Propensity Score Methods and Unobserved Covariate Imbalance: Comments on “Squeezing the Balloon”

M. Sanni Ali, Rolf H. H. Groenwold, and Olaf H. Klungel

Abstract. In their recent *Health Services Research* article titled “Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance,” Brooks and Ohsfeldt (2013) addressed an important topic on the balancing property of the propensity score (PS) with respect to unmeasured covariates. They concluded that PS methods that balance measured covariates between treated and untreated subjects exacerbate imbalance in unmeasured covariates that are unrelated to measured covariates. Furthermore, they emphasized that for PS algorithms, an imbalance on unmeasured covariates between treatment and untreated subjects is a necessary condition to achieve balance on measured covariates between the groups. We argue that these conclusions are the results of their assumptions on the mechanism of treatment allocation. In addition, we discuss the underlying assumptions of PS methods, their advantages compared with multivariate regression methods, as well as the interpretation of the effect estimates from PS methods.

The use of propensity score (PS) methods in observational studies of medical treatments to adjust for measured confounding has increased substantially during the last decade (Shah et al. 2005). The PS is defined as a subject’s probability of treatment given his or her characteristics. For groups of subjects with the same PS, measured covariates that were used to construct the score tend to be balanced across treatment groups (Rosenbaum and Rubin 1983). However, unlike random assignment of treatments in a randomized trial, covariates that were not measured (and thus not included in the PS model) will not necessarily be balanced when conditioning on the PS. Hence, imbalances in unmeasured covariates are not addressed by PS methods.

In a recent study, Brooks and Ohsfeldt assessed the balancing property of PSs with respect to unmeasured covariates (Brooks and Ohsfeldt 2013) and wondered why subjects with the same PS may receive different treatments. Essentially, they stated that for two subjects with the same PS, for example, a

PS of 0.65 (i.e., a probability of 0.65 for receiving treatment), of whom one received treatment and the other did not, there must be a reason why one received treatment and the other did not. They argued that, apparently, this reason was not included in the PS model (thus being unmeasured covariates, potentially confounders). Thereafter, they concluded that PS methods balance measured confounders at a cost of exacerbating any imbalance in unmeasured covariates that are independent of the measured covariates. Further extended, if these unmeasured covariates are confounders (related to both treatment and outcome), PS methods can exacerbate the bias in treatment effect estimates.

We do not agree with their main conclusions for reasons outlined below. In the following paragraphs, we focus on three topics: (1) the assumptions underlying PS methods; (2) the conceptual advantage of PS methods in contrast to classical regression techniques; and (3) the estimand (treatment effect estimate) obtained using multivariable regression and the different PS approaches.

THE ASSUMPTION UNDERLYING PS METHODS

Two main assumptions underlying PS methods are relevant to interpret the findings by Brooks and Ohsfeldt: exchangeability and positivity. As described in the original paper by Rosenbaum and Rubin (1983), PS methods rely on the assumption of strong ignorability or exchangeability, which is a stronger version of "ignorable" mechanisms coined by Rubin (1976, 1977). It can be stated formally as: $\{Y(0), Y(1)\} \perp X|Z$, where Z denotes a vector of (measured) covariates, X denotes treatment assignment (Yes/No), and $Y(0)$ and $Y(1)$ are the potential outcomes under control and treatment conditions, respectively. It means that conditional on (measured) covariates (Z), treatment assignment (X) is independent of potential outcomes ($Y(0)$, $Y(1)$). Intuitively, this assumption is equivalent to the colloquial notion of all confounders having been mea-

Address correspondence to Olaf H. Klungel, PharmD, Ph.D., Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, P.O. Box 80082, 3508 TB Utrecht, The Netherlands; e-mail: O.H.Klungel@uu.nl. M. Sanni Ali, D.V.M., M.Sc., is with the Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Utrecht, The Netherlands. Rolf H.H. Groenwold, M.D., Ph.D., is with the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht; Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht Institute for Pharmaceutical Sciences, University of Utrecht, Utrecht, The Netherlands.

sured (Hill 2008). This assumption of no systematic, unmeasured, pretreatment differences between treated and untreated subjects that are related to the outcome under study is needed not only for PS methods but also for ordinary methods to adjust for confounding (e.g., multivariate regression methods) to get an unbiased treatment effect estimate.

