Reviewer's report

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

Version: 1 Date: 21 August 2008

Reviewer: Miranda Tromp

Reviewer's report:

In this paper the SAIL databank is introduced which holds many datasets in the field of health and social care. Purpose of the current study is to develop and implement an accurate matching process to enable assignment of a unique identifier which would make SAIL ready for record linkage research studies.

The creation of such a databank is very promising and offers many possibilities for future research. However, my main concerns for this manuscript are with the method section from which it is not clear what has been done exactly, which complicates the interpretation of the reported results.

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. 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.

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. 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? What is meant by ‘derived NHS number’? 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?

3. The second research question addresses the probability threshold in PRL. The matching algorithm MACRAL – how are the probabilities calculated for PRL? MACRAL makes sequential passes, are matching probabilities than determined for each pass? 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?

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

5. The method section should include a paragraph on the outcome measures by which the matching process is assessed. Definition of sensitivity and specificity should be given. 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?. Also error rates are reported in the result section, please give definition of ‘error’. The part in the discussion section on outcomes in record linkage should become part of this method section.

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.

2. It is not clear what is understood under a ‘valid’ NHS number.

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.

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.

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.)

6. The second paragraph of the conclusions contains no conclusion of this study. One closing paragraph on future work would be adequate.

Level of interest: An article of importance in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I declare that I have no competing interests.