Reviewer's report

Title: An open access medical knowledge base for community driven diagnostic decision support system development

Version: 0 Date: 02 Dec 2018

Reviewer: Steven Babin

Reviewer's report:

In general, this paper reflects research that is well-done and an absolutely needed first step. Too often, machine learning algorithms having unintended biases and unfortunate consequences. With machine learning results that clinician subject matter experts would judge as obviously wrong, I definitely agree with the authors' assessment of the importance of knowledge bases for complementing ML approaches in future diagnostic decision support systems.

Major Compulsory Revisions:

1) Utility and Discussion, line 155: The sentence says Subfigure (a) shows the query interface with the used symptoms for an ebola patient;" I don't know what "used symptoms" means here. Did you intend to say "usual" symptoms? If not, please explain what is a "used symptom."

2) Utility and Discussion, lines 166-176: This a very important part of your method and you will now see that I am not following it as currently written.

I think you are saying that m(D) is the rank, defined by the equation in line 173, that equals the sum of the size of the set of observed symptoms with non-zero weights and the product of the maximum weights of this set of symptoms So, the rank m(D) equals the number of symptoms in the observed set plus a product of maximum weights (a nonzero number less than or equal to one). I can see why you want to rank the subsets by size in the first term. The second term and the fact you are adding to the first term is difficult to follow as written. I understand that this product term is nonzero and cannot be greater than unity. Because the product term is a product of multiple weights, then this implies there is more than one maximum weight, so how does one determine which set of multiple weights are maxima; which nonzero weights are maxima under what conditions? How is adding their product to the set size in the first term an effective means of utilizing these weights? How is q determined? Are you ranking here with a preference for small rank or large?
Then, for q sufficiently large (how large is sufficient?), the greedy algorithm is trying to maximize the difference between the rank for the union of the set of all diseases with set of the observed diseases and the rank defined in the equation above. Maximizing the difference in rank means you are trying to find the smallest rank in the equation on line 173, correct? As I understand it, set-covering typically means you want to find the minimum overlap, so is this what your greedy algorithm is doing? Why are you doing this only for sufficiently large q and what do you mean by sufficient? What don't you always use the greedy algorithm, regardless of q?

As should now be obvious, I am not following the text in lines 166-173. Please provide more detailed explanation here, as this seems to be an essential part of your method. I have read your cited reference (Reggia et al) but am afraid I did not find answers to my questions.

Minor Essential Revisions:

1) Background, line 54: I think you meant to have the word "to" in between the word "similar" and "the" so it would read "similar to the"

2) Background, line 55: Don't need a comma between fast and non-conscious; having one makes it a bit confusing.

3) In Word, sometimes having a hyperlink from a citation in the text to the item in the References is a bad idea due to the flakiness of Word. Word loses track of it and gives you an "Error! Reference source not found." Unfortunately, it is a Microsoft bug so the best thing is not to hyperlink it until it is an online PDF. It the current version of the paper, this error means the reader has no idea which reference you are citing. This error occurs twice: on line 123 in Construction and Content and on line 228 under Performance Comparison.

4) Construction and Content, line 128: most non-physicians would not know what you meant by "differential" unless you call it "differential diagnosis"

5) Utility and Discussion, lines 140-141: "In its current state, the knowledge base provided robust basis for DDSS development and delivered comparable results established DDSS;" Don't
understand the last few words unless you meant to say "delivered comparable results to established DDSS"

6) Utility and Discussion, lines 163-164: "A diagnosis can be either a single medical disease, a medication, or can contain multiple disease explanations." Even though you explain it with an example, it would be clearer if you said medication side effect instead of just medication in that sentence.

7) Appendix D: There are some misspelled words in the tables. Are these just typos in the manuscript or would these misspelled words have any impact on the algorithm results? In Table 3, it should be hydatid instead of hytatid. In Table 4, adolescent instead of adolcescent, maculopapular instead of maculoppular, gynecomastia instead of dynecomastia, dizziness instead of dizziness, abdominal instead of absolinal, consolidation instead of onsolidation, and ulcers instead of ulvers. In Table 5, larvae instead of larve, rickettsial instead of ricketisal, mimicking instead of mimicing, striae instead of strie, and dyspnea instead of dyspneax.

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.