In a large randomized clinical trial (RCT) where subjects are assigned to treatment or control by flip of a fair coin, the probability of being assigned to treatment (i.e., the PS) is 0.5. This means that approximately half of the subjects will be treated while the other half will not. The randomization implies that, on average, measured as well as unmeasured covariates will be balanced between the two treatment groups. In contrast, in observational studies, treatment assignment is a non-random process. Nevertheless, PS methods help researchers mimic randomization by creating a sample of subjects receiving the treatment that is comparable on all measured covariates to a sample of subjects not receiving the treatment (Austin 2011, 2007; Jo and Stuart 2009). Hence, randomization in a RCT and mimicking random assignment of treatment in PS analysis (e.g., by PS matching) are sufficient to generate group of subjects with the same PS yet receiving different treatment modalities. This does not require differences in unmeasured covariates. However, randomization implies that unmeasured covariates are balanced as well, but PS methods will not guarantee balance of unmeasured covariates.

Brooks and Ohsfeldt considered a deterministic treatment assignment model rather than a random treatment assignment model given measured and unmeasured covariates. Particularly, in their simulation set-up, treatment status depends on a number of covariates. Given that the values of these covariates are known, the treatment status is fixed, that is, it no longer depends on any random process. Thus, if treatment status of a subject as well as the values of certain covariates (e.g., three out of four covariates) is known, this will restrict the values of the other covariate(s). In fact, the set-up of the data generation implies that among the treated subjects, those who have low values for certain covariates will have high values for the other covariates (and vice versa). Consequently, by design, balancing part of the covariates between the treatment groups (by PS matching) will result in an imbalance in the other covariates. Brooks and Ohsfeldt demonstrated an “exacerbation” in the imbalance of unmeasured covariates when using PS methods to balance measured covariates compared with the full unweighted sample. This finding could in part be explained by the fact that they not only assumed unmeasured covariate variation that is unrelated to measured covariates as a requirement for PS

methods but also imposed this thought in generating their data (in particular, the simulation of treatment status). Moreover, in their simulations, Brooks and Ohsfeldt demonstrated that the "exacerbation" in imbalance of unmeasured confounders was only detected when the measured and unmeasured covariates are uncorrelated. Hence, if one attempts to measure as much covariates as possible, it seems more likely that risk factors correlated to unmeasured covariates or proxies for unmeasured covariates included in the PS model, for example, using high dimensional propensity score (hdPS) method (Schneeweiss et al. 2009; Rassen and Schneeweiss 2012), could lessen the exacerbation or improve balance of unmeasured covariates.

Apart from the assumption of no unmeasured covariates (confounders), another requirement for identifying a causal effect is that both treated and untreated subjects exist at all levels of confounders in the population under study, commonly known as the positivity assumption (Cole and Hernán 2008). In terms of the PS, this would mean that there is sufficient overlap of the PS distributions between treated and untreated subjects. For example, in case of PS matching (where treated and untreated subjects with the same PS are matched), this requirement is definitely met. The absence of sufficient overlap of PS between treatment groups (i.e., violation of the positivity assumption) can increase both the variance and bias of causal effect estimates (Petersen et al. 2010). Therefore, processing data, for example, using PS matching to generate two group of subjects with the same PS receiving different treatment modalities assures validity of positivity, without requiring systematic differences in an unmeasured covariate that is unrelated to measured covariates.

