Author's response to reviews

Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Authors:

Andrew J McMurry (mcmurry.andy@gmail.com)
Britt Fitch (britt.fitch@gmail.com)
Guergana Savova (guergana_savova@childrens.harvard.edu)
Isaac S Kohane (isaac_kohane@harvard.edu)
Ben Y Reis (benreis@gmail.com)

Version: 3 Date: 30 June 2013

Author's response to reviews: see over
June 28, 2013

Dear editorial office of BMC informatics and medical decision making:

We are pleased to submit our revised manuscript “Improved de-identification of physician notes through integrative modeling of both public and private medical text”. We have carefully considered the reviewers’ comments and revised the manuscript accordingly. Each comment is addressed in detail below.

Kind Regards,
Andrew McMurry, MS
Reviewer's report

Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 2 Date: 28 June 2013

Reviewer: Murat Kantarcioglu

Reviewer's report:

In this paper, authors discuss their machine learning based techniques for removing identifying information in medical notes. One novel aspect of this work is that they leverage the existing medical literature to model non-identifying medical terms. After this step, they use standard natural language processing and machine learning techniques to build models to classify terms in medical notes as PHI or non-PHI.

Major Compulsory Revisions:

From the discussion in the paper, it is not clear to me how the proposed system compares against the results of machine learning based techniques given in i2b2 challenge. I suggest authors to provide detailed comparison to existing machine learning based techniques in the context of i2b2 challenge data.

Revision

Background and results sections have been updated to use recall, precision, and F-score metrics to simplify the comparison to other i2b2 challenge submissions. Specifically, the background section describes how our approach is favorable with regards to recall albeit with lower precision. The results section was updated to specify how we would have placed had we competed in the original challenge.

Minor Essential Revisions:

Overall, I believe that the paper could be improved in terms of presentation. It is hard to follow the paper since many important terms are given in supplemental tables. I suggest authors to incorporate more information from table 1 and supplemental doc file to make the paper more self-contained.

Revision

The paper was revised to use standard IR terms (recall, precision, F-score) to make it easier for the reader to compare the results. Table 1 is now limited in scope to the “Different Type of PHI” section, and no longer used as a scoring measure. Supplemental Figure S1 is now in the main body of the manuscript. Similarities and differences between public and private medical texts have are now summarized in the body of report (Figure 6). Lastly, the list of HIPAA identifiers are mapped to the list of I2b2 challenge identifiers in the supplementary doc in Table S3.

Discretionary Revisions:

There is some existing work that shows that (e.g., A.C. Solomon, R. Hill, E. Janssen, S. Sanders,Privacy and De-Identification in High Dimensional Social Science Data Sets, The 32nd Annual IEEE Symposium on Security and Privacy, Oakland, California, May 22-25, 2011) certain combination of medical terms can be used to uniquely identify individuals. I suggest authors discuss how such attacks may affect their proposed system.

Response

The authors internally discussed this potential vulnerability at length with the i2b2 data providers and determined that the risk is at or near zero for several reasons: 1) individual medical terms
cannot be combined because individual medical terms are not reported; 2) even if one were able to learn a single medical term that term could not be combined because each token is annotated individually and pooled among all note instances; 3) no words are shared only aggregate word features of synthetic names generated by the i2b2 annotation team. In total, the authors believe strongly that it would be nearly impossible to recreate the original physician notes using the provided model of randomly selected tokens and aggregate features.

Also, the discussion about not using n-grams is not very convincing. I believe that a single machine with multiple cores can easily build models with n-grams on reasonably sized data sets (e.g., few TBs).

Response
The authors agree that the original argument was not convincing. The greater concern for using n-grams is that the training model could overfit section headings and physician writing styles in the same way that CRF (conditional random field) models often do. In our experience, overfitting a single set of training instances is the most difficult challenge to de-identification. We were therefore cautious to avoid any features that we believed could overfit the training examples and eager to share our annotated examples. Furthermore, the authors strongly want to encourage sharing training and test datasets, and we were not convinced that this would be possible if we shared long sequences of annotations (such as entire sentences).

