PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Severe Acute Kidney Injury Predicting Model based on transcontinental databases: a single-center prospective study
AUTHORS	Liang, Qiqiang; Xu, Yongfeng; Zhou, Yu; Chen, Xinyi; Chen, Juan; Huang, Man

VERSION 1 – REVIEW

REVIEWER	Joseph R. Ledsam Google Inc
REVIEW RETURNED	06-Aug-2021

GENERAL COMMENTS	This paper addresses an important topic, that many AKI models lack external/prospective validation. The authors develop and evaluate a model for AKI prediction in ICU setting. The model predicts severe AKI (defined as AKI II and III) in next 48 hours. The authors report AUCs of 0.86 for validation and 0.84 for the prospective period. While the topic is important, there are several concerns. More information is needed on the methodology, and more information is needed to justify some of the research decisions. This is especially important in the discussion of deleting patients/missing data in the prospective period, which needs to be much more clearly described and justified. Secondly, the ratio of true and false positives should be described. Finally, the code should be made available for review. Major comments 1. Exclusion criteria Why did the authors exclude AKI 1? This creates the flaw that the model may often be predicting the progression from AKI 1 to more severe forms of AKI. That is a much easier, and very different, task to many of the models the authors cite. As many patients in ICU would be expected to progress from AKI 1 to 3 due to the severity of their conditions, it seems like this is an easier task. The authors should report the performance in patients who had no AKI separately to the patients that had

	AKI 1 at the time of prediction. When patients using RRT within 48 hours of admission, how many patients who went from normal or AKI to severe requiring RRT were excluded? Why were patients in ICU for fewer than 3 days excluded? This seems like the group of patients in ICU who are least sick, and in whom it is important to ensure the model is not making too many false positives. 2. More methodological detail needed I understand how the data were chosen to include, but how did the model cope with missing information for some patients? More information is needed here. Inevitably data will be missing for patients as not all patients have all labs recorded. How frequent were predictions? I think the discussion mentions daily - why were more frequent predictions not made? How was data grouped? Did the model have access to data for only previous days (i.e. model that predicts on day t, only sees data up to and including day t-1) or would that model have seen data from day t as well? The the authors say “Comorbidity included hypertension, diabetes, cardiopathy, liver disease, and malignant tumors” are these just examples, or is this an exhaustive list? If an exhaustive list, why were these chosen and not others that may be risk factors for AKI? “unifying the unit and diagnostic codes” can the authors say more about how they did this? “We deleted more than 50% of the missing values and replaced the remaining missing values with multiple interpolations” What do the authors mean by deleting >50% of the missing values? Why were these deleted? “There were more negative data than positive data, so we randomly sampled the negative datasets and constructed a new data subset with a sampling ratio of positive and negative data of 1:5.” Why was this needed? Shouldn’t the models be trained on representative data? Did the authors check if any patients represented in the internal validation dataset were also included in the prospective validation? What would the results have been if these were excluded? 3. The description of deleting patients / missing data is confusing “We sampled the 20% predicted data every month and deleted samples with more than 50% missing values to ensure data correctness. The criteria to terminate prediction were A. a positive diagnosis; B. Transfer out of ICU or death with negative diagnosis.” This is confusing. It is not clear to me why examples were

	deleted, and why discharge or a positive diagnosis would result in deletion. This should be much better described, as it seems like it reduces the clinical applicability of the results: Why was it necessary to delete patients? What was this criteria based on? What would the results have been without this deletion? It's it particularly important to predict accurately on patients who died, as those deaths may be preventable? And if the authors mean a positive diagnosis for AKI, isn't that what this study cares about? Why were only 20% sampled? How did the authors arrive at this percentage? 4. The ratio of false positives to true positives should be described In the discussion, the authors mention this about a previous study: "A total of 55.8% of severe AKI patients were predicted within the first 48 hours, although each accurate prediction was accompanied by two mispredictions". But I can't seem to find the same number for the present study from the authors. How many false positives were predicted for every true positive? This figure can help understand how clinically useful the model is. 5. Is the code available or not? "The model code can be obtained by email if readers need it, but we cannot guarantee that all the code will be provided" Is the code available or not? This is a very unusual code availability statement. If the code is not available, why not? If it is, please make it publicly available. Can the code be provided for review? Minor comments The abstract says 358 positive results predicted and 344 patients diagnosed with Severe AKI, but doesn't say what the crossover is in the abstract. Better to instead say sensitivity, specificity, PPV, NPV, and AUROC. The accuracy is also not needed. Severe AKI in 9% of patients - this seems low, given that the prospective data collection occurred during 2020, when one would expect ICUs to have a large number of COVID patients, up to 50% of which develop AKI in hospital, and perhaps more in ICU. Why was the incidence much lower? In the discussion authors suggest their visualization graphic is a benefit over a previous work that compared performance with clinicians. Can the authors say more about the visualisation tool? How was it developed? Is there any evidence it improves predictive performance of humans? In Figure S3, why is sepsis also being flagged? Does a similar model also flag sepsis and other scores that was not included in this manuscript?

