Reviewer’s report

Title: A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records

Version: 0 Date: 19 Jul 2019

Reviewer: Kenney Ng

Reviewer's report:

The manuscript presents a rule-based natural language processing (NLP) method that automatically labels brain image (CT and MR) radiology reports with cerebrovascular and other neurological phenotypes. The method is developed, evaluated, and validated using two UK National Health Service data sets.

The manuscript addresses an important challenge on how to efficiently and effectively make use of clinical information contained in free text radiology reports. The NLP approach presented is appropriate; the use of multiple real world clinical data sets partitioned into development and validation subsets is fine; and the metrics used to evaluate performance are appropriate. The manuscript is organized and generally well written and easy to understand.

However, several important issues need to be addressed in order improve the manuscript to a level suitable for publication:

First, since the main focus of the paper is on the validated natural language processing algorithm, a much more detailed description of the rules (including some specific rule examples) used to do the final phenotype classification/labeling is needed in the "Natural Language Processing" section. The current one sentence description ("document-level labels on the patient's type of stroke discussed in the report are assigned based on the entities and relations present in the text") is not adequate. These labeling rules are the key components of the algorithm and from the current description, the reader can not get a clear idea of how this critical step is actually done. In addition, adding a high level system workflow diagram of the various NLP processing stages would be useful.

Second, an analysis of the errors and when and why the NLP rules failed (including a few informative examples) would be a very interesting and enlightening discussion to expand the "Discussion" section and to add clarity to the NLP algorithm.

Third, in the "Data Sets" section, there should be a ”standard table 1” that describes the key characteristics of the four data sets: ESS development and validation sets and NHS Tayside development and validation sets. It is important to show the similarities and differences between
the different data partitions and data sets. Also include in this table the prevalence of the 24 phenotypes in the data partitions. Related to this, additional clarity on how radiology reports were selected to be annotated and how the training and validation data sets were created are also needed (several specific detailed comments below touch on these issues).

More detailed comments at the (page, line) level follow:

Page 2, line 23: remove "was developed and tested" (repeated phrase).

Page 2, line 25: change "tested" to "evaluated".

Page 2, line 33: change "unselected" to "unseen".

Page 2, line 43: Include summary statements about the performance of the other 21 phenotypes that are not highlighted in the results subsection.

Page 2, line 51: Include a brief description of the application of the method on the large report data set (n=110K) to justify this "scale" conclusion.

Page 2, line 58: Maybe include "phenotyping" as keyword.

Page 3, line 31: Include references for the "inter/intra observer variation" statement.

Page 3, line 35: Provide some examples of "standard automated methods" and maybe a few references.

Page 4, line 37: what does "unselected" mean here?

Page 4, line 39: The number of patients is mentioned for ESS and the number of reports is mentioned for NHS Tayside. Should be consistent and report both number of patients and number of radiology reports from each of the two different data sources.

Page 4, line 49: Elaborate on how the development and validation data sets were partitioned. Random, stratified, matched?

Page 4, line 56: Were the 1692 reports all the brain imaging reports or only a subset? If a subset - how was it selected? Is this the "enriched" subset?

Page 5, line 12: Explicitly clarify whether one radiology report can contain multiple phenotypes. Include statistics on the min, max, and average number of phenotypes per report.

Page 5, line 14: remove "chose" to fix grammar.

Page 5, line 14: Describe how the enrichment selection was performed.
Page 5, line 23: Mention the size of the subset used to compute the kappa. Why not use all 1692 reports instead of just 366?

Page 5, line 35: Rephrase: move "at its core" to the end of the sentence.

Page 5, line 56: Provide some statistics on the size of these lexicons.

Page 6, line 8: Rephrase
from: "We developed the EdIER first in the ESS, and validated the system on separate, novel reports from the same dataset."

to: "We developed the EdIER first on the ESS development data set and validated the system on separate, novel reports from the ESS validation data set."

Page 6, line 12: What were the major changes to the rules, dictionaries, etc. that resulted in improved performance?

Page 6, line 35: Describe how the 700 reports were selected (i.e., randomly from the 1062 annotated notes in the NHS Tayside data set? or some other way?).

Page 7, line 16: How were these 366 reports selected from the larger sets?

Page 7, line 17: There is a missing right parenthesis ")

Page 7, line 45: Rephrase
from: "developed our model in 362 expert-read reports"

to: "developed our model on 362 expert-annotated reports"

Page 7, line 46: Rephrase
from: "tested the final EdiER model in 700 unselected expert-read NHS Tayside reports"

to: "tested the final EdiER model on 700 unseen expert-annotated NHS Tayside reports"

Page 8, line 16: Missing period (.) after "(table 4)".

Page 8, line 19: Are these prevalences in the 110K report data set expected or unexpected?

Page 9, line 17: Including the phenotype prevalence in a new "Table 1" that describes the data set partitions will shed light on this issue of different rates across the data sets.

Page 10, line 5: The first key message should not focus on image data types - but rather on the radiology notes that accompany the images.
Page 13, Figure 3: Include a flow branch for the reports that are not selected to be read and annotated by the experts. There seem to be many of these reports. Include a description on how the reports read and annotated by the experts are selected from the larger set.

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

No

**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.

Yes

**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 am able to assess the statistics

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

Needs some language corrections before being published

**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.

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