Author’s response to reviews

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

Authors:

Hossein Estiri (hestiri@hms.harvard.edu)

Jeffrey Klann (jeff.klann@mgh.harvard.edu)

Shaun Murphy (snmurphy@partner.or)

Version: 1 Date: 16 Apr 2019

Author's response to reviews:

Please find our point-to-point responses to the reviewers’ comments:

Carrie Daymont, MD MSCE (Reviewer 1):

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.
AUTH: This is a valid concern regarding the evaluation of the method. Delineating between what is plausible or implausible tends to be very difficult, with no gold standards. Where the exact threshold for implausibly high systolic BP is (whether it is 560, 490, 514, or 1200) remains unknown. However, a set of linear thresholds are needed to evaluate the algorithms. The list of silver standards is our best effort to address this need. We agree that a 490 value for systolic BP flagged as false positive may not be 100% true, but the algorithm would likely not pick that value if the distance and frequency of its neighboring data points do not warrant assignment to the implausible cluster.

AUTH: Regarding the insight from data distributions, we did not explicitly set the threshold based on the distribution of lab values in RPDR. We first did a literature search, then used expert judgement to adjust/validate the thresholds. RPDR distributions were used to validate the thresholds. We have edited the text to emphasize this point. As we report later in the results section (3.2), 9 out of 50 labs did not have any implausible records in RPDR – Systolic BP is one of the 9 labs. Moreover, the algorithm is completely unsupervised, so the silver standard thresholds does not impact the way the algorithm works.

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.

AUTH: The value for K (the number of clusters) is automatically identified by the kluster algorithm. The only parameter to be set is alpha, and it has to be set once for all types of measurements. We show that as alpha becomes smaller, the specificity increases. Due to the large amount of data, we focused on reducing the false negatives. So, based on the size of the data in a repository, the user can alter the value for alpha, which would loosen or tighten the level of comfort at the institution for having false positives. We have addressed this in the implementation considerations section (4.2).
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.

AUTH: In the revised section on the limitations and directions for future research, we have clarified the primary goal for this study and what would be needed to apply the proposed methodology for other use cases: “Finally, the primary use case for this method is intended for an algorithmic screening of EHR laboratory observations for potential implausible values that would not be suggested for secondary use. The methodology can be applied to other use cases, such as identifying exceptionally well-performing clinicians or detecting unusual patient management actions, would require identification of outlying but still physiologically plausible values. However, because the evaluation criteria may be different for other use cases (we focused on false positives for the detection of implausible observations in large scale EHR repositories), further work may be needed to adjust the α in order to optimize performance.”

3) Overall the manuscript would benefit from improved organization.

AUTH: We have reorganized the manuscript significantly. We have moved background on clustering into the Background section, tightened and clarified the methods, and moved some information from the results to the discussion. We have also removed extraneous detail and clarified the language where possible.
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.

AUTH: We have removed the word “truthful” (although this was a direct reference to the definition of plausibility in reference 3).

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.

AUTH: We have added the recommended description to lines 14-15.

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.

AUTH: We have revised the sentence on lines 20-21 to include vital signs.

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.

AUTH: We consolidated the background in outlier detection to the Background section. Here, we start with broad categories (parametric vs. model-free and distance- and density- proximity methods) and focus in on the methods relevant to this study.
Page 6 Line 12: It would be useful to have a reference for this statement.

AUTH: The statement is based on the findings of this study. We have updated the text to clarify this.

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.

AUTH: Thank you for the suggestion. We moved the algorithmic background to the Background and focus on our particular method in the Methods. Here, we now start with a conceptual-level approach, then a data description, then implementation details.

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.

AUTH: More information about the data is provided in the revision under the data section.

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.

AUTH: We have clarified that kluster is an R package. We have also included a note to page 10 (Implementation) that the algorithm is developed in R and tested using high performance computing cluster provided by the Partners HealthCare’s Enterprise Research Infrastructure & Services.

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.

