**Reviewer’s report**

**Title:** Customization Scenarios for De-identification of Clinical Notes

**Version:** 0  **Date:** 23 Aug 2019

**Reviewer:** Shelley Rusincovitch

**Reviewer's report:**

Summary: This manuscript describes the results of a machine learning model to remove personally identifiable information from clinical notes. The authors describe the results of this model with 6 datasets created from previously existing data sources (i2b2 challenges, PhysioNet, and MIMIC-III). Then they extrapolate their findings to a broad set of conclusions intended to inform selection of de-identification tools by health systems.

As a reviewer, my specialty is informatics and the secondary use of EHR data for analysis, and I have experience in leading machine learning projects using clinical notes (which included the design to meet compliance with appropriate use of these data). I would recommend that a machine learning expert also review this paper to assess the strength of the quantitative methods and results.

********************************************************************************

0. This manuscript is highly relevant to the mission of how we can better use healthcare data to understand our patients and enable insights relevant to their care. The unstructured narrative within clinical notes (such as discharge summaries, progress notes, and procedure reports) is extremely important and rich in context often absent in other parts of the patient medical record. But use of clinical notes can be difficult because of the necessary compliance controls to protect the privacy of our patients, their caregivers, and members of their care team. In my experience, many people struggle with understanding this complex landscape and how to implement solutions that meet compliance requirements, and they lack practical information when they select or design de-identification processes.

I believe that this manuscript could be an important contribution to the journal and provide important information for readers in the field. However, I have several concerns about the manuscript, described in 7 points below.

********************************************************************************

1. 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

**********************************************************************
2. 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.

**********************************************************************
3. 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.

**********************************************************************
4. 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.

5. 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?


6. 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 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?


7. 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.

*********************************************

Conclusion: Thank you for the opportunity to review this paper. I respect the work of the authors described here and hope that this feedback is useful.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I recommend additional statistical review

Quality of written English
Please indicate the quality of language in the manuscript:

Acceptable

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds
or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests. For full disclosure, Robert Califf is the director of my organization, the Duke Forge, and he is also an advisor and advisory board member of Verily Life Sciences, which is part of Alphabet, Inc. The authors of this manuscript are affiliated with Google Inc., which is also part of Alphabet, Inc. However, I personally do not have an affiliation to Alphabet, Inc., and I do not declare a conflict of interest.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal