The Good, the Bad, and the Ugly: Reminders About Propensity Scores

Lauren Strand, MS, **Blythe Adamson, MPH**, Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington, Seattle, WA, USA; **Joseph Delaney, PhD**, Department of Epidemiology, University of Washington, Seattle, WA, USA; **Anirban Basu, PhD**, Comparative Health Outcomes, Policy, and Economics (CHOICE) Institute, University of Washington, Seattle, WA, USA

Lauren Strand, MS

KEY POINTS . . .

Propensity scores are a tool useful when analyzing observational data and have advantages over traditional multiple regression in certain situations.

The benefits of propensity scores include bias reduction and containment of dimensionality (i.e., many covariates).

Cautiously approach causal inference with propensity scores methods and consider the assumptions required, such as the unverifiable (no unmeasured confounding).

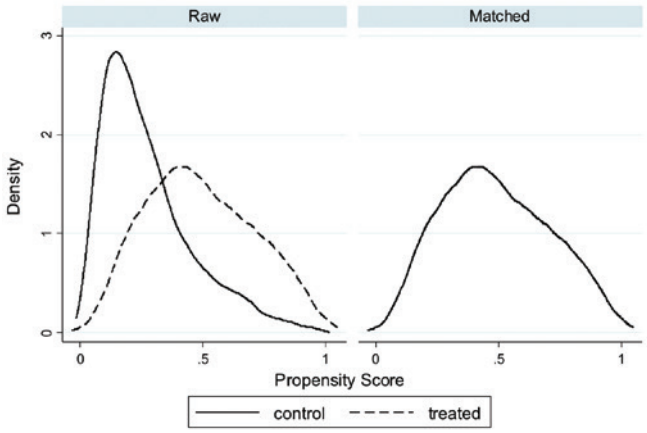
Propensity scores (PS) are ubiquitous. The methods Rosenbaum and Rubin first pioneered in the health sciences in the early 1980s have steadily increased in popularity [1]. PS are used to analyze observational data to produce inferences about the effects of a binary intervention after controlling for confounders observed in the data. While PS may be a helpful tool, they are not a magic wand for inference nor a panacea for poor quality data or study design. Here we review PS methods and describe the advantages and limitations.

PS Methods

A propensity score is the probability that a subject with a vector of characteristics X is assigned a treatment. This probability is estimated using the data at hand. The estimated PS, often referred to as a "balancing score," can be used to balance the distribution of baseline characteristics X across treated and untreated study participants. PS are often calculated in a logistic regression of the intervention against all covariates that are associated with the outcome of interest [2]. Best practices for generating PS typically involve including interactions and second-order polynomials of the covariates. Since the PS model is not used in prediction of who will get treatment in the future, overfitting the model is not only accepted but also necessary to create the correct balancing score. It is essential to check that the support of the distribution of the estimated PS has "overlap" across the two intervention groups. An example of good overlap is shown in the figure. Occasionally, some trimming or truncating may be employed to ensure the same distribution support across the two intervention groups. Analysts should also check the balance of covariates across intervention and non-intervention groups conditional on the PS (after the matching or weighting procedures).

There are many ways in which the estimated PS can be used to create balance across the intervention groups: using the PS as a covariate, blocking on PS percentiles, matching, and inverse probability weighting. Including the score directly in the regression is no longer used [3], and IPW may be the most preferred of the PS-based methods in terms of bias and consistency. Weights are derived by taking the inverse of the probability of receiving treatment for individuals who are treated and the inverse of the probability of not receiving treatment for individuals who are untreated. Matching methodologies also used include exact matching, or matching treated and untreated individuals within some "neighborhood" (bandwidth) of the PS (e.g., Mahalanobis distance matching, kernel matching). Doubly robust methods, where regression techniques are combined with PS methods, can increase robustness of inference, but this often comes with decreased efficiency [4,5]. >

Figure. Distribution of propensity scores among 835 statin initiators (treated) and 1554 non-initiators (control) in the Multi-Ethnic Study of Atherosclerosis before and after matching. *Analysis by Lauren Strand.*



The Good

The main advantage of PS methods is reduction of the dimensionality of the matching problem that regression-based estimators face when there are many confounding factors to consider. This is especially true when the exposure is common and the outcome is rare. While traditional multiple regression modeling tries to find the best fit to the data with all those covariates, PS methods rely on trying to match a single scalar quantity—the estimated PS—to solve the matching problem. Because it is not necessary to obtain the correct specification for estimating the propensity scores, PS methods can produce robust inference on intervention effects compared to regression models under certain conditions. Moreover, PS methods can readily identify situations where comparison between two intervention groups is unwarranted because there is not enough overlap in the distribution of PS.

The Bad

PS methods produce the incremental effects of an intervention, marginal over the distribution of other covariates in the model. In contrast, coefficients on the intervention variable in a non-linear model (e.g., a logistic regression) produce a conditional effect, which is conditioned at the mean of the distribution of other covariates. It is important to remember this distinction when comparing results [6].

In finite samples, estimated PS may not perform well as balancing scores. Conditioning or matching on PS balances the covariate means across treatment groups easily, but not necessarily higher order moments of distributions (i.e., variance and skewness). Consequently, if the data-generating process is non-linear, PS matching may still be subject to inefficiency and/or bias [4]. Some authors have recently suggested the PS matching should not be used at all, citing examples where in already relatively balanced data, matching reduced efficiency and increased imbalance [7]. Perhaps the most important limitation of PS methods is the strong ignorability/no unmeasured confounding assumption. PS methods do not correct bias in estimates due to unmeasured confounders or account for measurement error in the outcome. Since researchers can never be sure if all confounders have been measured in available data, PS results should be interpreted as adjusted effects rather than causal effects.

The Ugly

In most health care applications that have directly compared PS methods to regular regression methods, PS methods yielded estimates that did not differ substantially from those of multiple regression models. Stürmer et al. found that only 13% of reviewed papers had an effect estimate in PS analysis that differed by more than 20% from traditionally modelled estimates [8]. In addition, even if the results differ across analytic approaches, it cannot be assumed that PS-based estimates are less biased [7].

Attempts to mitigate the no-unmeasured confounding assumption have been revitalized recently with Schneeweiss et al. using high-dimensional algorithmic derivation of PS in big data to identify mild correlates of these confounders [9]. While this methodology is promising, questions remain about whether any big data source possesses sufficient correlates of the unmeasured confounder.

Conclusion

Given their widespread use, analysts and consumers of scientific literature should be cognizant of the advantages and limitations of PS.

References

[1] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70(1):41–55 [2] Brookhart MA, Schneeweiss S, Rothman KJ, et al. Variable Selection for Propensity Score Models. Am J Epidemiol 2006;163(12):1149–1156. [3] Austin PC, Grootendorst P, Normand SLT, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. Stat Med 2007;26:754–768. [4] Basu A, Polsky D, Manning WG. Estimating treatment effects on healthcare costs under exogeneity: is there a "magic bullet"? Health Serv Outcomes Res Methodol 2011;11(1–2):1–26. [5] Kang JDY, Schafer JL. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Stat Science 2007;22:523–539. [6] Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. Stat Med 2010;29(20):2137–2148. [7] King G, Nielson R. (2016). Why Propensity Scores Should Not Be Used for Matching. Retrieved October 16, 2017, from j.mp/PScore. [8] Stürmer T, Joshi M, Glynn RJ, et al. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006;59(5):437–447. https://doi.org/10.1016/j.jclinepi.2005.07.004. [9] Schneeweiss S, Rassen JA, Glynn RJ, et al. High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. Epidemiology (Cambridge, Mass.), 2009;20(4):512–522. ■