Revision (discussion section)
"However, common and rare word sequences can vary considerably across the different types and formats of physician notes and journal publications. We chose instead to err on the side of caution and use a single token model rather than n-grams or conditional random fields."

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.
Reviewer’s report

Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 2 Date: 28 June 2013

Reviewer: Joel Martin

Reviewer’s report:

General comments

The authors present a new system to address de-identification of medical records and apply it to the I2B2 de-identification dataset. They achieve somewhat competitive results by using a novel source of information: the frequency of words that appear in non PHI sources.

1. Is the question posed by the authors well defined?

The question is well-defined with a clear measure of success.

2. Are the methods appropriate and well described?

The methods are appropriate and fully replicable.

3. Are the data sound?

The experimental data appear to be sound. They are reported differently than the original I2B2 experiments. It is time consuming for the reader to compare. I believe it is the authors’ responsibility to directly compare with the past results on the data. They make a brief mention early in the paper, but there are no quantitative comparisons. This makes it difficult to know the value of their novel feature or the practical value of their system compared to those that already exist.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?

Yes.

5. Are the discussion and conclusions well balanced and adequately supported by the data?

In a local sense, the discussion and conclusions are balanced and supported by the data. However, after reading the paper, I am not clear about the true value of frequency as compared to other features. For example, Szarvas et al. (J Am Med Inform Assoc 2007;14:574–80) successfully used frequency from the training set for de-identification. That system appears to have performed better than the current one. What is the incremental advantage, if any?

I am also concerned that some of the advantage of word frequency came because the dataset created artificial words that permuted the order of syllables to create terms like: Girresnet, Diedreo A. Since those new constructions would be highly unlikely in external sources, frequency gives an unnatural advantage. Will the feature be as useful for realistic datasets?

Overall, I am concerned that the paper does not adequately defend the value of the novel feature. They do show that two features (including this one) are sufficient for very high performance. That
is valuable and very interesting. However, they do not exclude some obvious uninteresting explanations of that performance, such as the artificial names.

Also, I am not sure what to do with the fact that other deidentification systems perform better than Scrubber on the same dataset. This is especially so because the authors did not provide any easy way to compare their system to the earlier ones. Since I did do the comparison, I know that I would use one of the other systems rather than Scrubber in a practical application. Maybe Scrubber is better/interesting because it is simpler, has fewer features, or does better on some alternative measure of success. Maybe, but the paper does not make this clear.

6. Are limitations of the work clearly stated?

I mention what I consider to be significant limitations of the empirical comparisons above. The authors do not discuss these. The authors do mention some other limitations.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?

Yes.

8. Do the title and abstract accurately convey what has been found?

The title and abstract are accurate.

9. Is the writing acceptable?

The writing is clear.

**Major compulsory revisions:**

i. show a quantitative comparative evaluation against previously reported systems that do better than Scrubber. Just quote the previously reported results.

*Revision*

Background and results sections have been updated to use recall, precision, and F-score metrics to simplify the comparison to other i2b2 challenge submissions. Specifically, the background section describes how our approach is favorable with regards to recall albeit with lower precision. The results section was updated to specify how we would have placed had we competed in the original challenge.

ii. explain clearly why Scrubber or the external frequency feature is still an important contribution

*Response*

The method reported here was developed for other note types such as pathology and radiology reports and clinical summaries. In a small sample annotated by a single human expert (20 cases, 3K PHI tokens, 66k total tokens) the NCI annotator reported terrific performance of PHI recall (98.6%) but did not provide us metrics for precision or specificity. Due to funding limitations we were unable to continue that study, however, the annotator was pleased to report that she could still easily read the scrubbed output. The performance was satisfactory for their IRBs.

iii. explain the interaction between the special words in the dataset and the use of frequency

*Revision*

Validation section has been added to the results section, which shows that the result is
not just an artifact of the i2b2 surrogate words.

iv. explain why Scrubber is an advance over, rather than just different from, other systems that us frequency within the dataset

*Revision*
Background section has been revised to describe the key differences between this method and other de-identification systems.

