Author’s response to reviews

Title: Automated Detection of Altered Mental Status in Emergency Department Clinical Notes: A Deep Learning Approach

Authors:
Jihad Obeid (jobeid@musc.edu)
Erin Weeda (weeda@musc.edu)
Andrew Matuskowitz (matuskow@musc.edu)
Kevin Gagnon (kevinmgagnon@gmail.com)
Tami Crawford (crawfot@musc.edu)
Christine Carr (carrc@musc.edu)
Lewis Frey (frey@musc.edu)

Version: 1 Date: 11 Jun 2019

Author’s response to reviews:

Manuscript MIDM-D-19-00084
Title: Automated Detection of Altered Mental Status in Emergency Department Clinical Notes: A Deep Learning Approach Jihad Obeid
Authors: Jihad Obeid; Erin R. Weeda, PharmD; Andrew J. Matuskowitz, MD; Kevin Gagnon; Tami Crawford, PhD; Christine M. Carr, MD; Lewis J. Frey, PhD.

Editor,
BMC Medical Informatics and Decision Making

Dear BMC Medical Informatics and Decision Making Editor,

Thank you for your response on May 31, 2019 to our submission entitled “Automated Detection of Altered Mental Status in Emergency Department Clinical Notes: A Deep Learning Approach”. We would like to thank the reviewers for their insightful comments. We have addressed their concerns and believe the revised manuscript is much improved. The following provides point-by-point response to the comments along with descriptions of our modifications in the revised manuscript.

Reviewer 1
Introduction:
1) Please clarify that which types of diseases are mainly focused in this manuscript. Is it only pulmonary embolism or other diseases as well?

Response:
This is a good question. And the answer is a nuanced “no”. The inclusion criteria for the study group were that patients must have had ED visits with AMS ICD-9 codes 799.5x and/or ICD-10 codes R41.x. In order to ensure that the model is exposed to patients with thromboembolic conditions, both patient groups were enriched with patients with ICD codes for venous thromboembolism (ICD-9 453.x, ICD-10 I82.x) and/or pulmonary embolism (ICD-9 415.1x, ICD-10 I26.x). This was accomplished by including all patients that match these conditions within 60 days of the ED visit in the AMS group, which made up about 5% of that population as well as a balanced proportion in the non-AMS control group.

We added this clarification to the methods section.

2) Line 71: What do you mean by pipeline? Is it ML?

Response:
In this particular context the word “pipeline” is referring to NLP pipelines, i.e. a series of software packages typically used in NLP, examples include tokenization, part-of-speech tagging, named entity recognition, negation and mapping to ontologies. Several references were cited in the manuscript regarding such pipelines [1–3]. We did not include a full description herein, since it is out of scope for this paper.

3) Line 75: Please explain name entity recognition.

Response:
Named-entity recognition is a subtask of NLP that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, medications, labs, and other medical codes. They are typically machine learning based and use the conditional random fields (CRF) algorithm. These are also explained in detail in the cited references [2,3]. A full description is out of scope for this paper.

Methods:
4) Line 104: Give ref. or small details on Epic EHR System.

Response:
A reference for Epic was added [4].

5) Line 122: Explain briefly about REDCap.

Response:
REDCap (Research Electronic Data Capture) is an online data entry system widely used at academic institutions, which allows users to create data entry forms and manage the data in a secure environment. A reference to REDCap was also provided [5]. The text was updated.

6) Lines 136-137: Please explain briefly.

Response:
In order to calculate the inter-rater reliability (or inter-rater agreement), which is the degree of
agreement among clinicians who reviewed the records, we randomly picked a set of 100 notes out of the total notes reviewed by clinicians. We picked only 100 out of 1130, because this is a labor intensive task and 100 was enough to allow us to assess consensus between reviewers.

7) Table 1 is confusing. Please redraw or clarify.

