Reviewer's report

Title: Rare Disease Knowledge Enrichment through a Data-Driven Approach

Version: 1 Date: 08 Dec 2018

Reviewer: Daqing He

Reviewer's report:

I have read the authors' responses to the comments on the previous version. The revision has been extensive and mostly addressed the concerns raised by the reviewers on the previous version. As one of the reviewers on the previous version, I really appreciate the authors' efforts.

Now I have the following relatively minor comments for the authors to further improve the manuscript.

1. page 4 line 78. With the revision of the text to indicate that all clinical notes were used in the study, it is now clearer than the previous version on what data was used in the study. However, it would be much clearer if the authors can include the example types of EMR notes mentioned in the responses in the manuscript. That is, my suggestion is to add into the manuscript the text from the responses "including Consultant Notes (CON), … for each patient".

2. page 8 line 164. So the ground truth was collected by merging the two files mentioned in the manuscript. I am not an expert on medical knowledge, should there be some form of manual check to make sure that the ground truth is truly correct?

3. page 8 lines 172, 173. To me, it is more common to calculate precision, recall and F1 measure as a way to evaluate a classification system. precision = TP/(TP+FP), recall = TP/(TP+FN), F1 = 2xPrecisionxRecall/(Precision+Recall). Furthermore, with F1 measure, it is possible to calculate statistical significant differences among different models. Consequently, the question on Figure 3 is whether there was significant differences among the three graphs.

4. page 10, line 198. it states that "the density for the HPO-Orphanet+ graph was in between HPO-Orphanet graph and EMR graph", but table 3 shows that the density for the HPO-Orphanet+ graph, HPO-Orphanet graph and EMR graph is 0.005, 0.006 and 0.007, respectively. So HPO-Orphanet+ graph has the lowest density value. Did I miss anything?
5. page 11 table 4 and page 12 table 5. Both tables list 15 different candidates for each graph respectively. I am confused on the meaning of these candidates. Are they the ones generated from the graph as the ones that are most similar to a given rare disease such as Hodgkin lymphoma? Since there is ground truth on the correct association, should they be checked on whether the candidates are correct or not?

6. so figure 3 shows some performance score of the system, and figure 4 shows an interface of the system. Have the authors conducted any evaluation, even just initial ones, with domain experts to see if the results generated by the system make sense or useful?

Overall, the manuscript has been greatly improved. Really appreciate the authors' effort.

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.

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:

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’

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.