PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Variation in Model Performance by Data Cleanliness and Classification Methods in the Prediction of 30-day ICU Mortality, a US Nationwide Retrospective Cohort and Simulation Study
AUTHORS	Iwashyna, Theodore; Ma, Cheng; Wang, Xiao Qing; Seelye, Sarah; Zhu, Ji; Waljee, Akbar

VERSION 1 – REVIEW

REVIEWER	Hans Flaatten ICU Dep of Anaesthesia Haukeland University Hospital Bergen, Norway
REVIEW RETURNED	07-Jul-2020

GENERAL COMMENTS	In this study using data from a large clinical database of 355.823 ICU admissions different methods to calculate expected mortality using 3 different ways to vary laboratory values at day 1. They found little differences for three different methods to adjust for missing values (AUC, but suboptimal results when just excluding missing values. It is an interesting, but somewhat difficult topic, that seems to be well performed, although this reviewer has no detailed insight in all methods used. My first question is regarding use of these models. Obviously, this a model that will work in retrospect on cohorts of patients where it is possible to impute values in a sensible way. To use this for prognostication at a lower level would be difficult, so then its clinically relevance is of course disputable (although no-one use such a score alone, but in addition to clinical judgment of patient's prognosis. Please comment. I also find description of the main aim of the study to differ w/r to what is described in the method section. In the introduction (page 4 line 41-43) you claim to study prediction of 30-day mortality from day-1 laboratory values. In the method (page 6, lie 9-11) you describe use of patient demographic, laboratory values and diagnosis to predict 30-day mortality. Please comment and change accordingly (you have as I see used more than laboratory values alone, which is obvious from the section Predictor variables on page 6). On page 6 you describe what I see as the main message of this study, the six different methods you have used to impute values. (This is probably also the most important findings). Hence I would

	suggest that you rewrite the introduction and focus more on variation of imputation of missing data For the un-informed readers, I would also explain better the curves produced with the partial dependence plots since they are not immediately intuitive. From table 2 the proportion of missing laboratory values are displayed. I am surprised with the vey high number of missing values in dataset A (BUN 84%) but falling to 13% for the same variable using dataset B and C (BUN 13%), the same pattern with Hematocrit. Could you explain these huge differences for the reader? The figures are a bit difficult to read since line-numbers and text is superimposed on axis text, but this is probably not the authors fault. Last, in a manuscript dealing with prognosis, I would have liked to see a discussion about using fix time data (like laboratory values at Day-1), compared to dynamic values; values that changes over time. Temperature is often included in prognostic score, but iy see a temperature curve of one patient over time it varies a lot and give other signals to its use as a prognostic factor. The same can be said about laboratory values as well.

REVIEWER	Ben Gibbison University of Bristol. UK
REVIEW RETURNED	08-Jul-2020

GENERAL COMMENTS	The manuscript is useful and on the surface (I am not a statistician) appears well thought through and conducted. It suffers somewhat from imprecision in language and unclear writing that make the reader have to "work hard" to understand what the authors are saying. P5 L 11. Readers come from outside the USA. What are: "US News and World Report to Medicare and Medicaid" P11/12 They say that both datasets C and D are "the cleanest" in separate paragraphs. They should be clear what they are talking about - or is this just a "typo"? P12 L9: Should be "values" (pl.) P12 L53: should read "..missing data leads to a..." P13 L12: I do not understand "use-case-specific" There are a number of non-standard words in the discussion e.g. "canonical" and "desiderata" that do not help understanding and should be replaced with clear English. P14 L 56: The sentence beginning "Thus we considered only one outcome...." needs revising. Again, I don't understand what it is saying. The discussion needs more clarity and the evidence that supports their findings and those which it is contract to needs to be more clear. P13 L43: The APACHE was not designed to predict things"by hand"... The conclusion is somewhat nebulous and sounds a bit like the "more research is needed...." P15 L13: They mean "different" not "alternative"

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

Reviewer Name: Hans Flaatten
Institution and Country: ICU Dep of Anaesthesia, Haukeland University Hospital, Bergen, Norway
Please state any competing interests or state 'None declared': None declared

In this study using data from a large clinical database of 355.823 ICU admissions different methods to calculate expected mortality using 3 different ways to vary laboratory values at day 1. They found little differences for three different methods to adjust for missing values (AUC, but suboptimal results when just excluding missing values.

