Reviewer’s report

Title: Medical record linkage in health information system by approximate string matching and clustering

Version: 2 Date: 1 August 2005

Reviewer: catherine quantin

Reviewer’s report:

General

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

The author must respond to these before a decision on publication can be reached. For example, additional necessary experiments or controls, statistical mistakes, errors in interpretation.

2.2.3. Page 8, line 5: the authors should give an argument for the choice of 0.01 in the formula 1 – “0.01 times the number of additive characters”.

The authors should give a specific reply to this point : I did not find a specific reply in the answer to Mr Yancey.

2.3. Results section:
This section should begin with a figure describing the different sub-populations and the relation between them, and in particular the sample sizes and characteristics of each sub-population of record couples.

I admit that the presentation of these results is now more detailed. However, I still consider that a figure would help the readers in understanding how all the results are intricated (or nested)

2.4.2. The argument of absence of gold standard is not acceptable in so far as it is always possible to make a validation of administrative data either on a subsample (by manual validation with medical records) or by linkage of hospital data with validated medical data, such as a registry for patients with in-hospital records, as previously done by Quantin et al, (Methods of Information in Medicine, 1998 and 2005).

The authors could have tried to perform such a validation, even by using simulated data if a clean database was not available. But I understand that this would require an extensive additional work. The paragraph is clearly rewritten and one of the proposed references was cited (why not for the second one ?).

Minor Essential Revisions (such as missing labels on figures, or the wrong use of a term, which the author can be trusted to correct)

The author can be trusted to make these. For example, missing labels on figures, the wrong use of a term, spelling mistakes.
7.5 Table 1: This table is composed of two parts. The first part should be clarified. The title of the rows and columns may be misleading: it is difficult to understand that total correspond to the number of couples, and even more to understand the meaning of 100%.

The authors should divide this table into two separate tables so that the presentation of the first part would be clarified.

Discretionary Revisions (which the author can choose to ignore)

These are recommendations for improvement which the author can choose to ignore. For example clarifications, data that would be useful but not essential.

Please note that both the comments entered here and answers to the questions below constitute the report, bearing your name, that will be forwarded to the authors and published on the site if the article is accepted.

What next?: Accept after minor essential revisions

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

Quality of written English: Acceptable

Statistical review: No

Declaration of competing interests:

I declare that I have no competing interests.