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: 3 Date: 9 January 2009

Author's response to reviews:

Health Information Research Unit (HIRU)
Centre for Health Information Research & Evaluation (CHIRAL)
School of Medicine
Swansea University
Grove Building
CHIRAL suite
Singleton Park
Swansea
SA2 8PP
8th January 2009

BMC Medical Informatics & Decision Making

Dear Editor,

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

Please find attached a second revision of the 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 have amended the author details so that only two are marked as having made an equal contribution, we have inserted the correct wording in relation to not having any competing interests, we have removed the title from the image file for Figure 1, and we have followed the instructions to crop the figure. (This has
been cropped to A5. If there is any problem with this, please accept our apologies and use the version submitted on 9th December. The content has not been amended.)

Thank you for the second set of comments from Reviewer 1. Again, we found these to be very constructive and helpful to us in further improving the paper. We have addressed the reviewer’s comments in the revised manuscript and a point-by-point description is given below. Responses are shown in italics.

Reviewer 1: Miranda Tromp

Minor essential revisions

1) I still feel the background section is too elaborate and should be focussed more on the specific subject of this paper. In addition, the conclusion paragraph is also quite elaborate and would benefit from making it shorter and more to the point.

Some sections of the background section have been removed, and some have been amended, to improve the focus and make it less elaborate. The section on the description of the split-file approach has been retained as this was added in response to a request from Reviewer 2 previously. We are conscious that, due to the nature of our work, our readership will include non-experts in the area of record linkage, and so some detail on the principles and values of record linkage have been retained. Also, we have retained a brief introduction to the work of HIRU and to the NHS for the benefit of international readers. The conclusion has also been shortened and focused.

2) It is still not clear to me from table 2 how the last line in the table 'Overall combining valid NHS, DRL & PRL (50%)' can have more records matched when compared to the former last line 'Valid NHS plus PRL 50%'. The only difference is the adding of DRL and DRL is exact matching based on the five variables. How can these additional 13 and 3 record pairs not have been included in the PRL 50% matching?

A brief paragraph has been added to the discussion section – ‘In table 2, the slightly greater numbers of records matched by the sequential process compared to PRL at the 50% threshold are most likely due to rare occurrences of duplicate records on the NHSAR. In those cases, records with an exact match on all five variables would be matched by DRL, but would not be matched by PRL as the highest score/second highest score would be <2 (as set out in Figure 1).’

3) Table 2 only mentions the number and percentage of matches records of the specific databases, information on error rate is mentioned in the text. Can the error rate be incorporated in the table?

As the calculation of the error rates depends on cross-checking using the validated NHS number, we have not added the error rates to Table 2, as the samples in this analysis also included records without (valid) NHS numbers. But,
we have amended Table 1 so that the error rates are incorporated there. Table 1 shows the results of assessing the accuracy of accepting the NHS number supplied in routine NHS data as the basis of a unique identifier. We have extended this table to include results on a GP dataset, as well as a PEDW dataset, and we show the results at various PRL thresholds (not just for DRL and PRL at the 50% threshold), with the corresponding error rates. The corresponding sections in the text have been amended accordingly.

Thank you once again for this reviewer’s constructive comments – we hope that these have now 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.

We 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