Reviewer’s report

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

Version: 0 Date: 06 Feb 2019

Reviewer: Carrie Daymont

Reviewer's report:

This manuscript describes the creation and evaluation of a novel method to detect implausible values in laboratory and vital sign data in electronic health records. This is an important topic, and methods like these could be very useful to people in both research and clinical contexts. I do have concerns about some aspects of the manuscript that I believe need to be addressed. My primary concerns are:

1) Much of the assessment of the method focuses on specificity, but the "silver standards" used to define plausibility are very wide. The upper limit for a systolic blood pressure is 560, so a method that identified an SBP of 490 as implausible would then be considered to have identified a false positive, which does not ring true. Even more fundamentally, the authors say that the silver standards were developed in part, based on "insight from data distributions." If these were distributions of these measurements in the RPDR (which seems likely since an alternate source of data was not described) then the number/values of outliers in the evaluated data impacted the definition of outliers, and the evaluation of sensitivity/specificity is not really meaningful. Additionally, I could not identify values for k and alpha that are identified by the authors as appropriate for use across this spectrum of laboratory and vital sign values. If k and/or alpha have to be determined individually for each type of measurement, I am not sure how that would be an improvement on identifying upper and lower limits for each type of measurement.

2) Related to #1, it was not clear to me what specific uses the authors envision for this method. For example, many of the potential uses for an outlier-detection in the methods section, such as identifying exceptionally well-performing clinicians or detecting unusual patient management actions, would require identification of outlying but still physiologically plausible values, but this method seemed to be designed/tested to identify completely implausible values.

3) Overall the manuscript would benefit from improved organization.
Additional more detailed comments are below. There were two sets of line numbers on the manuscript I received. The line numbers below refer to the larger numbers closer to the text.

Introduction

Page 4 Line 9 Plausible means possibly true, it is not synonymous with truthful.

Page 4 Line 14 I'm not sure that manual is the most clear description of these procedures; I would suggest describing that you mean that separate limits would need to be determined for each type of measurement here, rather than in the next paragraph.

Page 4 Line 20 This paragraph could be more clear regarding the reasons you chose to evaluate both vital signs and laboratory results, and how they are similar/different. "We focus on laboratory results" makes it sound as if you do this to the exclusion of vital signs, which is not accurate.

Methods

Page 5 Line 15: This sentence implies that these are the only types of outlier detection methods, but in later paragraphs you discuss other methods.

Page 6 Line 12: It would be useful to have a reference for this statement.

Overall the organization of the methods section makes it difficult to follow. It would be helpful to start with a less detailed overall approach, then detail the data, and then detail the primary clustering algorithm and provide some information on the alternate outlier detection methods.

Page 7 Line 18: Both this manuscript and the reference describing the RPDR contains very little information regarding which patients and which data are included in the RPDR, and what processing/cleaning those data undergo prior to being made available to researchers. The information provided in section 2.2 does not include this essential information, and also does not include information on the source of vital sign data.
Page 9 Line 20: The software used for the analyses should be included, and it should be made clear that kluster is a package in R.

Page 10 Lines 7-8: It is not clear what is meant by experimented. If the algorithm were run on all available data for all evaluated measurements, the choice of alpha may be specific to these data, with no evidence that it will generalize to other data/measurements.

Results

The first paragraph of the results section describes methods.

Page 12 Line 12: It would be helpful to have some description of what was different about troponin and cholesterol compared to other tests, and to include figures depicting the distribution of a few chosen tests, including these 2. Also, you say that you achieved a sensitivity of >0.85 in all but two tests, but the max sensitivity I could find for Troponin T was 0.6577.

Page 12 Line 16: It was not until I looked at the table that it was clear you meant that the specificity for troponin was 100% for all algorithms.

Page 12 Line 21: Which alpha and which k were used for the final clustering approach? In Table 3 it is not clear to me that the same alpha was used for all analyses.

Page 13 Line 3: The size of many of the differences in specificity is extremely small; if the authors believe these are meaningful differences it would be important to explain why.

Discussion

Page 16 Line 3 The development of the kluster procedure was described in a different publication
Figures

In Figure 1, it seems the labels upper/lower implausible range should be upper/lower plausible range.

In Figure 2, it appears to show analysis of two-dimensional data, but it seems that the laboratory and vital sign data are unidimensional.

Figures 4-5 It would be helpful to explain to readers why you evaluated the square root of 1-sensitivity/specificity.

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.

I recommend additional statistical review

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