

Reviewer's report

Title: Automated De-Identification of Free-Text Medical Records

Version: 1 **Date:** 3 February 2008

Reviewer: Pierre Zweigenbaum

Reviewer's report:

This paper presents a program for de-identifying free-text medical records and its evaluation on nursing notes. It has the following strong points:

- a good section on related work;
- a thorough description of PHI detection methods;
- a precise description of corpus preparation and corpus contents and features;
- the re-identified gold standard corpus is made available;
- a very informative presentation of results, including a breakdown according to the different types of PHI and an error analysis;
- a honest and lucid discussion of results, with pros and cons of the implemented method.

I have found no weak point in the paper, which I enjoyed reading.

Major Compulsory Revisions

None.

Minor Essential Revisions

1. Examples of regular expressions would be welcome. They would help more NLP-oriented readers to get a more precise idea of the kind of strings which can be matched.
2. p. 27 refers to "The initial version of the de-identification software". Earlier in the paper I thought I had understood that the software described in the paper was an improvement over an initial version. But this sentence seems to mean that the older version is the one which was evaluated on the test corpus.

Discretionary Revisions

3. The UMLS is described as a "collection of standard medical dictionaries": "vocabularies" would seem more appropriate. Can all vocabularies included in the UMLS be called "standards"?

4. p. 20, Locations:

How was the list of locations compiled? How large is it?

5. p. 26, why not measure precision on the test corpus too? The effort needed to check for false positives seems much lower than that needed to check for false negatives, given the estimate of about 474 instances of PHI per 100,000 words, which predicts about 1,400 PHIs in the whole test corpus (about 100 per reviewer).

6. p. 33, "It should be note*d*" is the only typo I could spot in the whole paper.

7. p. 40, the URL "<http://www.physionet.org/physiotools/deid/>" is correct but the link in the PDF is incorrect (it includes the next words).

What next?: Accept after minor essential revisions

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