Author's response to reviews

Title: The SAIL Databank: building a national architecture for e-health research and evaluation

Authors:

David V Ford (d.v.ford@swansea.ac.uk)
Kerina H Jones (k.h.jones@swansea.ac.uk)
Jean-philippe Verplancke (p.verplancke@swansea.ac.uk)
Ronan A Lyons (r.a.lyons@swansea.ac.uk)
Gareth John (gareth.john@hsw.wales.nhs.uk)
Ginevra Brown (g.brown@swansea.ac.uk)
Caroline J Brooks (c.brooks@swansea.ac.uk)
Simon Thompson (simon@chi.swan.ac.uk)
Owen Bodger (o.bodger@swansea.ac.uk)
Tony Couch (tony.couch@hsw.wales.nhs.uk)
Ken Leake (ken.leake@hsw.wales.nhs.uk)

Version: 2 Date: 18 June 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
18th June 2009

BMC Health Services Research

Dear Editor,

Re: Submission of revised paper – The SAIL Databank: building a national architecture for e-health research and evaluation

Please find attached a revised copy of a paper entitled: ‘The SAIL Databank: building a national architecture for e-health research and evaluation’. The paper is included in Word (.doc) and there are two separate Word files with Figures 1 and 2. The titles of these additional files are:
We found the reviewers’ comments very 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: John Parkinson

1. Major compulsory revisions

None

2. Minor essential revisions

2.1 The abstract does not currently give the population/nation cover of SAIL and this is also not stated clearly or early in the main part of the paper.

A sentence has been added to the abstract and to the section headed ‘Datasets incorporated’.

2.2 On page 4 the authors’ have not included any statement about the potential for disclosure of any dataset where its population cover is known in a geographic sense when other data, particularly with dated, timed events can be accessed by researchers. The paragraph on p.4 beginning ‘However, anonymisation alone..’ has been expanded to include this issue.

3. Discretionary Revisions

3.1 The dataset has been set up as an “anonymised” dataset and there is no discussion about an ability to validate data that could have been enabled were a pseudonymised approach to have been taken. On the one side this can be taken as a strength, on the other a potential weakness. Perhaps there should be some discussion on this point.

As the topic of validation is covered in another publication (reference 27), we have not amended the text, but an explanation is provided here. It might have been an easier process if we could have used re-identification to verify and validate, but it is a part of our mode of operation that we do not access any identifiable data, or the keys to re-identify. The method we have developed has given us the opportunity to demonstrate a reliable process using only anonymised data, but admittedly, it is more difficult to assess errors - false positive and negatives.

3.2 In the results section there is no mention about data encryption on transfer? This may or may not be relevant. We have included information in the text (From data source to databank) about HTTPS, the NHS Wales National Switching service and the Blowfish algorithm for encryption.

3.3 There appears no mention of what fields are used in MACRAL. These are: first name, surname, gender, postcode, date of birth. The text (From data source
to databank) has been amended to include these.

3.4 In the discussion para 1 the authors might like to review whether it is the work of the OSCHR E-Health Records RB or of the UKCRC-CfH RCP.

Much of this work is in the developmental stages at present with some roles and responsibilities yet to be assigned. We have amended the text to say it’s under the work of the OSCHR E-Health Records RB, rather than by that board.

Reviewer 2: Fortunato Castillo

1. Major compulsory revisions

1.1 The authors chosen format for publication seems to be inappropriate for the message that is to be conveyed. The current format draws a somewhat artificial distinction between Methods and Results. The format of the paper is an important consideration and we have been in discussion with the journal editors on this matter. Having considered a number of other possible formats, we were advised by the journal to use this format and so we have kept to this.

1.2 Page 6, Paragraph 1, Line 4. The independent internal audit is an extremely important element of this article and there needs to me much more detail here. The following information needs to be provided: • Who carried out the audit (what was the name of the international company?)
• What was the controls audit framework that was used?
• Would it be possible to provide the full audit report as an appendix to this manuscript?

