Invited Commentary

Invited Commentary: Advancing Propensity Score Methods in Epidemiology

J. Michael Oakes and Timothy R. Church

Received for publication December 13, 2006; accepted for publication December 15, 2006.

Every epidemiologist knows that unmeasured confounding is a serious analytic problem, but practically speaking, there seems to be little one can do about it. In this issue of the Journal, Stürmer et al. (Am J Epidemiol 2007:165:1110–18) offer a novel solution that combines propensity score matching methods with measurement error regression models. They call this technique “propensity score calibration” (PSC) and assess its strengths and limitations with simulated data. Their analyses demonstrate that PSC greatly improves inference when the critical assumption of surrogacy holds, but when surrogacy does not hold, PSC estimation can exacerbate bias relative to uncorrected propensity score models. The benefits of propensity score methods (and PSC) lie not only with potentially improved effect estimation but with conceptualization and practice as well.

bias (epidemiology); cohort studies; confounding factors (epidemiology); epidemiologic methods; models, statistical; propensity score calibration; research design

Abbreviations: IV, instrumental variable; OR, odds ratio; PSC, propensity score calibration.

In this issue of the Journal, Stürmer et al. (1) have contributed yet another important paper on the use of propensity score methods in epidemiology. They use simulation to reveal some of the benefits and pitfalls of adjusting a propensity score model to account for unmeasured covariates/confounders. The paper is noteworthy for its novelty, methodological insight, and clarity. Here we highlight central aspects of the paper and offer some thoughts on the use and advancement of propensity score methods in epidemiology. Our target audience is not expert methodologists but rather those who wonder whether this alternative technique may be helpful in their own research.

THE PROBLEM

Adjustment by stratification, restriction, and especially regression is frequently used to control for confounding by measured variables. But what of the unmeasured confounders (and/or mismeasured confounders that yield residual confounding)? While discussing instrumental-variable (IV) methods, Hernán and Robins (2, p. 360) recently asked and answered this dreaded question with abundant clarity: “Can you guarantee that the results from your observational study are unaffected by unmeasured confounding? The only answer an epidemiologist can provide is ‘no’."

Beyond hope and disregard, improved theory, better measurement, sensitivity analysis, bounding, imputation, IV techniques, marginal structural models, and more recently Bayesian methods have been the principal tools for addressing unmeasured confounding. Stürmer et al. (1) have advanced propensity score methods to do the same, and in their paper they show when and why the new technique could be useful.

THE ADVANCE

Though use is rapidly increasing, epidemiologists have been slow to adopt propensity score methods (3, 4). Technical details may be found elsewhere, but the basic idea is to 1) use standard logistic regression to estimate the probability of exposure (i.e., the propensity score) for each subject in your data, 2) construct sets of subjects who have similar propensity score estimates in hopes that the sets (often matched
pairs) are exchangeable in every way but for the exposure itself, and 3) estimate the causal effects of exposure by, say, conditional logistic regression contrasting persons actually exposed and those not actually exposed. The key is getting the propensity score model right, since errors yield improper matches and thus biased effect estimates. Effect estimates will be biased if there are omitted or poorly measured predictors of exposure that undermine the exchangeability requirement; the same holds true if the predictors themselves are affected by the exposure. For this reason, economists have long considered propensity score methods useful only for estimates with “observable selection,” relying instead on IV estimation for overcoming the far more common problem of “unobservable selection” (5).

Stürmer et al. have shown the extent to which this limitation of propensity score methods can be overcome by combining standard measurement error models and subsample validation techniques. The basic idea is to treat the estimated propensity scores as a mismeasured variable—mismeasured because of the omission of important predictors of exposure. Adjustment of the error-prone propensity score is done by exploiting better information in a validation subsample and, in Stürmer et al.’s paper (1), the application of regression-calibration techniques to adjust the propensity score estimates in the full set of data. The corrected, or calibrated, propensity score is then used for analyses of outcome measures as usual. Standard errors of effect estimates are typically calculated through resampling techniques (e.g., bootstrapping), but other approaches may be considered (6).

In their paper (1), Stürmer et al. assess the performance of the propensity score calibration (PSC) technique by simulation. They begin with reasonable values for some arbitrary disease-exposure-confounder relations and then alter the parameters to assess the impact. Importantly, they relax the key assumption of surrogacy (discussed below).

Their table 2 presents results obtained when surrogacy holds. For point estimate evaluation, one should appreciate that the true effect (the odds ratio (OR) for the exposure) holds. For point estimate evaluation, one should appreciate parameters to assess the impact. Importantly, they relax the key assumption of surrogacy (discussed below).

