PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Prediction Model of In-hospital Mortality in Intensive Care Unit Patients with Heart Failure Machine learning-based, retrospective analysis of the MIMIC-III database
AUTHORS	Li, Fuhai; Song, Yu; Fu, Mingqiang; Han, Xueting; zhou, Jingmin; Ge, Junbo

VERSION 1 – REVIEW

REVIEWER	Joseph Rigdon Wake Forest School of Medicine, United States
REVIEW RETURNED	14-Nov-2020

GENERAL COMMENTS	I congratulate the Authors on great work. My comments to follow are mostly stylistic. 1. Abstract: some abbreviations (e.g., HF and GWTG-HF) – do we need to spell them out somewhere? 2. Abstract, Methods: What % of the sample was used for model derivation, and what % for model validation? 3. Abstract, Methods: perhaps clarify that XGBoost and LASSO were used for variable selection, but then the selected variables were entered into a logistic regression model, which was then validated, and also turned into a nomogram for clinical use 4. Strengths and limitations: Perhaps rephrase "We developed the first in-hospital mortality prediction nomogram selected by XGBoost model using logistic regression analysis" to "We developed the first in-hospital mortality prediction nomogram using logistic regression with included variables selected by the XGBoost algorithm". 5. Introduction, page 4, line 25: What limitations do current in-hospital mortality algorithms have in clinical practice? 6. Introduction, page 4, lines 48-57: Perhaps clarify that XGBoost and LASSO were used to select variables, which were then entered into a logistic regression model. 7. Methods, page 5, Study patients, line 40: Perhaps add the specific ICD-9 codes to the text that indicate a diagnosis of HF? 8. Methods, page 9, end of page: Is the U statistic approach the same as the ROC test outlined in this paper? http://europepmc.org/abstract/med/3203132 If so, I suggest to cite the paper. 9. Results, page 11, selected variables: I'm curious, when using XGBoost, did you experiment with hyperparameter tuning at all? Also, how did you select the lambda for LASSO? Was it the value that minimized mean cross-validated error? 10. Results, page 15, first line: There was no statistically significant difference between … (add the word difference). 11. Discussion, page 16, last line: Change to "To avoid

	shortcomings such as over-fitting, and predictor variables with skewed distributions, ” 12. Conclusion, first sentence: Capitalize We 13. I might suggest the TRIPOD checklist (https://www.ahajournals.org/doi/full/10.1161/circulationaha.114.014508), rather than (or in addition to) the STROBE checklist.

REVIEWER	Brent Richards Gold Coast Hospital and Health Service Australia
REVIEW RETURNED	30-Nov-2020

GENERAL COMMENTS	Thank you for the opportunity to review this paper. The subject chosen is worth reviewing, and the methodologies chosen reasonable. However there are a number of opportunities to improve this manuscript. The discussion around the variance in hospital mortality for ICU vs general patients requires clearer explanation – namely that it is due to underlying severity of illness, as well as patient selection. This latter bias in inherent to any ICU-based prediction scores. The MIMIC database is spread across a number of years (2001 – 2012), which both needs noting, and needs to be recognized as an unmeasured confounder. Treatment of heart failure did change over that period. However dates in MIMIC-III have been scrambled for privacy preservation, and thus any mortality changes across years will be obscured. The choice of exclusion criteria, which are quite strict, are not otherwise discussed. Although they appear somewhat reasonable, the total number of patients left for analysis are quite small, thus making conclusions more tenuous than is presented. The somewhat unexpected and counter-intuitive finding of LVEF not being a predictor may in part be due to the smaller numbers. I would also note that LVEF needs further discussion, as it is a predictor in other papers. A similar discussion would also be relevant for atrial fibrillation. Some other parameters than need further thought and discussion include: whether heart rate as a predictor was simply due to atrial fibrillation, or separate; whether BMI was predictive (as compared to height and weight separately), and the role of calcium (in part whether this was due to the ion itself, or was a surrogate for carrier proteins e.g. albumin. Key to managing Health data is attention to missing data imputation. Unfortunately this has not been done well here as described. The mean (or median) is used, however it is not clear whether this is for each patient, the patient cohort, or the MIMIC dataset overall. No consideration or discussion is given to other simple techniques e.g. LOCF or NOCB, nor current more contemporary techniques, e.g. MICE or cluster analysis. Although using these techniques will require re-analysis, as this work is code-based, this is achievable. MIMIC-III is a single centre dataset, and thus comes with the limitations of this. As such the transferability of the techniques and findings are clearly limited, and this needs to be recognized in the discussion. Utilising a nomogram to assist with interpretation is a good step to help clinicians – noting that on a single dataset across more than a decade, and a relatively small patient number, it presents an option for further exploration rather than a definitive answer in itself. Research reproducibility is critical, and data science research on fixed data sets is the clear place where this can and should occur. In addition, the data agreement for using the MIMIC-III database

	requires researchers to publish code along with their paper. However, no link is provided in the paper as yet. This will need to be provided, both as part of the data use agreement and to assist future researchers to build on this work.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1
1. Abstract: some abbreviations (e.g., HF and GWTG-HF) – do we need to spell them out somewhere?