We agree this is very important and further information has been incorporated into the text in the Methods section. Having been in consultation with Bentley-Jennison, unfortunately we are not able to append the full audit report to the manuscript. Hopefully sufficient information has now been provided in the text.

1.3 The authors outline a fairly complex journey for the SAIL dataset that begins at the Data Provider Organisation, through Health Solutions Wales and the Health Information Research Unit and out to researchers. The workflow needs to be represented more clearly. Although it is described within the text this should also be illustrated in diagrammatic form.

We have illustrated the workflow in Figures 1 and 2. Figure 1 shows the SAIL databank system including the journey of a dataset from the Data Provider Organisation, through Health Solutions Wales and HIRU and on to access by researchers. It also summarises the measures applied at each stage. Figure 2 shows the stages involved (including the splitting of the demographic data from the clinical data) from data provider via HSW and on to HIRU for construction of the databank.

1.4 Although the authors describe their record linkage technique that enables related person-based records to be linked both within and between datasets. They do not provide any evidence that they have actually linked any of the 500
million person-related records. If this has been achieved then we need to see details. Please provide the figure for the total number of people whose records have been collected in SAIL and some indication of the current success that has been achieved in linking across all the datasets that are listed in Table 1.

The details of the matching process and its efficacy are the subject of another paper (reference 27). Through the method that has been developed whereby an Anonymous Linking Field (ALF) is assigned to each individual-level record (and within HIRU we use an encrypted version of this, an ALF_E) in the databank we are able to link the many disparate datasets, as stated in the ‘At HIRU’ section of ‘From data source to databank’. The matching success is consistently efficient with matching results with specificity values >99.8%, and sensitivity rates between 95% and 100% for a variety of datasets (Results – addressing the objectives). A sentence has been added to the first paragraph of ‘From databank to utilisation by researchers’ to clarify and emphasise record-linkage. A sentence has been added to the abstract and to the section headed ‘Datasets incorporated’ to provide information on the numbers of people whose records have been incorporated into SAIL.

1.5 Page 6, Paragraph 1, Line 9. The authors make reference to wide consultation with regulatory and government agencies but provide no evidence of this. This should be provided, possibly in the form of a table to indicate the bodies with whom the authors have consulted, the nature of this consultation and key points that have emerged. Relevant correspondence with these agencies should be included in the appendix to this article. The text indicated has been expanded to provide details of the main organisations that were involved in the consultation, and their key areas of engagement. We have done this rather than include copies of correspondence (which may not have been provided for wider circulation) in an appendix, and we hope that this provides sufficient information.

1.6 Page 8, Paragraph 2, line 3. How is the process of due diligence supported by HIRU? Please explain.

The text has been amended to provide an explanation.

1.7 The Information Governance Review Panel is referred to however its Terms of Reference are not provided. How does the IGRP assess each proposal for IG compliance, SOPs and data management policies? How is disclosure risk assessed? The Terms of Reference need to be provided along with the Standard Operating Procedures and Data Management Policies. This is an important point and the text under ‘From databank to utilisation by researchers’ has been expanded to include an explanation on the operation of the Collaboration Review System. We have not included the documentation as an appendix as we are considering the possibility of using it in a future publication on systems for assessing proposals for the re-use of routine data. We hope that the information we have provided will be sufficient.

1.8 Page 12, Par 2, Line 4. What does the phrase “do not have facilities to allow data to be transferred out” actually mean? Please re-phrase.
An explanation has been added.

1.9 Page 12, Paragraph 2, Line 6. Please define what is meant by a “secure terminal”. An explanation has been added.

1.10 Page 12, Paragraph 2, Line 6. The concept of the HIRU data haven is introduced but not defined. Please define. An explanation has been added.

11. Page 12, Paragraph 2, Line 7. How do the HIRU team scrutinise the statistical outputs? Do they make use of any tools or guidelines or is this purely subjective human assessment?
An explanation has been added to the text.

1.12 Page 12, Paragraph 2, Line 8. Please provide the “Access Agreement” as an appendix to the article.
We would be happy to provide a copy of the data access agreement for the reviewer to see, but do not wish this to be included as an appendix.

