Reviewer’s report

Title: Reducing bias through directed acyclic graphs

Version: 2 Date: 15 May 2008

Reviewer: Jay Kaufman

Reviewer’s report:

Review of “Reducing bias through directed acyclic graphs”

The aim of this manuscript is to describe, explain and demystify a procedure for covariate selection that is outlined in Pearl 2000. The authors are explicit that they do not present new results here, but see the purpose of the paper as clarifying work of Pearl and others in order to help it become more widely applied. Since the work is not new, there are two important criteria for the evaluation of this work, 1) is the original work accurately described in this review, 2) is the new paper actually clearer and easier to understand, so that it represents an advance over the original and over other subsequent reviews and syntheses that are already published. I have some concerns about both points. Most importantly, a review article by Greenland, Pearl and Robins was published in 1999, and is cited in the current manuscript (reference #9). This article contains the following simple algorithm for covariate selection in order to reduce confounding bias:

A set of variables S is sufficient to control for confounding in the relation between exposure X and outcome Y if, upon adjustment for variables in S, there is no unblocked backdoor path between X and Y.

Two graphical criteria will ensure that this condition is satisfied:

1) Every unblocked backdoor path between X and Y is intercepted by a variable in S, and

2) Every unblocked path between X and Y that is induced by adjustment for the variables in S is intercepted by a variable in S.

The authors then go on to note that condition 2 may be difficult to operationalize confidently, and so they propose the following equivalent criterion:

2*) If every collider on the backdoor path between X and Y is either in S, or has a descendent in S, then S must also contain a non-collider on that path.

Given the simplicity with which this algorithm was expressed in 1999, it is an open question whether the authors of the current manuscript have achieved a greater degree of clarity with their six-step program. For example, the recipe in the current manuscript requires the a priori selection of a confounder adjustment set, for which the 6 steps program will then reveal whether or not causal effects estimated from that model will yield exposure effects that are unconfounded. The
algorithm in the 1999 article does not require this a priori selection of an adjustment set, however, making it considerably more convenient. The authors never explained on what basis they picked their two chosen variables as candidates for adjustment, for example.

The current manuscript is further compromised by a large number of incomplete or inaccurate statements. It is a concern, therefore, that the review may generate more confusion than the original. Some examples of text that is questionable:

Abstract (page 2):
“Using the simple 6-step DAG approach to confounding and selection bias discussed decreases the likelihood that the chosen statistical model yields a biased estimate of effect.”

As noted by previous reviewers, confounding bias is not a binary occurrence, such that we would distinguish between a model for which the effect is confounded and a model for which the effect is unconfounded. Rather, it is always a matter of degree – we wish to reduce the extent of confounding bias through optimal choice of covariates. Therefore I would change the language from referring to the “likelihood that a model yields a biased estimate of effect” to something about reducing the degree of bias, since confounding is only one source of estimation bias, and since confounding is not completely eliminated if there is measurement error, etc. These assumptions are mentioned later in the manuscript, but the unqualified statement here in the abstract doesn’t refer to any such assumptions or caveats.

Page 3: “There are many nuances to the definition of cause and for the purposes of this manuscript, we define it in counterfactual terms: ‘Had the exposure not been present at that time, the disease would not have occurred’.”

This sentence seems to assume that all exposures are binary and all outcomes are binary, and therefore there are no causal effects that involve changing the mean or the distribution of an outcome. Pearl makes no such restrictions in his book. I should also note that this restrictive definition is inconsistent with the more general one given a page later, which is that an effect occurs when “changing an exposure will affect the outcome”.

Page 4: “In the causal directed acyclic graph (DAG) approach, each variable is connected to other variables by an arrow directing causation.” This is incorrect. It is certainly permissible in a DAG to have nodes that are not connected by directed arcs. In fact, nodes that are not connected by directed arcs are statistically independent, and so the absence of an arc is highly informative.

Page 5: “…very little has been published on how to approach the problem of choosing the correct subset of covariates.” As noted above, the 1999 review has a very clear and simple algorithm.
At the bottom of page 5, the authors cite the Pearl book in general, with no page numbers. It seems to me odd to cite specific sections of the book in reference 4, 13 and 18, and then to cite the whole book in reference 14, especially, because the sentence in which this citation occurs refers to specific material in the book. I would cite the book once, and then reference specific pages each time that single numbered reference is invoked.

Page 6: Here the authors refer to causal “maps”. I do not know what a causal “map” is. It is not defined in the paper, nor is it defined in the Pearl book.

Page 6: "1) any approach to confounding (including the one described below) requires that one knows and can measure all variables causally related to exposure and outcome”. This is incorrect. See Pearl 2000, figure 3.5 (page 81) for a counterexample.

Instrumental variables approaches are also counterexamples.

Page 6: “Step 1: The covariates chosen to reduce confounding…should not be descendents of X.” This is true for confounding via “open backdoor paths”, but again, there are counterexamples such as the one cited above (page 81 in Pearl's book).

Page 6-7: The authors state that a confounder has the following characteristics: "First, the covariate must be associated with the exposure and with the outcome, but cannot be affected (i.e. caused) by exposure.” For a counterexample, see Robins J. The control of confounding by intermediate variables. Stat Med. 1989 Jun;8(6):679-701. The authors allude to this issue later (i.e. confounding by factors affected by exposure), but this later mention doesn’t excuse the erroneous statement here.

Page 7: “Thus, if a covariate is associated with an exposure, and the exposure cannot cause or be a marker for a cause of the covariate, then the covariate must cause (or be a marker for a cause) the exposure.” This is true for relations in the superpopulation, but it is not necessarily true of relations in a sample. See, for example Hernán MA, et al A structural approach to selection bias. Epidemiology. 2004 Sep;15(5):615-25. The Greenland et al 1999 review cited above also refers to associations in samples that are non-causal.

Page 9: Step 5 instructs the reader to “strip all arrowheads from lines”, and then Step 6 instructs readers to “delete all lines emanating from the covariates”. I can’t understand this logic, as once the arrowheads are eliminated in Step 5 it becomes impossible to know in Step 6 what is emanating from what, since there is no longer any information on directionality.

Page 11: “At the present time, there is no algorithm and the six-step process should be repeated until a subset of covariates is found such that X is dissociated from the outcome after the 6-step process is completed.” Again, see the simple 2 step procedure outlined in Greenland et al 1999, which would seem a more parsimonious and less ambiguous approach.
Page 11: “(i.e. more degrees of freedom if less covariates). “ This is not grammatical English.

Page 13: “However, one would still not want to adjust for a covariate (or is a marker for a covariate) that lies along a causal pathway and therefore the DAG approach remains an important step in the process.” This is incorrect. The whole justification for techniques such as g-estimation is to adjust for confounding by variables that lie on the causal path between exposure and outcome. Also, note that the parenthetical phrase is non-grammatical.

Figures: Some greater care should be taken in the arrangement of the nodes and arcs to avoid crossing of lines. I realize that this can be difficult, but the current situation (e.g. thick dashed line in Figure 3b) can probably be improved without too much effort.

In summary, this is an admirable project, but is not contributory in its present form. In order to warrant publication the authors need to 1) clarify what this article contributes beyond the more general and parsimonious algorithm in Greenland et al 1999, and 2) clean up the incorrect statements that are listed above.

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I have no financial or personal interests to declare.