*Response*
Physician notes vary considerably with respect to type, format, and style. In the authors’ experience, these differences significantly impact the ability to reproduce de-id performance in other hospital settings. The top submissions in the i2b2 challenge (Szarvas, Aramaki, MITRE) are impressive but the methods use features that may be specific to the features in discharge summaries: section headings, sentence positions, and longer phrases of tokens.

**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.
Reviewer's report

Title: Improved de-identification of physician notes through integrative modeling of both public and private medical text

Version: 2 Date: 28 June 2013

Reviewer: Oscar Ferrandez

The paper describes an interesting approach that tackles the de-identification problem from a perspective mainly based on the probability distributions of non-PHI/PHI words. A decision tree classifier is built for this purpose, using public medical texts for learning non-PHI words and private physician notes for PHI terms.

Major Compulsory Revisions

1-PHI types

Authors refer to 8 PHI types (table 1). HIPAA specifies 18 different PHI identifiers, are these 18 HIPAA identifiers covered by the 8 types in table 1? Could the authors provide a mapping between types?

Revision
As requested, we have provided a mapping of the 18 HIPAA types to the 8 PHI types in the i2b2 Challenge (Table S3)

2-Features

I would like to see how a bi- or tri-gram language model would affect the results. Authors commented on calculating term frequencies using two or more words can lead to exponential time complexity; however I do believe there are efficient ways to process this information with reasonable computational cost.

Response
The authors agree that the original argument was not convincing. The greater concern for using n-grams is that the training model could overfit section headings and physician writing styles in the same way that CRF (conditional random field) models often do. In our experience, overfitting a single set of training instances is the most difficult challenge to de-identification. We were therefore cautious to avoid any features that we believed could overfit the training examples and eager to share our annotated examples. Furthermore, the authors strongly want to encourage sharing training and test datasets, and we were not convinced that this would be possible if we shared long sequences of annotations (such as entire sentences).

Revision (discussion section)
"However, common and rare word sequences can vary considerably across the different types and formats of physician notes and journal publications. We chose instead to err on the side of caution and use a single token model rather than ngrams or conditional random fields."

As the feature set is not very big, and the authors are using Weka, apart from the results shown in table 3 for each kind of features, I would like to know the meaningfulness of each feature, for instance, this could be obtained processing information gain, which should be very easy using the
Weka framework.

Revision
Per your request, Information Gain is now provided for each feature in the supplementary material (Table S4). This includes both the original i2b2 data as well as the validation dataset which uses real names from Medicare and the US patent office.

3-i2b2 corpora

Please note that as said in your reference [13], the i2b2 corpora was resynthesized with surrogates that, in most cases, could not be found in dictionaries, making their de-identification difficult for techniques based on dictionaries (“For patients, doctors, locations, and hospitals, we created surrogates by permuting the syllables of existing names from dictionaries such as the U.S. Census Bureau names database […] the generation of surrogates that could be found in dictionaries and did not make any effort to eliminate such surrogates from the data. Still, most of the generated surrogates could not be found in dictionaries, e.g., Valtawnprinceel Community Memorial Hospital and Girresnet, Diedreo A.”). This fact makes the term frequency for these identifiers very low and the approach proposed here could be favored by this fact. I would like the authors to address this issue.

Validation section has been added to the results section, which shows that the classifier performance is not just an artifact of the i2b2 surrogate words.

4-Patterned features

Is the approach using the regular expression set provided in [15]? Please clarify this when the feature set is described, also I would appreciate few comments about the usefulness of the regexs integrated into the approach and if some tweaking of these regex was performed.

