Reviewer’s report

Title: An exploration of mortality risk factors in children with non-severe pneumonia using clinical data from Kenya

Version: 0 Date: 07 Aug 2017

Reviewer: Leo Celi

Reviewer's report:

This review was performed with Rodrigo Deliberato MD PhD.

This paper aims to use a variety of machine learning (ML) techniques as a variable selection method to help determine when patients with moderate pneumonia have high mortality and therefore require inpatient care. The authors seek simple decision rules to improve WHO guidelines in treating pneumonia, and found three: patients 2-11 months, respiratory rates higher than 70 breaths/min, and presence of comorbidities.

The manuscript is well written and has a valid concept. Regarding the methodology, overall, the authors used the best practice in validating their models, and they treat the models with a good amount of respect and skepticism. However we believe that some issues need addressing.

Major Comments:

1. The data seems too clean. How did the authors check the quality of the data that was analyzed? Please describe what measures were taken to evaluate the fidelity of the data. This is the most crucial step in data modeling.

2. We agree in principle that ML can be used in this way for variable selection. However, ML does best when it is allowed to produce non-linear and higher-order models, which the authors did not report testing for or examining.

We recommend two approaches:

a. Either test for interactions or nonlinearities in the ML models

b. Use a ML tool specifically designed for this process: Causal Falling Rule Lists (Fang & Rudin 2012)
3. Page 7: they say variables "were selected", then list 10 or so. Earlier, they say they have 350 variables. Which is it? If the investigators used statistical methods to go from 350 to 10, what were they? ML is really good at variable selection in high dimension, and a lot of models benefit from adding in variables that are only weakly predictive.

4. How different would their results be if they ran a simple logistic regression of mortality on their data with feature selection applied to the 350 candidate variables? How much accuracy do they gain compared to the WHO guidelines? How many lives might it save? And at what financial cost?

5. Their variable importance procedures are not all strictly comparable, but the idea of ranking the importance of predictors across multiple measures is a strong point of the paper. The correlations in Table 3 provide confidence that the results are solid.

Minor Comments

1. We disagree with the authors' assessments of logistic regression as being biased. It's not biased in the presence of collinear predictors; it just doesn't have an analytic solution. Please re-phrase.

2. It would seem that the effect of one of the 3 highly discriminative factors on mortality - respiratory rate greater than 70/minute - is modified by age. Were interaction terms evaluated in the logistic regression?

3. Abstract in the first page: 'Change the word "three" for "four" machine learning modelling.'

4. Line 162: there is a missing parenthesis.
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.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

Yes

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:

Acceptable

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 have no conflict of interest relevant to this paper.

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