LETTER TO THE EDITOR

Remarks on the method of propensity score

From: Judea Pearl,
UCLA Computer Science and Statistics,
Los Angeles, *CA*, *U.S.A.*

Dear Editor,

I read with great interest Donald Rubin's paper 'The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials' (2007; **26**(1):20–36) [1], as well as the Letter To The Editor by Ian Shrier (2008; **27**(14):2740–2741) [2], and Author Reply by Don Rubin (2008; **27**(14):2741–2742) [3].

Shrier's Letter posed an important question that remains unanswered in Rubin's reply. I here venture to answer this question and to clarify related issues concerning the interpretation of propensity scores (PS) and their role in causal inference.

Shrier's question was whether, asymptotically, the use of PS methods as described by Rubin may actually *increase*, not decrease, bias over and above a crude, unadjusted comparison between treated and untreated subjects. The answer is: Yes, and the M-graph cited by Shrier (see also [4, 5]) provides a simple example; the crude estimate is bias-free, while PS methods introduce new bias.

This occurs when treatment is strongly ignorable to begin with and becomes non-ignorable at some levels of e_i. In other words, although treated and untreated units are balanced in each stratum of e_i, the balance only holds relative to the covariates measured; unobserved confounders may be highly unbalanced in each stratum of e_i, capable of producing significant bias. Moreover, such imbalance may be dormant in the crude estimate and awakened through the use of PS methods.

There are other features of PS methods that are worth emphasizing [6].

First, the PS e_i is a probabilistic, not a causal concept. Therefore, in the limit of very large sample, PS methods are bound to produce the same bias as straight stratification on the same set of measured covariates. They merely offer an effective way of approaching the asymptotic estimate which, due to the high dimensionality of X, is practically unattainable with straight stratification. Still, the asymptotic estimate is the same in both cases, and may or may not be biased, depending on the set of covariates chosen.

Second, the task of choosing a sufficient (i.e. bias-eliminating) set of covariates for PS analysis requires qualitative knowledge of the causal relationships among both observed and unobserved covariates. Given such knowledge, finding a sufficient set of covariates or deciding whether a sufficient set exists are two problems that can readily be solved by graphical methods [4, 6, 7].

Finally, experimental assessments of the bias-reducing potential of PS methods (such as those described in Rubin [1]) can only be generalized to cases where the causal relationships among covariates, treatment, outcome and unobserved confounders are the same as in the experimental study. Thus, a study that proves bias reduction through the use of covariate set X does not justify the use of X in problems where the influence of unobserved confounders may be different.

In summary, the effectiveness of PS methods rests critically on the choice of covariates, X, and that choice cannot be left to guesswork; it requires that we understand, at least figuratively, what relationships may exist between observed and unobserved covariates and how the choice of the former can bring about strong ignorability or a reasonable approximation thereof [6].

REFERENCES

1. Rubin D. The design *versus* the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine* 2007; **26**:20–36.
2. Shrier I. Letter to the editor. *Statistics in Medicine* 2008; **27**:2740–2741.
3. Rubin D. Author's reply (to Ian Shrier's Letter to the Editor). *Statistics in Medicine* 2008; **27**:2741–2742.
4. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. *Epidemiology* 1999; **10**(1):37–48.
5. Greenland S. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology* 2003; **14**:300–306.
6. Pearl J. *Causality*: *Models*, *Reasoning*, *and Inference* (2nd edn). Cambridge University Press: New York, 2009, forthcoming.
7. Pearl J. Comment: graphical models, causality, and intervention. *Statistical Science* 1993; **8**(3):266–269.

LETTER TO THE EDITOR

Propensity scores and M-structures

Arvid Sjölander*, †

Department of Medical Epidemiology and Biostatistics, *Karolinska Institutet*, *Stockholm*,
Nobels väg 12, *171 77 Stockholm*, *Sweden*

SUMMARY

In a recent issue of *Statistics in Medicine*, Ian Shrier [*Statist. Med.* 2008; **27**(14):2740–2741] posed a question regarding the use of propensity scores [*Biometrika* 1983; **70**(1):41–55]. He considered an 'M-structure' illustrated by the directed acyclic graph (DAG) in Figure 1. In Figure 1, z is a binary exposure, r is a response of interest, x is a measured covariate, and u_1 and u_2 are two unmeasured covariates. Shrier stated that for the M-structure, '... it remains unclear if the propensity method described by Rubin would introduce selection bias or not'. In the same issue, Donald Rubin [*Statist. Med.* 2002; **27**(14):2741–2742] replied by clarifying several key points in the use of propensity scores. He did not, however, discuss the original question posed by Shrier. Given the popularity of both propensity score methods and graphical

*Correspondence to: Arvid Sjölander, Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Nobels väg 12, 171 77 Stockholm, Sweden.
†E-mail: arvid.sjolander@ki.se, arvid.sjolander@meb.ki.se