PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Development and validation of a novel computer-aided score to predict the risk of in-hospital mortality for acutely ill medical admissions in two acute hospitals using their first electronically recorded blood test results and vital signs: a cross-sectional study
AUTHORS	Faisal, Muhammad; Scally, Andrew; Jackson, Natalie; Richardson, Donald; Beatson, Kevin; Howes, Robin; Speed, Kevin; Menon, Madhav; Daws, Jeremey; Dyson, Judith; Marsh, Claire; Mohammed, Mohammed

VERSION 1 – REVIEW

REVIEWER	Anthony Bleyer MD Wake Forest School of Medicine, Medical Center Blvd., Winston-Salem, NC
REVIEW RETURNED	21-Mar-2018

GENERAL COMMENTS	It was with great pleasure that i read this manuscript. We are similarly interested in this issue, so i feel i have a good understanding of what was done. I think that this provides exciting new information that will soon be applied to many hospitals throughout the world. I would like to congratulate the authors on a very nice job. I am not an expert in statistics, but here are some thoughts i had: 1) First, i would like to see a direct comparison with the NEWS. Could you do ROC curves with the NEWS on the same graph as the ROC curve for the new score and show how they compare? 2) Instead of imputing values for the missing values, could you please just have an indicator variable instead that states these labs were missing? I think this would show that these patients actually have a better survival. 3) If greater than 90% of blood tests are within 24 hours, i would not include records up to 4 days, but just up to 24 hours. You might want to compare what would happen if you took them out. I see this as a minor point and leave it to the discretion of the authors. i think if you do the first 24 hours it would have better clinical applicability and you already have enough data.. 4) Could you say a little more about how you brought all of the variables into one model. Anthony Bleyer MD

REVIEWER	Aiman Tulaimat Cook County Health and Hospitals System Chicago, IL, USA
REVIEW RETURNED	10-Apr-2018

GENERAL COMMENTS	In this study the authors developed and validated a model to predict mortality of hospitalized patients based on the initial set of

	vital signs and laboratory data in the electronic medical record. The model is well reasonably accurate and reliable. Major comments: 1. Many of the patients that die in hospitals have do not resuscitate orders. Was this information available in the medical record? It crucial to know if the model is accurate in patients that accept resuscitation. 2. There was no subgroup analysis. I know the model already includes gender and age, but we do not really know how it performed in women, younger, or older adults. We also do not know how it performs in patients with longer admission versus shorter ones. It is also useful to perform subgroup analysis based on the chief complaint or major diagnosis: shortness of breath, bleeding, sepsis, urinary tract infection, etc. 3. There are many cyclical phenomena in medicine: respiratory illnesses in winter, new staff joining at a certain time of year, staff patterns during the day. The authors should run analysis by time of admission in the day, day of the week, and month of the years. 4. Keeping the model in the supplement reduces its exposure and the way it is written is not very useful either. I suggest grouping similar individual variables from one system together (respiratory rate and oxygen next to each other; AKIs, urea, and creatinine together, etc). I suggest that the authors create a table with the variables of the model in the first column, grouped by system, the next column is the coefficient associated with the variable, then two example patients (one dead, one alive) in the following columns and how the logit is calculated and then converted into probability. 5. How did you select threshold of 0.08 and what are the tradeoffs associated with it. Minor comments Please add a table to the supplement with the age, gender, length of stay, mortality for the patient that were excluded. Results: Page 11, line 14: were temperature, blood pressure, and oxygen saturation higher in patients that d Results: page 11, line 17: were respiratory rates and pulse lower in patients that died? Figure S6: Is the predicted probability the probability of death and the observer proportion the observed death? Figure S8: is the threshold the same thing as the predicted mortality?