Response: Thanks for the reviewer's reminding. In the revised version, the abbreviations in the abstract were spelled out in line27, line 31-32, line 37-38, and line 45, where the changes were marked in red.

2. Abstract, Methods: What % of the sample was used for model derivation, and what % for model validation?

Response: We have listed the number and percentage of derivation and validation group respectively in lines 35-36.

3. Perhaps clarify that XGBoost and LASSO were used for variable selection, but then the selected variables were entered into a logistic regression model, which was then validated and also turned into a nomogram for clinical use.

Response：Yes, we entirely agree with the reviewer. This is exactly the process we developed our model. We modified our expression to make it clearer in line 39-40.

4. Strengths and limitations: Perhaps rephrase "We developed the first in-hospital mortality prediction nomogram selected by XGBoost model using logistic regression analysis" to "We developed the first in-hospital mortality prediction nomogram using logistic regression with included variables selected by the XGBoost algorithm."

Response：Thanks for the reviewer's valuable suggestion. I modified our sentence in the revised version in line 55-56.

5. Introduction, page 4, line 25: What limitations do current in-hospital mortality algorithms have in clinical practice?

Response：We thank the Reviewer for this insightful question. Among all, the accuracies of these methods are unsatisfactory is the most important reason that limitations current in-hospital mortality algorithms have in clinical practice. We have listed in the in the revised version in line 82-83.
6. Introduction, page 4, lines 48-57: Perhaps clarify that XGBoost and LASSO were used to select variables, which were then entered into a logistic regression model.

Response：We thank the reviewer' excellent suggestion. In the revised version, we clarified this in line 94-96.

7. Methods, page 5, Study patients, line 40: Perhaps add the specific ICD-9 codes to the text that indicate a diagnosis of HF?

Response：The specific ICD-9 codes used in our study are as follow:428.0 Congestive heart failure, unspecified; 428.1 Left heart failure; 428.9 Heart failure, unspecified; 428.2 Systolic heart failure; 428.3 Diastolic heart failure;402 Hypertensive heart disease; 428 Heart failure; 404 Hypertensive heart and chronic kidney disease;398.91 Rheumatic heart failure (congestive).

8. Methods, page 9, end of page: Is the U statistic approach the same as the ROC test outlined in this paper? http://europepmc.org/abstract/med/3203132 If so, I suggest to cite the paper.

Response：We thank the reviewer for the suggestions. We cited this paper in the revised version(line 207).

9. Results, page 11, selected variables: I'm curious, when using XGBoost, did you experiment with hyperparameter tuning at all? Also, how did you select the lambda for LASSO? Was it the value that minimized mean cross-validated error?

Response：It is an honor to receive this comment.In our study, the XGBoost algorithm was conducted by R package Coxboost.Parameters tuned by optimCoxBoostPenalty.

Tuning parameter (λ) selection in the LASSO model used 10-fold cross-validation via minimum criteria. The partial likelihood deviance (binomial deviance) curve was plotted versus log(λ). Dotted vertical lines were drawn at the optimal values by using the minimum criteria and the one standard error of the minimum criteria (the 1-SE criteria).

10. Results, page 15, first line: There was no statistically significant difference between … (add the word difference).

Response：We apologize for our carelessness. In the revised version, we added the word "difference" in line 317.

11.Discussion, page 16, last line: Change to "To avoid shortcomings such as over-fitting, and predictor variables with skewed distributions, "

Response：We appreciate the reviewer on these excellent suggestion, which is of help for us to improve the quality of our paper. In the revised version, we modified the sentence(line 359-360).
12. Conclusion, first sentence: Capitalize We

Response：We apologize for our carelessness. In the revised version, we capitalized "we" in line 412.

13. I might suggest the TRIPOD checklist (https://www.ahajournals.org/doi/full/10.1161/circulationaha.114.014508), rather than (or in addition to) the STROBE checklist.

Response：We thank reviewer for the suggestions.we checked our manuscript according to the TRIPOD.