It is an interesting, but somewhat difficult topic, that seems to be well performed, although this reviewer has no detailed insight in all methods used.

1. My first question is regarding use of these models. Obviously, this a model that will work in retrospect on cohorts of patients where it is possible to impute values in a sensible way. To use this for prognostication at a lower level would be difficult, so then its clinically relevance is of course disputable (although no-one use such a score alone, but in addition to clinical judgment of patient's prognosis. Please comment.

This is a great point. The approach to missing data in prospective application requires careful thought, but it is not impossible. Fundamentally, most imputation approaches take some form of "development" data and use it to develop statistical rules (or, more crudely, clinical rules like "if missing impute normal"), and then apply them to the parts of the data where missingness occurs. One could do similarly in prospective application—develop the missingness imputation rules in the development data set and apply them at the bedside prior to applying the formal "risk prediction" rule. We agree completely with the reviewer's point that no-one uses such a score alone without supplementary clinical judgment, or at least we hope this will remain true!

2. I also find description of the main aim of the study to differ w/r to what is described in the method section. In the introduction (page 4 line 41-43) you claim to study prediction of 30-day mortality from day-1 laboratory values. In the method (page 6, lie 9-11) you describe use of patient demographic, laboratory values and diagnosis to predict 30-day mortality. Please comment and change accordingly (you have as I see used more than laboratory values alone, which is obvious from the section Predictor variables on page 6).

We corrected the introduction to be consistent with the variables described in the methods section. On page 4 line 41-43 and page 6 line 9-11, we also added "comorbidities" as variables included in the prediction of 30-day mortality to be consistent with our description in Predictor Variables on page 6.

3. On page 6 you describe what I see as the main message of this study, the six different methods you have used to impute values. (This is probably also the most important findings). Hence I would suggest that you rewrite the introduction and focus more on variation of imputation of missing data

We believe both the imputation and prediction approaches were of co-equal importance, and we have modified the introduction to reflect this.

4. For the un-informed readers, I would also explain better the curves produced with the partial dependence plots since they are not immediately intuitive.

We have provided additional clarification for the partial dependence plots in the results section.

5. From table 2 the proportion of missing laboratory values are displayed. I am surprised with the vey high number of missing values in dataset A (BUN 84%) but falling to 13% for the same variable using dataset B and C (BUN 13%), the same pattern with Hematocrit. Could you explain these huge differences for the reader?

Thank you for your careful attention to these differences between datasets. We have added an explanation for the large number of missing values in dataset A for BUN and hematocrit compared to datasets B and C.

6. The figures are a bit difficult to read since line-numbers and text is superimposed on axis text, but this is probably not the authors fault.

We have provided updated figures that, we hope, improve readability.
7. Last, in a manuscript dealing with prognosis, I would have liked to see a discussion about using fix time data (like laboratory values at Day-1), compared to dynamic values; values that changes over time. Temperature is often included in prognostic score, but iy see a temperature curve of one patient over time it varies a lot and give other signals to its use as a prognostic factor. The same can be said about laboratory values as well.

We agree this is a very interesting additional project – but it exceeds the scope of the present manuscript and we do not, unfortunately, have those data clean and readily available. We have noted this as a limitation.

Reviewer: 2
Reviewer Name: Ben Gibbison
Institution and Country: University of Bristol. UK Please state any competing interests or state ‘None declared’: None declared

1. The manuscript is useful and on the surface (I am not a statistician) appears well thought through and conducted. It suffers somewhat from imprecision in language and unclear writing that make the reader have to "work hard" to understand what the authors are saying.

We revised the manuscript throughout for greater clarity and precision.

2. P5 L 11. Readers come from outside the USA. What are: "US News and World Report to Medicare and Medicaid"

We deleted this from the manuscript to suit a broader audience.

3. P11/12 They say that both datasets C and D are "the cleanest" in separate paragraphs. They should be clear what they are talking about - or is this just a "typo"?

