Reviewer’s report

Title: Evaluating bias due to data linkage error in electronic healthcare records

Version: 3
Date: 27 January 2014

Reviewer: Dick Heasman

Reviewer’s report:

I am very sorry that the authors have not fully implemented suggestions 1. and 3. from my previous review. It seems to me that a real opportunity has been missed here, to demonstrate that when data linkage (note here that PII is not data linkage) is used to construct an analysis dataset, the results are likely to be biased regardless of the data linkage method used. I am fairly confident that the study could have done this. Maybe it is still not too late.

On reflection I have upgraded the level of interest, as it now seems to me that what the study sets out to do has importance in its field. However, the study as it currently stands is scientifically unsound. Please see the attachments for a full explanation of this conclusion.

==========

Some comments on author’s response to reviews:

1. In the context of our study, we feel that deriving optimal thresholds based on the known match status of each record pair would result in the HW method performing better than it would in a real linkage situation (i.e. when the match status is not known).

I have stated in my review why I think this is not necessarily true. But more importantly, why does this matter? It would do the study no harm at all evaluate the bias of the best that data linkage could do, especially as it also evaluates the bias of the best that PII can do.

2. If we were to use joint match probabilities within the HW method, we would be comparing a new method (i.e. PII) with a method that is not used in practice (i.e. HW with joint match probabilities).

Not so. HW with joint match probabilities is used fairly frequently, and would be recognised by an experienced data linker as the appropriate method to use in this particularly straightforward and restricted case. Of course, it is not usually called that. It is called rule-based (deterministic) matching.

3. One of the aims of this paper is to demonstrate the limitations associated with the HW method.
That is not stated in the paper. I argue strongly in my review that it should not be an unstated aim of the paper either. The paper should stick to the much more interesting stated aim of demonstrating the limitations associated with data linkage itself, whatever method is used to carry it out.

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

As before.