10% in the matched cohort. In the matched cohort, the hazard ratio for all-cause mortality, referring to the comparison of patients without and with diuretics, was 1.31, suggesting a protective effect of diuretics (95% confidence interval: 1.11–1.55). *Table 2* summarizes the typical steps in the analysis of an observational study by propensity scores, exemplified the study of Ahmed *et al.*

Table 2 Some typical steps of propensity analysis, exemplified by the observational study of Ahmed *et al.*[17]

Step	Task	Method used in example
1	Identify confounding variables	19 relevant covariates measured at baseline
2	Estimate propensity scores as the probability of receiving experimental treatment	Logistic regression of diuretic treatment (yes/no) at baseline on 19 covariables including clinically meaningful interactions
3	Match experimental to control patients	Matching algorithm: '5 to 1 digit matching on propensity score'
4	Evaluate success of matching	Compute standardized differences, compare with values before matching
5	Compare mortality of treatment groups	Cox regression stratified for matched pairs, adjustment for confounding variables
6	Interpretation	'No-diuretic patients have 1.3-fold mortality compared with diuretic patients with equal baseline characteristics'

Final remarks

In conventional regression or matching approaches used to adjust for confounding, the EPC rule limits the number of potential confounders that can be adjusted for, and thus the adequacy of the adjustment may be limited. However, propensity analyses do not have this limitation.[18] Still, Glynn *et al.*[7] pointed out that simulation studies comparing results from propensity analyses and conventional regression did not supply clear evidence in favour of or against one of these approaches. The additional estimation step required when computing propensity scores may contribute to increased variance of the final treatment effect estimate, and this may negate the favourable effect from using a more parsimonious treatment effect model.

A disadvantage of propensity analysis is that effects of covariates used for adjustment are not obtained at all, or, if these covariates have been included in the treatment effect model, have no meaningful interpretation. It is also not meaningful to evaluate the predictive accuracy of treatment effect models involving propensity score adjustment, e.g. by Harrell's popular survival c-index,[2] or by the proportion of explained variation,[19] because the purpose is not prediction but the estimation of a minimally biased treatment effect.

In summary, propensity analyses may be useful if the number of potential confounders is high such that conventional regression approaches are not feasible. They can help to identify subgroups lacking an overlap of propensity scores, among which an evaluation of treatment effects would not make sense. Neither propensity score methods nor conventional multivariable regression methods address bias resulting from unmeasured confounders; only studies involving random allocation to experimental or control groups yield estimates of treatment effects that are unbiased with respect to unmeasured confounders.

Acknowledgement

The authors are grateful to Karen Leffondré, Bordeaux, France; Rainer Oberbauer, Linz, Austria; Daniela Dunkler, Vienna, Austria; and two anonymous reviewers for valuable comments on an earlier version of the manuscript.

Funding

This work was supported by the European Community's Seventh Framework Programme (grant agreement number HEALTH-F2-2009-241544 to G.H.).

Conflict of interest: none declared.

References

1. Hill HA, Kleinbaum DG. Bias in observational studies. In: Gail M, Benichou J, eds. *Encyclopedia of Epidemiologic Methods*. Chichester: Wiley; 2000. p94–100.
2. Harrel FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*. New York, NY: Springer Inc.; 2001.
3. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. 2nd ed. New York: Wiley; 2000.
4. Cox DR. Regression models and life-tables (with discussion). *J R Stat Soc Ser B* 1972;**34**:187–220.
5. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983;**70**:41–55.
6. Tamburino C, Di Salvo ME, Capodanno D, Marzocchi A, Sheiban I, Margheri M, Maresta A, Barlocco F, Sangiorgi G, Piovaccari G, Bartorelli A, Brigori C, Ardissino D, Di Pede F, Ramondo A, Inglese L, Petronio AS, Bolognese L, Benassi A, Palmieri C, Patti A, De Servi S. Are drug-eluting stents superior to bare-metal stents in patients with unprotected non-bifurcational left main disease? Insights from a multicentre registry. *Eur Heart J* 2009;**30**:1171–1179.
7. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;**98**:253–259.
8. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Stürmer T. Variable selection for propensity score models. *Am J Epidemiol* 2006;**163**:1149–1156.
9. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor VM. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. *Pharmacoepidemiol Drug Saf* 2005;**14**:227–238.
10. Piazza N, van Gameren M, Jüni P, Wenaweser P, Carrel T, Onuma Y, Gahl B, Hellige G, Otten A, Kappetein AP, Takkenberg JJM, 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**:580–588.
11. Rosenbaum PR. Sensitivity to hidden bias. In: Rosenbaum PR, ed. *Observational Studies*. New York: Springer; 1995. p87–135.
12. 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**:62–67.
13. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968;**24**:295–313.
14. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;**11**:550–560.
15. 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**:734–753.
16. Austin PC. The relative ability of different propensity score methods to balance measured covariates between treated and untreated subjects in observational studies. *Med Decis Making* 2009;**29**:661–677.
17. Ahmed A, Husain A, Love TE, Gambassi G, Dell'Italia LJ, Francis GS, Gheorghiade M, Allman RM, Meleth S, Bourge RC. Heart failure, chronic diuretic use, and increase in mortality and hospitalization: an observational study using propensity score methods. *Eur Heart J* 2006;**27**:1431–1439.
18. D'Agostino RB Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998;**17**:2265–2281.
19. Schemper M, Henderson R. Predictive accuracy and explained variation in Cox regression. *Biometrics* 2000;**56**:249–255.