Reviewer's report

Title: Effects of Aggregation of Drug and Diagnostic Codes on the Performance of the High-Dimensional Propensity Score Algorithm: an Empirical Example

Version: 1 Date: 22 August 2013

Reviewer: Rolf Groenwold

Reviewer's report:

This manuscript deals with the impact of aggregating confounder information when using the high-dimensional propensity score algorithm to control for confounding. Different methods are compared in empirical data on celecoxib / NSAID use and the risk of gastrointestinal complications among patients with rheumatoid arthritis or osteoarthritis. Well written manuscript that contributes to the literature on the (high-dimensional) propensity score.

Major revisions:

1. Unfortunately, there is relatively little confounding present in the data. After adjustment for basic covariates the effect estimates do not change that much. Could the authors reflect on the power of their study to identify differences between the different methods. I think they could speculate somewhat more extensive in the discussion part of the manuscript. Obviously, aggregating confounder information in small datasets is a trade-off between inducing residual confounding (because the relations between confounder and exposure and confounder and outcome are implicitly considered to be the same for the different components of the aggregate) and non-positivity due to rare confounders (‘empty cells’).

2. Clearly, the prevalence of the aggregate confounder will be higher than the prevalence of its components. Given that the hd-PS algorithm ‘prioritizes variables … by their potential for confounding control based on their prevalence … ’, do the authors agree that the hd-PS algorithm theoretically favors aggregation of confounder information?

3. One of the advantages of the hd-PS algorithm is that it can deal with a large number of confounders. It therefore seems somewhat counterintuitive to aggregate confounders in order to reduce the number of confounders.

Minor revisions:

4. From the abstract it is not directly clear to me, what is considered the reference. Is it the result from the RCT, or the result from the base scenario (no aggregation)?

5. Results: Please add confidence intervals for the reported relative risks.

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

Quality of written English: Acceptable

Statistical review: Yes, but I do not feel adequately qualified to assess the statistics.

Declaration of competing interests:

'I declare that I have no competing interests'