ADVANTAGE OF PROPENSITY SCORE METHODS

Brooks and Ohsfeldt further claimed that "the conceptual advantage of PS-based methods relative to standard regression appears to hinge on the assumption that balancing measured covariates between treated and untreated subjects leads to unmeasured covariate balance between treated and untreated subjects." This thought is shared by several researchers applying PS methods (Shah et al. 2005). However, the primary goal of PS methods is to create a balance of measured covariates between treatment groups, and PS methods may not seem to be superior to multivariable regression methods with respect to adjustment for unmeasured covariates. Nevertheless, Rosenbaum argued that

“PS matching can lead to a reduction in both sample variability and the estimated treatment effect’s sensitivity to potential omitted variables” (Rosenbaum 2005). Furthermore, confounding by variables unmeasured in the main study can be addressed using variables measured in another validation study via PS calibration (Stürmer et al. 2005).

It is worth mentioning the potential advantages of PS methods compared with conventional regression methods, although both methods give similar effect estimates, as demonstrated in most published empirical and simulation studies (Shah et al. 2005; Stürmer et al. 2005). PS methods provide an effective way of controlling for covariates in case of a limited number of events, thereby overcoming the “dimensionality problem,” where the introduction of a new balancing covariate in the regression model increases the minimum necessary number of observations (outcome events) in the sample (Cepeda et al. 2003; Glynn, Schneeweiss, and Stürmer 2006). This is particularly common in pharmacoepidemiology, where outcomes are often rare compared to a large number of covariates available for estimation of (adverse) drug effects (Glynn, Schneeweiss, and Stürmer 2006). Cepeda et al. (2003) proposed a helpful guideline on when to use PS methods to effectively improve estimation (fewer than eight outcomes per included covariate). PS methods in general and PS matching in particular could also help to reduce “the dependence of causal inference on hard-to-justify but commonly made statistical modeling assumptions” and allow for a simple, transparent analysis (Glynn, Schneeweiss, and Stürmer 2006; Ho et al. 2007). In addition, unlike most multivariable regression models, assumptions regarding model specifications, variable functional form, normality, and linear projections beyond the observed data are not required particularly when matching on PS (Glynn, Schneeweiss, and Stürmer 2006; Zanutto 2006).

THE TREATMENT EFFECT ESTIMATE (THE ESTIMAND)

When addressing the bias in the treatment effect estimate, Brooks and Ohsfeldt compared effect estimates from different approaches, including regression and several PS methods. Furthermore, they concluded that PS methods can exacerbate the bias in treatment effect estimates when there are unmeasured confounders related to measured confounders. We agree that PS methods may not reduce bias from unmeasured confounders, but we would like to point out a potential drawback of direct comparison of estimates from

PS methods and regression analysis where covariates are included in the adjustment model.

In studies in which non-linear models (such as logistic regression or Cox proportional hazard model) are applied, effect estimates may differ between PS methods and regression adjustment methods not only due to differential adjustment for confounding but also due to noncollapsibility (Greenland, Robins, and Pearl 1999). Noncollapsibility is the phenomenon that when estimating the treatment-outcome association using an odds ratio (OR) or hazard ratio (HR), the conditional OR or HR does not equal the marginal OR or HR in the presence of non-null treatment effect, even in the absence of confounding and effect modification (Greenland, Robins, and Pearl 1999; Austin 2007; Martens et al. 2008). Adjusting for covariates that are predictors of the outcome will change the treatment effect estimate, even when there is no confounding present. Obviously, the number of covariates included in the adjustment model can be very different when comparing PS matching with regression analysis where covariates are included in the adjustment model and thus a direct comparison may be flawed.

