Reviewer's report

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

Version: 0 Date: 17 Oct 2018

Reviewer: Daqing He

Reviewer's report:

This article talks about a study of enriching rare disease knowledge by mining electronic health records. Specifically, the authors focused on identifying phenotype-disease association using association run mining technique. They tested on Wilson's disease and Hodgkin lymphoma, and compared the performance of their model against a clinical diagnostic tool called the Phenomizer.

Main limitations

1. There is no detail review of the existing literature. The second paragraph in background section serves as a quick review of some related work. However, only a few work was mentioned, and none of them are directly related to enriching phenotype-disease association. In addition, even though there might not be many related work on phenotype-disease association, there are many existing work on gene-disease association, therefore authors should review these works in order to present the innovations of this article. In addition, the following related work on rare disease are missing in the related work


2. I am not an expert on phenotype-disease association, so my question might be too naive. Is it a strong association between phenotype and diseases, particularly some rare diseases? Why it is important to mine phenotype-disease association for rare diseases? Authors should present clearer motivations for taking this approach.

3. on page 3. this work essentially utilizes the co-occurrence information between a phenotype and a disease for identifying the association. This is a common method in data mining domain, but there usually has a window size of 20 or 50 words to limit the distance between two words when examining their co-occurrence relationship. In this paper, there is no window size, but the whole clinical note is used. I do not know how long a typical clinical note is, but if it is like a normal document, it would be a too large context for identifying co-occurrence relationship based on data mining studies. Therefore, the authors should make it clearer on this aspect, and probably discuss the motivations of using the whole note rather than a window size of clinical note for obtaining co-occurrence relationship.

4. page 6. DCG is a measure designed for modeling relevance scores at multiple grade level (i.e., more than two). If there is just binary relevance, it is actually common to use average precision (AP) to model the quality of the ranked list. AP has the advantage of having maximum value at 1, whereas DCG can increase indefinitely. In addition, current paper does not present the meaning of p in DCG clear enough. Maybe an example to illustrate the idea is better.

5. page 11. The Phenomizer is a generic clinical diagnostic tool, so it can be a baseline, but it should not be the only baseline for comparison. The authors should select from the related work one or several recently published rare disease diagnostic tools to use as baselines. Only through this, it makes sense to claim the innovation of the proposed method.

Overall, this paper have some major flaws that need to be fixed before it can be published.
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.

Unable to assess

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 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