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: 4 Date: 6 October 2009

Reviewer: Anita CJ Ravelli

Reviewer’s report:

Title
Creating the linked dataset for the PIAC cohort study: record linkage between several datasets using statistical linkage keys

Authors: Rosemary Karmel et al.

Version 4 Date October 4 2009
Reviewer: Anita CJ Ravelli
Ms 15700 49224230154

The authors have revised the paper and the main text is now on the description on the stage 3 linkage and the objective of the paper is defined and the text is reduced. The current paper is however more a descriptive report about a large deterministic linkage project with several obstacles in combination with a description of a set of used cutoff rules rather than answering a scientific question.

It is not clear from the paper how and if in other situations the linkage “stepwise deterministic” descriptions or the “three suitable criteria” in other databases or health care domains could be used. I won’t use it based on the given information.

My main comments raised in the previous revision are not solved

1. In this chosen example stage 3 a probabilistic strategy was the preferred methods dealing with 100 to 128 possible match keys is a lott of work in a deterministic approach without knowing exactly if you are doing. In a probabilistic approach for each match key a vector could be defined with a weight. So it is a description what happened which is relevant for the users of the database and for the audience to see what is happening but not to prove that this the wright way.

2. The suitable criteria (p10) are unclear, why are they chooses this way? If the criteria 2 and 3 were different would this have effect the linkage outcomes of this study? Are the three suitable criteria robust enough to be used in other situations, where is this stated?
If the final database is used to answer questions how would this use be different if different suitable criteria were set (an external validation step)?

If further needed I have several remarks throughout the whole paper but because of the above I do not think that another revisions would make a difference.

**Level of interest:** An article of limited interest

**Quality of written English:** Needs some language corrections before being published

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

**Declaration of competing interests:**

I declare that I have no competing interests