Reviewer's report

Title: An efficient record linkage scheme using graphical analysis for identifier error detection

Version: 4 Date: 5 December 2010

Reviewer: Stuart M Speedie

Reviewer's report:

The authors have responded to the issues raised by the reviewer in his first review in a largely satisfactory manner. However some issues still remain but these are now issues that the authors can address without the need for additional review.

- Major Compulsory Revisions
None

- Minor Essential Revisions
1. The authors repeated refer to records of individual patients. However technically they do not have unambiguous evidence that any single cluster corresponds to the records of a single person, even though the likelihood is high. They need to adjust the wording in the Abstract and their Results to reflect this ambiguity.
2. Page 9, Sec. 4, First Paragraph, second sentence – syntax needs correction.
3. Page 11, Sec. 6 First Paragraph, first sentence – syntax needs correction.
4. The authors concentrate most of their discussion in the methods and results sections on the issue of “collision” in clusters that might contain more than one individual's records. Their efforts in disambiguation address the issue of specificity as they have defined it. Equally important is the issue of records from a single patient being assigned to more than one cluster. This is most likely to occur when there is only a single identifier available such as in the LIMS and Microbiology data. Table 3 indicated that there are a relatively large number of such records and this there appears to be a relatively high potential to erroneous data entry that could lead to this problem. The Fuzzy Logic investigation is an attempt to investigate the nature of this problem on a small subset of records. Their findings indicate that the extent of this type of problem may be larger than that for collisions. The lack of any major effort to deal with this problem detracts from the attractiveness of their approach. The discussion needs to address this issue to a greater extent than it does now so that the reader can fully understand the benefits and challenges of the approach they are describing.

- Discretionary Revisions

After reviewing the revisions, the title no longer seems to accurately reflect the
substance of the article. The article claims to describe an “efficient” algorithm. Efficiency is usually measured relative to other approaches. The article does report that the approach is apparently fast but does not, in the Discussion, review and draw any conclusions about its relatively efficiency. The discussion needs to address this in greater detail or the title should be modified. Also the method that is reported and used addresses only certain types of errors – those that result in collisions. The title gives the impression that the algorithm detects all such errors but it does not appear to do so.

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

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

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

Declaration of competing interests:

I declare that I have no competing interests