Author’s response to reviews

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

Authors:

Timothy Tuti (ttuti@kemri-wellcome.org)
Ambrose Agweyu (AAgweyu@kemri-wellcome.org)
Paul Mwaniki (PMwaniki@kemri-wellcome.org)
Niels Peek (Niels.Peek@manchester.ac.uk)
Mike English (MEnglish@kemri-wellcome.org)

Version: 1 Date: 29 Aug 2017

Author’s response to reviews:

Reviewer #1: 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 – 1) how many negative cases were resampled in each model? 2) how were "nearest neighbours" determined? 3) are they positive cases or negative cases? 4) and how was the information from nearest neighbours combined with the information from the original negative cases? 5) how much of the information? 6) which covariates? 7) why those covariates? this technique seems like it could introduce a lot of bias, especially as it has may have a lot of moving parts (i.e.8) it could be done in so many different ways?

We do agree with the reviewer that more details on SMOTE technique need to be reported. Part of our reasons for omitting these details within the main manuscript was to try to keep the message palatable to a more clinical/epidemiological audience. In addressing the reviewer's comments, we have added Supplementary Table 3 with more details on our implementation of SMOTE technique and the way we dealt with possible bias. The covariates used in SMOTE sampling techniques are all those eligible for analysis as defined in lines 151-166. These variables were identified a priori guided by clinical expert opinion and literature review as indicated in line 152-153. The total number of negative cases resampled was not model specific as it was within cross-validation pre-processing, so cases generated was similar across all models. While there exist alternative ways of performing class balancing when the outcome is a rare event, there isn’t an agreed upon gold standard. However, SMOTE is a common preferred choice because it achieves better classifier performance (in ROC space) than alternative choices.

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?

The decision to combine the variables was made a priori based on previous studies that have demonstrated the clinical overlap of signs for malaria, dehydration, anaemia and pneumonia with increased risk of mortality with each of the variables, and further motivated by the desire to provide simplified criteria for clinical decision making in low resource settings. Notably, the variable representing comorbidity ranked higher than the individual constituent variables in almost all the five modelling techniques.

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.
We support our use of PLS-DA model on the basis that it had a AUC score higher than logistic model (Fig 2), and sensitivity greater than all other models when considering its performance in both imputed and complete case analysis [Table S4 in additional file 2], we used it for comparison of variable importance across the alternative criteria for pneumonia classification. Our choice to preferring AUC over accuracy, is also guided by evidence from published studies showing AUC to be statistically consistent and more discriminating measure than accuracy: we have explained this in line 338-343 and added a reference to report expounding on this. Given our interest was mortality outcome, sensitivity and AUC were the primary performance measures considered.

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.

We would beg to differ. The bottom chart on Additional file 5 is a plot of p-values from DeLong’s test comparing the AUC curves based on classification criteria with the reference category being AUC based on the new WHO classification criteria for pneumonia. They [p-values] are all 0.35 and above.

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.

We agree with the reviewer that the complete case analysis in the previous version of the manuscript, the logistic model was doing better than the other models. However, our analysis strategy was on using the full dataset together with imputation as the primary analysis and our assessments were based (and consistent) with this strategy.

Additionally, in efforts to address some of the other reviewers’ comments, this version of the report has introduced some interaction terms as evidenced in Table 3. The performance in
supplementary table 4 show logistic regression not to perform better than the other models in full dataset analysis as well as complete case analysis. Our evaluation of good performance is based on a good AUC score, coupled with the smallest drop in sensitivity (due to public health concerns being central) between the full dataset analysis and complete case analysis. This is detailed in line 338-343, 404-407 and supplementary table 4.

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).

We agree with the reviewer that there were some missing steps as to how the analysis population was reached at. We have updated the flow diagram to reflect our analysis population, with lines 158-163, 236-246, 248-255 expounding on the final analysis population

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

The reference and the corresponding text has been updated.


We do agree with the reviewer. The misplaced reference (due to manuscript reformatting) has been removed.
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.

