Reviewer’s report

Title: Evaluating data linkage methods for linking routine electronic healthcare records

Version: 1 Date: 3 October 2013

Reviewer: David Cromwell

Reviewer’s report:

This paper describes an interesting evaluation of how errors in the linkage of datasets can affect the results of medical research studies. The study provides information on two things: (1) the comparative performance of different linkage methods, and (2) the size of bias in the results that can occur when different types of linkage errors arise. The study uses data on admissions to a PICU submitted by two English hospitals to a national clinical audit and data on PICU-acquired BSI as collected in the hospital laboratories.

The research tackles a difficult problem, and one that is growing in importance with the increasing linkage of health care datasets. The research design is robust and appears to have been carefully conducted. There are various weaknesses in the paper, which I suspect come from the difficulty in writing concisely about a complex study rather than methodological weaknesses of the research. The paper would be improved if the following issues were addressed.

Major Compulsory Revisions

First, there are various ways in which datasets are linked. In many circumstances, there are records for an individual in both datasets regardless of the outcome being studied such as when GP records are linked to hospital records of patients who undergo surgery. The situation in this research is different because records are only linked when an outcome is observed, ie. There was a positive BSI result. Laboratory records of negative culture results are not included. Consequently, the quality of the linkage has a direct result on the results calculated. The paper should make this particular type of linkage more explicit in the text. If possible, the authors should comment on how the algorithms might perform in other type of linkage situations (eg, if the laboratory dataset had records for all child blood tests regardless of BSI positive status).

Second, I found the description of the methods hard to follow. It would be helpful to have a diagram that illustrates the flow of data manipulation from the original data linkage, the creation of the simulated data, and then the testing of the new linkage approach using fewer identifiers. The text also needs clarifying. I assume that the simulated data were created using selection with replacement given that some datasets had a 70% prevalence of BSI when the observed rate was 7%. This needs to be stated. I also do not understand the sentence (pg 5) “non-random by outcome (children with PICU-acquired BSI were 5 times more
likely to have error
than children without PICU-acquired BSI).” when children without a
PICU-acquired BSI had no records to be linked.

Third, it would be helpful if the paper could provide some way of relating the rate
of linkage (patient based) in Table 1 to the crude rate of BSI per bed-days (in
Table 2). On first reading, I was expecting the crude rate based on the
gold-standard records to be 10% (datasets 10-12 have 10% matched (linked)
records to a positive BSI in the lab data), and it took me a while to recognise my
oversight.

Minor Essential Revisions
1. The results part of the abstract should contain some quantitative results rather
than qualitative statements.
2. Please provide a reference for, or define, Soundex
3. As the results are expressed as per bed-days, it would be helpful to have
some length of stay information. I assume there was only one BSI infection
counted per admission, although this is not stated
4. I do not find Figure 1 easy to understand. It would be helpful to have some
further explanation
5. The red lines on Figures 3 and 4 should be labelled. It would also be helpful if
the graphs could be more clearly linked to the datasets.

Discretionary Revisions
1. I would have preferred to see absolute rates for the hospitals rather than the
difference but I realise this would have allowed the hospitals to be identified and
may not be feasible

**Level of interest:** An article of outstanding merit and interest in its field

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

'I declare that I have no competing interests'