Author's response to reviews

Title: The SAIL databank: linking multiple health and social care datasets

Authors:

Ronan A Lyons (r.a.lyons@swansea.ac.uk)
Kerina H Jones (k.h.jones@swansea.ac.uk)
Gareth John (gareth.john@hsw.wales.nhs.uk)
Caroline J Brooks (c.brooks@swansea.ac.uk)
Phil Verplancke (p.verplancke@swansea.ac.uk)
David V Ford (d.v.ford@swansea.ac.uk)
Ginevra Brown (g.brown@swansea.ac.uk)
Ken Leake (ken.leake@swansea.ac.uk)

Version: 2  Date: 9 December 2008

Author's response to reviews: see over
Dear Editor,

Re: Submission of revised paper – The SAIL Databank: linking multiple health and social care datasets

Please find attached a revised copy of a paper entitled: ‘The SAIL Databank: linking multiple health and social care datasets’. The paper is included in Word (.doc) and there is a separate Word file with a single page figure.

We found the reviewers’ comments constructive and helpful to us in clarifying the work described in the paper. We have addressed the reviewers’ comments in the revised manuscript and a point-by-point description is given below. Responses are shown in italics.

Reviewer 1: Miranda Tromp
Major compulsory revisions

1. Although the purpose of this work was to develop and implement an accurate matching process, the matching process is not described into sufficient detail. Additional detail on the matching process has been added to the methods section.

The purpose of the matching is to assign a unique identifier which can be used in future studies, but detail on this identifier are lacking - the constitution of the ‘ALF’ is not described. The ALF takes the form of a unique 10-digit number assigned at the individual record level.

2. The first research question was to check the accuracy of NHS number as a unique identifier. It is not clear from the method section how this had been established. The method section on how the first question was addressed has been re-written to make it clearer.
Is the linkage performed on the five linking variables using the combination of DRL and PRL and is the agreement on NHS number checked afterwards?

*Linkage was performed using DRL and separately using PRL to be able to compare the outcomes as shown in the revised Table 1. Agreement on the NHS number was checked afterwards.*

What is meant by ‘derived NHS number’?

The derived NHS number means the NHS number that would be allocated to that individual record on the basis of matching their record, on the five matching variables (first name, surname, date of birth, gender and postcode) to the same set of variables on the NHS Administrative Register (NHSAR).

Table 1 mentions a ‘ID generated from record linkage’, it is not described in the method section what this is. What does ‘different’ mean in table 1?

Table 1 has been revised and the text has been amended to ‘Results of comparing the NHS number allocated by the record linkage process with the NHS number submitted in the dataset’ in order to make this clearer and more consistent.

3. The second research question addresses the probability threshold in PRL. The matching algorithm MACRAL – how are the probabilities calculated for PRL?

*Additional detail on how the probabilities are calculated using a Bayesian approach have been added to the methods section.*

MACRAL makes sequential passes, are matching probabilities than determined for each pass?

*The operation of MACRAL has been clarified. In the second research question, we sought to determine the optimum way of operating the algorithm. We ran it in different ways: using DRL, or PRL at each threshold level and the numbers and percentages of records matched by each method are shown in the revised Table 2. We finally ran it using the algorithm to run sequentially from DRL (with exact matches on all five variables), through to PRL at the 50% threshold. This is shown as the final row of figures in the revised Table 2. Matching probabilities are assigned to each record in the dataset as a result of this process.*

How is the threshold determined? Figure 1 states that individual weighted scores are applied in the PRL, but this is not mentioned in the method section. The linking algorithm should be clear from the description in the method section and should not be partly explained in a figure. How are the scores related to the threshold value?

*Further detail has been added to the methods section. The thresholds are determined based on the posterior odds of the match being correct. The posterior odds are the product of the prior odds and the likelihood ratio for each pair of variables to be matched. The likelihood ratios are based on the particular variable and its distribution on the NHSAR.*

4. The method section should contain a clear description of the various datasets
(NHSAR, PEDW, GP, PARIS) used in this study as this information is now divided among the introduction, methods and discussion section.

This has been corrected: a section on datasets has been added to the methods section.

The NHSAR is the reference dataset, than for each dataset it should be described if all records/patients are expected to have a match within NHSAR.

The expectation of records having a match on the NHSAR has been added to the method section.

Furthermore the presence of the linking variables and NHS number should be described consistently.

The text has been amended to refer to the linking variables as ‘the set of matching variables’.

5.

The method section should include a paragraph on the outcome measures by which the matching process is assessed.

The way in which the matching process is assessed has been added to the sections on questions one and two in the methods section.

Definition of sensitivity and specificity should be given.

The definitions of sensitivity and specificity as used in this paper have been added to the methods section.