We have added N at the top of the table, together with an explanation of this count as dealt with in earlier comment. Some of the variables do not have missing values because we have been working with the hospitals to involved to improve quality of documentation. This is detailed in line 126-140, with reference materials on our continuing work with these hospitals expounding further on improving data quality for information making.

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 coefficient 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?

We have updated the justification for machine learning techniques in line 174-189. Our critic here labours on the following:

While the goal of machine learning approaches is to “learn” from data of all sorts where no rigid pre-assumptions about the problem and data distributions, traditional models’ goals are to focused on analysing and summarizing data where tight assumptions about the problem and data distributions exist. Additionally, generalization in machine learning techniques is pursued empirically through training, validation and test datasets whereas in traditional modelling approach, generalization is pursued using statistical tests on the training dataset. In a machine learning approach, redundancy in features (variables) is okay, and often helpful, with algorithms being designed to handle large number of features. Traditional modelling approaches often require independent features and prefer use less number of covariates (sparser models). When the decision boundary for classification is not a linear function of the features (variables), GLMs might not be ideal for feature selection. We have added a reference to published works that explain in details what this entails in line 181 and 184, and elaborated a bit on feature selection in the supplementary table 1.
11. Statistical methods, line 182: what is a "10-fold internal cross validation"? please explain further, in particular why it is "10-fold".

We have added a brief explanation of the 10-fold internal cross validation in the additional file 2, supplementary table 1. We have avoided putting it in the main manuscript in order to keep the message as clear as possible to towards a clinical audience, and to prevent extending the manuscript word count further.

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.

We have simplified the statements in line 217-225 as requested by reviewer.

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?)

It is not different to multiple imputation by chained equations rather seeks to expound on how the chained equations were arrived at. In accordance with the reviewer’s previous comment to simplify the language (12), we have omitted this sentence from the manuscript to make it easier to understand by clinical audience. Additional details, necessary for reproducibility can be obtained from the attached additional file 4.

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.
The 11,633 was a typo which we have rectified. We have also added a Venn diagram referenced in the additional files, showing how the classifications are distributed and it is referenced by line 254-255.

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 8167 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.

We do agree with the reviewer that there were conflicting details regarding number of cases included in logistic regression in table 2. We have removed line 264 to make it clear that all non-severe pneumonia cases included in the final stage of figure 1 were used in table 2.

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%?

From our findings, in non-severe pneumonia cases, less than 12 months old with respiratory rate ≥ 70 breaths/minute and with very low WAZ, roughly one in every 7-14 patients in this population stands the risk of mortality. The key point is that decision analysis is not trying to predict the “correct” threshold in the future, but using the provided range of probabilities where there is a predicted net benefit, clinicians can decide what would be a reasonable decision cut-off on admitting non-severe pneumonia cases based on the local context i.e. resources available at the hospital level, the socio-economic considerations for the patient in question, the public-health gain they see getting from admitting the patients (if they choose to…).

17. Figure 3 - what are the units of "net benefit" - please also explain how net benefits were calculated in your methods section.
The unit of net benefit is true positives, and we have provided a reference to a key report in the methods section that details its calculation. We have included this in the main manuscript in line 210-211.

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?

We have added an explanation of “Accuracy” and how it differs from AUC in the supplementary table 4. AUC is considered to be a statistically consistent and more discriminative measure than accuracy and that is why we focused on it rather than accuracy (Liu et al 2003, “AUC: a statistically consistent and more discriminating measure than accuracy.”). DeLong’s test compares two AUC-ROC curves (in our case the Logistic regression AUCROC curve versus alternative) to see if they are significantly different, which is what additional file 5 illustrates.

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

We have added line 119-122 to explain this population in the 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).

Malnutrition is an important risk factor for mortality as described in the background. We have used weight-for-age Z score to represent malnutrition in the analysis. We have included a statement to clarify this action in the methods section under quantitative variables. In general, it is linked to susceptibility of children having an increased number of comorbidities.
20. Additional file 4 - please show the AUC values and their 95%CI, as well as the p-values for their comparison.

We have added these details to Table S4 additional file 2.

