Published in final edited form as:
Am J Epidemiol. 2007 ; 165(10): 1122–1123.

Propensity Score Calibration and its Alternatives

Til Stürmer, **Sebastian Schneeweiss**, **Kenneth J. Rothman**, **Jerry Avorn**, and **Robert J. Glynn**

We appreciate the thoughtful commentary of Oakes and Church (1) on our paper (2) and their conclusion that propensity score calibration may be helpful when some confounders are unmeasured. We agree that usual applications of propensity scores only control for confounding by "observable selection" but we see much closer links between instrumental variables (3–5) and propensity score calibration than those described by Oakes and Church. Indeed, the gold-standard propensity score estimated in the validation study hopefully better approaches the true, but unknown, propensity of treatment than the error-prone one and thus performs as an approximate instrument under assumptions similar to surrogacy (6,7).

Propensity score calibration is no panacea for missing data on confounders – there is no substitute for having good data on important confounders for every subject. Propensity score calibration was developed in a pharmacoepidemiologic analysis of claims data that lack information on a variety of confounders (8). Using data from a validation study, we obtained an estimate of the association between nonsteroidal anti-inflammatory drugs and short-term all-cause mortality in older adults that was more plausible than the naïve estimate.(9) We now briefly respond to the 6 issues raised by Oakes & Church (1):

1. The low precision of the estimation with a cohort of N=1,000 is due to the very low expected number of outcomes (N=10). We would not call this low precision an anomaly because the median OR is still unbiased.
2. The scope of our simulations does not yet allow us to propose a sharp criterion to decide whether the surrogacy assumption is valid. The assessment of surrogacy is dependent on having outcome data in the validation study. With such data available, other methods, including imputation, are promising alternatives to propensity score calibration (10). Unfortunately, validation studies do not always contain outcome information. In such settings, propensity score calibration might be the best possibility for bias reduction. Important violations of surrogacy could be explored by considering factors measured in the validation study individually in combination with literature estimates of their independent effect on the outcome.(11)
3. We did not address how closely the validation sample needs to be representative of the main study and there clearly are dangers in estimating the measurement error model in an external validation study (6,9). This will be an important judgment that investigators will have to make when applying propensity score calibration.
4. Should the estimation of the measurement error model be included in the bootstrap method? The usual implementation of regression calibration takes the estimation of the measurement error model into account (12) but provided variance estimates that were too small compared with the empirical variance over simulations. We

Corresponding Author: Til Stürmer, MD, MPH, Associate Professor of Medicine, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Harvard Medical School, 1620 Tremont Street, Suite 3030, Boston, MA 02120, USA, Phone:+1 617 278-0627, Fax:+1 617 232-8602, eMail: til.sturmer@post.harvard.edu.

therefore used conditional mean imputation, matching, and the bootstrap for matched pairs to implement propensity score calibration, resulting in variance estimates that were close to the empirical ones (2).

5. Because we match subjects, exposed subjects for whom no unexposed match can be found, owing to non-overlap, are automatically excluded from the analysis. Non-overlap will tend to increase with propensity score calibration, because the gold-standard propensity score is at least as strongly associated with the exposure as the error-prone one. Investigators should carefully assess exposed subjects excluded from estimation, because the estimate might not be generalizable to them. (13)

6. Design aspects of validation studies need more attention. In pharmacoepidemiologic research based on routinely collected data, the scope of covariates that one would like to control, beyond those already contained in the administrative data, might include e.g., smoking, body mass index, physical activity, activities of daily living, and cognitive function.(9) But certainly some potential confounders and their measurements will always be elusive.

References

1. Oakes JM, Church TR. Advancing propensity score methods in epidemiology. Am J Epidemiol. 2007
2. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Performance of propensity score calibration – a simulation study. Am J Epidemiol. 2007
3. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996; 81:444–55.
4. Brookhart MA, Wang PS, Solomon DH, Schneeweiss S. Evaluating short-term drug effects using a physician-specific prescribing preference as an instrumental variable. Epidemiology. 2006; 17:268–75. [PubMed: 16617275]
5. Glynn RJ. Commentary: genes as instruments for evaluation of markers and causes. Int J Epidemiol. 2006; 35:932–4. [PubMed: 16854935]
6. Carroll, RJ.; Ruppert, D.; Stefanski, LA. Measurement error in nonlinear models. Chapman/Hall; London: 1995.
7. Buzas JS, Stefanski LA. Instrumental Variable Estimation in Generalized Linear Measurement Error Models. J Am Stat Assoc. 1996; 91:999–1006.
8. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. J Clin Epidemiol. 2005; 58:323–37. [PubMed: 15862718]
9. Stürmer T, Schneeweiss S, Avorn J, Glynn RJ. Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. Am J Epidemiol. 2005; 162:279–89. [PubMed: 15987725]
10. Stürmer T, Schneeweiss S, Rothman KJ, Avorn J, Glynn RJ. Comparison of Performance of Propensity Score Calibration (PSC) and Multiple Imputation (MI) to Control for Unmeasured Confounding Using an Internal Validation Study. [abstract]. Pharmacoepidemiol Drug Saf. 2006; 15:S39.
11. Schneeweiss S, Glynn RJ, Tsai EH, Avorn J, Solomon DH. Adjusting for unmeasured confounders in pharmacoepidemiologic claims data using external information: The example of COX2 inhibitors and myocardial infarction. Epidemiology. 2005; 16:17–24. [PubMed: 15613941]
12. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. Am J Epidemiol. 1990; 132:734–45. [PubMed: 2403114]
13. Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. Basic Clin Pharmacol Toxicol. 2006; 98:253–9. [PubMed: 16611199]