	"Marine F" in the discussion should be "Flechet, M et al". Marine is a first name.

REVIEWER	Horng Ruey Chua National University Hospital, Internal Medicine, Nephrology
REVIEW RETURNED	13-Aug-2021

GENERAL COMMENTS	Dear colleagues, thank you for having me review this manuscript. The key strength of this study is the real-time application and prospective validation in the institution over a period of one year, beyond the usual derivation and internal validation that limit translation to clinical practice. The chief limitation is a prediction window of < 48 hours, in which, the features used in prediction could be too proximate to the event of interest, making the model less useful for clinical risk management. (see comments below). After reading the script, it is still unclear to me, what's the proximity of the features to the event. For example, if a serum creatinine fell within 4 hours of an AKI-defining creatinine, would the former still be included as a feature in prediction? This needs to be clarified in the revised manuscript. Major suggestions: 1. We need details on how the investigators coded the disease codes / comorbidities. These were standardised as ICD10 codes, but there are multiple permutations of diagnosis codes. For instance, chronic kidney disease may be called chronic renal… or renal impairment….etc. How did the investigators recode these diagnoses for purpose of data analysis, and how were these transformed in the real-time data feed for prospective validation? What do the authors mean by "unifying the diagnostic/diagnosis codes"? 2. Data coding – I refer to the use of Pearson correlation coefficient to analyse characteristics before selecting variables with high correlation. We need details on how this was actually performed. What's the threshold to include or exclude features? Please pardon me but I don't often see such interim steps when it comes to machine learning techniques. 3. Missing values – this is a common issue in all machine learning studies. I don't quite understand what the authors mean when they say "we deleted more than 50% of missing values"; what happened to "remaining missing values"? Do they mean they handled missing values in different ways – some by imputation and some by null? Please clarify. 4. 2 training datasets were merged into 1 (which 2? The SHZIU and MIMIC?) 5. Random sampling of negative dataset – the authors decided to trim the non-AKI patients randomly to reduce the imbalance of the case to control numbers. I think this is a reasonable strategy when it comes to retrospective analysis, but could the authors share in discussion, how would that be processed when it comes to real-time data feed and prospective validation?

	6. We need more details about the selection of features. In this study, from my understanding, the prediction window (i.e. final feature included seems to be within 48 hours of the AKI-defining creatinine, but was there a time-gap between the feature to be included and the AKI-defining creatinine?) If there was no time-gap, that would be a major limitation as the features selected were too proximate to the event to be predicted, making it less useful for risk prediction. In addition, what was the time period from which features were selected for prediction? 1 day? 2 day? 3 days? Please kindly add these details. (it seems to me the feature window was 7 days, since Figure S2 illustrates most features to be 7D max or min or std). The authors should comment on these aspects in discussion, as risk prediction would serve us best if a reasonable lead time is created to try prevent a complication of care. 7. Analysis – "non-normally distributed data was log-transformed and deleted if it was still non-normal in distribution"? Really? But the authors said they used non-parametric tests? Please clarify. 8. It is interesting to note that in the Amsterdam database, many of the comorbidities such as cancer, cirrhosis, cardiomyopathy, diabetes, hypertension, were missing. Could the authors comment how did that affect the validation of the severe AKI prediction algorithm in that database? If we refer to Table S1 – it seems that "comorbidities" were all removed in the final model? Please clarify this in the methodology and feature selection. 9. The risk prediction threshold was said to be 0.6 (in methodology) but specified to be 0.423 in the results (line 177). Please clarify. 10. Line 194: "we deleted 267 patients among 94 patients with creatinine baseline more than 3.0 mg/dL" Please clarify. These details are best described in an enhanced patient flow diagram please. 11. The authors should comment on the sensitivity and positive predictive value. Usually, such analysis scores really well in accuracy (due to an imbalanced outcome for which non-AKI are usually much higher than AKI). However the interest is in predicting AKI (versus predicting non-AKI). Therefore the sensitivity and positive predictive value help us decide on the clinical performance and validity with regard to false alerts and recall. These should be included in the discussion. Minor suggestions 1. Grammar may need correction in many areas of the script. E.g. Abstract "many diagnosis models lack of external…."; "the prediction model of severe AKI exhibits promises as a clinical application….". The abstract does not read well, and may I suggest reviewing the grammar with appropriate rectifications

