Response to Invited Commentary

Stürmer et al. Respond to “Propensity Score Methods in Epidemiology”

Til Stürmer1,2, Sebastian Schneeweiss1,3, Kenneth J. Rothman1,4,5, Jerry Avorn1, and Robert J. Glynn1,2,6

1 Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA.
2 Division of Preventive Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA.
3 Department of Epidemiology, Harvard School of Public Health, Boston, MA.
4 Department of Epidemiology, Boston University School of Public Health, Boston, MA.
5 Research Triangle Institute, Research Triangle Park, NC.
6 Department of Biostatistics, Harvard School of Public Health, Boston, MA.

Received for publication January 29, 2007; accepted for publication January 31, 2007.

Abbreviation: PSC, propensity score calibration.

We appreciate the thoughtful commentary of Oakes and Church (1) on our paper (2) and their conclusion that propensity score calibration (PSC) may be helpful when some confounders are unmeasured. We agree that usual applications of propensity score methods control only for confounding by “observable selection,” but we see much closer links between instrumental variables (3–5) and PSC 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 propensity score and thus performs as an approximate instrument under assumptions similar to surrogacy (6, 7).

PSC is no panacea for missing data on confounders—there is no substitute for having good data on important confounders for every subject. PSC 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 use of nonsteroidal antiinflammatory drugs and short-term all-cause mortality in older adults (9) that was more plausible than the naïve estimate. Below, we briefly respond to six issues raised by Oakes and Church in their commentary (1).

The low precision of the estimation with a cohort of 1,000 was due to the very low expected number of outcomes ($n = 10$). We would not call this low precision an anomaly, because the median odds ratio is still unbiased.

The scope of our simulations does not yet allow us to propose a sharp criterion for deciding 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 PSC (10). Unfortunately, validation studies do not always contain outcome information. In such settings, PSC 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).

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 parameters of 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 PSC.

Should estimation of the parameters of the measurement error model be included in the bootstrap method? The usual...
implementation of regression calibration takes estimation of the measurement-error model parameters into account (12), but in our study it provided variance estimates that were too small compared with the empirical variance over simulations. Therefore, we used conditional mean imputation, matching, and the bootstrap for matched pairs to implement PSC, which resulted in variance estimates that were close to the empirical ones (2).

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

Lastly, 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 over-the-counter drug use, smoking, body mass index, physical activity, activities of daily living, and cognitive function (9). Certainly, however, some potential confounders and their measurements will always be elusive.

ACKNOWLEDGMENTS

This project was funded by a grant (ROI 023178) from the National Institute on Aging.

Dr. Til Stürmer does not accept personal compensation of any kind from industry but has received salary support from unrestricted research grants from the pharmaceutical industry to the Brigham and Women’s Hospital. Dr. Stürmer does not have a conflict of interest regarding the content of this manuscript. Dr. Robert J. Glynn has received grant support from AstraZeneca, Bristol-Myers Squibb, Merck & Company, Novartis International AG, and Pfizer Inc.

REFERENCES