21. Line 368: "based on relevance" - this need explaining in more detail - how is "relevance" determined here?

We have added more details explaining that it is based on clinical relevance (as indicated by the techniques applied which allowed for the ranking of the importance of clinical signs and patient characteristics as risk factors for pneumonia mortality) in line 388-392.

22. Line 399: "microbiological or microbiological" - please correct

We have corrected the grammar in this in line 422.

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).

While the two variables were not among the most important variables, we felt it necessary to include them in our discussion as they were both significantly associated with mortality in the adjusted logistic regression model (table 2).

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

We have corrected the wording to be more reflective of our meaning in line 436-438.
Reviewer #2: This is an interesting piece of work on the important matter on whether a WHO guideline on pneumonia appropriately identifies children at risk of dying in SSA. It is overall a well-written piece with some interesting findings. The analysis uses a variety of advanced statistical methods that, regretfully, I am not able to assess the adequacy of and I thus strongly recommend to have a statistician review the manuscript.

These are my comments:

Definitions
I suggest to help the reader by clarifying some of the key concepts in the paper, for example it is sometimes confusing what is meant by pneumonia - I believe it would be easier to follow what has been done if stating whether referring to WHO 2013 definition, earlier WHO definition, clinician's diagnosis etc throughout the manuscript. Also, please clarify that mortality refers to in hospital mortality (line 33 and others).

In the methods section, under the paragraph describing study participants, we have qualified the use of the term pneumonia by stating the term pneumonia to refers to children with a documented clinical diagnosis of pneumonia, and the terms non-severe and severe pneumonia refer to those for whom WHO, under the 2013 revised definitions, recommends outpatient and inpatient care respectively.

Comorbidities
Abstract background talks about malaria, diarrhoea and malnutrition as dangerous comorbidities but in analysis you are focusing on malaria, diarrhoea/dehydration and anaemia. Pls provide explanation or rephrase.

Diarrhoea is, by far, the most common cause of dehydration requiring hospitalization in children. The terms are therefore considered jointly. The abstract has been revised accordingly. Malnutrition is an important risk factor that is represented by weight-for-age Z score in the analysis. We have included a statement to clarify this in the methods section under quantitative variables.
How is "pallor" defined? Was there any assessment of Hematocrit?

Pallor was defined clinically in line with the existing WHO guidelines. Hematocrit was not routinely performed. We have acknowledged the challenge of limited diagnostic capacity in the discussion.

What is the reason for clustering different comorbidities into one single variable? My understanding is that pallor causes the significance of this variable since dehydration and malaria were insignificant?

The decision to combine the variables was made a priori based on previous studies that have demonstrated the clinical overlap of signs for malaria, dehydration, anaemia and pneumonia with increased risk of mortality with each of the variables and further motivated by the desire to provide simplified criteria for clinical decision making in low resource settings (line 144-148). Notably, the variable representing comorbidity ranked higher than the individual constituent variables in almost all the five modelling techniques.

In table 2, why have malaria and diarrhoea not been presented separately? Or was dehydration used as a proxy for diarrhoea?

We suspect the reviewer meant to ask why diarrhoea and dehydration are not presented separately. Our response to this is captured in an earlier comment clarifying that diarrhoea is, by far, the most common cause of dehydration requiring hospitalization in children. The terms are therefore considered jointly.

Sampling

Since children were only included if born after 2011 there could not have been any 4 and 5 years old in the beginning of the study. This should have increased the proportion of younger children, which in turn could contribute to the fact that younger children seemed to be at higher risk of death. I think this needs to be highlighted in the discussion of results.

We agree with this observation and have inserted a statement to this effect in the discussion line 449-453.
Also, as a potential limitation is that enrolment of children was only done from hospitals and that this group may be different from children with pneumonia that are treated at health centres or even at home. While they may be less severely sick also other factors could have influence such as tendencies to seek earlier for younger children, better socioeconomics etc. Pls discuss.

We agree with this observation and have inserted a statement to this effect in the discussion in line 455-460.