Reviewer: 2
1、 The discussion around the variance in hospital mortality for ICU vs general patients requires clearer explanation – namely that it is due to underlying severity of illness, as well as patient selection. This latter bias in inherent to any ICU-based prediction scores.

Response：We thank the Reviewer for this valuable comment. We agree with the Reviewer that the variance in hospital mortality for ICU vs general patients requires clearer explanation, and the ICU patient selection may bias our prediction scores. We discussed this in line 358- 359 and 399-402.

2、 The MIMIC database is spread across a number of years (2001 – 2012), which both needs noting, and needs to be recognized as an unmeasured confounder. Treatment of heart failure did change over that period. However dates in MIMIC-III have been scrambled for privacy preservation, and thus any mortality changes across years will be obscured.

Response：We thank the reviewer for the suggestions. Indeed, from 2001 to 2012 is really a long time, during which the treatment of heart failure had significantly changed, which may weaken the application of our model. We discussed this comment in the limitation part(line 402-404).

3、 The somewhat unexpected and counter-intuitive finding of LVEF not being a predictor may in part be due to the smaller numbers. I would also note that LVEF needs further discussion, as it is a predictor in other papers. A similar discussion would also be relevant for atrial fibrillation.

Response：A lot of variables were reported to correlate with mortality in heart failure patients, such as gender, age, BMI, smoking, LVEF, NYHA classification, diabetes mellitus, chronic obstructive lung disease, low systolic blood pressure, serum creatinine levels, not receiving beta-blockers, and not receiving ACEIs/ARBs1.Whereas, our study showed that LVEF was not a predictor of in-hospital mortality of ICU-admitted HF patients. This was also observed in previous small number HF cohorts2-5. This may be partly attributed to our relatively small sample size of participants and the duration of hospitalization is a relatively shorter time(line368-375).

Our study found that the HF patient who had comorbidity of atrial fibrillation had higher in-hospital mortality. Whereas atrial fibrillation is not an independent impact factor that strong enough to affects the outcome after adjusting for other covariates. This is consistent with the previous reports6-8(line375-378).

4、 Some other parameters than need further thought and discussion include: whether heart rate as a predictor was simply due to atrial fibrillation, or separate; whether BMI was predictive (as compared to height and weight separately), and the role of calcium (in part whether this was due to the ion itself, or was a surrogate for carrier proteins e.g. albumin.

Response：We thank the Reviewer for these insightful suggestions. Whether heart rate was a predictor, the heart failure mortality prediction models showed differently. Some models believe that it affects prognosis strongly8, 9 and some models disagree6, 10, 11. In our model, the heart rate did no be recruited into the final model. BMI also encountered the same situation. Our study failed to prove that BMI was a predictor of in-hospital mortality of ICU-admitted HF patients. This may be due to the different populations studied and our relatively small sample size. Whether “the obesity paradox” bias our result is hard to say, because critical care-related outcome12 and heart failure 13both have “the obesity paradox” respectively. We would therefore like to explore these confusions in future studies. Both hypocalcemia and hypercalcemia were reported to associated with an increased short-term mortality risk in heart failure patient14. Our study showed that hypercalcemia indicated an adverse outcome. Free serum calcium ions, a very important electrolyte, plays a major role in excitation, contraction, and relaxation coupling of the myocardium. The alterations of serum calcium homeostasis may adversely affect the prognosis of heart failure patients. Besides, the amount of calcium-binding proteins are significantly altered altered in end-stage heart failure15(line379-line393).

5、 Key to managing Health data is attention to missing data imputation. Unfortunately this has not been done well here as described. The mean (or median) is used, however it is not clear whether this is for each patient, the patient cohort, or the MIMIC dataset overall. No consideration or discussion is given to other simple techniques e.g. LOCF or NOCB, nor current more contemporary techniques, e.g. MICE or cluster analysis. Although using these techniques will require re-analysis, as this work is code-based, this is achievable.

Response：We thank the Reviewer for this valuable comment. We apologize for the oversight. A feature of our data is a relatively small sample size and a relatively large number of variables.The number of missing values in our data(<25%) is not large and the missing pattern is missing at random.In this situation,single imputation with mean or median did not introduce much bias16.

6、 MIMIC-III is a single centre dataset, and thus comes with the limitations of this. As such the transferability of the techniques and findings are clearly limited, and this needs to be recognized in the discussion. Utilising a nomogram to assist with interpretation is a good step to help clinicians – noting that on a single dataset across more than a decade, and a relatively small patient number, it presents an option for further exploration rather than a definitive answer in itself.

Response：We thank the Reviewer for this valuable comment.The limitations that the Reviewer pointed out including a single centre dataset, a relatively small patient number did exist in our study. We discussed this comment in the limitation part(line 405-409,413-414).

