Reviewer's report

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

Version: 1 Date: 17 October 2008

Reviewer: Marco Eichelberg

Reviewer's report:


- 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.

- 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.

- 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.

- 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.

- 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.

- Discretionary revision: The technical environment in which the processing
- 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.

Level of interest: An article whose findings are important to those with closely related research interests

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.