RMD Open
Rheumatic & Musculoskeletal Diseases

VIEWPOINT

Three handy tips and a practical guide to improve your propensity score models

Sytske Anne Bergstra,[1] Alexandre Sepriano,[1,2] Sofia Ramiro,[1,3] Robert Landewé[3,4]

To cite: Bergstra SA, Sepriano A, Ramiro S, *et al*. Three handy tips and a practical guide to improve your propensity score models. *RMD Open* 2019;**5**:e000953. doi:10.1136/rmdopen-2019-000953

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/rmdopen-2019-000953).

Received 14 March 2019
Revised 9 April 2019
Accepted 11 April 2019

[1]Department of Rheumatology, Leiden University Medical Center, Leiden, The Netherlands
[2]NOVA Medical School, Universidade Nova de Lisboa, Lisbon, Portugal
[3]Zuyderland Medical Center, Heerlen, The Netherlands
[4]Amsterdam Rheumatology and Immunology Center, Amsterdam, The Netherlands

Correspondence to
Dr Sytske Anne Bergstra; s.a.bergstra@lumc.nl

ABSTRACT

Real-world data are increasingly available to investigate 'real-world' safety and efficacy. However, since treatment in observational studies is not randomly allocated, confounding by indication may occur, in which differences in patient characteristics may influence both treatment choices and treatment responses. A popular method to adjust for this type of bias is the use of propensity scores (PS). The PS is a score between 0 and 1 that reflects the likelihood per patient of receiving one of the treatment categories of interest conditional on a set of variables. At least in theory, in patients with similar PS, the treatment prescribed will be independent of these variables (pseudorandomisation). But researchers using PS sometimes fail to recognise important methodological flaws which can lead to spurious conclusions. These include perfect prediction of treatment allocation, untied observations and lack of generalisability due to oversimplification of complex clinical scenarios. In this viewpoint we will discuss the most commonly encountered flaws and provide a stepwise description on the estimation and use of PS, such that in future publications these flaws can be avoided.

VIEWPOINT

Real-world data are almost routinely collected in rheumatology and are now available to investigate 'real-world' safety and efficacy of medical interventions. However, treatment in observational studies is not randomly allocated. In other words, a specific patient may receive a specific treatment (and not another one) due to some specific personal or disease characteristics. This means that differences in patient characteristics that are predictive of disease severity may guide both treatment choices as well as treatment responses and may thus lead to confounding by indication. Therefore, crude comparisons between treatment effects are insufficient and methods should be applied to adjust for this bias, in order to obtain valid results. An increasingly popular method to address this is the use of propensity scores (PS).

The PS is a score between 0 and 1 that reflects the likelihood per patient of receiving one of the treatment categories of interest. This likelihood is estimated by binomial or polynomial regression analysis and is conditional on a set of pretreatment variables that together reflect to some extent the factors the prescriber considers when making a treatment choice, and that at the same time influence the outcome (eg, disease activity, physical functioning, imaging findings, and so on). At least in theory, in patients with similar PS, the treatment prescribed will be independent of the added variables (pseudorandomisation). To adjust for confounding by indication, the PS can be used for stratified sampling, matching or as a covariate in regression analyses.[1 2] But the process of estimating the PS is not straightforward and many authors do it inappropriately. In this viewpoint, we highlight three major issues often overlooked (or under-reported) by authors, using examples from the literature, and provide a practical step-by-step guide on how to estimate a PS using Stata, a commonly used statistical package.

