Statistical nugget – propensity scores

Turkiewicz A[1], Cook JA[2], Ranstam J[3]

1 Clinical Epidemiology Unit, Orthopaedics, Clinical Sciences Lund, Lund University, Lund, Sweden

2 Centre for Statistics in Medicine, University of Oxford, Oxford, UK.

3 Lund University, Dept of Clinical Sciences, Lund, Sweden

The authors report no conflict of interest.

A propensity score (PS) can be used to minimize confounding bias in observational studies, where the treatment assignment (or more broadly exposure) is not random.[1–3] In its most basic form the PS is the estimated probability of being treated (as opposed to not treated) given the measured covariates. It can be estimated by fitting a regression model (typically logistic regression) with the treatment received (or exposure) as outcome. There are four common ways in which a PS is used to estimate the effect of the treatment: matching, stratification, adjustment and inverse probability of treatment weighting. For example, it could be used to seek to compare use or not of oral antibiotic and mechanical bowel preparation following left-sided colonic rectal resections.[4]

One of the most common implementations is PS matching where each treated subject is matched with one or more untreated (or unexposed) subjects with a similar value of PS. How "similar" is defined varies, but is usually based upon the width of so called caliper (i.e. the difference in the PS between two subjects).[5] In practice, matching often results in exclusion of study participants that have no match which in turn can result in bias. On occasion, the proportion excluded can be high which limits generalizability.[6] Recent reports suggest that matching on PS should not be used to adjust for confounding in observational studies, and may in fact increase confounder imbalance leading to estimates with greater bias.[7]

Stratification is usually implemented by dividing the study sample into subsets based upon the PS; however, this can lead to large heterogeneity within strata. Adjustment is typically done by including the main confounders and the propensity score in the analysis model for estimation of the exposure-outcome association. However, this procedure may give little gain, as it generally leads to the same estimate as standard adjusted regression if the same confounders are used in the regression and in the estimation of PS.

Alternatively, PS can be used to derive weights, so called "inverse probability of treatment weights" to adjust the analysis to account for differences in the propensity to receive treatment. The simplest approach is to derive such weights as 1/PS for the treated and 1/(1-PS) for the untreated. These weights are then used in the statistical model for estimating the association between treatment and outcome.[8] Extreme weights can lead to bias and methods for stabilizing weights exists.[9] The inverse probability of treatment weighting can take into account time-dependent confounding and selection bias.[8,10]

Using a PS approach does not necessarily mean that the estimated effect can be interpreted as causal. Using PS for causal inference, such determining the right treatment decisions (using a laparoscopic or open surgical approach) requires that certain assumptions are fulfilled. For example, that there are no unmeasured confounders and that each subject has a positive probability of receiving each treatment. It is important to remember that using the PS approach does not necessarily remove potential bias from unmeasured confounders. Further, the choice of variables to include in the model for estimation of PS is not a trivial one. Only true confounders (to remove bias) and covariates predictive of the outcome (to increase precision) are typically useful for inclusion in the PS model. However, not including relevant variables when estimating PS easily leads to bias. Including variables affecting treatment but unrelated to outcome or variables that are effects of exposure and outcome (so called colliders) generally lead to bias. In surgery, the pre-operative (such time from diagnosis until surgery and ASA grade) and operative variables (such as anaesthetic and surgical approach) can typically be safely used to assess operative and post-operative exposures and confounders. Statistical hypothesis testing should not be used for selection of variables for inclusion in a model for estimating PS.[9]

A situation when the use of PS can be particularly advantageous is when the outcome is rare, the exposure is common and there is a large number of confounders to adjust for. Standard regression methods may not be suitable, as adjustment for many covariates with rare outcome is unreliable. The PS approach, where the confounders are included only in the model for exposure, allow for adjustment for more variables. Only the corresponding PS is used in the final model for exposure-outcome association.[11]

Gorge Box stated that "all models are wrong but some are useful" and this applies also to using PS. They can help in obtaining useful inferences when applied with care to suitable data.

References

1 Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. [Oxford University Press, Biometrika Trust]; 1983; **70**: 41–55.

2 Cook JA, Ranstam J. Statistical models and confounding adjustment. *Br J Surg*. 2017; **104**: 786–787.

3 Adamina M, Guller U, Weber WP, Oertli D. Propensity scores and the surgeon. *Br J Surg*. 2006; **93**: 389–394.

4 Golder AM, Steele CW, Conn D, MacKay GJ, McMillan DC, Horgan PG, *et al.* Effect of preoperative oral antibiotics in combination with mechanical bowel preparation on inflammatory response and short-term outcomes following left-sided colonic and rectal resections. *BJS Open*. 2019 Oct 16; **3**: 830–839.

5 Austin PC. Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat*. 2011 Apr; **10**: 150–161.

6 Norlén O, Stålberg P, Zedenius J, Hellman P. Outcome after resection and radiofrequency ablation of liver metastases from small intestinal neuroendocrine tumours. *Br J Surg*. 2013; **100**: 1505–1514.

7 King G, Nielsen R. Why Propensity Scores Should Not Be Used for Matching. *Polit Anal*. Cambridge University Press; 2019 Oct; **27**: 435–454.

8 Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006 Jul; **60**: 578–586.

9 Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015; **34**: 3661–3679.

10 Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiol Camb Mass*. 2000 Sep; **11**: 561–570.

11 Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*. 2003 Aug 1; **158**: 280–287.