Reviewer's report

Title: De-identification of Primary Care EMR Free-text Data in Ontario, Canada

Version: 1 Date: 6 April 2010

Reviewer: Pierre Zweigenbaum

Reviewer's report:

- General comments

This paper presents an experiment where an existing de-identifier (DE-ID) designed in the United States and tested on nursing notes is adapted to the context of Ontario and tested on primary care EMRs.

A serious effort was performed to collect relevant knowledge to update the system: names of potential patients and doctors, locations, etc.

The adapted system was tested on a dataset which is different (primary care) and more varied (multiple types of documents by multiple authors in multiple different geographic locations) than the original system.

Three large datasets were manually tagged to build a reference corpus of de-identified notes, part of which was used to develop the system, and another part to evaluate it.

Therefore this work, though not designing a new methodology, successfully presents the further adaptation and test of an existing method to a new context.

- Major Compulsory Revisions

None.

- Minor Essential Revisions

1. A range of previous work on de-identification is cited in the Background section (3-12), but is not described in more detail. A summary of the different classes of methods used in these former studies would better put the present work in context. This is all the more important as adaptation methods depend on the approach implemented in the initial system: adapting a machine-learning-based system would be different from adapting a lexicon- and pattern-based
system such as DE-ID.

2. Other teams have attempted to re-use DE-ID:


The present paper should compare its methods and results to these different attempts.

3. Note that name removal based on known names of patients and hospital staff is also implemented in DE-ID, though in a slightly different way. This should be mentioned when presenting the initial name removal phase.

- Discretionary Revisions

How representative is Practice Solutions® EMR of Ontario's practices?

On p. 8, Names,

"'Ambiguous' names are only removed if they occur beside a first name or a last name, or if there is no immediate word preceding or following such as 'Dr.', ...":

shouldn't it be "if there is *an* immediate word..."

p. 9, "in order to detect those entries which had a high likelihood of being a name":

Does "those entries" refer to those of the shorter list? Which action was taken for these?

p. 10, "As well, it included medications, common medical acronyms, parts of the body and words that are commonly used in a clinical context.":

Can you please give an example?

p. 13, repeated "can distort"
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