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: 1 Date: 24 February 2009

Reviewer: Anita CJ Ravelli

Reviewer’s report:

Major Compulsory Revisions

The PIAC cohort is an import cohort for Australian researchers and for other researchers interested in the care of elderly people. There is a lot of work done by the authors to create this cohort. The uses of a linkage key together with additional variables in different health care databases makes it an interesting setting for the research question.

<probabilistic record linkage>

1. The objective of this study is “a stepwise deterministic method for linkage datasets using in SLK in the conjunction with other set when probabilistic record linkage cannot be used (p5 introduction). The authors give two rationales for there statement probabilistic record linkage cannot be used: difficulties for setting weight cut-off points (threshold) and no possibilities for clerical review. Can the authors give references for these rationales? I am not convinced yet that this rational is true and that probabilistic linkage was not the preferred method for this study question.

The authors mentions the problems of the difficulties in setting the weight cut-off points (p10 methods) and therefore no probabilistic method was used. Are there in this study (at the start) mi and ui weights calculated (with or without the SLK key’s) and was there a threshold value calculated?

2. The nice thing about this study is that the researches have the possibility of a statistical linkage key “SLK” based on real patient data and additional variables. The use of the key give a high linkage results (90-95% according to table 6). For readers / countries who are thinking to start using these linkage key’s it is still interesting to see the results of using the key and than the results of the use of the other possible linking variables after the linkage key.

I think that probabilistic records linkage is the method to help the researcher to interpret the results of the different match keys (appendix A.1) where now the results are difficult to understand and to generalise to other situations.

<validation studies>

3. What is the accuracy and the sources of bias when the stepwise deterministic approach is used compared to other methods? To evaluate the linkage accuracy
three validation studies are described (this validation could be added to the abstract method section). Is it true that the validation studies were only between the databases and the NCI (death) database? This gives the reader only information on this kind of linkage but not on the main health/population based database linkage between the ACAP and the RAC/EACH/CACP database?

<The tables could be revised and simplified>:

4. Table 1: description of the seven different databases to be linked (ACAP, RAC/EACH, CACP, AccMISS, HACC, VHC, NDI) with the (short) name, number, years and the program.

The used linkage strategy (Table 2) is a strategy of both cleaning/ preparation linkage phases (linkage of one databases of two different years) and linkage between different databases. This could be simplified given the objective; more focus on the results of the latest “the results of the linkage of the different databases”. Maybe only the results of the lasted should be given?

The before estimate of the number of matches between the different databases is not given, but the authors must have had before estimates. The before change of a match between the ACAP database and the NDI (death certificates) must have been different from the match between the RAC/EACH database and the ACAP database. How is this incorporated in the stepwise deterministic linkage strategy?

Table 5 and 6 give more basis information and are for me more clear. Is it a suggestion that these tables can be put more in front (changes with the current table 3 and 4, to help the reader to get the picture? Could the new table 3 (old 5) give information on the data availability of the linkage key and the additional extra variable information for each of the seven databases (to help to understand later the stepwise results)?

Table 4 (old 6) give the information “results of the linkage between the different databases”. The results of the study is that with the SLK linkage key 90-95%? of the cohort members (ACAP) n=105,077 could be linked to the RAC/EACH/NDI database, 90-95%? To the NDI database, and ?% to the HACC database, and ?% to the VHC database. With the additional steps xx% could be linked correctly, xx% possible.

<bias/ missing data>

5. Can the authors give information on the influence of missing data in the various datasets on the outcome of the linkage possibilities of those records and the possible outcomes for the study question of the PIAC cohort?

Minor points

In the introduction some basic references articles on probabilistic medical record linkage techniques were missing (Fellegi 1969, Jaro MA 1995).

In conclusion, the paper is rather long and not yet precise enough on the study
question. The paper would be strengthened if the text and focus is only on the objective of the study and not with the extra details on the PIAC cohort and PIAC specific questions.

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

**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'