Author’s response to reviews

Title: Characterization of clinical patterns of dengue patients using an unsupervised machine learning approach

Authors:

Gleicy Hair (ghairbioengineer@gmail.com)

Flavio Nobre (flavio@peb.ufrj.br)

Patricia Brasil (patricia.brasil@ini.fiocruz.br)

Version: 1 Date: 30 May 2019

Author’s response to reviews:

Technical Comments:

1. Please include the email address of all authors in title page.

Thank you. We added this information in the title page as requested.

2. Kindly rename Summary to Abstract.

Thank you. We renamed it as requested.

3. For all research involving human subjects, informed consent to participate in the study should be obtained from participants (or their parent or legal guardian in the case of children under 16) and a statement to this effect should appear in the manuscript. If no consent to participate was obtained, please state this under the heading "Ethics approval and consent to participate", and explain why.

The informed consent was waived by Research Ethical Committee of Fundação Oswaldo Cruz since this study was based on data that is routinely collected at Instituto Nacional de Infectologia Evandro Chagas (INI)/FIOCRUZ. The data were anonymized and de-identified prior to analysis. We added these statements at the Ethics approval and consent to participate section in the end of the Manuscript (lines 452 and 458).
4. The initials of the authors in the Authors' contributions section do not match the names of the authors of the paper. Can you please clarify why and correct this?

Thank you. They were accidently shuffled. We reordered it.

5. Figure titles (max 15 words) and legends (max 300 words) should be provided in the main manuscript.

Thank you for pointing this out, we have now included the figure title and legends in the manuscript.

Reviewer 1, Major Comment

6. The authors analyzed the clinical profiles of 523 confirmed dengue cases using self-organizing maps (SOM) and random forest algorithms to identify groups of patients with similar patterns. They identified four natural clusters and found that age appeared as the key features response for splitting the data into four clusters. As well, other variables such as abdominal pain or tenderness, clinical fluid accumulation, mucosal bleeding, lethargy, restlessness, liver enlargement and increase hematocrit were identified to contribute to the severity of dengue patients.

There are some comments:

1) Introduction: The authors described "Classic statistical methods used to evaluate warning signs and determine risk criteria for severity in dengue patients cannot handle the complexity of the clinical profiles of the disease...(7-9)" (line 44-45). Please specify the exact problems faced in clinical practice. The failure of the classic statistical methods for prediction, is this mentioned in reference 7-9?

Thank you for your suggestion. We restructured this paragraph, mentioning some of the problem faced in clinical practice. We also updated the references with better description of these issues. Please refer to lines 70-72.

2) Study population: please specify inclusion criteria and exclusion criteria, especially if patients had some background comorbidities? That is because patients with underlying comorbidity may also have clinical characteristics the same with variables to be identified. For instance, a patient with liver disease may have a AST or ALT >1000, but possibly not caused by dengue infection.
Patients with comorbidities were not included in this study. We added the inclusion/exclusion criteria information in the methods section, subsection “Study Population and eligibility criteria”.

3) Study population (line 68), "patients admitted in three pediatric hospitals in RJ...", Since the patients were enrolled from three pediatric hospitals, a majority of patients were children?

Not necessarily. The age shifting in Brazil started in 2008. Before that dengue was an adult disease. In the beginning of 2007, the Instituto Nacional de Infectologia Evandro Chagas (INI) of FIOCRUZ started a project to study dengue infection in children in collaboration to some pediatric hospitals in the city. FIOCRUZ also supports the assistance to several communities including free diagnosis and admission to suspicion of dengue infection either for children or adults. Therefore, the selected cases for this study included either adults or children which is a strength of this study since it helped us to identify if there were differences in the dengue clinical profile by age. We have clarified it by creating a variable that described the number of children and adults in this study. We added this information in the methods (lines 125-127) and in the results sections (lines 260-263) describing how many adults (> 18 years old) and children (< 18 years old) were included in our analysis.

4) The basic information of subjects, such as age and gender distribution of study population, should be told. And please change the median to mean in age variable and present more demographic information in table.

Thank you for your suggestion. We added this information to Table 1.

5) Methods: Please explain how to determine the sample size.

This is a cross-sectional study design that considered retrospective data of patients who were assisted at the reference centers for suspected cases of dengue between the years 2007 and 2013. All cases that matched the inclusion criteria were selected for this analysis.

6) Results: line 162-164. The basic clinical data and outcomes should be provided.

Since we used an unsupervised approach, our analysis does not consider an outcome. However, we made a comparison of our results with the specialist classification. The specialist classification is described in Table 1 and the clinical variables used in our model are described in the Table 2.
7) Table 1, Why fever as a clinical symptom was not included?
Because 100% of the patients had fever, therefore this information would not be helpful in finding natural groups.