Also a definition of ‘successful match’ should be given (as presented in Table 2). In result section this is described as ‘percentage agreement’, but on what – on NHS number?, on 5 linking variables?, on linking variables using fuzzy matching?.

The definitions have been clarified in the text and in the revised Table 2.

Also error rates are reported in the result section, please give definition of ‘error’.

The definition of error, as used in this paper, has been added to the methods section.

The part in the discussion section on outcomes in record linkage should become part of this method section.

This has been moved as requested.

Minor essential revisions

1.

The background section is quite elaborate; I would try to focus more direct on the problem that this study tried to solve.

The background section has been amended accordingly.

2.

It is not clear what is understood under a ‘valid’ NHS number.
This has been clarified in the methods section. A valid NHS number is one that has been checked for correct format via a check-digit algorithm.

3. PRL is applied here in addition to DRL full on all five variables (according to figure 1), describing the results for DRL and PRL separately in table 2 is then misleading. Exact agreement on all five variables would be included in PRL as well if not used in addition to DRL.

The description of the way in which the second question was addressed was previously misleading in the text. This has been corrected to show that by operating the algorithm in different ways i.e. DRL, or PRL at defined thresholds, different levels of matching is achieved.

4. In the result section, the processing speed of the linkage is introduced. If reported as a result, this should be introduced in the method section as an outcome measure, although I doubt the relevance of this measure with current processing speed. This has been removed.

5. Use consistent naming throughout the text; introduce datasets in method section with appropriate abbreviation which can be used further on. (GP is first mentioned in Table 2, PARIS is introduced in method section but indicated as social services in table 2.) This has been corrected throughout the text.

6. The second paragraph of the conclusions contains no conclusion of this study. One closing paragraph on future work would be adequate. This has been added as suggested.

Reviewer 2: Marco Eichelberg
Minor essential revision: The URL in [4] does not work anymore. Replace by working URL or remove reference. This has been removed as it is no longer available.

Minor essential revision: The computation of the "fuzzy matching odds" mentioned in Figure 1 is an important part of the matching algorithm but is not explained anywhere. A description of this is needed to make the overall approach understandable to the reader. The description may be added in the main text or in the figure as per the authors' preference. This has been added to the methods section and is also covered in addressing reviewer one's queries above.

Discretionary revision: Table 2 describes the changes in sensitivity but not the changes in specificity caused by varying the probability threshold. This seems to be an important omission, because one would expect a significant decrease in
specificity if thresholds are lowered to a 50% cut-off threshold. I would encourage the authors to add (and discuss) these figures because they put the benefit of a slight increase of the percentage of records successfully matched into perspective of the "cost" of an increased false positive rate. If the authors do not have access to these figures (false negative rates), they should at least discuss the possible negative implications of lowering the threshold too much - as it stands, Table 2 would indicate that the lower the threshold, the better the overall result, which is simply not true. As all analyses were conducted on anonymised data it was not possible for us to check the actual source of any error, which could be done by reviewing individual clinical notes. We have added a paragraph discussing the effects of different thresholds to the discussion.

Discretionary revision: I would encourage authors to add a brief description of the split-file approach to anonymisation mentioned on page 5. This would improve the overall understandability of the paper to the reader.
A paragraph has been added as requested.

Discretionary revision: An "exact match" is defined in the paper as a match of name, gender, date of birth and postal code. It would be interesting to know how many cases of people with the same values exist in the reference database (NHSAR) and the PARIS database, i.e. how many false positive matches could be caused by exact matches.
It would seem possible though very unlikely, but we have not been able to address this at this stage.

Discretionary revision: The paper does not discuss the relative importance of false positive matches versus false negatives. In patient care, false positives are high risk (because they may lead to incorrect diagnoses or treatment) whereas false negatives may hide information from the medical doctor (which he/she would not have seen anyway otherwise). For research applications, this may be different and it would be of interest to read the authors' opinion on this.
We acknowledge the importance of false positives and false negatives, particularly where the information may be used to inform patient care. For research purposes, analysis can be run and re-run, including and excluding records matched at lower PRL thresholds to ensure consistency of results. We have added a paragraph to the discussion on this issue.

Discretionary revision: The technical environment in which the processing speeds have been measured should be described.
The data on processing speed has been removed as it was identified by reviewer one that this was not of particular significance.

Discretionary revision: An additional figure showing the different sources of data and furthermore showing what data is matched to what other data would be beneficial for the overall understanding of the reader.
A paragraph on the datasets has been added to the methods section to increase clarity.
Thank you for the reviewers’ comments – we hope that these have been addressed to their satisfaction. If you have any queries, please contact me and I will endeavour to provide you with any additional information you require.

I look forward to hearing from you.

Thank you.

Yours sincerely,

Dr Kerina Jones
Research Development Fellow
HIRU

Tel: 01792 602764
Email: k.h.jones@swansea.ac.uk