7、 Research reproducibility is critical, and data science research on fixed data sets is the clear place where this can and should occur. In addition, the data agreement for using the MIMIC-III database requires researchers to publish code along with their paper. However, no link is provided in the paper as yet. This will need to be provided, both as part of the data use agreement and to assist future researchers to build on this work.

Response：Thanks for the reviewer’s reminding. We have uploaded the code used in our study in the supplementary.

VERSION 2 – REVIEW

REVIEWER	Joseph Rigdon Wake Forest School of Medicine, United States
REVIEW RETURNED	15-Feb-2021

GENERAL COMMENTS	1. Abstract, line 47: not clear what is meant by “prediction effectiveness”? C-statistic? Accuracy? Same question throughout the manuscript when this term is mentioned. 2. Introduction, line 81: what are the range of accuracies of current methods? 3. Introduction, lines 90-95: save this content for the methods? 4. Methods, line 185: which lasso model was to be selected? The one that minimizes the misclassification error? 5. Methods, line 199: citation for decision curve analysis? 6. Tables 2 and 3: why are there NA values in the multivariable regression model results? Were these variables with NA values included in the final models for creation of the nomogram? 7. Methods, model comparison section (lines 309-335): why refit the models on the total sample? Shouldn’t we just report the findings on the validation samples, as these results are likely most representative of what we will see in practice?

	8. Discussion, line 384: change to “heart rate did not appear in the final model”

REVIEWER	Brent Richards Gold Coast Hospital and Health Service Australia
REVIEW RETURNED	02-Feb-2021

GENERAL COMMENTS	Thank you for your resubmission, and thoughtful responses. The paper as presented is substantially improved. I would note that you have addressed a number of my comments directly back to me, with these not clearly reflected in the final submitted paper. It is important the journal reader is made aware of your detailed and deep understanding and reflections, not just me. As such I’ll revisit these with some suggestions. Both referring clinicians and Intensive Care clinicians often select those patients most likely to benefit from ICU therapy for admission to ICU, rather than admitting all very ill heart failure patients. This creates inherent bias when comparing ICU mortality and hospital mortality. Therefore a simple statement addressing this may be useful, such as ‘A decision to admit to intensive care varies depending on both clinician expectations and resource availability, so both factors will add unmeasured variance to outcome studies’. Lines 402-404 do not clearly address the date range potential outcome variance for MIMIC. I’d suggest adding to line 409 a statement such as ‘Given the 12 year date range of MIMIC-III, there may also be unmeasured outcome variance over time’. I’d also note this additionally means that any algorithm from MIMIC would not necessarily reflect 2021 best practice. For the missing data, a slightly more detailed statement regarding this is needed. Is the imputed mean value of the overall patient group, or for that individual patient? It would therefore read ‘ …..missing values were replaced with the mean for the patient group’. I still do not appear to have access to a supplement that includes the SQL and (presumedly) python code used for extraction and analysis. This would be of great value to other researchers, and is key to reproducibility. I would like to see this is present prior to publication. Thank you again for your time and efforts.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 2

1.Both referring clinicians and Intensive Care clinicians often select those patients most likely to benefit from ICU therapy for admission to ICU, rather than admitting all very ill heart failure patients. This creates inherent bias when comparing ICU mortality and hospital mortality. Therefore a simple statement addressing this may be useful, such as ‘A decision to admit to intensive care varies depending on both clinician expectations and resource availability, so both factors will add unmeasured variance to outcome studies.’

Response:

We appreciate the reviewer on these excellent suggestions. We add this statement in lines 361-363 in the revised version（marked copy）.

2.Lines 402-404 do not clearly address the date range potential outcome variance for MIMIC. I'd suggest adding to line 409 a statement such as 'Given the 12-year date range of MIMIC-III, there may also be unmeasured outcome variance over time'. I'd also note this additionally means that any algorithm from MIMIC would not necessarily reflect 2021 best practice.

Response:

We thank the reviewer for this valuable comment. We add this statement in lines 421-422 in the revised version（marked copy）.

3.For the missing data, a slightly more detailed statement regarding this is needed. Is the imputed mean value of the overall patient group, or for that individual patient? It would therefore read ' …..missing values were replaced with the mean for the patient group'.

Response:

We thank the reviewer for the suggestions. We add this sentence in line 195 in the revised version（marked copy）.

4.I still do not appear to have access to a supplement that includes the SQL and (presumedly) python code used for extraction and analysis. This would be of great value to other researchers, and is key to reproducibility. I would like to see this is present prior to publication.