What is surrogacy and how might it be assessed? Stürmer et al. and Carroll et al. (7, p. 36) offer technical definitions and useful examples. Surrogacy is the state of conditional independence: \( f(y|X, W) = f(y|X) \). In what may be more familiar terms, given a traditional linear model \( E(Y) = Xb \), a covariate \( W \) is a surrogate for covariate \( X \) when \( W \) is not correlated with \( Y \) after adjustment for \( X \). That is, when it comes to \( Y \), there is no information in \( W \) beyond that contained in \( X \). Surrogacy implies nondifferential measurement error between an outcome (which in this case is the probability of exposure) and covariates. It is a vital assumption. Such a state will be found in idealized randomized trials but may be difficult to assess in observational studies (8). In Stürmer et al.’s simulations (1), surrogacy is violated when the direction of confounding in the measured covariates differs from the direction of confounding in the unmeasured covariates. This is because such directional difference tends to introduce differential measurement error, a condition where errors are correlated with the estimated propensity score. Surrogacy would seem to be routinely violated, and effect estimates exaggerated, when estimating contextual effects through aggregate proxies of microlevel data, as in the recent stream of papers on neighborhood effects (9). Indeed, surrogacy is at risk whenever survey responses are employed (10). It is noteworthy that surrogacy appears to be related to the “exclusion restrictions” found in IV estimation.

CONCLUSION

PSC may be a helpful technique for addressing unmeasured confounding, especially when IV techniques are not feasible or desired. Yet we trust that Stürmer et al. agree that it is no panacea, and more work is needed before researchers should routinely adopt the technique. Setting aside the need for direct comparisons with other techniques such as IV estimation and multiple bias modeling (11, 12), at least five questions about PSC merit immediate attention.

First, why the anomalous results when the main cohort size is 1,000? What sort of “specification error” might cause this? We presume that better results would be obtained if the proportionate size of the validation study exceeded 100 subjects (a default level of 10 percent of the main study size).

Second, what is the best measure or measures with which to assess surrogacy? We agree with Stürmer et al. that the pseudo-\( R^2 \) value may be best, but given the critical nature of the assumption, confirmatory work, especially with respect to meaningful cutoffs, is a must. It seems that the methods employed in assessing identification in IV techniques might also be useful here.

Third, how closely must validation subsamples represent main study subjects in order for benefits to obtain? There is a great deal of room between perfectly representative subsamples (as employed here) and disparate useless ones. Practicing epidemiologists must know the degree of exchangeability (a stronger term than the “transportability” used by some) required in subsamples before they can be
confident in the PSC estimates. Some recent work in economics (13) may turn out to be a useful addition to the epidemiologic literature.

Fourth and relatedly, we have some concern that the reported coverage probability of the bootstrap confidence interval obtained using PSC may not be accurate, especially in situations where the true odds ratio is not equal to 1. This possibly stems from the fact that these intervals appear to have been constructed by bootstrapping only the full study estimation procedure instead of including the validation study estimation inside the bootstrap loop. The concern is not great, since the reported coverage probabilities are not wildly inaccurate, and they certainly exceed those of the conventional estimator using only the full study data. Nevertheless, attention seems appropriate.

Finally, how does PSC perform when the overlap between exposed and unexposed subjects is incomplete and/or imperfect? Consideration of unmatched subjects is essential if we are to minimize off-support inferences, which is akin to comparing apples to oranges (14–16). Does PSC exacerbate or mitigate the problem, which seems especially common in areas such as social epidemiology?

Another issue, more logical in nature, concerns the design of the validation subsample. For PSC to work, confounders not measured in the main study need to be measured with precision in the subsample. But how are we to know what to measure? The case of variables that are merely too expensive to measure in a main study is straightforward, but better etiologic understanding is needed if we are to know what we do not know.

In summary, while certainly no panacea, the benefits of propensity score methods (and PSC) lie not only with potentially improved effect estimation but with conceptualization and practice as well. First, setting aside outcome measures and predicting exposure is vastly superior to naïve modeling of outcomes directly, since the latter invites regression screening and inflated type I errors (17, 18). Second, propensity score matching methods follow naturally from counterfactual reasoning (19). Getting from counterfactuals to regression estimates is far more complicated and thus less transparent. Propensity score methods tend to remind analysts that models, especially less transparent regression models, substitute assumptions for data. Recall that if we had excellent data (on counterfactuals), simple cross-tabulations would suffice. Third, propensity score matching methods permit the use of diagnostics for assessing inferential support. Although high-dimensional matching through propensity scores requires strong assumptions, the ability to exclude aberrant (nonmatched) cases is virtually impossible in conventional regression approaches, at least as typically practiced. Finally, the fact that the method requires investigators to consider the experimental analog to their observational study may be benefit enough. By requiring them to appreciate that exchangeability is obtained with both idealized experiments and idealized propensity scores, the method makes the link between experiments and observational designs abundantly clear—namely, ignorable treatment/exposure assignment mechanisms.

ACKNOWLEDGMENTS

Conflict of interest: none declared.

REFERENCES