Author’s response to reviews

Title: Customization Scenarios for De-identification of Clinical Notes

Authors:
Tzvika Hartman (tzvika@google.com)
Michael Howell (mdhowell@google.com)
Jeff Dean (jeff@google.com)
Shlomo Hoory (hoorys@google.com)
Ronit Slyper (rslyper@google.com)
Itay Laish (itaylaish@google.com)
Oren Gilon (ogilon@google.com)
Danny Vainstein (dannyvain@google.com)
Greg Corrado (gcorrado@google.com)
Katherine Chou (kic@google.com)
Ming Jack Po (mingp@google.com)
Jutta Williams (juttawms@google.com)
Scott Ellis (scellis@google.com)
Gavin Bee (gavinbee@google.com)
Avinatan Hassidim (avinatan@google.com)
Rony Amira (ronya@google.com)
Genady Beryozkin (genady@google.com)
Idan Szpektor (szpektor@google.com)
Yossi Matias (yossi@google.com)

Version: 1 Date: 29 Sep 2019

Author’s response to reviews:

Comment from Reviewer 1
Q1. In error analysis of a fully customized system (page 13), several false positive cases were concluded as some names can be a person's name or a medical term. Can you address some possible solutions, for example, using medical terminology resource?

A1. This is a great idea, thanks! We added to the last paragraph of the fully customised system, p.13: "which could be corrected using heuristics based on medical terminology."

Comment from Reviewer 2
Q1. I did not find any novelty in the manuscript.

A1. As seen in the related work section, there are not many works in the area of medical notes de-identification that confront the problem of large variations in notes between corpora (note type, or institutional convention). Because of the sensitive nature of the problem, this may be a real show stopper for automatic de-identification systems. This is the only work that I am aware of that confronts the problem by a systematic evaluation on a significant number of publicly available datasets. Our second contribution is evaluating across practical deployment scenarios for health organizations, which is also novel.

Q2. I found the method did not explain clearly.

A2. We added some explanation about the definition of PHI in the first paragraph of the datasection. Other than that, we would need more detailed feedback.

Q3. The authors did not present and explain the results in a clear way. For that, the authors need to re-organize and re-write the manuscript clearly.

A3. Need more detailed feedback to respond.

Q4. The language needs to repair.

A4. Need more detailed feedback to respond.

Comment from Reviewer 3

Q1. The definition of de-identification seems to be ambiguous. On page 4, the authors provide this definition: "De-identified clinical datasets are created by labeling all words and phrases that could identify an individual, and replacing them with surrogate data or context-specific labels."

This is an admirably straightforward definition, but because there is no citation, it was initially unclear if the authors' definition was intended to meet the standard of de-identification defined by HIPAA. Later on page 7, the manuscript introduces the HIPAA Safe Harbor method (although a reference is again not provided), but interestingly, this sections describes that all of the source datasets actually go
beyond the 18 elements specified by Safe Harbor (1).

The abstract describes results on page 2 as: "Fully customized systems remove 97-99% of personally identifying information." It would be helpful to describe how the removal of personally identifying information relates to the process of de-identification, the extent to which these may be slightly different concepts, and how the decisions made for source data preparation relate to de-identification.

The manuscript also references the related concepts of anonymization and protected health information (PHI), in addition to personally identifying information (PII).

I believe that these concepts are especially important because the authors intend this paper to assist health organizations "...in order to choose the [de-identification deployment solution] that best matches their resources and target performance level" (abstract, page 3). A very important basis of such a choice is understanding how a given solution does (or does not) meet a specific standard of compliance. Therefore, if it is possible to define more of these foundational concepts, the reader may be better able to perform such an assessment and choice.

(1) I used the HHS Website for my own reference here, "Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule." https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

A1. These are great points and we agree with the need for accuracy. The definition on page 4 is informal, to keep it simple for the reader. Similarly, we referred to "personally identifying information" in the abstract, risking vagueness in order to reduce acronyms and make our work more accessible to the reader.

For the remainder of the paper, we are striving for accuracy. We use "PHI" throughout the paper, and avoid "PII". We introduce the formal definition at the beginning of the data sources section on page 7; we added a citation there by your suggestion. In addition, we added a paragraph at the end of the discussion section to clarify how such a system might fit HIPAA law in practice.

Q2. What is basis for recall and precision results?
On page 11, the authors describe that the presence of PHI is used as the basis for recall and precision results. However, it was unclear to me how the analysis could be based on PHI, given that in the in the description of source data on page 8, the datasets are described as being de-identified text (not containing PHI).

