Reviewer's report

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

Version: 1 Date: 1 June 2005

Reviewer: catherine quantin

Reviewer's report:

General

This paper addresses very important questions regarding record linkage.

-----------------------------------------------------------------------------------------------

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

1- GENERAL COMMENTS:
This paper presents the reviewer with three major problems:
- First, all the sections should be reorganised as detailed in the specific comments. Especially, the main objective of the study should be clearly stated at the end of the background section.
- Second, the interest of choosing arbitrary weights in the computation of the global similarity value is not explained in the paper. This choice makes the proposed method data-dependent and so the generalizability of the results is questionable.
- Third, this paper should be revised by a native English speaker: some expressions sound unusual (wife name, sexes, birth dates)

2- SPECIFIC COMMENTS:
2.1 Background section:
This section is too long (almost 4 pages) and provides too much details on methods previously published by other authors. This section should be reduced up to 2 pages, and should stress more specifically on the limits of the previous methods and the justification of the complementary approach (in particular, building clusters of duplicates) proposed in this paper. Furthermore, this section should provide a clear but concise justification of the objective of the present study.

For instance:
2.1.1 Page 3, last paragraph and page 4, first paragraph: the explanation of the complexity problem and the description of the four blocking methods are too detailed for this section and should be placed in (and adapted to) the methods section.
2.1.2 Page 4, last paragraph and page 5, first and second paragraphs: the description of the Porter-Jaro-Winkler algorithm (basic formula and improvements) is too detailed for the background section and should be copied to the methods section, with addition of the final formula, which is unexpectedly absent from this paper (to understand how the value 1 is obtained according to the initial algorithm page 8 line 4).
2.1.3 Page 5, at the beginning of the second paragraph (lines 11-14). Again, this part is too developed for the background section and should be copied to the methods section. In this section, this explanation should indeed be illustrated with an example and with the formula of the global
similarity value (weighted mean of atomic similarities).

2.2 Methods section:
2.2.1 Page 7, line 6 (from the bottom): I do not really see the point of using these weights rather than the method proposed by Jaro-Winkler based on the Expectation Maximization Algorithm.
2.2.2 Page 5, end of the second paragraph: the limits of the weights obtained by the Expectation Maximization algorithm are not very convincing. The fact that non-literal data fields, such as dates or numbers, have to be converted into strings before processing is not a so important drawback as this step can be easily achieved during the normalization phase, as suggested by the authors.
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.

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

2.4 Discussion section:
2.4.1 The choice of the canopy clustering method should be fully explained in the first part of discussion section, by comparing the advantages of this method with the other methods published in the literature.
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).
2.4.3 The authors should clearly specify their intention to develop a weight computation which is not domain dependent, using for example the Expectation Maximization algorithm.

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

Background section:
1. Page 5, lines 6 and 7: it is difficult to admit that both Most of typographical errors are not located at the beginning of the string, but rather in its body and The use of inverted keys allows comparing strings with an error on the very first characters, which seems to be common in patients database (page 7, lines 19-20).
2. Page 7, line 23: this sentence should be rewritten as: their characteristics such as date of birth and gender are equal.
3. Page 7, line 24: the second condition, required for the comparison of two records sharing the same canopy, should be clarified.
4. Page 8, third paragraph: as the definition of weights is somewhat arbitrary, their detailed description should be extracted from the text and provided as a Table (or an appendix). The author should justify the choice of arbitrary weights instead of probabilistic algorithms classically used for data linkage.
5. Page 9, line 7: would be impossible without should be replaced by is rended possible by.
6. Page 9, line 3-5 (from the bottom): adding-up all the figures of the incomplete graphs leads to 3 206 individuals rather than 4 677, as proposed in the first line of page 10: the discrepancy between
these results should be clarified.

7. Table 1: This table is composed of two parts. The first part should be clarified. The title of the mono 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%. As $28,517 + 9,566 = 38,083$; the point of coding the now rest is not obvious.

Discretionary Revisions (which the author can choose to ignore)

**What next?:** Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions

**Level of interest:** An article of importance in its field

**Quality of written English:** Needs some language corrections before being published

**Statistical review:** No

**Declaration of competing interests:**

I declare that I have no competing interests