PHARMACOEPIDEMIOLOGY AND DRUG SAFETY (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/pds.968

EDITORIAL

On principles for modeling propensity scores in medical research

It is clearly important to document how a proposed statistical methodology is actually used in practice if that practice is to be improved. This target article by Weitzen *et al.*, reviewing the way propensity score methods are used in current medical research, therefore, is an important contribution, and I am delighted to have been invited by the editorial board to discuss it.

I am a firm believer in the utility of propensity scores for application in observational studies for causal effects, not as a panacea for the deficiencies of observational studies, but as a critical tool contributing to their appropriate design. The target article reveals that many published articles employing propensity scores for medical research are not taking full advantage of the technology, and some may even be misusing it. A possible reason, as indicated by the authors, may be confusion between two kinds of statistical diagnostics: (i) diagnostics for the successful prediction of probabilities and parameter estimates underlying those probabilities, possibly estimated using logistic regression; and (ii) diagnostics for the successful design of observational studies based on estimated propensity scores, possibly estimated using logistic regression. There is no doubt in my mind that (ii) is a critically important activity in most observational studies, whereas I am doubtful about the importance of (i) in most of these.

At the outset, it is essential to realize that observational studies should be designed in analogy with the way randomized experiments are designed. This is a theme that can be traced to classical work[1] in observational studies, and a theme I have recently emphasized in the context of the tobacco litigation.[2] When we design a randomized experiment, we cannot try one randomization and see the answer, then try another randomization and see the answer, and continue until we find an answer that is 'satisfactory' for publication. Randomized experiments are designed blind to the answer, and this is one of the most important features of randomized experiments. It is a feature that can also be shared with observational studies, although sadly, observational studies are often not conducted this way.

Randomized experiments are designed to have balance between treatment and control groups, often within blocks (i.e. within strata, subclasses or matched pairs) on all covariates. Blocking assures balance on the observed covariates used to create the blocks, and randomization implies balance (at least on average) on all other covariates, both observed and unobserved. Due to the absence of randomization in observational studies, we cannot force balance on unobserved covariates, but we must attempt to balance the observed ones (at least on average), and propensity score technology, often combined with blocking on especially important covariates, is an important tool for achieving this balance in observed covariates. In this sense, propensity score technology is the observational study analog of randomization in experiments; randomization is superior in a critical way, however, because it achieves this average balance on all covariates, both observed and unobserved, whereas propensity score methods only operate on observed covariates.

If this balance is achieved in an observational study—that is, if the treatment and control groups have very similar distributions of the observed covariates within blocks (subclasses, matched pairs etc.) of the propensity score (perhaps crossed by blocks on critical covariates)—then it really makes no difference, for estimation of effects controlling for these covariates, as to how this balance was achieved. Within blocks balanced on propensity scores, future model-based adjustments for distributional differences between treatment and control groups (e.g. using linear covariance, relative risks, proportional hazards) will typically have only minor effects on point estimates, although they can have important effects on estimated precisions, and therefore, on interval estimates.

*Correspondence to: Professor D. B. Rubin, Chairman, Department of Statistics, Harvard University, Cambridge, MA 02138, USA.
E-mail: rubin@stat.harvard.edu

Received 7 January 2004
Revised 27 January 2004
Accepted 28 January 2004

Just as in a randomized experiment, covariate balance is assessed without any access to the outcome variables. Also, the resulting blocking is common and fixed by design for all outcome variables, just as in a randomized experiment. For example, in a randomized experiment estimating the effect of hormone replacement therapy on heart disease, cancer rates etc., the randomization and the blocking are the same for each of these outcomes.

Therefore, if propensity scores are to be used in design to balance covariates, it is the distributional balance of those covariates that is achieved within blocks (strata, subclasses or matched pairs) that is the critical diagnostic tool. This advice, however, does not completely solve the problem of which distributional diagnostics to examine. If the yet to be observed outcome variables (or, e.g. their logits) are thought to be approximately linearly related to the covariates, then it is critical that the means of the covariates be assessed for balance. If, it is thought instead that the logs of these covariates are approximately linearly related to the outcomes, then the means of the logs should be assessed for balance. If, it is thought that such dependencies might also involve second order terms and interactions among covariates, then the variances and correlations of the covariates should also be checked for balance within blocks. If some outcomes are thought to be linearly related to a covariate x and others linearly related to $\log(x)$, then the means of both x and $\log(x)$ should be assessed for balance. And so forth. The topic of exactly how to do this assessment is important, but beyond the scope of this brief discussion.

