Author's response to reviews

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

Authors:

Karen Tu (karen.tu@ices.on.ca)
Julie Klein-Geltink (julie.klein-geltink@ices.on.ca)
Tezeta F Mitiku (teze_81@yahoo.com)
Mihai Chiriac (mihai.chiriac@ices.on.ca)
Joel Martin (Joel.Martin@nrc-cnrc.gc.ca)

Version: 2 Date: 7 May 2010

Author's response to reviews: see over
Cover Letter/Response to Reviewers

Thank you for the opportunity to revise our manuscript. I have taken the Reviewers comments and pasted them below, my responses are directly underneath in blue letters. Included are the clean and track changes versions of the manuscript.

Reviewer 1

Major Compulsory Revisions
Page 4: clarify the used to define PHI for the purposes of this study. It is unclear if the authors use the HIPAA definition or a Canadian one. For example date of birth is PHI, but age may not be (at least in US). Ethnicity may not be PHI either in US

We did not use the HIPAA definition for PHI but instead only intended to remove the PHI that would not affect the clinical content. ICES is a ‘prescribed entity’ under privacy legislation that has the capacity to link to administrative data holdings (including dates and reasons for hospitalization, dates and reasons for physician visits) that are linked by a healthcard. We are permitted to take individual level data and strip identifying information in-house or partition identifying information to be stored separately away from the main bulk of the data as per standard ICES policies and procedures that are reviewed regularly by the privacy commissioner of Ontario. Thus the aim was only to remove the ‘stronger’ identifying information as described in the manuscript and retain the identifying information that would affect the clinical interpretation of the data. For instance HIPPA would require dates of hospitalization be removed but the date of hospitalization could have clinical relevance when looking at things such as events that have occurred in a given time period, length of stay, mortality rates post discharge, readmission rates… Ethnicity is not part of PHI thus we have deleted this from the paragraph in the introduction describing HIPPA. We have described the prescribed entity status and the goal to preserve clinical content in the introduction and have acknowledged the limitations for generalizability of our work in the discussion. Nonetheless from the other papers I’ve read and referenced on this topic preservation of clinical content has not been a priority and I think that this issue is important as the intent for obtaining EMR data and de-identifying it is to render the data useable for research purposes.

Page 6: More detail on the creation of the reference set is needed. How were they manually tagged. How many evaluators did this? What was inter rater agreement?

The reference set was tagged first by one of the project personnel. All of the tagged records were run through the program which generated a list of words
that were removed, false positives and false negatives. The lists were reviewed in detail and any word that appeared to be incorrectly removed by the program was reviewed by using a simple word search function to identify where it appeared in the text. If a tagging error was made the tag was corrected. This was repeated several times for each data set until we believed there were no more tagging errors and then the data was run through all the steps described. This has detail has been added to the Methods.

Page 9: Provide more detail on what an eponym is and an example for readers unfamiliar with the term

An eponym is the name of a person after which an item is named or thought to be named after. In medicine it is often the name of the physician, discoverer or patient with the disease, clinical sign or syndrome. I have added a descriptive sentence to the Methods.

Minor Essential Revisions
The first paragraph in the background should include some references to support the assertions made.

The relevant references have been added.

Page 4 - although a number of references are provided about de-identification studies it would be helpful to have a more thorough discussion about some of the more pertinent articles

This has been added to the introduction.

Reviewer 2

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.

This has been added to the introduction.

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


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

We have added a paragraph to the discussion section to compare these other attempts to ours.

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.

This has been added in the Methods section under Names subtitle on page 9 of the manuscript.

Discretionary Revisions

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

Practice Solutions has approximately 50% of the market share in Ontario with no clear second place EMR vendor, this has been added to the methods

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..."

This has been corrected.

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?

'Those entries' refer to the names that are also on the list of the most common English words, the sentence has been fixed to better reflect what was done.

p. 10, "As well, it included medications, common medical acronyms, parts of the body and words that are commonly used in a clinical
Can you please give an example?

Examples for each have been added where the source of the list was not provided.

p. 13, repeated "can distort"

Thanks this has been corrected.

p. 16, ref 8, Uzuner

Thanks this has also been corrected.