REVIEWER	Kenneth R Hess, PhD Department of Biostatistics, UT MD Anderson Cancer Center Houston, TX USA
REVIEW RETURNED	29-May-2018

GENERAL COMMENTS	1. Title: “score” should be “model” since score implies a discrete set of values that can computed easily and are not directly tied to predicted probabilities. 2. Page 3, Abstract, Materials: A brief description of the model development should be added. 3. Page 3, Abstract, Results: Calibration results should be added. 4. Page 8, paragraph 1: Please explain for readers the choice not to perform variable selection for other main effects once highly significant interactions were included in the model. Many researchers perform variable selection during model development in the belief that parsimony improves prediction performance on independent data. 5. Page 8, paragraph 1, line 5: Some discussion should be added

	around the choice to using simple data transformations instead of using more general methods such as splines or fractional polynomials to model non-linear covariate effects. 6. Page 8, paragraph 3, line 4: Some discussion should be added around the interpretation of scaled Brier score values. Is between 15% and 20% usefully accurate? 7. Page 8, paragraph 3, line 7: Given Figure S6 it is clear that graphical calibration analysis was also performed. Please describe here the methods used to generate these figures. The plot for training data prior to re-calibration should also be shown. Consideration should be given to moving these plots to the main manuscript and dropping the HL assessment altogether. 8. Page 9, paragraph 2, line 5: “higher risk” should be ‘higher predicted risk”. 9. Page 11, top: Given that plots of the validation data were used during model development, the resulting “validation” does not represent true external validation where researchers remain completely blinded to the validation data until the final model developed on the training data has been locked down. This fact needs to be added to study limitations and the term “external validation” used with appropriate qualification. 10. Page 12, paragraph 1: I don’t think the right question is whether the current model works well when median imputation is used on the samples omitted due to missing data but rather whether a model developed including the imputed data performs better than the current model. Frankly, this is probably more important for smaller datasets. 11. Page 12, paragraph 2: “positive predictive value” should be “positive and negative predictive values”. Again since the validation data are used in selecting the cut-off this does not represent true external validation. The cut-off should be based only on the training data and then tested on the validation set. 12. Table 1: It is confusing that the percentages for the “N” column are column percentages while the percentages for the “Died” column are row percentages. 13. Table 3: Please change the column header “Discrimination” to “Discrimination Slope” and add a footnote describing how it is estimated. 14. Table 4: “PPV” should be “PPV %” and “NPV” should be “NPV %”.Consideration should be given to adding a column indicating the proportion of patients with predictions > cut-off. 15. Figure S2 and S3: Outliers should not be omitted from these plots. 16. Figure S4 and S5: These are not scatter plots. They appear to be line plots generated by connecting the proportions dead computed for six equal-sized groups and plotted at the midpoint of the resulting intervals. Given the size of the training data set it would seem that 20 groups would still have enough patients to reliably estimate the proportion died in each interval and give more granular information on the pattern of the covariate effect. 17. Figure S6: Please add a superimposed histogram or rug plot to show distribution of predicted probabilities. 18. Figure S7: Please revise figure legend to clarify that these are based on CARM predictions for patients with imputed values that omitted during model development and validation. 19. The authors should explicitly reference the TRIPOD guidelines and verify for readers that these were followed.

VERSION 1 – AUTHOR RESPONSE

Reviewer(s)' Comments to Author:

Reviewer: 1
Reviewer Name: Anthony Bleyer MD
Institution and Country: Wake Forest School of Medicine, Medical Center Blvd., Winston-Salem, NC
Please state any competing interests or state 'None declared': I am actually doing research on the same thing right now at our hospital.

Please leave your comments for the authors below It was with great pleasure that i read this manuscript. We are similarly interested in this issue, so i feel i have a good understanding of what was done.
I think that this provides exciting new information that will soon be applied to many hospitals throughout the world. I would like to congratulate the authors on a very nice job. I am not an expert in statistics, but here are some thoughts i had:
1) First, i would like to see a direct comparison with the NEWS. Could you do ROC curves with the NEWS on the same graph as the ROC curve for the new score and show how they compare?
Response: We have now added the ROC curve for the NEWS score (see Figure S9 in supplementary material).

