Author’s response to reviews

Title: Interpreting Patient-Specific Risk Prediction Using Contextual Decomposition of BiLSTMs: Application to Children with Asthma

Authors:

Rawan AlSaad (ralsaad@sidra.org; rawan.alsaad@hotmail.com)
Qutaibah Malluhi (qmalluhi@qu.edu.qa)
Ibrahim Janahi (ijanahi@sidra.org)
Sabri Boughorbel (sboughorbel@sidra.org)

Version: 1 Date: 11 Jul 2019

Author’s response to reviews:

** I included a Response to Reviewers Letter in the revised submission. Please check my responses there. A copy of this letter is below:

Dear Dr. Brzan,

Thank you for giving us the opportunity to submit a revised draft of our manuscript. We appreciate the time and effort that you and the reviewers have dedicated to providing your valuable feedback on our manuscript. We have carefully reviewed the comments and we have been able to incorporate changes to reflect most of the suggestions provided by the reviewers.

The main points raised by the reviewers requiring further explanations mainly revolve around two issues:

1. Adding demographic and statistics description of the dataset

Table 1 was added to the manuscript describing the statistics and demographics of the dataset. Table 1 can be found on page 6.

2. Accompanying the AUC with a measure of variability.

Table 2 was extended to include 95% CI in addition to the AUC.
Table 2 can be found on page 6.

Here is a point-by-point response to the reviewers’ comments and concerns.

Response to Reviewer 1:

1. The title is not specific and too broad. The study essentially deals with asthma prediction. It would be good to tune down the title.

Thank you for this suggestion. While we agree with you that asthma prediction was used to demonstrate the proposed approach, we believe that the title reflects well the paper focus and contribution on: a) level of interpretability: patient-specific (compared to population-level interpretability), b) machine learning technique: BiLSTMs, and c) interpretability technique: contextual decomposition. Tuning down the title to asthma prediction might misinform the reader that this technique is specific to asthma prediction, while it’s actually developed to work with several clinical outcomes prediction using EHR data.


T is an integer representing the total number of visits for each patient.

We have accordingly modified the text to clarify this point. This modification can be found in: section: Methods, subsection: Long Short Term Memory Networks, page 3, line 135.

2.2. The number of visits for patients is different. How to deal with the dimension difference in the model?

Deep learning libraries assume a vectorized representation of the data for time-series prediction problems. In our case, since the number of visits for each patient is different, we transformed the data such that all patients will have the same sequence length. This is done by padding the sequence of each patient with zeros so that all patients will have the same sequence length, equal to the length of the longest patient sequence. This vectorization allows the implementation to efficiently perform the matrix operations in batch for the deep learning models. This is a standard approach when handling sequential data with different sizes.
2.3. Is time stamp considered or only sequence considered?

Only sequence is considered. Time stamp is not considered.

We have revised the text accordingly to clarify this point. This modification can be found in: section: Methods, subsection: Dataset and Cohort Construction, page 5, line 274.

3. Dataset.

3.1. What is the sensitivity and specificity of using ICD-9 code to define case and control patients?

You have raised an important point here. There is currently no consensus on approaches to defining asthma or assessing asthma outcomes using EHR-derived data (Nissen et. al., 2017). As an estimate, Nissen et. al. concluded that identifying asthma cases in electronic health records is possible with high sensitivity, specificity or PPV (>80%), by combining multiple data sources, or by focusing on specific test measures.

Our study is based on retrospective EHR data, and defining case and control patients using ICD-9 codes is the most common approach in similar EHR-based studies. It would have been interesting to explore this aspect. However, in the case of our study, it seems slightly out of scope because estimating the sensitivity and specificity of this approach on our dataset will require having at least two pulmonologists independently assessing the EHR data of all the patients in our dataset (11,071) against the ICD-9 code inclusion criteria. After that, the sensitivity and specificity values can be estimated based on their assessment of TP, FP, TN, FN patients. It would be interesting to explore this aspect in the future with the help of pediatric pulmonology experts.

3.2. What is the demographics and statistics (e.g., average number of diagnosis) for the patients.

Table 1 was added to the manuscript describing the statistics and demographics of the dataset. Table 1 can be found on page 6.
4. What is the proportion of case and control in the training, validation, and test sets?

The same proportion of cases (56%) and controls (44%) was maintained among the training, validation, and test sets.

We have modified the text accordingly to clarify this point. This modification can be found in: section: Methods, subsection: dataset and cohort construction, page : 6, line: 305.

5. Page 6. What d_art is chosen? How to generate p_art?

d_art was chosen to be a synthetic diagnosis code which does not exist in the ICD-9 codes list.

p_art includes one synthetic diagnosis code, which is d_art.

We have revised the text accordingly to explain this point. This modification can be found in: section: Results , subsection: Validation of Contextual Decomposition for BiLSTMs, page: 7, line: 357.

6. Page 7. How many cases and controls in the subset of 5k patients?

The subset of 5k patients are all cases (children who developed asthma at school-age). This is mainly because in this experiment we are interested in assessing the matching between the contextual decomposition scores and the logistic regression coefficients. The coefficients reflect the impact of each variable on the odds ratio of the observed event of interest, and the event of interest here is children developing asthma at school-age (cases).

We have revised the text accordingly to emphasize this point. This modification can be found in: section: Results , subsection: Validation of Contextual Decomposition for BiLSTMs, page: 8, line: 468.

Response to Reviewer 2:

1. The approach described uses a sliding window approach to find the most predictive subset of visits, as the authors noted this is a greedy algorithm and usually the number of visits is low (it was defined as less then 10), but it would make to sense to suggest/propose an alternative approach in cases where the number of visits is not low, eg. a longer time period and patients with a chronic condition (toward the end of the manuscript, they describe examples with 19 and 14 visits).
• We totally agree with you that a more efficient approach would be required for a larger number of visits over a longer time period, and this has been highlighted in the submitted manuscript. However, we believe that addressing this issue as a future extension of the paper would be more appropriate, because it requires more time and effort to optimize the design of the search algorithm and more experiments to validate it.

• Note: in the paper, the number of visits was actually defined to be in tens (<100) not only 10 visits. It was mentioned that “since the total number of visits doesn't exceed tens usually, going through all possible combinations of consecutive visits is still computationally feasible”.

2. The authors choose a scenario where the proportion of cases is high (~55%), that might not be a general scenario, which might be a less balanced scenario for prediction and it should be noted.

• Yes, we agree with you that the proportion of cases is relatively high in our cohort, and this might not be the case for other cohorts or different diseases. Therefore, we have revised the text accordingly to emphasize this point. This modification can be found in: section: Methods, subsection: dataset and cohort construction, page: 5, line: 260.

• Note: please note that we have used the AUC metric for evaluating the performance of our models, which is less sensitive to data imbalance compared to other metrics (e.g. accuracy).

3. Results were presented as average AUC values values for 10 iterations and in addition to the mean one would expect that there were some additional information provided about the AUC, maybe confidence intervals, so the reader might get a sense of the variability of the AUC.

We agree with this suggestion, and we have modified the results table accordingly to include 95% CI. This modification can be found in: section: Results, Table: 2, page: 6

We hope that these responses have sufficiently addressed the reviewers’ comments and look forward to hearing from you in due time regarding our submission. We remain available to respond to any further questions and comments you may have.

Sincerely,

Rawan AlSaad

Machine Learning Group

Sidra Medicine