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: 18 May 2009

Reviewer: Noralou P Roos

Reviewer's report:

As I stated when I first read the paper - it is long and much too technical and it needs major, major revisions. I also said

The reader would be better served by taking one or two data sets and following these examples through the linkage process. Thus, the same examples would be used for Client Identification within data sets. This would follow into the Results section.

Thus, the paper would be considerably shorter but better focused. A considerable effort at rewriting is needed. The results are quite interesting but they get submerged in the detail and the multiple data sets.

While the authors have followed one suggestion and moved some of the material in the methods section to appendices - the above major reservations I had about the paper remain. It is turgid and very poorly written. It needs to have someone from outside the research team - who understands the area - but who can write more clearly - be responsible for a major editing.

I would not recommend publishing this revised draft.

Some specific comments

Page 4 - give examples re the blocking variables and match variables Page 9 - what does this mean? "a measure of match key uniqueness?"

How do you know in advance - "an estimated false match rate for links...no more than0.5%)?

... with match rates expected to range (why expected? How do you get these %?) "was the recognition that, for the PICA study, the cohort of interest was those.... It is unclear how this fact influenced their work Page 13 - Reliable SLK-581 data .... How do you know if these were reliable when the only criteria = not Jan 1 or July 1????

Page 14 -- ...was due to relevance rather than poor reporting... ie does this mean the individual had no event on the data set or ???

Page 15 - where does SLK - 581 come from??

"match keys with measure A> 99.85 -- what is measure A???

Incidentally -while you do give a table of abbreviations - the continual use of so
many also contributes to the unreadability/incomprehensibility of this paper