2) Instead of imputing values for the missing values, could you please just have an indicator variable instead that states these labs were missing? I think this would show that these patients actually have a better survival.
Response: We are not keen to do this, because we set out to develop a risk score for those patients who have the necessary data items based on usual care. The reasons why clinicians do/do not undertake a specific blood test are complex and where the mechanisms for that missingness are not random. We included a simple medium imputation exercise to offer a preliminary insight into how CARM performs in those patients with some missing blood tests (most frequently albumin) and note that the NHS AKI algorithm also adopts this approach which is now widely accepted. Furthermore, all our blood test results are included as continuous covariates and so adding a value for "missing data" is not feasible without turning the continuous covariate into categories (eg quartiles) which is likely to undermine the accuracy of our model. We further acknowledge in the discussion that further work on how to best address the issue of missing data is required.

3) If greater than 90% of blood tests are within 24 hours, I would not include records up to 4 days, but just up to 24 hours. You might want to compare what would happen if you took them out. I see this as a minor point and leave it to the discretion of the authors. I think if you do the first 24 hours it would have better clinical applicability and you already have enough data.
Response: We would prefer to leave the model as is because the impact is likely to be small and our current performance statistics for CARM although good are therefore slightly more conservative.

4) Could you say a little more about how you brought all of the variables into one model.
Response: We have clarified this more in the methods. "The primary rationale for using these variables is that they are routinely collected as part of process of care without considering the statistical significance of any given covariate.". We did however use an automated search for two-way interactions which lead to improvements in the calibration of the model but not its discrimination.
Reviewer: 2
Reviewer Name: Aiman Tulaimat
Major comments:
1. Many of the patients that die in hospitals have do not resuscitate orders. Was this information

available in the medical record? It crucial to know if the model is accurate in patients that accept resuscitation.
Response: This is an interesting suggestion, but this information is not available in our data set.

2. There was no subgroup analysis. I know the model already includes gender and age, but we do not really know how it performed in women, younger, or older adults. We also do not know how it performs in patients with longer admission versus shorter ones. It is also useful to perform subgroup analysis based on the chief complaint or major diagnosis: shortness of breath, bleeding, sepsis, urinary tract infection, etc.
Response: This is an interesting suggestion but our first aim is to produce an overall model because the variables in CARM are routinely collected for the majority of patients. Once we have our CARM score peer-reviewed we will then be able to undertake subgroup analyses which are clinically relevant based on prior evidence from the literature and feedback from our clinical staff. We have used a similar approach to develop a Computer Aided Risk of Sepsis (CARS) score [1].
1 Faisal M, Scally A, Richardson D, et al. Development and External Validation of an Automated Computer-Aided Risk Score for Predicting Sepsis in Emergency Medical Admissions Using the Patient's First Electronically Recorded Vital Signs and Blood Test Results. Crit Care Med 2018;46:612–8. doi:10.1097/CCM.0000000000002967

3. There are many cyclical phenomena in medicine: respiratory illnesses in winter, new staff joining at a certain time of year, staff patterns during the day. The authors should run analysis by time of admission in the day, day of the week, and month of the years.
Response: This is an interesting suggestion, but the aim of this study is to develop an externally validated model. However, we will consider this in subsequent studies once the CARM score has been peer-reviewed.

4. Keeping the model in the supplement reduces its exposure and the way it is written is not very useful either. I suggest grouping similar individual variables from one system together (respiratory rate and oxygen next to each other; AKIs, urea, and creatinine together, etc). I suggest that the authors create a table with the variables of the model in the first column, grouped by system, the next column is the coefficient associated with the variable, then two example patients (one dead, one alive) in the following columns and how the logit is calculated and then converted into probability.
Response: We have now added a table as suggested although we emphasize that our score is for computer based implementation, not pencil and paper.

5. How did you select threshold of 0.08 and what are the tradeoffs associated with it.
Response: We have now added positive and negative log likelihood ratio and selected this threshold based on the positive likelihood ratio (LR+) being around 4 and negative likelihood ratio (LR-) around 0.20. Of course we acknowledge that such decisions are complex and subject to local contextual factors and this is why we have provided a range of values for the reader to consider.