Three eye-catching misunderstandings in PS estimation

The perfect PS

A common misunderstanding is that researchers aim for perfect prediction of treatment allocation, using regular model building techniques and measures for model fit (eg, area under the curve or c-statistic). For instance, in 2012 the effect of adherence to three of the 2007 EULAR recommendations for the management of early arthritis on the occurrence of new erosions and disability was assessed.[3] Since the impact of recommendations on treatment delivered in clinical practice cannot be investigated in randomised controlled trials, the authors appropriately decided to calculate

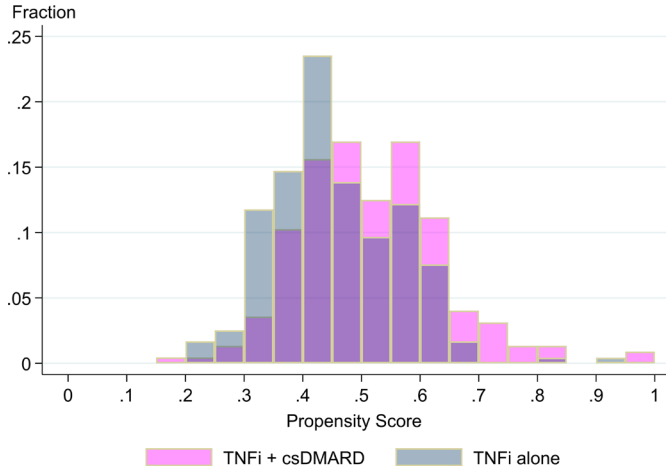

Figure 1 Propensity score distribution at baseline for two treatment groups. Untied observations fall outside the area of common support (0.20; 0.70) and should therefore be trimmed. Used with permission from Sepriano *et al.*[14] csDMARD, conventional synthetic disease-modifying antirheumatic drug; TNFi, tumour necrosis factor inhibitor.

a PS to adjust for potential biases related to being treated according to the recommendations or not. For PS estimation, the authors selected all variables related to recommendation adherence (the main predictor of interest). Furthermore, the authors built the PS model using an automatic process of selecting variables, with statistical thresholds for inclusion of variables into the model. The quality of the model was then assessed by Hosmer-Lemeshow tests for goodness of fit and c-statistic for discriminatory ability. The authors concluded that the PS model had a good discriminative ability, with a c-statistic of 0.77. However, the aim of a PS is to efficiently control for confounding, and not to predict treatment allocation. Hence, measures of model fit are inappropriate to judge the validity of the model or to select variables, since these measures judge a model on its ability to predict treatment allocation, instead of its ability to control for confounding. Instead, we should aim for a perfect balance of measured covariates across treatment groups and variable selection should be based on content knowledge.[1 2 4] In PS models the best balance (between treated and untreated) is achieved by adding variables that, based on content knowledge, are expected to be related to the outcome (eg, new erosions), or to both the outcome and predictor (eg, following recommendation). Variables that are only related to the predictor should be avoided since they lead to decreased precision of the effect estimates.[4]

Untied observations

Especially when authors aim for perfect predictability, as in the example above, 'untied observations' often occur. These are patients for which we can almost perfectly predict which treatment they will receive. In a proper PS the range of predicted probabilities should cover the entire possible spectrum from 0 to 1, and for each predicted probability a sufficient number of patients that are treated and non-treated should be present.[2] One way to think about this is to see PS as an advanced matching technique. It enables us to 'match' for many variables at the same time, by compressing those variables into a single score.[2] Untied observations are patients without a 'match,' which should be deleted. Alternatively, one could trim (ie, delete) patients without a 'match,' and patients with a very low probability of receiving one of the treatments. For example, all patients with a PS<0.05.[5] When many observations are deleted obviously the data only apply to the resulting selected patient group, which means that the data are less generalisable (figure 1).

Two or more than two treatment choices?

Most frequently PS refer to binomial treatment decisions. But in rheumatology there are many scenarios in which multiple treatment options are considered in individual patients. In a previously published study, the clinical outcomes of patients with rheumatoid arthritis (RA) treated according to daily clinical practice were compared after 1 year of treatment in patients who received treatment with either abatacept or tocilizumab.[6] The authors describe that in daily practice abatacept and tocilizumab are prescribed to patients with RA with uncontrolled disease despite treatment with conventional synthetic disease-modifying antirheumatic drugs and argue that treatment assignment of either abatacept or tocilizumab may be non-random, that is, different types of patients are being treated with either drug, which will most likely lead to a biased comparison of the outcome. Therefore, they apply PS matching to handle this potential bias.

