Reviewer's report

Title: A Clustering Approach for Detecting Implausible Observation Values in Electronic Health Records Data

Version: 0 Date: 06 Mar 2019

Reviewer: Steven Johnson

Reviewer's report:

The paper describes a clustering-based method for identifying outliers and implausible values in EHR laboratory data. This is an important area since we know that EHR data contains errors, but they are hard to find manually. As the authors point out, creating rule-based procedures for finding implausible values is time consuming, error prone and doesn't scale. The authors propose a hypothesis that for large data sets, implausible values should be sparse. The paper is organized and well written.

The primary issue that I see with this paper is that there is a difference between an outlier and an implausible value. An implausible value is data that doesn't make sense given what we know of how the world is supposed to work. For example, if the EHR contains a patient record for a 6-year old child that lists them as married, that is an implausible value. However, if the EHR has a patient record for a 120 year old man, that is an outlier. It might not be likely, but it is not an impossible value. We probably need to look at other information to decide.

So the question is, can we distinguish implausible values from outliers using the authors proposed clustering technique? Maybe, but the paper doesn't explain that well enough. For the lab values that are identified as implausible in the paper, can the authors explain more about how they determined their cutoffs and whether the value is theoretically possible, and therefore an outlier? For example, for LOINC 10839-9, there were 39,000 implausible values. Why? What was the underlying cause? It seems like you have the data to explain that.

Page 6, line 13. The transition to discussing clustering is abrupt. Doesn't clustering also use density or distance metrics? You need to explain a little better how the proposed clustering approach is better than the methods described earlier.

Page 10. What does db(x) refer to?
Page 10-11. The authors do a nice job of explaining the tradeoffs of increasing sensitivity and specificity.

Page 13, line 3. It's not clear at all from Table 2 that "...the clustering approach produced overwhelmingly better specificity...". In fact, it looks pretty good. Are the differences meaningful?

Detecting implausible data is very important. Overall, I believe the authors' approach leads to better outlier detection. But the authors need to explain a little better how their approach is better than CAD. On Page 15, there is the start of an explanation for why clustering is better when they say "...[implausible] observations can be found anywhere across the distribution of data..." You should expand on that line of thinking and show how that is better than just having cutoffs at extreme low or high values. Were there any LOINC codes that showed that behavior?

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/](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