**Reviewer’s report**

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

**Version:** 0  **Date:** 14 Oct 2018

**Reviewer:** Rafael Berlanga Llavori

**Reviewer's report:**

This paper presents a method for enriching existing knowledge resources (KRs) about rare disease with data extracted from electronic medical records (EMRs). The proposed method allows decision support systems to take into account common diseases that can be misdiagnosed for rare ones. Basically, the method consists in extracting phenotype-disease co-occurrences from EMRs, then filtering out irrelevant ones and keeping those with highest interestingness (e.g., support, lift, etc.) These co-occurrences along with rare disease knowledge resources like Orphanet are combined to define a bipartite disease-phenotype graph for diagnosis decision making. Experiments are conducted to demonstrate the quality and impact of the enriched KRs in misdiagnosis. Authors report improvements in the increment of exploratory power as well as in sensitivities/specificities in diagnosis tasks.

One of the most confusing points of the paper is the introduction of D3N. A D3N is mainly aimed at giving a ranking a diseases given a set of observed symptoms/phenotypes. It must be said that authors provide no reference to existing D3N approaches. Broadly speaking D3N aims at measuring the impact of each factor during the diagnosis decision making via a variety of methods like Bayesian estimation and/or machine learning methods (Jiang et al. 2017). In the proposed method, authors define a rather simple binary similarity measure (Jaccard) for comparing diseases in terms of their phenotypes. This similarity can be useful for evaluating potential misdiagnosis, but not as a D3N since it requires assigning/learning weights to the phenotypes and their attributes.

Another confusing point is the comparison to Phenomizer. Which is the purpose of this comparison? Since the evaluation is performed by using the EMR data as ground truth, it is obvious that the proposed method is going to work much better than Phenomizer. Indeed authors agree with this as in the conclusions they say that "the comparison actually indicate the graph of diagnosis support between ...". Apart from comparing to Phenomizer, authors must evaluate sensitivities/specificities for the three proposed configurations, namely: EMR data only, HPO-Orphanet data only and the combined graph (Table 3). This comparison is neutral to the underlying ranking method (D3N vs. Phenomizer), providing thus proper evidence of the benefits of enriching KRs with EMR data.
Other minor comments are the following ones:

- In the introduction (line 45), D3N needs some citation and a brief description.

- (Line 73) Association rules were not proposed in [18], this is a reference to a tool implemented in R (arules). Please cite the former work of R. Agrawal & R. Srikant (Fast algorithms for mining association rules, VLDB 1994).

- (Line 74) Please provide the citations to the original papers that presented these interestingness metrics.

- (Lines 103 - 104) From my opinion, the D3N is the phenotype-disease bipartite graph not the graph of diseases similarities.

- Figure 3, please indicate the point in the x-axis at which support is below the given threshold (5E-06). Probably a better method to retain good co-occurrences is to take the elbow in the curve as reference point.

- (Line 131) "bipartite graphs derived based on" - > "bipartite graphs based on"

- (Line 187) "The increment of the average degree for combined graph" This is not true according to Table 3 where EMR graph has a higher average degree than the combined one. Some error?

- In future work it would be interesting to study the extension of Phenomizer with EMR-derived knowledge, and compare if this enrichment improves its performance.

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.

No

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