However, since daily practice data were used, eligible patients could have likely received other treatments than only abatacept or tocilizumab. In theory one could select two of the available treatment options and apply a binomial PS to adjust for confounding by indication (eg, treatments A and B and ignore that patients could also have received C or D). Within the sample of patients starting one of the two selected treatments (ie, A and B), the binomial PS would be valid. However, this would be a gross simplification of the true clinical scenario, in which the rheumatologist had many more treatment options to choose from (ie, C and D). Therefore, external validation falls short, and any generalisation of these data to the whole population of patients with a given disease is not valid. Obviously, this is an important limitation, since one of the main strengths of testing treatment effects with observational data compared with clinical trials is the inclusion of a less selected population, potentially resulting in better generalisability. Therefore, as an alternative, a 'multiple PS' should be considered to account for multiple treatment options simultaneously to better reflect reality.[7]

Estimating PS step by step

When the decision has been made that a PS would be appropriate to adjust for confounding by indication in an observational study, several steps are required to calculate, evaluate and use the PS appropriately. We will provide a stepwise description for the estimation of binomial PS, including a syntax example in Stata in online supplementary file 1. Previous publications have provided a description on how to perform multiple PS.[8 9] For PS estimation in SAS, SPSS and R similar steps can be followed using the software-specific syntax. In SPSS, the command 'Propensity Score Matching' is available from the 'Data' tab. In SAS, the 'PROC PSMATCH' procedure is available. In R, users can calculate the binomial PS using logit or probit regression with the 'glm' command. A tutorial for estimating PS in R is available online.[8]

Step 1: select variables

For the estimation of both binomial and multinomial PS, the first step is the selection of variables to include in the PS. Extensive literature is available regarding variable selection for PS models.[4] In short, only variables that are measured before treatment assignment should be included, since variables that are measured after treatment assignment cannot possibly act as confounders (of the treatment allocation process). The highest precision is achieved by adding all variables related to the outcome of the study (eg, disease activity). Variables that are only related to the exposure (eg, treatment), but not to the outcome, decrease precision and should not be included. Ideally, these variables are selected based on subject matter knowledge. However, especially when a large number of pretreatment variables have been collected and the relationship with the outcome is unclear, regression analyses may be used to identify all available pretreatment variables with an association with the outcome. For example, when a continuous measure has been used as outcome, linear regression may be used to select all variables with associations at $p<0.10$ with the outcome.[9]

For steps 2–8 a Stata syntax example is available in online supplementary file 1.

Step 2: assess the standardised differences between variables before calculating the PS

This step is not relevant for variable selection or for further analyses, but it provides insight into the initial comparability of the binomial outcome groups by using standardised differences.

Step 3: estimate the PS

Step 4: check the level of balance between treatment and control groups

After obtaining the PS we check the level of balance between treatment and control groups. This can be done by (1) splitting the sample in strata and testing whether the means of the PS are similar within strata across treatment groups (step 5a); and (2) by visual analysis of a density plot of the distribution of the PS in the treatment groups before (figure 1) and after defining the area of common support (step 5b).

Step 4a: check the distribution of the PS in each quintile and per treatment strata

It is common to first split the data in quintiles and investigate the balance across the quintiles. If balance is not achieved, the number of strata can be increased.

Step 4b: find the area of common support

This can be done by creating a histogram similar to figure 1. The area of common support is the range in which the PS for the two groups overlap. The minimum and maximum values defining this range can be used in step 6.

Step 5: graph the PS distribution within the area of common support

Create a similar histogram as in step 4b, but now excluding any data outside the 'area of common support.'