Minor comments
Please add a table to the supplement with the age, gender, length of stay, mortality for the patient that were excluded.
Response: We have now added the table as suggested (see table S3 in supplementary material).
Results: Page 11, line 14: were temperature, blood pressure, and oxygen saturation higher in patients that d
Response: We have now corrected this.

Results: page 11, line 17: were respiratory rates and pulse lower in patients that died?
Response: We have now corrected this.

Figure S6: Is the predicted probability the probability of death and the observer proportion the observed death?
Response: Yes – we have now made it clear in the figure legend.

Figure S8: is the threshold the same thing as the predicted mortality?
Response: Yes – we have now made it clear in the figure legend.

Reviewer: 3
Reviewer Name: Kenneth R Hess, PhD
1. Title: “score” should be “model” since score implies a discrete set of values that can computed easily and are not directly tied to predicted probabilities.
Response: We have used a statistical model to produce a risk of mortality and so prefer to describe it as a risk score which ranges from 0 to 100%.

2. Page 3, Abstract, Materials: A brief description of the model development should be added.
Response: We have now added the brief description of the model development.

3. Page 3, Abstract, Results: Calibration results should be added.
Response: We have added the calibration slope results in abstract.

4. Page 8, paragraph 1: Please explain for readers the choice not to perform variable selection for other main effects once highly significant interactions were included in the model. Many researchers perform variable selection during model development in the belief that parsimony improves prediction performance on independent data.
Response: There are two broad classes of statistical model: explanatory and predictive models [1]. An explanatory model seeks to identify causal processes and the magnitude of the influence of specific variables on an outcome. In a predictive model, the aim is to extract the maximum useful information from all relevant available data, without focusing on the specific roles or influence of individual variables. These two approaches require very different modelling strategies. In this paper we adopt a predictive modelling approach and have two sentences in the methods to make this clearer – “The primary rationale for using these variables is that they are routinely collected as part of process of care and their inclusion in our statistical models is on clinical grounds as opposed to the statistical significance of any given covariate. The widespread use of these variables in routine clinical care means that our model is more likely to be generalisable to other settings.”
[1] Galit Shmueli. To Explain or to Predict? .Statistical Science 2010, Vol. 25, No. 3, 289–310

5. Page 8, paragraph 1, line 5: Some discussion should be added around the choice to using simple data transformations instead of using more general methods such as to model non-linear covariate effects.
Response: We found in other study (under revision) that logistic regression performs just as well as splines or fractional polynomials methods. A key reason for this may be that nonlinear and non-additive signals are not strong enough to make splines or fractional polynomials methods advantageous.

6. Page 8, paragraph 3, line 4: Some discussion should be added around the interpretation of scaled Brier score values. Is between 15% and 20% usefully accurate?
Response: Interpretation of the scaled Brier score is similar to R2. We have now added this in the manuscript.

7. Page 8, paragraph 3, line 7: Given Figure S6 it is clear that graphical calibration analysis was also

performed. Please describe here the methods used to generate these figures. The plot for training data prior to re-calibration should also be shown. Consideration should be given to moving these plots to the main manuscript and dropping the HL assessment altogether.
Response: We have now added the calibration plot before baseline risk correction, but the plots remain in the supplementary material as only 5 illustrations are allowed by journal. We have now dropped the HL results and added the description of graphical calibration analysis as follows: Calibration is the relationship between the observed and predicted risk of death and can be usefully seen on a scatter plot (y-axis observed risk, x-axis predicted risk). Perfect predictions should be on the 45° line. The intercept (a) and slope (b) of this line gives an assessment of 'calibration-in-the-large'. At model development, a=0 and b=1, but at validation, calibration-in-the-large problems are indicated if a is not 0 and if b is more/less than 1 as this reflects problems of under/over prediction.