	please. Other segments – for example; Introduction – "decrease in urine volume lags the onset of AKI", "common defect in these studies"; methodology – "patients who used RRT"; data collection – "basic and primary diseases (what were "basic" diseases?)", "comorbidity included hypertension, diabetes…."; line 202 – "more details see in supplementation….". Would suggest use online tools such as Grammarly application to revise the text. 2. Abstract – best to provide improved clarity, with regard to the prediction window of the model; was the prediction window a minimum of 48 hours prior to severe AKI or was it within 48 hours of a severe AKI? 3. Abstract – was the model derived from the SHZJU and MIMIC databases with external validation in Amsterdam database, or was it derived from all 3 databases? Please clarify. 4. Please enhance the patient flow diagram to help us understand (in the derivation cohort at least – how many AKI were selected and how many non-AKIs were selected from the 3 databases). (There is patient flow diagram provided but the exclusion box needs to be more informative, as the authors had deleted many non-AKI patients in the analysis). 5. TABLE on model performance – "sepcificity" is spelt wrongly. Thank you for the study.

VERSION 1 – AUTHOR RESPONSE

Reviewer 1:

LOA 4: We added a sentence in limitation in Line 265: "As a result of the study design, we deleted patients with ICU hospitalization of less than 48 hours, which may result in the exclusion of most relatively mild patients and may reduce false positives.".

LOA5: We added a sentence in Line 119. We deleted variables missing more than 50%. Variables missing more than 30% but less than 50% are listed to clinicians who determine the potential correlation between these variables and AKI. We carry out multiple interpolation for

these variables but clinicians require to be retained, and the others deleted. Variables missing less than 30% are fill in multiple interpolation.

LOA 6: We changed our statement: “When a patient has the following conditions, AKI prediction system will end the patient’s prospective prediction” in Line 146.

LOA 7: We have uploaded the public part of code as a supplementary file including “AKI diagnosis code”, “model train code”, “data partition code”.

LOA 8: We have carefully revised the abstract: “358 positive results were predicted, and 344 patients were diagnosed with severe AKI, with the best sensitivity of 0.72, the specificity of 0.80, and the AUROC of 0.84.” in line 40.

LOA 9: Revised a grammatical mistake in Line 234.

Reviewer 2:

LOA 10: We add a sentence in Line 113, “We use a method similar to the forward incremental method in the multivariate logic regression model, that is, the combination of embedded feature selection and forward addition for feature selection. First of all, all variables are trained in the model, then list by variables importance. variables are added to the model one by one according to the variable importance. a variable is retained if it causes the AUC growth to be greater than 0.01, otherwise delete it.”

LOA 11: We have made changes in Line 125: “The SHZJU and MIMIC databases training sets were mixed into a new training set”.

LOA 12: Amended the Figure 1.

LOA 13: Deleted a sentence in Line 172.

LOA 14: We add an explanation in Line 280: “The diagnostic performance of severe AKI is good with the sensitivity as high as 0.85 in model construction and external validation. however, the sensitivity decreases to 0.72 in the prospective validation, and the overall PPV effect is general. Our model seems to be superior to diagnostic non-AKI patients rather than AKI because of the proportion of positive data that we include. A large number of negative data will increase the specificity and reduce the sensitivity. In the retrospective study, we reduced the proportion of negative data by randomization but retain all date in prospective phase with the sensitivity decreases. We believe that such results are still acceptable and need to be viewed by the reader as a whole.”

LOA 15: Fixed some syntax errors.

