Reviewer’s report

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

Version: 0 Date: 18 Jul 2017

Reviewer: Tim Colbourn

Reviewer's report:

Overall an interesting and important paper. I have a number of major comments that need addressing however, and there are also numerous aspects of the methods, results and interpretation that need clarification or justification. I've divided my comments into major and minor comments below.

Major comments

1. Statistical methods, lines 204-208: "The SMOTE technique was used to oversample the minority class elements, which eliminated the possibility of information loss. This was achieved by combining the features of existing instances with the features of their nearest neighbours to create additional synthetic instances." This needs to be more clearly explained - how many negative cases were resampled in each model? how were "nearest neighbours" determined? are they positive cases or negative cases? and how was the information from nearest neighbours combined with the information from the original negative cases? how much of the information? which covariates? why those covariates? this technique seems like it could introduce a lot of bias, especially as it has a lot of moving parts (i.e. it could be done in so many different ways?).

2. Table 2 - why not include diarrhoea and anaemia in the model separately given that the other two co-morbidities included in the "Presence of co-morbidity" variable - malaria and dehydration - are each not significantly associated with mortality. This has major implications for the discriminant power of your "Presence of co-morbidity" variable in all your models - perhaps it would be better to model the four co-morbidities separately?

3. Sensitivity analysis, lines 318-320: "Given that the PLS-DA model had the highest ROC score (Fig 2), and its sensitivity was higher than all other models when considering its performance in both imputed and complete case analysis [see Table S2 in additional file 2]" This does not seem to be the case - the RF model is more accurate in both imputed and complete case analysis, and whilst the PLS-DA model is most sensitive in the
imputed analysis, it is only the third most sensitive of the five models in the complete case analysis. Your selection of the PLS-DA model therefore seems somewhat arbitrary.

4. Lines 340-342 (additional file 4 AUC results): "This is suggestive that differences in pneumonia classification criteria might not have a statistically significant impact in determining risk factors for this population." This does not seem to fit with what is shown in Figure 4 at all.

5. Lines 379-380: "all models had a higher AUC and sensitivity compared to the traditional logistic regression model" - this is not true for the complete case analysis as shown in Table S2. The "accuracy" of the Logistic model is also higher than that of the PLS-DA and Elastic net models for the imputed dataset according to Table S2 too. Given the AUC results also indicate that all models are approximately equal (not statistically significantly different) in terms of fit, it seems you may be over-interpreting the benefits of the machine learning approaches. Line 432 of your conclusion "performed much better than the traditional logistic regression" also does not seem warranted if statistically the machine learning models are no better fit than the logistic regression model.

Minor comments

6. The drop from 16162 total cases to 11318 analysed is quite a large drop - why were so many cases excluded? It seems from the methods section (lines 146-151) that this may be because severe cases were excluded, though looking at Table 1 this does not seem to be the case as the total in Table 1 already seems to be 11318. Please explain what happened to the rest of the 16162 cases in the abstract, and also more clearly in the methods section (the Figure 1 flowchart is insufficient given ambiguities in the numbers stated in the results section, see comments below).

7. Reference 1: New estimates for 2015 are now available - see Liu et al 2016 Lancet article

9. Table 1 - it seems odd that there are no missing values for Age, Respiratory Rate, Malaria, Acute malnutrition, Clinician pneumonia diagnosis, Penicillin, and Co-morbidity, when there are missing values for the other variables. Why is this? Also, including total N at the top of Table 1 would be useful.

10. Statistical methods, lines 159-170, critique of logistic regression: co-linearity can be tested for and covariates that are too closely correlated should not be included in a predictive model together. "Additionally, apart from model coefficients and significance tests, logistic regression models offer no guidance on feature selection from model outputs that can guide future intervention design" - what further guidance is required? Why is the size and significance of a co-efficient from a logistic regression model not sufficient? A better justification for use of machine learning is required in this paragraph, especially considering you are critiquing your own group's logistic regression work (under review at Lancet Global Health) here! Perhaps further elaboration on the unfamiliar term "feature selection" would be useful?

11. Statistical methods, line 182: what is a "10-fold internal cross validation"? Please explain further, in particular why it is "10-fold".

12. Line 205: it may be easier to say "cases who survived" (or negative cases) rather than "minority class elements". In general language could be simplified throughout the statistical methods section bearing in mind this is a medical journal and not a journal of statistical science.

13. Lines 211-212 - as above: "Out-of-Bag imputation method" - what is this? And is it different from "multiple imputation by chained equations (generating 10 imputed datasets) was performed under a Missing At Random (MAR) assumption"? (Is this "Out-of-Bag" sentence needed as well?)

14. Line 234 "11, 318 cases of non-severe pneumonia" - why is this different to the "11,633 non-severe pneumonia cases" stated on line 238 - why is the 11,633 number not shown in Figure 1 (and is this the total for Table 1? If so, please add it to the table). In general the flow of patients and numbers in each category needs to be made clearer. Table 1 should also show the overlap between the Clinician, WHO and Penicillin classifications of non-severe pneumonia, to make this clearer.
15. At the bottom of page 11 you state that "only non-severe pneumonia cases according to WHO criteria used" in the logistic regression model reported in Table 2 on page 12, and the footnote to Table 2 indicates that there were 11,318 cases included in the model. However, Table 1 indicates that there were only 8,167 cases of non-severe pneumonia according to WHO guidelines. Why the difference in numbers? again it is very hard to follow which cases were included in which analyses and why.

16. Page 15 and Figure 3: you need to make it clearer to the reader where the probability of death is derived from. You state: "for probability threshold of mortality above 6.5%, there is a net benefit for admitting non-severe pneumonia cases less than 12 months old with respiratory rate ≥ 70 breathes/minute and with presence of one or more comorbidity." How would a clinician know when the probability of mortality is above 6.5%?

17. Figure 3 - what are the units of "net benefit" - please also explain how net benefits were calculated in your methods section

10. Additional file 2, supplementary Table S3: how is "accuracy" different to AUC? please explain this in your methods section. Also please explain accuracy in Table S2, and why AUC is not shown in this table. Also why is AUC rather than 'accuracy' compared in Additional file 4?

18. Line 330 and Figure 4 (and Table S3): "the ideally diagnosed and managed population" - this needs to be introduced and explained in your methods section

19. Line 331: it's not clear why you suddenly start discussing WAZ here (especially as it is not one of your top three variables).

20. Additional file 4 - please show the AUC values and their 95%CI, as well as the p-values for their comparison.

21. Line 368: "based on relevance" - this needs explaining in more detail - how is "relevance" determined here?
22. Line 399: "microbiological or microbiological" - please correct

23. Lines 395-411: it is not clear why you discuss the importance of pallor and of female here given these were not included in your three most important variables and you therefore (implicitly) recommend that clinicians do not focus on these characteristics when assessing non-severe pneumonia cases (i.e. these variables were not included in your decision analysis).

24. Line 415: "countless" - is this accurate?

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

No

**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.

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 recommend additional statistical review

**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 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.