Author's response to reviews

Title: The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa

Authors:

Chodziwadziwa W Kabudula (chodziwadziwa.kabudula@wits.ac.za)
Banjamin D Clark (benjamin@npongo.com)
F. Xavier Gómez-Olivé (F.Gomez-OliveCasas@wits.ac.za)
Stephen Tollman (Stephen.Tollman@wits.ac.za)
Jane Menken (menken@colorado.edu)
Georges Reniers (Georges.Reniers@lshtm.ac.uk)

Version: 4 Date: 9 May 2014

Author's response to reviews: see over
Dear Professor Nieminen:

Re: Submission of revised manuscript to *BMC Medical Research Methodology*

Thank you for giving us the opportunity to revise and resubmit the manuscript entitled “The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa”. We appreciate the constructive comments of the reviewers, and have revised the manuscript accordingly. Please find appended below our responses to each of the concerns.

Yours sincerely,

Chodziwadziwa Whiteson Kabudula.
For the Study Team

**Reviewer: Rainer Schnell**

0. My main objection to this paper is the lack of a clear question of the paper in the abstract and in the introduction. The paper answers a question, but this question is not stated clearly.

The paper is assessing the feasibility of record linkage between a Health and Demographic Surveillance System (HDSS) and a health facility in South Africa. We have restructured the background sections in both the abstract and the main manuscript (third paragraph, first sentence) to include this point.
1. Neither the abstract nor the conclusion point to the most interesting aspect of this study: The use of human fingerprints as gold standard for record linkage. I would prefer to see this feature in the title of the article.

   We have now included the use of fingerprints as gold standard for record linkage in the Methods section of the abstract and the conclusion in the main manuscript. The first sentence of the Methods section of the abstract now reads “Using a gold standard dataset of 623 record-pairs matched by means of fingerprints, we evaluate twenty record linkage scenarios ...”. The conclusion in the main manuscript now also starts with “Using records matched by means of fingerprints as the gold standard ...”.

2. The description of strategy 16 is incomplete. Please explain your steps more clearly.

   We have restructured the description of scenario S16 on page 9 as follows “...we create another scenario (S16 in Table 1), which first matches records deterministically using National ID number or a combination of telephone number and first name, and subsequently matches the remaining cases using the scenario that yields the maximum sensitivity and PPV among the first 15 probabilistic linkage scenarios”.

3. In order to interpret a logistic regression, the model fit must be checked. Significant coefficient don't have a valid interpretation if the model does not fit. Please provide goodness of fit measures for your logistic model.

   We have added Pseudo $R^2$ and Wald $\chi^2$ statistic as measures of goodness-of-fit to Table 3.

4. The p values of the chi-square test in annex 1 seem to be neither the values for a test of independence (would be wrong here) nor the values for a test of goodness of fit for the observed distribution. So please explain your computations and check the values.

   In Annex 1 we use the chi-square test to compare the distribution of background characteristics between records matched by means of fingerprints and records matched using conventional personal identifiers using the approach in Scenario S6, Scenario S16 and Scenario S17 separately. The null hypothesis is that there is no difference in the distribution of background characteristics between records matched by means of fingerprints and those matched using conventional personal identifiers. We have included the following caption “p-value using chi-squared test computed separately for records in each scenario compared to records matched by fingerprints” at the bottom of Annex 1 to make this point clear.
5. n should be used instead of N in all tables.
   Done. We have replaced N with n in all tables.

6. The conclusion on page 3 should be the same conclusion on page 18. The conclusion on page 18 is too general.

   We have revised the conclusion in the abstract and on page 18 to the following text: “Using records matched by means of fingerprints as the gold standard, we have demonstrated the feasibility of fully automated probabilistic record linkage using identifiers that are routinely collected in health facilities in South Africa. Our study also shows that matching statistics can be improved if other identifiers (e.g., another household member’s first name) are added to the set of matching variables, and, to a lesser extent, with clerical review. Matching success is, however, correlated with background characteristics that are indicative of the instability of personal attributes over time (e.g., surname in the case of women) or with misreporting of those attributes (e.g., age).”

Reviewer: Katie Harron

1. page 9: "using the best of the 15 probabilistic linkage scenarios". It is not clear how "best" is defined in this context. Please clarify.

   “best of the 15 probabilistic linkage scenarios” is defined in terms of sensitivity and PPV. We have added the following caption “**The best of the 15 probabilistic linkage scenarios is the one that yields the maximum sensitivity and PPV” at the bottom of Table 1 and restructured the description of scenario S16 in the manuscript on page 9 to “...we create another scenario (S16 in Table 1), which first matches records deterministically using National ID number or a combination of telephone number and first name, and subsequently matches the remaining cases using the scenario that yields the maximum sensitivity and PPV among the first 15 probabilistic linkage scenarios” to clarify this.

2. It is not clear why adding another household member's first name would improve the matching - can this be explained?

   The success of any record linkage is influenced by the amount, accuracy, completeness and consistency of information available for the linkage process. In our case another household member’s first name improves the linkage success because it is an extra independent piece of information that discriminates between matches and non-matches (unlike the household member’s surname) and is well reported at the health facility (unlike telephone or national ID number)
Discretionary revisions:

1. Although sensitivity and PPV are useful for comparing different linkage approaches, it would also be very useful if some actual results were compared. This would give the reader a sense of how results of analysis might differ from those where no linkage error was present (i.e. based on the gold-standard data).

The results that you might compare depend on the purpose of the linkage – i.e. you could look at rates of a particular condition, or the association between an outcome and a covariate. It would be particularly interesting to focus on a covariate or group that might be affected by the linkage, e.g. women / former refugees / poorly educated / older respondents. This would enable us to see the real benefit (or not) of investing the extra time and effort in manual review. I think this would add greatly to the interest of the paper and wouldn't require too much additional work to include.

   We fully agree with the reviewer about the added value of studying the implications of record linkage error for analytical studies and we intend to extend our work in that direction in the future. Unfortunately, the dataset from the health facility that we are using for this study does not contain any health outcomes, and we therefore limited ourselves to comparing the socio-demographic background characteristics of linked and unlinked records.

2. It would be helpful to give a sense of the additional time associated with manual review. The numbers of records in this study are fairly small – but manual review can quickly become unfeasible for large numbers of records in population-based data sources. This point could be brought out more in the discussion.

   We have included a sentence under the Results section on page 14 indicating the number of records that were reviewed for scenarios S17 and S20 (which allocated to clerical review 20% of record pairs around the threshold value of scenario S16) and the time it took. The sentence reads, “...for scenario S17, 1131 record pairs were reviewed and it took the two reviewers an average of 5 hours each to complete the task whereas for scenario S20, 3492 record pairs were reviewed in an average of 15 hours per reviewer.”.

Additional Editorial Requirement: □

1. We recommend that you copyedit the paper to improve the style of written English.

   Done. The manuscript has been copyedited.