Comments from the reviewers and the reply to each suggestion

Editor(s)' Comments to Author (if any):

Comment 1-*Please revise the title of your manuscript to include the research question, study design and setting. This is the preferred format of the journal.*

Reply：**I am so sorry; This has been amended.**

Comment 2-*Please ensure that all competing interests for authors are declared, including paid employment with companies***.**

Reply：**We have instead stated that our study has no competing interests.**

Comment 3-*Please move the Patient and Public Involvement subsection to the Methods***.**

Reply: We have made the adjustment of the Patient and Public Involvement subsection in Line 151. “The information of cases in three databases was in a state of complete desensitization in the process of building the model. During the prospective study, all the patients signed an informed consent form at the beginning of admission to ICU. The real-time data discussed and used by the study members only, and were not made public during the study period. The study was evaluated and approved by the Ethics Committee of the Second Affiliated Hospital of Zhejiang University School of Medicine as study number 2019-078. All

data were anonymized before the authors accessed them for the purpose of this study. According to the inclusion criteria and exclusion criteria, we selected 58492 patients from three databases who met the requirements of the study, including 6461 patients from the SHZJU-ICU database, 36690 patients from the MIMIC database, and 15341 patients from the AmsterdamUMC database."

Comment 4-*In your ethics statement, please indicate whether data were anonymized before the authors accessed them for the purpose of this study***.**

Reply: We ensure that all data are processed anonymously before analysis, which is one of the important principles for the construction and desensitization of the three databases. We had explained it in Line 157 **"All data were anonymized before the authors accessed them for the purpose of study".**

Reviewer #1:

Comment 1*- Why did the authors exclude AKI 1? This creates the flaw that the model may often be predicting the progression from AKI 1 to more severe forms of AKI. That is a much easier, and very different, task to many of the models the authors cite. As many patients in ICU would be expected to progress from AKI 1 to 3 due to the severity of their conditions, it seems like this is an easier task. The authors should report the performance in patients who had no AKI separately to the patients that had AKI 1 at the time of prediction.*

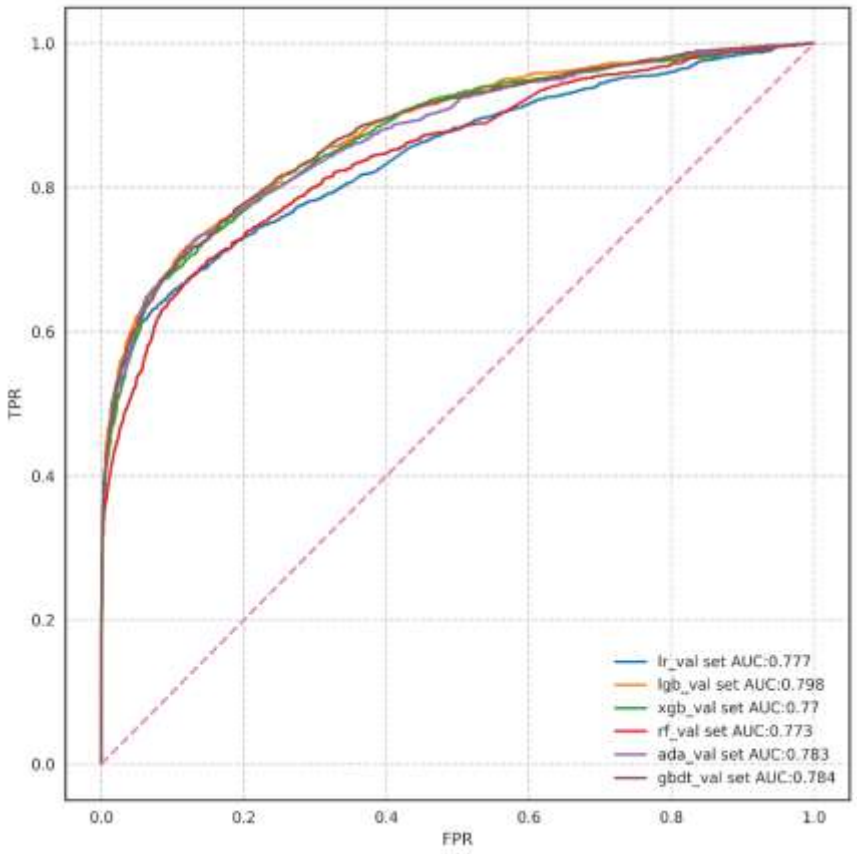
Reply: In the initial research stage, we included the AKI stage I in the construction of the prediction model, but deleted this part in the end. We consulted many clinicians who believe that the AKI I defined by the KDIGO standard is very common in clinical practice. AKI I patients are rarely change the treatment plan. However, the use of diuretics or CRRT is considered necessary if the patient progresses to AKI II/III. Machine learning was used to predict AKI to guide the follow-up guidance including the timing of early diuretics or CRRT use, while the prediction of AKI I was of little significance in clinical guidance. Therefore, although we completed the AKI I prediction in the retrospective study, it was not carried out in the follow-up prospective study.