Are the recall/precision results perhaps based on the surrogate information that was used to replace the actual PHI in the datasets? But if PHI was in fact used, describing the permissions for appropriate use of patient data (such as approval by an Institutional Review Board) would be an important element to include.

A2. Great question, we see now where we were unclear. We added the following text to the first paragraph of the dataset section:
"These datasets were de-identified by replacing PHI with plausible but realistic surrogate information;
we evaluate our systems on this surrogate PHI."

Q3. How well can results be expected to generalize across all clinical note types? Clinical notes can cover an enormous continuum of purposes and settings, and it might be intuitive to assume that different note types will have very different characteristics. For example, a procedure report written by a radiologist perhaps looks very different than an advance care planning note written by a licensed clinical social worker.

It might be helpful to consider whether the context of the available datasets can be a basis for any insight about the variability of different note types, and whether the findings of this analysis will be broadly applicable to the entirety of notes within a given healthcare organization.

A3. We have answered this question in the paper, although we chose not to emphasize the comparison between note types too much. On page 4, the last paragraph beginning with "A blocking factor", we present the problem and say that "free-text clinical notes vary widely with note purpose and institutional conventions, and include PHI in ways that are challenging to identify and redact". In addition, in Table 1 on page 10, we specify the type of medical notes in each or the used datasets. In the off the shelf section on pages 14-15 we show how a model trained on one dataset performs on the other. For example train on mimic-discharge, test on mimic-echo.

Q4. How well can automated systems meet the standard for de-identification?

This manuscript seems to be especially relevant for evaluating solutions that remove (various levels of) personally identifiable information within clinical notes, but the reader may benefit from more consideration of where these solutions may (or may not) be accepted as meeting the standard of Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

For example, the authors write on page 4: "...a de-identification system [based on machine learning] cannot *guarantee* performance on all medical text it will ever encounter." It seems undeniable that automated systems will outperform human annotators, but it would be helpful if possible to describe established precedence where annotated systems have been accepted as meeting standards of de-identification.

The authors also examine this basis on page 19, where they describe use cases, "...e.g. when adding an extra layer of security while working with trusted research collaborators, or minimizing exposure of PHI to human annotators who will complete the de-identification task." These are important and relevant applications, although it might be confusing to the reader whether datasets prepared for these purposes will be intended to meet the HIPAA definition of de-identification, or whether a different compliance/regulatory basis might be applicable.

A4. We added a paragraph at the end of the discussion section to clarify how such a system might fit HIPAA law. Automated de-identification can also play a role in a HIPAA-compliant data release, with the additional step of "Expert Determination," wherein a human expert in the field determines if the de-identification
process has ensured that "the risk is very small that the information could be used ... to identify an individual."

There is an option to discuss the recall/precision level of datasets such as I2B2-2014. As described in [9], its basic performance level for I2B2-2014 is two annotators + "multiple checks". From [7], two annotators achieve recall of 94% and precision of 97% which is similar to the levels achieved by the machine learning model. However, we do not know what "multiple checks" entails, and implying anything < 100% recall seems problematic.

Q5. How does this paper relate to recent findings in anonymization?
A recent piece in Nature Communications describes "Using our model, we find that 99.98% of Americans would be correctly re-identified in any dataset using 15 demographic attributes. Our results suggest that even heavily sampled anonymized datasets are unlikely to satisfy the modern standards for anonymization set forth by GDPR and seriously challenge the technical and legal adequacy of the de-identification release-and-forget model." (2) (I have no relationship with this research team.)

Anonymization is not the same as de-identification, of course, but does this work have potential relevance to this paper?


A5. Thanks for making this connection. We are aware of this work that was published after submission of our paper. I think it goes beyond the scope of the paper. One of the issues it raises is about the delicate nature of surrogation..

Q6. Will a recent lawsuit be relevant to these findings?
A lawsuit was recently filed against Google and University of Chicago because of data sharing practices (3) I A lawsuit was recently filed against Google and University of Chicago because of data sharing practices (3). I have no particular knowledge of this lawsuit beyond what is available in the news, but it appears that this lawsuit may pertain to the de-identification of clinical notes, which is also the subject of this paper. Will this be relevant information for the reader?


A6. Pending litigation. We can't discuss it at this time.

Q7. Will the basis of the manuscript support open science and collaboration with other researchers?
I didn't see any indication that the code, documentation, or other artifacts of analysis are available to others to review or reuse, nor that the prepared datasets might be made available for other research use. Where possible, making such work available and accessible to other researchers can help to support transparency, reproducibility, and better science.
A7. We are going to release the labeling of Physionet following I2B2-2014 guidelines. Regarding the modelling code, our system is very similar to the NeuroNER open source code, we added a citation in the Text De-Identification System Architecture section.