In addition, while both PS and regression methods provide different effect estimates, the inferential goal of the research question determines which estimand is appropriate. For example, in the absence of noncollapsibility (e.g., with linear models), marginal structural modeling using inverse probability of treatment weighing, multivariable regression, covariate adjustment using PS, and PS stratification estimate the average treatment effect in the population, ATE, the treatment effect estimate obtained from RCTs (i.e., the average causal treatment effect if everyone in the population is treated vs if everyone in the population is untreated) (Robins, Hernán, and Brumback 2000; Hernán, Brumback, and Robins 2000; Fang, Brooks, and Chrischilles 2012). On the other hand, PS matching typically focuses on either the average treatment effect in the treated or the average treatment effect in the untreated, not on the ATE; hence, the target of the causal contrast being the population that is going to receive the treatment (Hill 2008; Rassen et al. 2012). This is particularly important when there is treatment-effect modification regardless of the presence of confounding (Fang, Brooks, and Chrischilles 2012; Stürmer et al. 2006). In the study by Brooks and Ohsfeldt, exacerbation in bias was not clearly evident when using PS to balance measured covariates compared with regression estimates even in the presence of independent variation in the unmeasured covariate (example, Scenarios 1 and 2

of Table 3) despite inappropriate comparisons of treatment effect estimates from different models.

CONCLUDING REMARKS

The authors should be commended for raising an important topic in PS methodology, namely the implications of balancing measured covariates using PS on the (im)balance of unmeasured covariates. However, we disagree with Brooks and Ohsfeldt's statement that systematic differences in the unmeasured covariates is required for PS algorithms to balance measured covariates between treated and untreated subjects. In addition, the impact of unmeasured covariate imbalance on the bias of estimated treatment effects, if any, cannot be inferred from their study. This is due to the fact that the estimands are different and direct comparison may not be appropriate. Therefore, the findings of Brooks and Ohsfeldt should be interpreted with caution and further research is needed to evaluate the balancing properties of PS with respect to unmeasured confounding using properly designed simulation settings.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The research leading to these results was conducted as part of the PROTECT consortium (Pharmacoepidemiological Research on Outcomes of Therapeutics by a European Consortium), which is a public-private partnership coordinated by the European Medicines Agency.

Funding: The PROTECT project has received support from the Innovative Medicines Initiative Joint Undertaking (www.imi.europa.eu) under grant agreement no. 115004, resources of which are composed of financial contributions from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. In the context of IMI Joint Undertaking (IMI JU), the Division of Pharmacoepidemiology, Utrecht University, also received a direct financial contribution from Pfizer. The views expressed are those of authors only and not of their respective institution or company.

Conflict of Interest: Olaf Klungel received unrestricted funding for pharmacoepidemiological research from the Dutch private-public funded Top Institute Pharma.

REFERENCES

- Austin, P. C. 2007. "The Performance of Different Propensity Score Methods for Estimating Marginal Odds Ratios." *Statistics in Medicine* 26 (16): 3078–94.
- . 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioural Research* 46 (3): 399–424.
- Brooks, J. M., and R. L. Ohsfeldt. 2013. "Squeezing the Balloon: Propensity Scores and Unmeasured Covariate Balance." *Health Service Research* 48 (4): 1487–507.
- Cepeda, M. S., R. Boston, J. T. Farrar, and B. L. Strom. 2003. "Comparison of Logistic Regression Versus Propensity Score When the Number of Events Is Low and There Are Multiple Confounders." *American Journal of Epidemiology* 158 (3): 280–7.
- Cole, S. R., and M. A. Hernán. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168 (6): 656–64.
- Fang, G., J. M. Brooks, and E. A. Chrischilles. 2012. "Apples and Oranges? Interpretations of Risk Adjustment and Instrumental Variable Estimates of Intended Treatment Effects Using Observational Data." *American Journal of Epidemiology* 175 (1): 60–5.
- Glynn, R. J., S. Schneeweiss, and T. Stürmer. 2006. "Indications for Propensity Scores and Review of Their Use in Pharmacoepidemiology." *Basic & Clinical Pharmacology & Toxicology* 98 (3): 253–9.
- Greenland, S., J. M. Robins, and J. Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 14 (1): 29–46.
- Hernán, M. Á., B. Brumback, and J. M. Robins. 2000. "Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men." *Epidemiology* 11 (5): 561–70.
- Hill, J. 2008. "Discussion of Research Using Propensity Score Matching: Comments on 'A Critical Appraisal of Propensity Score Matching in the Medical Literature between 1996 and 2003' by Peter Austin, Statistics in Medicine." *Statistics in Medicine* 27 (12): 2055–61.
- Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Non-parametric Pre-processing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15 (3): 199–236.
- Jo, B., and E. A. Stuart. 2009. "On the Use of Propensity Scores in Principal Causal Effect Estimation." *Statistics in Medicine* 28 (23): 2857–75.
- Martens, E. P., W. R. Pestman, A. de Boer, S. V. Belitser, and O. H. Klungel. 2008. "Systematic Differences in Treatment Effect Estimates between Propensity Score Methods and Logistic Regression." *International Journal of Epidemiology* 37 (5): 1142–7.
- Petersen, M. L., K. E. Porter, S. Gruber, Y. Wang, and M. J. van der Laan. 2010. "Diagnosing and Responding to Violations in the Positivity Assumption." *Statistical Methods in Medical Research* 21 (1): 31–54.
- Rassen, J. A., and S. Schneeweiss. 2012. "Using High-Dimensional Propensity Scores to Automate Confounding Control in a Distributed Medical Product Safety Surveillance System." *Pharmacoepidemiology and Drug Safety* 21 (S1): 41–9.

