Reviewers report

Title: A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease

Version: 1 Date: 14 Feb 2017

Reviewer: Adam Dunn

Reviewers report:

Thank you for the opportunity to review this manuscript. The authors describe the application of what is broadly a literature discovery and EHR co-morbidity mining pipeline to autoimmune conditions associated with celiac disease. The research is interesting but I felt that the paper should be restructured and rewritten to more accurately reflect what was done. Currently the major flaw is that it describes the contribution as the data-driven workflow but the methods and results relate to what was discovered in a specific application, rather than an evaluation of the performance and value of the data-driven workflow. Several of my comments below relate directly to this problem. The quality of the writing was high.

1. Perhaps modifying the title to more clearly explain that this is a case study application of a process for supporting the mining of EHRs by restricting the set of conditions to look for in the records. I also immediately thought that the most interesting aspect of this research would be in looking at the "attention" in the literature versus the "incidence" discovered from the EHRs. To do this properly, however, I would expect to see a much broader range of coorbidities examined.

2. Abstract: The background is long and doesn't match the methods and results, which describe only *how* the pipeline was applied to generate results for a specific condition and the results. There is no description of whether the approach produces a valid result or a statistical test confirming that the comorbidities identified through this process are either (a) the same as would be produced using an alternative method but automated (answering "why use the literature"); or (b) more accurate than other automatic approaches to identifying comorbidities from EHRs.

3. It wasn't until about half way through the manuscript that I realised that the method was used to find a very restricted set of co-morbidities. I thought it would use a MeSH subset of *all* conditions and attempt to find those. From what I understand, it only selects from auto-immune conditions and does not explain or test how the method generalises.
4. Background: I think the title on Page 4 is the wrong place to have the title.

5. Background: Reference [1] does not appear to support the statement that "Today in all major hospitals Clinical Data Warehouses gather information"

6. Background: It might be worth trying for a more directed background, where the first paragraph introduces the problem (identifying comorbidities from EHRs), the second paragraph evaluates the existing approaches to doing this, an optional third paragraph that reviews literature discovery and any examples that might have been used in concert with EHRs, a paragraph on comorbidities in celiac disease, the next paragraph explains the rationale behind the approach that has been chosen, and the final paragraph states the main aims and objectives of the paper (e.g. a case study for one condition and one class of comorbidities).

7. Background: The paragraph explaining what is known about comorbidities in celiac disease is useful and an important paragraph to include. It might be worth arguing that besides the apparent variability measured in different places, to the best of your knowledge there has not been a clear synthesis of these studies published. It would be fine to say that this study *adds* to this literature by identifying comorbidities in a set of 741 new patients, and using a novel approach to the identification. To reiterate, there is nothing wrong with this as an aim.

8. Methods: In the study population section, aim to be more precise so it is clear what the study data encapsulate. The total number of patients and documents could be written here.

9. Methods: In the manuscript, it is very hard to understand how the co-occurrence in the literature is used to support the phenotyping, or what the actual purpose of using the literature is. I thought perhaps the aim of the study was to examine how these are different, to support hypothesis generation, and to design more pragmatic trials. After reading the manuscript, it now seems as though there may have been no reason to use the literature at all, and all that was needed was to select 15 auto-immune diseases and leverage existing terminologies to support phenotyping.
10. Pointing the reader to FASTVISU and skipping over the process for identifying comorbidities confused me in the first instance - I didn't realise that this was a manual process until I read through it a second time. Perhaps an improved flow diagram might help. Then I wondered - why do you need the literature discovery at all? Why not just highlight anything that matches any of the synonyms from the terminologies in the software and let people confirm it manually?

11. Results: I don't think you need to capitalise words in the table title - sentence capitalisation is fine.

12. Results: When reporting Cohen's kappa, also report the raw agreement (at least). This is a key performance indicator in the process, because the FASTVISU software is enabling faster and more consistent labelling of comorbidities by humans. How long would it have taken them without FASTVISU and would the agreement have been lower? What if FASTVISU was set up with different (fewer) terms that were highlighted?

13. Results: Figure 3 is interesting, I believe it shows the location in the EHR where the comorbidity was identified. The implications of these results are also described in the discussion, and I thought this was a useful and important contribution.

14. Discussion: I think the literature discovery results are first described in the Discussion section and appear to be out of place.

15. Discussion: I think the comparison with other prevalence estimates should be included in the results (just the comparison not the explanation). I think this comparison is an important evaluation of the performance of the method in the context of the specific dataset used, and the discussion section should only explain/speculate about why the results are different.

Overall, I think the manuscript would benefit from a clearer structure, and clearer explanations of (a) the places in the EHR where comorbidities are extracted; (b) comparing the use of different terminologies and synonyms to undertake the phenotyping; (c) comparison of the prevalence in the population versus the "attention" in the literature. At the moment it only really describes the steps without explicitly measuring the performance of the method or comparing the approach to alternatives.
Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

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.

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.

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