Author’s response to reviews

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

Authors:

Rosemary Karmel (rosemary.karmel@aihw.gov.au)
Phil Anderson (phil.anderson@aihw.gov.au)
Diane Gibson (diane.gibson@canberra.edu.au)
Ann Peut (ampeut@gmx.com)
Stephen Duckett (Stephen.Duckett@albertahealthservices.ca)
Yvonne Wells (Y.Wells@latrobe.edu.au)

Version: 4 Date: 2 September 2009

Author’s response to reviews:

The paper has been revised considerably to improve focus and readability. This process has been guided by the comments from both reviewers. The response to particular comments is discussed below.

Reviewer 1 (NR):

Re: Shorter but better focused…

To focus the discussion better the following approach was taken.

1. Dropped some detail concerning the Australian aged care system. This included deleting most of the acronyms.

2. Focussed solely on one linkage process (stage 3) for
   o Description of the method (Methods section)
   o Discussion of results of the linkage strategy (Results section). Results from other stages are no longer discussed or presented (apart from very broad information in Table 2).

3. Added a picture illustrating the PIAC linkage to provide context for stage 3.

To shorten the paper a number of deletions have been made:

4. Dropped all of Appendix B (with the stage by stage results), and one of the long tables in appendix A (old Table A2)

5. Reduced two of the remaining three tables in Appendix A (new Tables A2 and A3) so that they illustrate the linkage algorithm rather than give all the results.
6. Dropped Tables 3 to 6 as providing unnecessary detail.

Re: Specific comments

Page 4: Examples of match variables: discussion on linkage methods has been revised, and so this comment no longer applies.

Page 9:
‘match key uniqueness’: wording and terminology (now using ‘discriminating power’) has been changed to make this clear.

‘estimated false match rate’: the method for calculating this is given in Appendix A, and this is referred to at the bottom of the list of measures.

Page 13 ‘reliable SLK-581’: the discussion has been changed as a result of point 2 above. This phrase is no longer used.

Page 14 ‘relevance’: the absence of data was due to it not being relevant. This is now discussed under ‘Quality of match data’, and the wording has been changed to make this clearer: ‘these data were available for just 27% of all RCCP clients’.

Page 15
‘where does SLK-581 come from’: the origin of SLK-581 is now given in the third paragraph under the ‘Current context’ heading in the Methods section.

‘match keys with measure A> ..’: this discussion has been re-worked and is now in para 3 under the heading ‘Data linkage’ in the Results section. This phrase is no longer used.

Use of acronyms: the number used in the paper has been greatly reduced (see Point 1 above).

Reviewer 2 (AR):

Major revision 1 (probabilistic linkage):

Statements like ‘where probabilistic linkage cannot be used’ have been removed or reworded in
- the Abstract (last para)
- Basic Strategy (first para)
- Conclusion (last sentence).

Major revision 2 (trade-off method):

This comment suggests that the description of the linkage strategy was not yet
clear enough. It is not apparent to the authors what is meant by the comment ‘if the wrong key is used first’. In our method, the keys are used in order of their discriminating power (measure A). A potential key (i.e. one that meets the criterion based on discriminating power measure A) is only excluded if its estimated false match rate (measure B) is too high or the trade-off is too low (measure C). Also, the linkage keys for each stage are identified separately. Consequently, using our three criteria we are highly unlikely to ‘miss an important key for a later step’. The last paragraph under the heading ‘Phase 3’ has been re-worded to make it clear that all keys meeting the three criteria are used for matching.

Also, the method does not use an ‘estimated level of matches’ to decide whether or not to use a particular key, with all matches identified by keys selected as suitable for use being accepted. A sentence to clarify this has been added to the first para under the Phase 4 heading.

Major revision 3 (statement of objective):

The authors agree with the reviewer that this objective was not clearly stated. The Abstract has been re-worked, and a statement to this effect has been included. A similar statement has also been included in the last paragraph of the Background section in the main paper.

Major revision 4 (Tables):

As stated in the response to Reviewer 1, many of the tables have been dropped or shortened (see points 4 to 6 for Reviewer 1).

Minor points

1. Abstract:
   - The abstract has been re-worked as a result of this comment.
   - The word ‘full’ has been dropped from references to probabilistic linkage.
   - The last sentence of the Conclusions has been changed.

2. Background:

Para 2: The authors were not quite sure what the reviewer meant by ‘refreshed’. The paragraph has been revised, but the quote has been retained.

Page 4: Description of probabilistic linkage: this has been revised and sentences on blocking and match variables have been removed as suggested.

Page 5 ref 15 Threshold…: The last sentence in the paragraph has been changed and no longer refers to setting thresholds, but to process validation.

Page 5 weights: wording has been changed to reflect this comment.
Page 5 ‘…more successful…’: this statement is in fact a quote. The authors think that the quotation marks may have been missing from the version sent to the reviewers, as this error was noticed after submission and then corrected. Unfortunately, we do not have the data to make this the objective of the paper.

3. Method:

Page 8 and page 18 probabilistic linkage and the need for clerical review: The authors believe that a proper treatment of this topic is beyond the scope of this paper. Consequently, references to the need (or otherwise) for clerical review in probabilistic linkage have either been dropped or revised.

4. Results:

The Results section has been revised as a whole. It has been shortened both in response to Reviewer 1 suggesting a more focused approach (by ‘concentrating on one or two datasets’), and in response to the comment in Minor Point 1 by Reviewer 2 relating to validation procedures. As a result of this process the authors retained a subsection on the quality of match data. The authors believe that the current Results section is now both more focussed and more readable.

5. Outcome tables:

As described above, most of the large outcome tables have been either dropped, or reduced.