Reviewer's report

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

Version: 1 Date: 8 August 2010

Reviewer: Stuart M Speedie

Reviewer's report:

This article addresses a critical issue in the use of patient records for a large variety of purposes from epidemiological research to the problems of health information exchange and summary care records. Both better research and better patient care is dependent on the ability to link an individual’s records from diverse information systems. The authors present a system for rapidly creating such linkages within large databases of clinical information using an alternative to more standard probability-based systems. The work has the potential to provide a highly accurate and efficient system of record matching that could have great utility in a number of settings.

However the authors need to address one major issue and a number of minor issues in order to this manuscript to be suitable for publication.

- Major Compulsory Revisions

1. The goal of record linkage algorithms is to link all records for a given patient to that patient. This article needs to acknowledge that its method represents one step in this process of doing so. The reviewer is given pause when the authors report that the Patient Administration (which is supposed to identify every patient receiving hospital services) contains 2.6 million records while their methodology produces over 3.5 million unique clusters which should correspond to individual patients if their algorithm does what they claim it does. The manuscript needs to thoroughly address and explain this difference otherwise it would appear that it is creating multiple unique clusters groups of which correspond to a single unique individual. It would seem that while they spend a great deal of effort to insure that their algorithm generates unique clusters of records, it does not take the next step of aggregating those records into collections that reasonably correspond to individuals.

- Minor Essential Revisions

1. Background Para 5: typo – should be “errors”
2. Background Paragraph 5: typographical errors generate two kinds of problems related to specificity and sensitivity as defined in this paper. Such errors can generate two groups of records for records belonging to the same individual, or one grouping of records belongs to two or more individuals. One would expect that records belonging to two or more different individuals would have at least
some unique identifiers associated with those individuals. This needs to be clarified in its expression.

3. Initial Record Linkage …: Pg the text does not adequately distinguish sets and clusters.

4. Table 2 – It would be more helpful only underline the data that is used for grouping at each stage of the process.

5. Table 2, last row: It appears there is a typo - cluster No labeled 3 should be 6.

6. Additional Material 1 – why is the check digit removed for hospital ID? This would seem to be throwing away useful information about the validity of this identifier.

7. General question: Were records without any identifiers discarded from the linkage process?

8. Trial of a splitting heuristic...: – The value of the logistic regression step is not clear to this reviewer. It appears that the described graphical/quality improvement method in this section could accomplish the same end. The paper needs to describe the independent contribution of the logistic regression as opposed to the proposed splitting heuristic.

9. Methods, Task Definition: The authors need to explain why the PAS is not used as a "gold standard" since by its definition, it should contain information on every unique patient seen in the hospital.

10. Data Platform used …: Please provide a definition of the distance measurements used.

11. Addtl File 1 – why were not the check digit problems treated in the same way. Why drop the digit from Hospital numbers – this would seem a quality clue that was ignored.

12. Identity Collision and its detection: Bullet 2 appears to be a subset of bullet 4 and not an independent set. It so they should be combined.

13. Identity Collision and its detection:: It is questionable whether or not the approach to creating simulated bad clusters is the best approach given the types of errors that are likely to occur. The simulated "bad" clusters are more likely represent errors that are unlikely to occur rather than typographical errors. Such an bad cluster is created by entering “Smith” instead of “Jones” for a forename. This may not yield a good discrimination for typographical errors - the errors that are likely to occur.

- Discretionary Revisions

1. Figure 3 is hard to understand – This reviewer suggests keep boxes in the same positions on both sides and converting the figure to black and white with non-color coding for the edge lines. It would also be helpful to visually indicate the clusters.

**Level of interest:** An article of importance in its field
Quality of written English: Acceptable

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 interest