- Rassen, J. A., A. A. Shelat, J. Myers, R. J. Glynn, K. J. Rothman, and S. Schneeweiss. 2012. "One-to-Many Propensity Score Matching in Cohort Studies." *Pharmacoeconomics and Drug Safety* 21 (S2): 69–80.
- Robins, J. M., M. A. Hernán, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (5): 550–60.
- Rosenbaum, P. R. 2005. "Propensity Score." In *Encyclopaedia of Biostatistics*, 2nd Edition, edited by P. Armitage, and T. Colton, pp. 4267–72. Boston, MA: Wiley.
- Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.
- . 1984. "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score." *Journal of the American Statistical Association* 79 (387): 516–24.
- Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika* 63 (3): 581–92. (with Discussion and Reply).
- . 1977. "Assignment to Treatment Group on the Basis of a Covariate." *Journal of Educational Statistics* 2 (1): 1–26. Printer's correction note 3, 384.
- Sebastian, S., A. R. Jeremy, R. J. Glynn, J. Avorn, H. Mogun, and M. A. Brookhart. 2009. "High-Dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data." *Epidemiology* 20 (4): 512–22.
- Shah, B. R., A. Laupacis, J. E. Hux, and P. C. Austin. 2005. "Propensity Score Methods Gave Similar Results to Traditional Regression Modelling in Observational Studies: A Systematic Review." *Journal of Clinical Epidemiology* 58 (6): 550–9.
- Stürmer, T., K. J. Rothman, and R. J. Glynn. 2006. "Insights into Different Results from Different Causal Contrasts in the Presence of Effect-Measure Modification." *Pharmacoeconomics and Drug Safety* 15 (10): 698–709.
- Stürmer, T., S. Schneeweiss, J. Avorn, and R. J. Glynn. 2005. "Adjusting Effect Estimates for Unmeasured Confounding with Validation Data Using Propensity Score Calibration." *American Journal of Epidemiology* 162 (3): 279–89.
- Stürmer, T., M. Joshi, R. J. Glynn, J. Avorn, K. J. Rothman, and S. Schneeweiss. 2006. "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." *Journal of Clinical Epidemiology* 59 (5): 437–47.
- Westreich, D., and S. R. Cole. 2010. "Invited Commentary: Positivity in Practice." *American Journal of Epidemiology* 171 (6): 674–7. discussion 678–81.
- Zanutto, E. L. 2006. "A Comparison of Propensity Score and Linear Regression Analysis of Complex Survey Data." *Journal of Data Science* 4: 67–91.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.