Reviewer's report

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

Version: 0 Date: 15 Oct 2018

Reviewer: Rui Zhang

Reviewer's report:

The authors applied a data-driven method to enrich the current knowledge resources on rare diseases. They first extracted and annotated the phenotypic and disease terms from clinical notes using a phenotype-disease annotation pipeline based on HPO and UMLS. Association rule mining was performed to mine the associations between diseases and phenotypes. The odds ratio was utilized to measure the significance of the association. Bipartite graph was generated using the selected disease and phenotypes. They further built a disease diagnostic decision support tool (D3N). They evaluated the final results in three ways: 1) evaluating the enrichment of phenotype-disease associations referring to the SemMedDB as gold standard; 2) compare three bipartite graphs 3) carried out two case studies to evaluate the performance on rare disease diagnostic decision support. The results show that their approach is able to enrich the existing knowledge resources on rare diseases. The paper is generally well-written. The data are presented well and convincing.

Major comments:

1. Using SemMedDB may have some limitations due to the low precision of semantic predications generated by SemRep. If authors do not have solutions or other alternative resources (such as other rare disease knowledge base) as gold standard, they should mention this as a limitation.

2. The reason why selecting Wilson's disease and Hodgkin lymphoma is not well described.

Minor comments:

1. In the first line of abstract, "from literature" should be changed to "from the literature"

2. In the 20th line of page 2: "percentile" -> "the percentile"

3. On page 7, line 17, "V is the number of vertices in graph" -> "the graph"
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.
Not relevant to this manuscript

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