8. Page 9, paragraph 2, line 5: "higher risk" should be 'higher predicted risk".
Response: We have now corrected this.

9. Page 11, top: Given that plots of the validation data were used during model development, the resulting "validation" does not represent true external validation where researchers remain completely blinded to the validation data until the final model developed on the training data has been locked down. This fact needs to be added to study limitations and the term "external validation" used with appropriate qualification.
Response: We have noted this as a limitation in the discussion.
10. Page 12, paragraph 1: I don't think the right question is whether the current model works well when median imputation is used on the samples omitted due to missing data but rather whether a model developed including the imputed data performs better than the current model. Frankly, this is probably more important for smaller datasets.
Response: We are very cautious about imputation in routinely collected data because the underlying causal mechanisms are complex (including test not deemed necessary by clinician, lab report lost in transit, patient did not consent, patient is contraindicated, etc). This is why we set out to develop a risk score for those patients who have the necessary data items based on usual care. The reasons why clinicians do/do not undertake a specific blood test are complex and where the mechanisms for that missingness are not random. We included a simple medium imputation exercise as a preliminary indication of how CARM performs in those patients with some missing blood tests (most frequently albumin). This medium imputation approach is now clinically accepted as part of the NHS AKI algorithm also adopts this approach for creatinine blood test results that are missing.

11. Page 12, paragraph 2: "positive predictive value" should be "positive and negative predictive values". Again since the validation data are used in selecting the cut-off this does not represent true external validation. The cut-off should be based only on the training data and then tested on the validation set.
Response: We have now corrected this. We define cut-offs exclusively based on the development dataset and tested on the validation dataset.

12. Table 1: It is confusing that the percentages for the "N" column are column percentages while the percentages for the "Died" column are row percentages.
Response: We have now corrected this.

13. Table 3: Please change the column header "Discrimination" to "Discrimination Slope" and add a footnote describing how it is estimated.
Response: We have now corrected this and added the description as suggested.

14. Table 4: "PPV" should be "PPV %" and "NPV" should be "NPV %".Consideration should be given to adding a column indicating the proportion of patients with predictions > cut-off.

Response: We have now corrected this.

15. Figure S2 and S3: Outliers should not be omitted from these plots.
Response: We removed outliers to aid the visualisation; otherwise the y-axis scale was stretched to accommodate the most extreme outlier. However, we considered all the data in the statistical models.

16. Figure S4 and S5: These are not scatter plots. They appear to be line plots generated by connecting the proportions dead computed for six equal-sized groups and plotted at the midpoint of the resulting intervals. Given the size of the training data set it would seem that 20 groups would still have enough patients to reliably estimate the proportion died in each interval and give more granular information on the pattern of the covariate effect.
Response: We have now changed the title to line plots but have retained the six group because the graphs are there for exploratory visualisation purposes only, to show the overall relationship between mortality and a given covariate. However our modelling strategy incorporates nonlinearity by transformation the continuous covariates without categorising the data.

17. Figure S6: Please add a superimposed histogram or rug plot to show distribution of predicted probabilities.
Response: We have now added a rug plot as suggested.

18. Figure S7: Please revise figure legend to clarify that these are based on CARM predictions for patients with imputed values that omitted during model development and validation.
Response: We have now made clear and revised the legend of figure S7.

19. The authors should explicitly reference the TRIPOD guidelines and verify for readers that these were followed.
Response: We have now referenced the TRIPOD checklist in the manuscript.

VERSION 2 – REVIEW

REVIEWER	Anthony Bleyer Wake Forest University School of Medicine Winston-Salem, NC, USA
REVIEW RETURNED	09-Jul-2018

GENERAL COMMENTS	Your revisions addressed my concerns.