Revision
The main manuscript has had the section titled “Regular Expressions” updated to answer questions regarding tweaking of regular expressions and effectiveness of use.

5-Results

In the summary-results section, authors state that the method obtains 99.4% sensitivity for patient name. Where is this number in table 3? With which set of features was this number obtained?

Moreover, I would also like to have the data about specificity for the 9-way classification, i.e. for each PHI type. Conclusions could be made in a more rigorous manner having this information available.

Revision
By request of multiple reviewers, the results of the risk-based approach (as favored by our IRB) was discarded in favor of the binary token classification model that is commonly reported in the i2b2 challenge. Table 1 is now limited in scope to the “Different Type of PHI” section, and no longer used as a scoring measure.

Response
The authors agree that IRBs have a different view of confidentiality risk than that of the i2b2 challenge.

6-Discussion
I want the authors to answer this question for me: in the big picture, why should I use your de-identification approach? What are the advantages of your approach with regard of other de-identification systems? I like the method and idea regarding non-PHI frequency terms extracted from public medical texts, however your approach does not obtain better results than the 2006 i2b2 challenge participants, and it seems that if I train a machine learning classifier or use other available de-identification systems (e.g., MIST) I would get better results (at least on the i2b2 corpora).

Response
Physician notes vary considerably with respect to type, format, and style. In the authors’ experience, these differences significantly impact the ability to reproduce de-id performance in other hospital settings. The top submissions in the i2b2 challenge (Szarvas, Aramaki, MITRE) are impressive but the methods use features that may be specific to the features in discharge summaries: section headings, sentence positions, and longer phrases of tokens. The method reported here was developed for other note types such as pathology and radiology reports and clinical summaries.

Importantly, annotations generated by this method were considered to pose nearly zero risk to patient privacy, which allows us to share the annotated feature sets with other NLP teams. We consider this a valuable contribution, especially given that the dataset is provided with open source Apache software.

Another interesting issue I would like the authors to address is the portability of your models to other work types (i.e., other clinical text apart from discharge summaries). The authors could consider the possibility of using some data from the iDASH repository (http://idash.ucsd.edu/idash-data-collections), such as the MTSamples. At the very least, I would suggest making reasonable comments on the efforts that should be done to adapt the approach to other work types, and some comments on the expected performance/applicability.

Response
In a small sample annotated by a single human expert (20 cases, 3K PHI tokens, 66k total tokens) the NCI annotator reported terrific performance of PHI recall (98.6%) but did not provide us metrics for precision or specificity. The performance was satisfactory for their IRBs. Due to funding limitations we were unable to continue that study, however, the annotator was pleased to report that she could still easily read the scrubbed output. Since we still do not know the exact number of false positives, we could not include this result in the body of this paper.

The MTSamples was a great suggestion and we investigated using this dataset. Unfortunately, the MTSamples data is not annotated for PHI and we did not have resources to annotate it ourselves.

The authors repeatedly requested other datasets for independent evaluation. For a full two years we requested data from related de-id efforts, including the Beckwith study, the MIMIC study, Geisenger health, and Kaiser Permanente. It is so difficult to obtain annotated datasets for de-identification that we made releasing our annotated dataset a top priority. We consider our annotated dataset to be a valuable contribution to de-id research.

- Minor Essential Revisions
- Discretionary Revisions
- Results
1-As the authors used the i2b2 corpus, could they say something about their approach and the top three systems participating in the i2b2 challenge? Maybe, some comments on the differences between your results and the top three i2b2 participants, reasons for such differences. (please take this into the context that your approach was evaluated after the challenge took place).

**Response**

The background should now make it clear that we were solving a different problem than that of the i2b2 challenge: our goal is to learn the properties of PHI and non-PHI tokens so that we can train the model to work on different types of physician notes or different medical centers. For comparison, the results have been modified to use the standard i2b2 IR metrics (recall, precision, F measure).

**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