8) Table 1, many variables need to be defined. For instance, dehydration (how to judge), rapid and week pulse (means pulse >100 times per min?), severe bleeding.... Otherwise, the patients cannot be properly classified.

These definitions are described in the World Health Organization guidelines published in 2009 (WHO, 2009) and updated in the guidelines published by the Pan American Health Organization (PAHO) in 2016. As our group follow the premises described in these guidelines to classify the patients, we added a line in the Clinical data and data pre-processing in the Methods section to refer the reader to the 2009 WHO guidelines for the most recent definitions. We also added a briefly description of some signs and symptoms as defined by these guidelines in the methods, subsection “Data preprocessing and Clinical classification” (lines 131 to 148).

9) Table 1, there are all clinical characteristics in this table, so the number of inpatients or outpatients, days after onset of symptoms and age, this information are not proper included in this table, or should be re-organized. Also, if it will make more sense for this article.

Thank you for your suggestion. We have now created two separated tables from Table 1. One table with the demographics (age, sex), classification and dengue serotypes, and other table with the frequency of the clinical variables divided by age group.

10) Table 3, the authors put forward that they identified four natural clusters, with the cluster 4 having the highest percentage of patients with severe dengue (24.5%), on what basis the risk characteristics were compared and identified. Have you noticed that Cluster 2 also had a considerable percentage of severe cases (11.3%)?

We acknowledge that Cluster 2 also had a considerable percentage of severe cases and we explained this fact it in our discussion lines 405-413. Besides sharing characteristics with cluster 1 (low risk) and cluster 4 (high risk), this cluster also grouped more patients in the beginning of the critical phase showing the gradual transition from dengue without warning signs to dengue with warning and severe dengue. We added a figure (Fig. 3) showing the distribution of the specialist classification with clusters.
The statistical parameters should be added in figure 3 and 4 to prove the difference between four clusters. How could you judge the age is the key feature?

Thank you for point this out. We modified the Figure 3 and 4 (now Figure 4 and 5) by adding the statistics comparing these differences among the clusters in both Figures.

In the paper, the age was considered as the key features for splitting the data into four cluster. That is to say, it may be the most important risk factor for prediction of severe dengue? If so, to what extent the age may impact on the severity of dengue patients? And which age groups are at most risk of developing severe dengue?

Although the difference in the age among the clusters showed to be statistically significant, we cannot define which age groups are most at risk due to our descriptive study design. Further studies considering a longitudinal approach would be more suitable to answer this question.

Several clinical factors have been identified for dengue severity, such as abdominal pain or tenderness, clinical fluid accumulation, mucosal bleeding, lethargy, restlessness, liver enlargement, by comparing one cluster with another. How to validate these results? And the reliability of the result data should be explained.

The main goal of this descriptive study was to identify natural patterns that we could compared with the clinician’s classification based on the WHO guidelines in order to give insights of which clinical factors should be carefully considered in a hyperendemic area for dengue such as Rio de Janeiro. These results do not replace either the well-established dengue cases management protocol described by WHO/PAHO or the clinicians experience. We added this line at our limitation’s paragraphs in our discussion section to not misled the reader to wrong conclusions. Future studies with a longitudinal approach have been conducted by our group in order to validate these results.

In conclusion (line 290-296), you should summarize the most important findings and the significance.

Thank you for your suggestion. We restructured this section.

Generally, the study was conducted to identify the possible risk factors among clinical characteristics for evaluating the severity of dengue patients by using an unsupervised machine learning approach, namely random forest algorithms and SOM. However, the
logic of the article is not very clear. You may should show how to produce the 4 clusters first, and to compare all the variables between 4 clusters by using chi-square test or Fisher exact test. Then, to select meaningful variables to represent and make multiple comparisons further. what is the new information provided for clinical practice? As the authors said "... performs well considering a combination of available variables regardless of the data structure...". The question is how to use this combination of variables? especially these results data are not validated. When a patient develops a specific symptom, like severe bleeding or liver enlargement, the situation may get worse and of course would attract the attention of a doctor. From this point, identification of risk factors or markers for dengue severity before the onset of specific symptoms seems more meaningful.

We are grateful for your many positive comments. We will make sure that all the required modifications are made to make it as clear as possible. We would like to emphasis again that this study is merely descriptive and that by exploring these results will give future direction(s) for other studies. We have modified the text mainly in the discussion and conclusions sections to ensure the reader understands that we are proposing a future research direction based on the lead literature.

16) All figures should be improved and make the table clearer to fit the requirements of scientific articles.

Thank you. We have improved all figures according to the Journal requirements.

Ps.: For the lines number, please refer to the file "manuscript_GMH_May30th2019_withtrackchanges.docx". Thank you!