Reviewer's report

Title: Evaluation of the Propensity score methods for estimating marginal odds ratios in case of small sample size

Version: 1 Date: 15 February 2012

Reviewer: JAY STEINGRUB

Reviewer's report:

The authors have conducted a series of statistical simulations and an analysis on a real dataset to determine how sample size (particularly small sample sizes) affects results of a propensity score analysis. While interesting, a general comment is that the paper is truly a statistical methods one and may not be of great interest (or appreciated) by a clinical audience of physicians.

While the methodology appears sound, there are some issues to be raised.

1. The authors only examine caliper matching and IPTW matching. While caliper matching is popular, with modern software, the analyst often has a variety of matching methods at his disposal. Moreover, other papers in the literature have examined these issues (small sample size, matching, and the quality of the estimates etc). The authors should examine the following papers: Frolich’s work in The Review of Economics and Statistics Feb 2004; 86(1):77-90 (which addressed ridge matching) and Dehija and Wahba’s classic paper in The Review of Economics and Statistics, Vol. 84, No. 1 (Feb., 2002), pp. 151-161 which did empiric work on studies with similar sample sizes as were simulated here.

2. While it is appropriate to note that the variances of the IPTW estimators were systematically smaller than the variances of the PS estimators, you do not state if this result is counterintuitive. In fact, it is not and this expected result is due to the construction of the estimators by the IPTW methodology. This should be noted and explained more in the paper.

3. One page 9, you note, “When the strength of the association between the covariates and the treatment/outcome decreased, IPTW estimators remained similarly or even less biased than PS-matching estimators, down to 60 subjects (Table 1). However, when the sample size was set at 40, the PS-matching estimators outperformed better than the IPTW estimators. Was the improvement at Sample Size = 40 clinically and/or statistically significantly better? You offer one possible explanation that this phenomenon is due to excessive weights given to marginal subjects and note that there is a way to correct for this (stabilized weights). Since this is an acknowledged flaw of the IPTW method in small sample sizes, why did you not adopt stabilized weights in your analysis?

4. You do a good job showing the results of different model specifications (ie, one variable all the way to the full model). You do not discuss balance testing
adequately though and that is an important component of doing a propensity score analysis. A paper to reference is: Hansen in Statistics in Medicine 2008; 27:2050–2054.

5. Page 14, you state: At last, not at least, such a difference in estimates according to the approach or the covariates points out the obvious need for first defining the approach to be used unless it appears likely that one may chose “to pick the winner”. This is a curious statement. Wouldn’t a different interpretation from your analysis (and from the literature) that since you can get different estimates based on the matching methods and how you specify the propensity score model, you should try many different models and matching schemes and report the results. In other words, if you see consistency in your results in a variety of analyses that is more comforting that you are observing a “true” effect than merely reporting one result derived from an a priori selected approach. If there is a lot of variation in the results based on the different propensity score methods available, that is an important finding to report (and explain if possible) as well.

There are some grammar issues that should be addressed (and the paper is long):

6. Page 6, line 6 “did not affect none” should be “did not affect any”

7. Page 9, (data no shown) should be (data not shown)

8. Page 9, “thinner” is not the right adjective to describe MSE

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.