Of course, at some point, this sort of assessment must terminate, because no matter how large the samples, the investigator will almost certainly not be able to achieve this balance for many covariates simultaneously, and higher order terms in minor covariates are clearly less important than means of important ones, and so scientific judgment must enter the process, just as it does when designing a randomized experiment. If balance cannot be achieved on important covariates in a particular observational data set, then the inescapable conclusion must be that inferences for the treatment-control effect cannot be reliably drawn in the study population, unless either the inferences are restricted to a subpopulation where such balance can be achieved, or heroic assumptions (e.g. linearity of outcomes on covariates—which justifies linear extrapolation, or an assertion of irrelevance of a previously considered covariate) are made. The conclusion that the dataset cannot support any valid conclusion about treatment effects may be the right one.

In rare situations, the individually estimated probabilities (i.e. the estimated propensity scores) themselves may be used in the process of estimating treatment effects, but this is not the usual case in the design of observational studies. If it is, the propensity score estimation has to be conducted far more carefully. One example occurs when the estimated propensities are to be used as inverse weights for weighting adjustments.[3,4] In such cases, the estimated probabilities can be very influential on the estimated effects of treatment versus control, and so the probabilities themselves must be very well-estimated. In such cases, diagnostics on the accuracy of the estimated probabilities are appropriate, although diagnostics on the estimated underlying (logistic) regression coefficients still are generally irrelevant.[5,6]

An application somewhat similar to such weighting adjustments involves the use of estimated propensity scores in a model-based adjustment relating the outcome to a set of covariates, where the scalar estimated propensity score simply takes the place of the full set of the covariates. This method can be effective when nonlinear terms in the propensity score are included in the model, but in general the method must be used cautiously, and the user must be confident that the propensity scores are well estimated, at least up to a monotone transformation. I was disappointed to see that many of the articles selected for review in the target article appeared simply to perform a linear covariance adjustment for the estimated propensity scores. Perhaps one reason this generally inferior—or often even mistaken—approach to the use of propensity scores occurs is because of the almost universal focus on analysis rather than design in epidemiology and the statistics of observational studies. In fact, I disagree with the characterization of propensity score technology as one technique in the class of so-called 'structural marginal models' in the authors' opening paragraph—this is an inapposite characterization that focuses propensity score technology on analysis rather than on design and diagnostics for design, which I feel are far more important uses.

In conclusion, I congratulate the authors for their careful review of the current use of propensity scores in medical research, and I hope that my comments are a further contribution to the goal of improving this practice.

Donald B. Rubin*
Department of Statistics
Harvard University
Cambridge, MA
02138, USA
E-mail: rubin@stat.harvard.edu

REFERENCES

1. Rubin DB, William G. Cochran's contributions to the design, analysis and evaluation of observational studies. In *W.G. Cochran's Impact on Statistics*, Rao SRS, Sedransk J (eds). Wiley: New York, 1984; 37–69.
2. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol* 2002; **2**: 169–188.
3. Czajka JC, Hirabayashi SM, Little RJA, Rubin DB. Projecting from advance data using propensity modelling. *J Bus Econ Stat* 1992; **10**: 117–131.
4. Imbens GW. The role of the propensity score in estimating dose–response functions. *Biometrika* 2000; **87**: 706–710.
5. Landwehr JM, Pregibone D, Shoemaker AC. Graphical methods for assessing logistic regression models. *J Am Stat Assoc* 1985; **79**: 61–71.
6. Rubin DB. Assessing the fit of logistic regressions using the implied discriminant analysis. Discussion of 'graphical methods for assessing logistic regression models' by Landwehr, Pregibone and Smith. *J Am Stat Assoc* 1985; **79**: 79–80.