Reviewer’s report

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

Version: 0  Date: 22 Nov 2018

Reviewer: Jiahai Lu

Reviewer's report:

Comments

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?

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.

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? 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.

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

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

6) Table 1, Why fever as a clinical symptom was not included?
7) 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 can not be properly classified.

8) Table 1, there are all clinical characteristics in this table, so the number of inpatients or outpatients, days after onset of symptoms and age, these information are not proper included in this table, or should be re-organized. Also, if clinical characteristics can be presented in three groups (Dengue without WS, Dengue with WS and severe dengue), it will make more sense for this article.

9) 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%)?

10) 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?

11) 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?

12) 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.

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

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

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

Yes

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

No

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 am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

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