1.13 Page 12, Paragraph 3, Line 6. The assessment areas are referred to by letters a – e however these seem somehow to relate to 1 – 7 in the methods section this is confusing and needs to be corrected.
Apologies for any confusion, but they were referred to by letters a to e in the Methods section as well as the Results section. It’s the objectives that were referred to by numbers.

1.14 Page 12, Paragraph 3, Line 9. The data haven is not described above, as stated. (see point 9).
Please see points 9 and 10.

1.15 Page 13, Paragraph 1, Line 6. The backup schedule may be clear to the authors, however this should be stated in the text of this article. This has been added.

1.16 The authors consistently refer to the NHS Information Governance Framework as “Information Governance”. This is ambiguous and should be corrected. This has been noted at first mention.

Reviewer 3: Mark McGilchrist
1. Major compulsory revisions

(1) [Major] The paper describes the creation of a national healthcare databank that is compliant with UK legislation and recommended approaches to information governance. While the description is generally adequate I feel an opportunity has been missed to offer the rationale for the choices that have been made. Others wishing to pursue similar ends are seeking guidance. In particular, what additional principles have been employed to guide the solution? A central architecture has been chosen but no analysis is given of why this is the correct choice in this situation. That other architectures are possible is indicated in the
section on future work. Architecture is a driver for the costs, institutional arrangements, types of risk that result (present and future), scalability and external discipline. I feel more discussion of this is merited.

This is an important point and the first section of the Discussion has been expanded to draw attention to the main points that influenced the choice of architecture and systems.

(2) [Major] a) The mechanisms employed focus almost exclusively on a particular type of risk - the maintenance of patient anonymity - without an adequate discussion of what constitutes anonymous data and who should determine this. b) The paper seems to suggest that disclosure risks are mostly with the analyst community, but there are many other personnel within HIRU in contact with parts of the data, e.g. database administrators and other IT personnel. What is the rationale for placing trust in certain types of staff? c) Many other types of risk are possible relating to control, ownership and transparency, but these are not discussed in detail. d) From the receipt of data from DPOs through to publication HIRU has a high level of control and perhaps even veto. While I acknowledge the presence of the IGRP what are the consequences of this control for the external research community and the data providers given that HIRU will also use the data for analysis, and help vet the requests of others? e) Are there circumstances in which HIRU might reasonably argue in favour of its own analysts for certain types of requests? f) Does the review panel have the strength to deny requests it feels are not merited, even from HIRU? g) Does the panel draw on outside advice for this? Do the original DPOs have a say? h) Are the deliberations of the panel public knowledge? i) Some discussion on transparency and the extent of shared control of the overall process would be useful.

We have added lower case letters within this comment to help address these important points.

a) We agree that definitions of anonymous vary. The definition of anonymous is given in the introduction, in accordance with the MRC data & tissues toolkit. We have also included in the text that we use measures to assess disclosure risk in recognition that anonymous data may still reveal identity if unique or low copy numbers of records are present in the dataset, and that we operate measures to minimise disclosure risk at various stages of the process from DPO to researcher.

b) The main issues of disclosure risk are with the analysts as it is they who are involved with data to be viewed and used for research and with results to be released for dissemination. However, the other types of personnel – technical staff and DBAs, etc. - are also required to agree to abide to the data access agreement and to behave responsibly with the data. Additional text has been added to the last paragraph of ‘From databank to data utilisation by researchers’ to emphasise this.

c) Additional text has been added to the discussion section.

d) A fuller explanation of the operation of the Collaboration Review System (CRS), including the IGRP and the role of HIRU in reviewing proposals has been
added.
e) The HIRU review includes resource implications of the proposal. That would be the main reason we would be unable to support an application provided all other factors were acceptable. In that case we may suggest alternative options or agree a delayed start.
f) Yes, the IGRP can deny requests and please see e). All proposals are treated equally – including proposals originating within HIRU.
g) Yes they can do so if they require further information or expertise.
h) They are not actively made public, but are available to applicants and other relevant parties on request. The views of the panel are shared among the panel members and with the HIRU team.
i) Additional information has been added where the CRS is described in Results – From databank to utilisation by researchers, and also in the discussion section, to address all the points relating to the operation of the CRS.