Response:
This table shows the breakdown of the labeled notes and the patients that these notes belong to. As expected, some patients had more than one note. We had 1,130 labeled notes for 858 patients. The table shows the number of notes and patients within each of the AMS group as well as the non-AMS group. Clarification was added.

8) The authors should provide an account on how the data were arranged in matrix form. What was the dimensions of the matrix?

Response:
That is correct.
For the BOW models, we generated a tf-idf matrix. This was 904 (the size of the training/cross-validation data) by 4,765 sparse matrix for the training data, i.e. a vocabulary size of 4,765 after lower casing, removal of punctuation and stop-words, and word stemming.
For the WE models, the input layer had a dimension size of 717, which is the size of the longest sequence of tokens +1.
We added this information in the methods under BOW-based classifiers and Deep learning model respectively.

9) Please provide explanations on features in your dataset.

Response:
In the BOW word frequencies were used as features and were normalized using term-frequency, inverse document frequency (tf-idf).
In order to construct the features for the deep learning models, the token sequences were padded with zeros (using pre-padding) to match the length of the longest document. These were then mapped to word embeddings as described.
This was clarified in the methods section.

Reviewer 2
I thoroughly enjoyed reading this paper and believe that it can make a contribution to the literature on automatic decision making in clinical settings. My criticisms are somewhat brief, but can be encapsulated in two main points.

My first criticism is treating the clinical labeled data as ground truth. This speaks to a larger problem in not spending more time in the beginning of the paper outlining the issues with the ICD codes. I didn't catch this until the bottom of page 15 and top of 16. I think the author's could improve the rationale for the paper by outlining this as one of the main motivations, particularly in that you find high reliability across raters, but low agreement with ICD codes.

Response:
We agree with the review and have applied this recommendation. We therefore moved the rationale for the motivation of not using ICD codes into the introduction. We also restructured the discussion to accommodate this change.

Comment:
The manuscript is written in a way that it seems as though the initial goal was to demonstrate the benefits of deep learning in this context. Although deep learning, and the additional word embedding, improve upon the bag of word specifications, it isn't by much with respect to random forests, notably with overlapping AUC CI's. One of my takes is that the simpler models are pretty sufficient, particularly that we get interpretation out of it. The variable importance from random forests leads me to believe that a set of single words were used by both the classifiers and probably the clinicians to arrive at their diagnosis conclusion.
- It is my preference that most of the deep learning discussion is taken out from the paper, and more space is given to the clinical implications. My viewpoint would be different if the CNNs did much better than the simpler models.

Response:
Regarding BOW vs WE, we believe there is still value in exploring various deep learning architectures using the BOW models as a baseline, which is what we set out to do in this manuscript; however, as pointed out by the reviewer, it was difficult to tease out the advantage of the CNN architecture over random forest, due to the fact that many of the models are almost at ceiling performance, since this is a very focused classification task on one type of clinical notes looking for a cluster of symptoms relating to AMS. In fact this bodes well for automating tasks such as chart reviews, which are often perceived as expensive and time consuming when performed by a human. On the other hand, if this task is performed by a trained deep learning model, it takes a fraction of a second to classify a document, with an accuracy high enough to be actionable in assessing the risk level in a patient with pulmonary embolism.
We agree with the reviewer that more discussion on significance needs to be stated in the manuscript, and have added the latter points regarding significance in the discussion.
Moreover, the reviewer is correct about the fact that BOW models tend to be more interpretable, in fact it is useful to use both types of models when addressing a given problem leveraging advantages of deep learning models as well as advantages of explanatory models.

In conclusion, we would like to thank the reviewers for their diligent critiques and thoughtful comments. We have addressed the reviewers’ comments and believe the revised version of the manuscript is significantly improved. We are excited to share our work on this important topic with the informatics community.

Thank you for your consideration.

Sincerely,

Jihad S Obeid, MD
Associate Professor
Division of Biomedical Informatics, Health Systems and Policy
Department of Public Health Science
Medical University of South Carolina
References


