Reviewer's report

Title: Event linkage of health and aged care services data: comparing event-based and name-based methods

Version: 2 Date: 25 October 2007

Reviewer: Catherine Quantin

Reviewer's report:

General
This article is particularly interesting for those who are dealing with the linkage of health data. The paper addresses the comparisons of the results of the event-based and person-based linkage strategies.

However, without access to named information of persons for whom matching health records would be undertaken, it will be very useful to carry out the linkage using health events information variables of those persons.

Nevertheless, it is quite difficult to really to read and understand this paper according to its current form and presentation.

The methodology needs to be clarified, so that to define with precision the main steps of the event-based linkage strategy.

The discussion section of the article should be deeply corrected: results should appear in the Results section.

--------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

1 - Page 5 (Methods Section): the 1st paragraph should be clarified. The number of included patients and the justification of the Western Australia choice are unknown. Moreover, it is difficult to understand how the fact that the “linkage system was extended in 2005 to include RAC data” is useful for the study, which concerns period 2000-2001.

2 - End of page 6 to the beginning of the page 7: we don't know how Name-based (N) linkage is applied to RAC, is it commonly applied or specially applied for this study? The authors of the article should clarify what is done by the dynamic system, and precise if the hospital data linkage is already systematically done. Regarding this study the authors should precise if the results of an already done linkage have been used or if they used a specific methodology which they applied to their own data?

3 - Pages 8, 9: The strategy of the authors is not well designed, conducted and explained. It is difficult to understand what the order of the linkage steps is. For example, the authors should precise if at the beginning, the rules used in the
Table 1 provide the possibility to partition the different datasets or if they come after the probabilistic matching so that to realise the linkage 1:1, using the deterministic rules. It is important to explain what the authors mean by 1:1 linkage: what is the interest of using deterministic rules to get the 1:1 linkage, how the deterministic rules are elaborated?

The authors of the article should then reorganize the redaction of these two pages, putting paragraphs and tables in the right order.

4 - Event-based (E) linkage, in page 8, second paragraph: If the authors effectively use the blocking techniques in the classical way before the linkage, so that to limit the number of comparisons to be done as explained in a, b, and c sections, they should mention it. In this case, the presentation of the method would be clearer if the authors have introduced the use of the blocking techniques to realize the partition of the datasets at the fifth line of the paragraph. The authors should use more classical terms such as collision instead of coincident records (cf a) and doubloons instead of “number of records being compared at any one time” (cf c). It is not clear what the authors mean by «to increase linkage variables variation»: is it the provided quantity of information of each variable (discriminate power)?

5 - Table 2: more explanation is needed in the text to precise that this table includes at the same time, the blocking techniques before linkage, and the successively probabilistic matching passes, using different variable definitions. It is also necessary to explain with more details the different passes: at the first pass the variables will be used as defined at the beginning, at the second transition, all matched records obtained from the first transition will be removed from the dataset, and for the concordance of event dates a possible two days variation between the dates recorded in the two datasets is accepted, and so one for the next transitions. Moreover, to clearly explain the difference between the blocking and the successive passes of probabilistic matching, more details in the text are needed.

6 - In page 9 second paragraph, line 3: The authors should justify why two types of E linkage have been used.

7 - In page 10, lines 6 – 7: the given explanation is quite evident for epidemiologists, but needs to be more detailed for non epidemiologists, this can be done by through a Table. It seems also necessary to explain that the specificity and false predictive value will depend on the possible number of links. In this study, as the number of hospital discharges is higher than the number of RAC admission or re-admission, the total number of links will then be more important when the linkage is based on the hospital point of view than on the RAC point of view. Even if the specificity and false predictive values based on the hospital point of view and/or the RAC point of view are very different, it will be interesting for readers to get those values.

8 - In page 15 last paragraph: the relationship between Table 2 and table 3 should be explained and inserted in the paragraph. In particular it is important to explain that the first line of the E linkage «constrained-SLA, with 2-digit postcode
matching » corresponds to « constrained-SLA » strategy, after adding all the links obtained from the successive passes. Then, 7 595 = 7418 + 177 (cf Table 3 addition of right links and false links) = addition of links obtained from the Table 2. It is needed to explain that removing 342 links to 7 595 obtained from the first line of Table 3 (E linkage), corresponds to a strategy without 2 digit post-code, and also in the Table 2 to the utilization of the 6 successive first passes, the seventh excluded.

9 - In page 18, last paragraph, page 19, page 20, first paragraph with Table 6, page 22 first paragraph until Table 7, the discussion should be completely rewritten, in order to insert the new results in the result section.

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

1 - In the whole text it is preferable to use false positive and false negative terms instead of miss matches and false matches.

2 - Page 3 (Background Section) second line : the abbreviation for Residential aged care is missed.

3 - Page 6: what do the authors mean by Hospital separations ?

4 - Last line page 6 first line page 7 : the authors should explain why the RAC linkage has been realized for 1990-2003 period when, 2000-2001 period was announced at the beginning of page 6.

5 - The sentence beginning at the last line of the page 10 and ending at the third line of the page 11 is quite impossible to understand, it should be re-written.

6 - Page12: « SLA-based constrained strategy » expression is not defined in the text. It only appears in the Table 2. I suggest defining this expression in page 8.

7 - First line, page 12 : the authors should precise before that only the passes with positive predicted values superior to 60% is retained in the Table 2.

8 - Sixth and seventh lines page 23 : the proposed conclusion « methods using broader regions when matching performed better than those using smaller regions » seems to be in contradiction with the obtained results particularly with the sentence in the second paragraph of the page 14 : « comparisons indicated that one of the most effective ways of reducing the number of false matches made under the E strategy would be to reduce the size of the geographic region used in matching».

It is right that the authors precise that “narrowing the geographic matching criteria would result in dropping many more true links than false links”. For better comprehension, it is important to remind these two explanations in the conclusion.
Discretionary Revisions (which the author can choose to ignore)

What next?: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

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

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 interests’.