Age and respiratory rate is presumably linked and thus it would be interesting to see how the raised respiratory rate of >70bpm relates to mortality in children older and younger than 12 months. Could separate analysis for different age groups be considered?

In responding to reviewer #3, we included an interaction term for age and respiratory term - among other terms- in efforts to address the possibility that age and respiratory might affect inpatient mortality in a non-additive way. This would help model this relationship without necessarily splitting the analyses based on these factors while addressing the concerns raised by the reviewer.

Minor issues:
The abstract seems to be more than 350 words?
We have reduced the number of words in the abstract to 350

Line 29: I am not a native English speaker but to me the word "provided" seem unnecessary? Pls revise.
We have fixed the grammatical error

Line 83: the word "are" should be removed
We have fixed the grammatical error
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 scepticism. 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 analysed? Please describe what measures were taken to evaluate the fidelity of the data. This is the most crucial step in data modelling.

   The data got to this level of quality due to previous and current work we are doing with these study hospitals as a clinical information network group. Lines 126-140 give a brief description to what was done to improve data quality with emphasis given to referenced work which goes into details on the measures taken to ensure data fidelity.

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)

   We do agree with the reviewer that the above-mentioned approaches would best demonstrate the power of ML. We opted to adopt option (a) as proposed by the reviewer and included six interactions that are reasonably supported by literature for this population. We have updated the report to reflect this from page 14 onwards. We also attempted to include interactions terms for all variables with each other but this become untenable due to limited computing power on our end. Based on the luck-lustre performance by most interaction terms, with the exception of age interacted on respiratory rate, we are confident that the current results are robust.

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.
Although the database includes up to 350 variables for each patient, no patient would have data on all 350 as it depends on diagnoses and treatments. Also, many relate to clinical and nursing care instituted after the initial clinical assessment. Only a restricted subset, which we included, represent clinical characteristics that were relevant to this analysis of risk factors to aid in clinical decision making.

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?

We were only interested in presenting clinical variables at the time of admission that would potentially offer clinical utility in determining risk profiles as decisions would have to be made on admission - so not all those associated with treatment and subsequent stay would be relevant for that decision. Additionally, while the database included 350 variables, the vast majority represented features tied to skip logics so that many variables are not relevant to the majority of patients, and relate to care provided after initial clinical assessment such as treatment prescribed, investigations requested and nursing observations undertaken. The WHO guidelines form the basis of the structured admission record form which captures the variables collected. Therefore, there are no additional candidate variables against which we are able to compare the performance of WHO signs.

Decision curve analysis helps to address the following two general problems associated with applying traditional decision-analytic methods to prediction models when it comes to lives that might be saved and the financial cost involved:

(a) Lack of data on costs or quality-adjusted-life-years, that are not found in the validation data set meaning that a prediction model cannot be evaluated in a decision analysis without further information being obtained.

(b) The challenge that often-required explicit valuation of health states or risk-benefit ratios for a range of outcomes by decision-analytic methods are prone to systematic biases and burdensome to elicit from subjects.
This is further elaborated on in Vickers et al 2006 report “Decision curve analysis: a novel method for evaluating prediction models., and we believe this is ideal for our scenario, as illustrated in our response to reviewer 1’s question 16.

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.

We thank the reviewer for this comment and we are glad that our efforts in demonstrating that these results are solid are clear.

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.

We have rephrased the wording the challenge of an analytic solution in this instance in line 176-186.

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?

We thank the reviewer for this insightful comment. We have added interaction terms in this revision in the subsequent ML models (as explained in Reviewer’s 3, Major Comment 2) and we are happy to report that the results remain largely robust to inclusion of these interaction terms.

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

The requested change has been made but in light of comments to cut the abstract word count, the count of the ML techniques used was removed.
4. Line 162: there is a missing parenthesis.
That has been rectified.

Editorial Requests

Please ensure that the Abstract does not exceed 350 words.
We have shortened our abstract to 350 words.

Consent for publication is required for manuscripts containing any individual person’s data in any form (including individual details, images or videos).
The manuscript has a consent for publication section at the end of the manuscript.