Reviewer's report

Title: Creating the linked data set for the PIAC cohort study: record linkage between several data sets using statistical linkage keys

Version: 3 Date: 12 July 2009

Reviewer: Anita CJ Ravelli

Reviewer's report:

Reviewers report II

Title “creating the linked data set for the PIAC cohort study: record linkage between several datasets using statistical linkage keys”

Karmel R et al.

Ms: 1570049224230154

Version: 2

Thanks for the improved version of the paper. Although I did not see information for each raised questioning in version 1 in general in the paper some points are solved however not all.

Major compulsory revisions

1. In this is study the authors use a determininistic medical record linkage (DRL) strategy with an self made extension called the stepwise method (A,B,C). In this developed stepwise method some principles of probabilistic records linkages are uses. The extensive tables from A3 and A4 onwards visualize that. However I do not agree with the last sentence of the conclusion “where probabilistic matching cannot be used.

For the readers of the BMC this Australian study is a good examples of practical difficulties researched have to cope with achieving their goals to understand health care using existing health databases with different data qualities.

Probabilistic medical record linkage (PRL) approach is the preferred methods for these situations. The authors did not used the PRL method and their argumentation not using the preferred method (no possibilities for clerical review) is not solid. However this method decision cannot be changes at this time and the articles is still of importance to other researchers. It shows the amount of work, the certainly and uncertainty of the different steps when using extended deterministic methods in large databases in health care. And it is in its way a nice example of the work that was done and could be reduced if by start a PRL method was chosen.

2. I’m not convinced that the estimated trade of (method C) is the best stop rule in all situations. If the wrong key is used first and the estimated level of matches
is reached the authors will stop and miss an important later key for a next step. The nice aspect of PRL is that if gives you beforehand a clue which of the possible keys is of more importance without, like the authors did, having to thing and write out and all possible keys.

3. In both the abstract and in the background the objective of the paper is not clearly stated possible “to described a used extended deterministic medical record linkage strategy for situations where there is a general unique identifier (>=95%) and several addition variables of different quality available in several large databases in health care.

4. The amount of tables and the information given in the tables especially in the appendix is too much to get and keep the main message and to review in detail.

Minor points

1. Abstract
  • the abstract method section consist of background sentences rather than the used and developed method: method A,B,C description and the validation procedures.
  • why is the meaning of the word full added to probabilistic linkage (abstract p2, background p6) can this be left out?.
  • Results (abstract) which measures?
  • Conclusions: last sentence ..leave out?

2. Background
  • Can the sentence “ SLK used to link data for statistical and research?? (p3) Purposes does not have to be 100% accurate be refreshed?• Is in the description of probabilistic record linkage (p4) for this paper the sentences on blocking and match variables needed rather than an overall description?
  • Threshold can be set (ref 15) (p5) however even in this approach …… . Validation of both deterministic and probabilistic medical record linkage algorithms is in general needed but should not be combined with the above mentioned point: setting thresholds.
  • Measures of similarities (p5) are called weights.
  • Can the authors prove the sentence “..-or more successful than- ? If they can do this with their data than this effort should be the objective of the paper. .

3. Method
  • It is not clear what the authors want to say with the sentence “That could be used in probabilistic linkage (p8) for clerical review” . PRL can be done without clerical review. This same point is set op p18
4. Results

5. The amount of outcome tables given to the readers if of too much detail to get the bigger picture. Possible the whole paragraph on quality of match data can be put in the appendix and the general message put in the paragraph linkage of the data.

**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