Thank you for your comment. On page 11, we refer to dataset D as being the cleanest dataset among the four datasets used in this study. On page 12, we compare dataset C to datasets A and B. Among these three datasets, dataset C is the cleanest. We revised the wording on page 12 to clarify the comparator datasets.

4. P12 L9: Should be "values" (pl.)

We corrected this typo, changing “value” to “values” in both cases.

5. P12 L53: should read "..missing data leads to a..."

Thank you. We made this correction.

6. P13 L12: I do not understand "use-case-specific"

We edited this sentence for clarity.

7. There are a number of non-standard words in the discussion e.g. "canonical" and "desiderata" that do not help understanding and should be replaced with clear English.

We revised the manuscript throughout and replaced terms such as “canonical” and “desiderata” with more common words.

8. P14 L 56: The sentence beginning "Thus we considered only one outcome...." needs revising. Again, I don't understand what it is saying.

Thank you – we revised this sentence.

9. The discussion needs more clarity and the evidence that supports their findings and those which it is contract to needs to be more clear.

We have rewritten several aspects of the “Relationship to Past Literature” to be clearer about the which findings our results support and contradict.

10. P13 L43: The APACHE was not designed to predict things"by hand"...

We deleted this phrase.

11. The conclusion is somewhat nebulous and sounds a bit like the "more research is needed...."

We do, of course, believe more research is needed before the full generality of our findings can be understood, but we have modified our conclusion to be more clear about the implications of this work.

12. P15 L13: They mean "different" not "alternative"

We corrected the sentence using “different.”

VERSION 2 – REVIEW

REVIEWER	Hans Flaatten De. of Intensive Care Haukeland University Hospital Norway
REVIEW RETURNED	09-Sep-2020

GENERAL COMMENTS	The authors have responded to my satisfaction, and I have no further comments.

REVIEWER	Ben Gibbison University of Bristol, UK
REVIEW RETURNED	27-Aug-2020

GENERAL COMMENTS	P4 L8 - This sentence makes no sense and needs revising. One standardized what? P5 L37 The word "on" needs separating from the next word P13 L16 I have no idea what this sentence means "the extent to which laboratory ascension and labeling practices coincide with their aspiration varies over time".

REVIEWER	AKANSHA SINGH Durham University, United Kingdom
REVIEW RETURNED	21-Oct-2020

GENERAL COMMENTS	The paper is well written and has highlighted key differences in three statistical methods (logistic, neural network, random forest) while handling missing data. The authors can extend this analysis in the future by increasing the proportion of missing to a very large extent, impute with these prescribed methods, and see whether random forest or neural network are still performing better. Authors can also assess the impact of missing data using deep learning algorithms such as the deep neural network (DNN); convolution neural network (CNN) in those scenarios.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1
Reviewer Name: Hans Flaatten
Institution and Country: ICU Dep of Anaesthesia, Haukeland University Hospital, Bergen, Norway
Please state any competing interests or state ‘None declared’: None declared

The authors have responded to my satisfaction, and I have no further comments.

Thank you for reviewing our manuscript and the revisions we submitted.

Reviewer 2
Reviewer Name: Ben Gibbison
Institution and Country: University of Bristol. UK
Please state any competing interests or state ‘None declared’: None declared

P4 L8 - This sentence makes no sense and needs revising. One standardized what?

Thank you - we have revised this sentence.

P5 L37 The word "on" needs separating from the next word

We made this correction.

P13 L16 I have no idea what this sentence means "the extent to which laboratory ascension and labeling practices coincide with their aspiration varies over time".

Thank you - we revised this sentence.

Reviewer 3 (stats review)
Reviewer Name: AKANSHA SINGH
Institution and Country: Durham University, United Kingdom
Please state any competing interests or state ‘None declared’: None

The paper is well written and has highlighted key differences in three statistical methods (logistic, neural network, random forest) while handling missing data. The authors can extend this analysis in the future by increasing the proportion of missing to a very large extent, impute with these prescribed methods, and see whether random forest or neural network are still performing better. Authors can also assess the impact of missing data using deep learning algorithms such as the deep neural network (DNN); convolution neural network (CNN) in those scenarios.

Thank you for reviewing the manuscript and providing feedback. We appreciate your comments and hope to include these approaches in future studies.