Step 6: report how many patients were trimmed

Step 7: report the average PS and the number of patients per quintile after trimming

Step 8: assess the standardised differences within quintiles

Standardised difference tests are preferred to examine whether baseline covariates are equally distributed across treatment groups. Standardised differences <0.10 are generally considered acceptable.[1]

Step 9: re-estimate the PS model if balance is not achieved

Start again with step 3 if balance is not achieved. Options to improve the model include dropping or recategorising variables, or including interaction terms, higher order terms or splines.[1 2]

Step 10: estimate the effect before applying the PS model

First, perform all analyses without taking the PS into account. This will provide crude results.

Step 11: estimate the effect after applying the PS model

Finally, the PS can be used for matching, stratified sampling, or covariate adjustment in regression analyses. Whereas matching and stratification are performed before doing further statistical analyses, covariate adjustment is incorporated into the analyses. Previous publications are available with a more detailed description of each of these methods for binomial or multiple PS.[1 7 9]

It has been shown that PS matching is more successful in reducing bias than stratification or covariate adjustment.[10–12] However, when multiple exposure groups are compared, matching may not be possible since this may result in small treatment samples.[7 9] Furthermore, depending on the planned analyses, covariate adjustment may be considered more appropriate.

Concluding remarks

A PS can only entirely adjust for confounding by indication when all relevant pretreatment variables are

included, which is illusionary. In practice, it is impossible to check whether residual confounding is present.[13] As such, PS are an aid to better interpret crude treatment differences found in observational studies, but can never replace proper randomised controlled trials. Nevertheless, it is certainly more robust to address treatment effects in observational studies using PS than fully ignoring the inherent confounding by indication. Therefore, the appropriate use, estimation and reporting of PS can provide an important contribution to the quality and interpretability of observational studies into treatment effects.

Contributors SAB drafted the work. All authors contributed to the design and interpretation of the manuscript. All authors revised the work critically and read and approved the final version of the manuscript.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Obtained.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No additional data are available.

REFERENCES

1. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46:399–424.
2. Garrido MM, Kelley AS, Paris J, *et al*. Methods for constructing and assessing propensity scores. *Health Serv Res* 2014;49:1701–20.
3. Escalas C, Dalichampt M, Combe B, *et al*. Effect of adherence to European treatment recommendations on early arthritis outcome: data from the ESPOIR cohort. *Ann Rheum Dis* 2012;71:1803–8.
4. Brookhart MA, Schneeweiss S, Rothman KJ, *et al*. Variable selection for propensity score models. *Am J Epidemiol* 2006;163:1149–56.
5. Brookhart MA, Wyss R, Layton JB, *et al*. Propensity score methods for confounding control in nonexperimental research. *Circ Cardiovasc Qual Outcomes* 2013;6:604–11.
6. Kubo S, Nakayamada S, Nakano K, *et al*. Comparison of the efficacies of abatacept and tocilizumab in patients with rheumatoid arthritis by propensity score matching. *Ann Rheum Dis* 2016;75:1321–7.
7. Bergstra SA, Winchow LL, Murphy E, *et al*. How to treat patients with rheumatoid arthritis when methotrexate has failed? The use of a multiple propensity score to adjust for confounding by indication in observational studies. *Ann Rheum Dis* 2018.
8. Bergstra SA, Winchow L-L, Murphy E, *et al*. How to treat patients with rheumatoid arthritis when methotrexate has failed? The use of a multiple propensity score to adjust for confounding by indication in observational studies. *Ann Rheum Dis* 2019;78:25–30.
9. Spreeuwenberg MD, Bartak A, Croon MA, *et al*. The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Medical care* 2010;48:166–74.
10. Austin PC. Type I error rates, coverage of confidence intervals, and variance estimation in Propensity-Score matched analyses. *Int J Biostat* 2009;5.
11. Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med* 2007;26:734–53.
12. Austin PC, Mamdani MM. A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006;25:2084–106.
13. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997;127:757–63.
14. Sepriano A, Ramiro S, van der Heijde D, *et al*. Effect of comedication with conventional synthetic disease-modifying antirheumatic drugs on retention of tumor necrosis factor inhibitors in patients with spondyloarthritis: a prospective cohort study. *Arthritis Rheumatol* 2016;68:2671–9.