Reviewer's report

Title: Development and evaluation of a de-identification procedure for a case register sourced from mental health electronic records

Version: 1 Date: 8 December 2012

Reviewer: Adrian Benton

Reviewer's report:

Major Compulsory Revisions

1) I would be curious to see how using a pre-existing de-identification system on your corpora intersects with the PIs that CRIS identifies. Assuming that your system is comprehensive, there should be few credible PIs identified by the other system that your system does not spot. One strong, state-of-the-art baseline is the MIST system:


This could be done over your entire corpus as well, and a potentially small subset of the "mismatches" could be evaluated by hand. MIST also allows the user to import dictionaries (scraped from the structured fields of your EHRs), so would be comparable to the method that you employ. Showing that your de-identification method outperforms a strong, existing baseline would give it greater value.

Minor Essential Revisions

2) Your method of populating a dictionary of PIs from fields in the records under consideration is very similar to:


Although the records are not EHRs, the method of drawing information from structured fields in the record to inform de-identification is similar in spirit.

2) You should also report precision and recall after including the 'potential' PIs in your calculation (enumerated in Table 9). Claiming 100% precision and recall over the PIs explicitly noted in the structured fields can be misleading, without reporting the precision and recall when considering 'potential' PIs as well.

3) How extensible is this system to other domains? I understand that this was tuned for the SLaM EHR corpus; I am just curious to note if it could be applied more broadly.
4) I assume that the test set that CRIS was evaluated on was disjoint from the set that it was developed on. This point could be made clearer.

5) In "Other approaches to de-identification" paragraph 1, you note that the algorithm presented in Tu et al de-identifies records at a much slower rate than your algorithm. This seems like a very minor concern when dealing with de-identification of EHRs, since de-identification would be performed on a single pass. With more computing power/parallelization/optimization the aforementioned algorithm could perform within a reasonable amount of time. Processing time should not be a major concern for this problem.

Discretionary Revisions

6) I am not very familiar with psychiatric case registers. Do they, on the whole, include all of these structured PI fields, or is this system only applicable to the SLaM corpus? How simple is it to generalize the CRIS system.

7) Although you have explicitly stated that identifying misspellings is not the aim of this paper, in other domains, identifying these non-exact matches is critical. I am surprised to read the reported precision and recall and cannot believe that there would be more typographical errors in the free-text fields. A method which determines the minimum edit distance to tokens in your current PI dictionary, checks to see if the token is a common word, and if the minimum edit distance exceeds a threshold and is not a common word de-identifies it may be sufficient. Consider evaluating such a modification to your de-identification algorithm.

8) I am curious to know why you chose to replace all PIs with "ZZZ..." tokens, and relative-related PIs with "QQQ..." tokens. It may be useful for researchers if there were different tokens for different types of PIs, especially if they are applying automated methods of analyzing the EHR free-text.

9) I applaud your security model and your awareness of the risks posed by automatically de-identified corpora, despite reporting 100% precision and recall.

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