AUTH: We have clarified the text. The algorithm was run on all available data for all evaluated measurements. The general finding about alpha was that it has a negative relationship with specificity. This was the case for all 50 observation types. We believe there is enough variation in data distribution that supports generalizability of what we found about values for alpha.
Moreover, we are not insisting on a singular value for alpha, as we mention in the Implementation Considerations that “Given the size of data and the emphasis on sensitivity versus specificity, the choice of $\alpha$ (the ratio for flagging a cluster as implausible) can vary. In large clinical data repositories, an $\alpha$ between $1/4,000$ and $1/6,000$ would provide good balance between true positives and false negatives. A smaller $\alpha$ is also computationally more expensive. When the size of the dataset is small, a small $\alpha$ will be more appropriate. Nevertheless, the value for $\alpha$ can be adjusted over time to address the institutional needs and comfort level.”

Results

The first paragraph of the results section describes methods.

AUTH: Moved the paragraph to the methods section.

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.

The 2 tests were Troponin (it was Troponin I) and cholesterol. We describe what was unusual in the two labs.

AUTH: We have added more explanations on why the 2 labs behaved differently. A series of example plots with different alpha values were provided in the appendix. We have updated the Appendix plots and also provided plots for the 2 exception labs in the manuscript – the new figure 4.

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.

AUTH: The description on the 2 exceptions has been updated.
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.

**AUTH:** k is observation specific and is automatically determined by the kluster algorithm in an unsupervised manner. Table 2 compares the best specificity and sensitivity for each observation type between the clustering and conventional approaches. The purpose in table 2 is not to identify a best alpha, rather to compare the two approaches. Based on tables 1.a and 1.b, best specificity almost always was obtained by alpha = 1/10,000. Most alpha assignments resulted in the best sensitivity for many observation types. We describe details in section 3.3 of the results.

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.

**AUTH:** Differences are small within approaches, but between the two approach, clustering results are much improved. Additionally, when we deal with very large datasets, as is the case in this research, a 0.001 difference can mean hundreds or thousands of records. A note has been added to section 3.3 to clarify.

Discussion

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

**AUTH:** Citation has been added in the discussion section as well.
Figures

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

AUTH: All figures have been updated accordingly.

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

AUTH: Correct. We used 2-dimensional graphics to simplify the description of the pipeline and processes. Certainly, the same procedure can be implemented on unidimensional data.

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

AUTH: It was for easier understanding of the visualization as the square root transformation helps to exaggerate smaller values. A note has been added to the manuscript.

Steven Johnson, PhD (Reviewer 2):

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.

AUTH: This is a reasonable concern. For outlier detection, there are conventional methods that are much less computationally expensive than the clustering method. A reason for comparing the clustering approach with the conventional anomaly detection (CAD) methods was to see if clustering can do a better job of detecting observations that are specifically implausible.

The sparsity assumption that we use in the clustering (the alpha) adds to the distance measure for identifying implausible observations. As such, an implausible observation is not only at distance with the majority of the observations from a value stand point, but also does not have many similar neighboring observations. Our results showed that this assumption has helped in detecting things that are more like implausible observations, besides outlierness. We have clarified this point in the discussion section.

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.

AUTH: We used expert knowledge to validate the cutoffs, but they still may be theoretically possible. For the LOINC 10839-9, we have provided a figure that shows there are still many observations above the upper plausible limit that if we increase the limit by 2-3 times, there will still be many observations that can be implausible. We set 20 as the upper limit for this lab. It turns out that there are many values even higher than 50. We tracked this down to a single cTI test at one hospital run between 2001-2008. The RPDR team suspects an import issue, likely where the units are not consistent with the other labs assigned this LOINC code. This would result in this small subset of the results appearing to have an implausible value because the normal range is actually different. We are working with the RPDR team to verify if this is the case. This has been added to the discussion.
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.

AUTH: We have edited the text, reorganized this, and added more explanation to make the transition smoother.

Page 10. What does db(x) refer to?

AUTH: Extracted data on observation x is stored as db(x). Text has been slightly edited to clarify.

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?

AUTH: Figures 5 and 6 show the difference much better. The sentence meant to refer to figures. Text has been edited to clarify. The difference is meaningful, especially as we are dealing with large datasets a very small improvement can mean hundreds of records being flagged or not flagged as implausible.

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?

AUTH: We have expanded a bit on the possibility of detecting other types of implausible records. However, having hard-coded cutoffs at extreme low or high values is still the most precise way to detect implausible records. The problem is, as you mentioned, that approach does not scale and there are not gold standard cutoffs for all types of observations.