(3) [Major] A discussion of what motivated DPOs to participate would also be helpful. Does the mechanism actively assist DPOs to improve their function and the coverage, completeness and quality of their data, or are the beneficiaries elsewhere? Do DPOs have concerns about loss of ownership and control of their data?
This is an important consideration and relationships with DPOs are of great value to us. Additional text has been added to the Discussion section.

(4) [Major] The authors aspire to a much larger facility presumably involving non-NHS person-specific datasets. A more comprehensive discussion of barriers to scalability is therefore necessary. Beyond the danger of increasing workload on this central facility from the perspective of formats, coding, semantics, quality, and completeness of data as acknowledged in the text, is there a recognised limit as to the kind and size of organisation that would contribute data to HIRU? Would HSW be an acceptable linker to non public-sector bodies? It would be nice to know what boundaries are contemplated for this facility.
Additional text has been added to the Discussion section to address this valuable point.

(5) [Major] a) It would also be interesting to know the balance between data requests where HIRU plays an active part in the study/analysis and where it acts as a disinterested party in the provision of data views to external researchers. The issue of HIRU as an interested or disinterested party is important when considering the maintenance of discipline in its processes and future developments. b) While an independent audit has been performed, how often will audits like this take place? c) In considering future changes to the facility have any limits been placed on this now by other parties as a safeguard?
We have added lower case letters within this comment to help address these important points.

a) In the early days HIRU played an active part in most requests for the
preparation of data views, but this is changing as more externally funded studies are commencing, with skills and resources for routine data analysis. The development of a secure remote access system will also influence this dynamic. Additional text has been added to ‘Utility of the databank’.

b) We are currently in discussion with Bentley-Jennison in connection with a review of the changes we have implemented since their visit last year. Audits are to be conducted once a year, but the frequency is also dependent on any relevant changes to regulations and legislation and to any significant changes we make to our systems. Additional text has been added to Results – ‘Independent Internal Audit’.

c) No limits have been imposed to date, but plans for future developments are subject to the approval of the National Assembly for Wales, the IGRP and our organisational management structures. Additional text has been added to Future work.

(6) [Discretionary] My reading of the paper is that no data from the databank (view) leaves HIRU, but I would like to be clearer about what actually can leave HIRU. Remote data access is mentioned as a future development but presumably at present external researchers must travel to HIRU to make use of the data. It would be useful to hear of feedback from external researchers with this approach.

Researchers have access to data views and are not able to transfer data out of SAIL, but results of analyses are permitted to leave after they have been scrutinised. This will also be the case with the Remote Access System being developed. However, we are also developing a method to package and transfer specified data to authorised projects and additional information on this has been added to future work.

(7) [Discretionary] In summary, I would like to have seen a clearer statement of why the databank and its procedures look the way they do with more space given to discussing control, ownership, transparency and conflicts of interest. However, the text is fairly clear about what SAIL is offering to those looking to do studies.

In answering earlier comments, additional text has been added at various points in the paper to address this within the scope of the paper.

The editor also provided us with guidance as follows:

"Authors need to respond to minor discretionary comments (Parkinson) and the major comments (Castillo and MacGilchrist). I suggest that the authors consider some additional diagrams to show workflows, and possibly the risk assessment processes. There probably should be additional references to previous work, studies, reports in the UK (and Wales in particular), and the authors may wish to consider some additional appendices, and (if appropriate) to highlight any aspects that may be discussed in more detail in later publications".

We have addressed the reviewers’ comments through explanations within this letter and amendments to the text of the manuscript. We have also added a second diagram and have described the risk assessment processes more fully.
We also have included more information on the internal audit and have broadened the discussion on types of risk involved in this area of work. We have added additional references and have expanded the future work section.

Thank you for the reviewers’ comments and the editorial guidance. The advice received has been particularly helpful and we hope that the points raised have been addressed to their and your 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