Response:

We apologize for this. But we have upload the code in the supplement material in the last version.We will upload the code again in the new revised version in the supplement material.

Reviewer: 1

1.Abstract, line 47: not clear what is meant by "prediction effectiveness"? C-statistic? Accuracy? Same question throughout the manuscript when this term is mentioned.

Response:

We apologize for this confusion. Prediction effectiveness includes discrimination and calibration. Discrimination was assessed by calculating the area under the curve (AUC) of the receiver operating characteristic (ROC) curve and C-statistic testing. Calibration curves were plotted to assess the calibration of the in-hospital mortality nomogram. The 95% confidence interval (CI) was calculated using 500 bootstrap resamples.

This has been clarified in the "Materials and methods" part in lines 203-208 in the revised version（marked copy）.

2.Introduction, line 81: what is the range of accuracies of current methods?

Response:

Thanks for the reviewer's reminding. We have listed the range of current methods' accuracy in lines 90-91 in the revised version（marked copy）.

3.Introduction, lines 90-95: save this content for the methods?

Response:

We appreciate the reviewer on this excellent suggestion. We have saved this content for the methods in the revised version.

4.Methods, line 185: which lasso model was to be selected? The one that minimizes the misclassification error?

Response:

We apologize for this confusion. This sentence means variables in the LASSO regression model were selected，does not mean the lasso model was selected.

5.Methods, line 199: citation for decision curve analysis?

Response:

Thanks for the reviewer's reminding. We have added the citation for decision curve analysis in the revised version in line 208（marked copy）.

6.Tables 2 and 3: why are there NA values in the multivariable regression model results? Were these variables with NA values included in the final models for creation of the nomogram?

Response:

We apologize for this confusion. NA means this variable was not identified as a significant mortality risk predictor by multivariate logistic regression.

No, these variables with NA values in table 2 did not been included in the final models for creation of the nomogram.

7.Methods, model comparison section (lines 309-335): why refit the models on the total sample? Shouldn't we just report the findings on the validation samples, as these results are likely most representative of what we will see in practice?

Response:

We thank the reviewer for this valuable suggestion. Just report the findings on the validation samples may be a good choice, whereas，refit the models in larger sample sizes may be better.

8.Discussion, line 384: change to "heart rate did not appear in the final model"

Response:

We thank the reviewer for this valuable suggestion. We have rewritten this sentence as the reviewer suggested in line 395 in the revised version（marked copy）.

VERSION 3 – REVIEW

REVIEWER	Joseph Rigdon Wake Forest School of Medicine, United States
REVIEW RETURNED	17-Mar-2021

GENERAL COMMENTS	1. Abstract, Results: report C-statistic to show benefit of new models vs. old? 2. Methods, lines 155-161: any references cited for choice of imputation? 3. Table 5: What data set is being used to evaluate the methods

	here? The combined data set, or just the validation set. I would recommend just the validation set. Also, why are the performance statistics (e.g., ROC area) changing from one comparison to the next? 4. Minor comment – recommend using “multivariable model” rather than “multivariate model” throughout (see this citation: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2012.300897)

VERSION 3 – AUTHOR RESPONSE

Reviewer: 1
1. Abstract, Results: report C-statistic to show benefit of new models vs. old?
Response:
We appreciate the reviewer on these excellent suggestions. The prediction effectiveness that we used to compare the new and old models includes discrimination and calibration. C-statistic is only one of the parameters. So we think reporting C-statistic here does not
appropriate.
2. Methods, lines 155-161: any references cited for choice of imputation?
Response:
We thank the reviewer for this valuable comment. We add the references in line 159 in the revised version(marked copy).
3. Table 5: What data set is being used to evaluate the methods here? The combined data set, or just the validation set. I would recommend just the validation set. Also, why are the performance statistics (e.g., ROC area) changing from one comparison to the next?
Response:
We thank the reviewer for the suggestions. In the present study, we used the combined data set to evaluate the methods. When compared with different models, the EmpowerStats, software we used in this study, automatically fits the best curve. So every time the ROC area is slightly different.
4. Minor comment – recommend using “multivariable model” rather than “multivariate model” throughout (see this citation: https://ajph.aphapublications.org/doi/full/10.2105/AJPH.2012.300897)
Response:
We appreciate the reviewer on this excellent suggestion. We apologize for this. We corrected this error in the revised version in lines 46,265, and 402(marked copy).

VERSION 4 – REVIEW

REVIEWER	Joseph Rigdon Wake Forest School of Medicine, United States
REVIEW RETURNED	15-Apr-2021

GENERAL COMMENTS	Great work! I have no further comments.