REVIEWER	Aiman Tulaimat Cook County Health and Hospitals System, Pulmonary and Critical Care Medicine
REVIEW RETURNED	20-Jul-2018

GENERAL COMMENTS	I would like thank the authors for the revisions. I have two asks: 1. I disagree with the authors on the issues of subgroups, seasonality, and time of admission. This should be easy to perform and I think that it will strengthen the conclusions. 2. I am still having trouble with calibration. With my humble understanding, the risk value cut off points in Table 4 represent the predicted risks and the PPV represents the observed risks. That means that for a predicted risk of 8% the observed risk was

	21%. This is probably what figure S7 B shows, that the model underestimates mortality. But then you have the next figure S7 C that shows a calibration slope of 1. I am sure there is a clear explanation that will help me and other curious readers. 3. Again on the topic of calibration. Why does the probably of death end at 30%. Did you not have enough subjects with predicted mortality 80%? This takes me again to the underestimation of mortality.

REVIEWER	Ken Hess UT MD Anderson Cancer Center Houston, TX USA
REVIEW RETURNED	20-Jul-2018

GENERAL COMMENTS	None - authors have adequately addressed my concerns.

VERSION 2 – AUTHOR RESPONSE

We are most grateful for the reviewers' comments which have led to some key changes that have clarified and improved the manuscript.

I would like thank the authors for the revisions. I have two asks:

1. I disagree with the authors on the issues of subgroups, seasonality, and time of admission. This should be easy to perform and I think that it will strengthen the conclusions.

Response: We would like to thank reviewer for pointing this. We have now added the subgroups analysis results such by sex, age, seasonality, longer vs. shorter length of stay admissions, day of the week, and 16 major disease groups based on Charlson comorbidity index (see table S4 in appendix).

2. I am still having trouble with calibration. With my humble understanding, the risk value cut off points in Table 4 represent the predicted risks and the PPV represents the observed risks. That means that for a predicted risk of 8% the observed risk was 21%. This is probably what figure S7 B shows, that the model underestimates mortality. But then you have the next figure S7 C that shows a calibration slope of 1. I am sure there is a clear explanation that will help me and other curious readers.

Response: In general, the positive predictive value (PPV) of any test indicates the likelihood that someone with a positive test result actually has the disease https://www.bmj.com/content/339/bmj.b3835. Predictive values are useful to the clinician as they indicate the likelihood of disease in a patient when the test result is positive (positive predictive value) or the likelihood that the patient does not have the disease when the test result is negative (negative predictive value). The predictive values of a test depend on the sensitivity and the specificity of the test, as well as the prevalence of the disease. The positive predictive value (PPV) is not related to calibration as it does not show the observed risks.

The CARM model was developed using one hospital data (development dataset) and externally validated in another hospital (externally validated dataset). Figure 7 (B) shows poor calibration because we used CARM without any re-calibration (difference in mortality in development and externally validation datasets (5.7% vs. 6.5%)) which also known as calibration-in-the-large problem [1]. When we corrected this baseline differences in mortality, figure 7 (C) shows perfect calibration. We have now added this description in the legend of figure 7.

1 Faisal M, Howes R, Steyerberg EW, et al. Using routine blood test results to predict the risk of death for emergency medical admissions to hospital: an external model validation study. QJM 2017;110:27–31. doi:10.1093/qjmed/hcw110

3. Again on the topic of calibration. Why does the probably of death end at 30%. Did you not have enough subjects with predicted mortality 80%? This takes me again to the underestimation of mortality.

Response: Calibration is the relationship between the observed and predicted risk of death and can be usefully seen on a scatter plot (y-axis observed risk, x-axis predicted risk). Perfect predictions should be on the 45° line. The intercept (a) and slope (b) of this line gives an assessment of 'calibration-in-the-large'. At model development, a=0 and b=1, but at validation, calibration-in-the-large problems are indicated if a is not 0 and if b is more/less than 1 as this reflects problems of under prediction (b>1) or over prediction (b<1). The calibration slope b=1 that shows no problem of under/over estimation of mortality. For visualisation purpose, we just restrict to 30% predicted risk as beyond this point there are few subjects.

VERSION 3 – REVIEW

REVIEWER	Aiman Tulaimat Cook County Health and Hospitals System, Chicago, IL, USA.
REVIEW RETURNED	25-Sep-2018

GENERAL COMMENTS	I am satisfied by the responses of the authors and the edits of the manuscript. I have no more comments. I look forward to see it published.