It is absolutely right in understanding of disease progression and prediction models. The percentage of severe AKI patients who develop from AKI I patients is indeed higher than that of non-AKI patients. We have included the change rate of creatinine and urea as an important parameter in the data analysis, and this factor has been included in the model. Making a

distinction between the AKI and severe AKI can give us a clearer understanding of the progress of the disease, but it is not the focus of this study.

The following supplement our prediction model including all AKI, and the overall effect is not as good as the severe AKI prediction.

Model	AUROC	Accuracy	Sensitivity	Sepcificity	PPV	NPV	Recall	precision	F1
Logistic regression	0.777	0.814	0.666	0.888	0.748	0.842	0.666	0.763	0.734
LightGBoost	**0.798**	0.829	0.706	0.891	0.763	0.858	0.706	0.838	0.719
GBDT	0.784	0.836	0.629	0.939	0.838	0.835	0.629	0.854	0.718
AdaBoost	0.783	0.838	0.620	0.947	0.854	0.833	0.620	0.748	0.704
Random Forest	0.773	0.814	0.651	0.896	0.757	0.837	0.651	0.757	0.700
XGBoost	0.770	0.831	0.589	0.951	0.858	0.822	0.589	0.858	0.699



Comment 2- *When patients using RRT within 48 hours of admission, how many patients who went from normal or AKI to severe requiring RRT were excluded?*

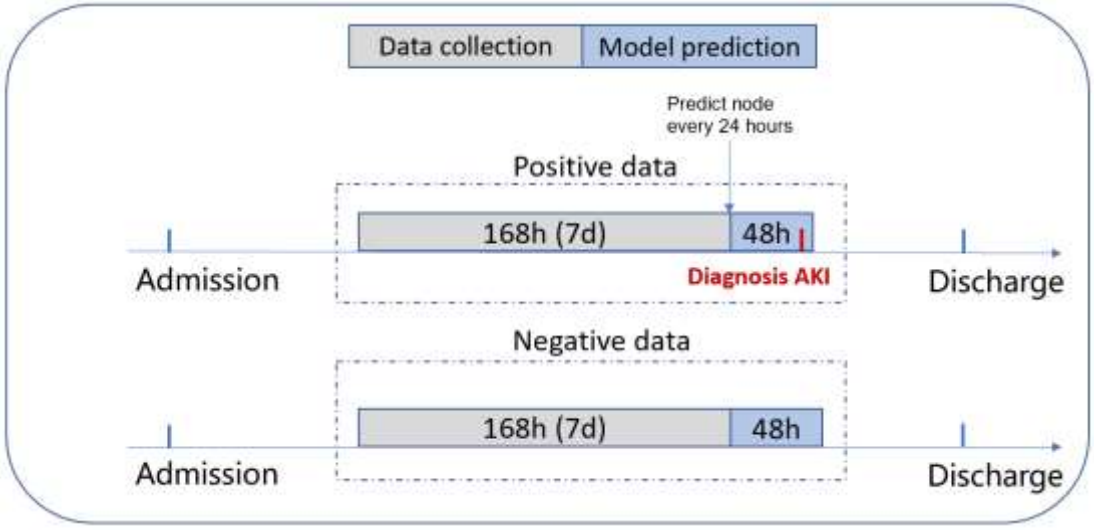
Reply: Firstly, our prediction is not meaningful if CRRT is required immediately after admission. Secondly, RRT would clear creatinine and affect urine volume, so we could not obtain accurate AKI grade. According to our statistics, a total of 2514 patients who underwent RRT within 48 hours of admission, including 345 SHZJU-ICU patients, 1517 MIMIC patients, and 652 AmsterdamUMC patients.

Comment 3- *Why were patients in ICU for fewer than 3 days excluded? This seems like the group of patients in ICU who are least sick, and in whom it is important to ensure the model is not making too many false positives.*

Reply: Thank you for your advice and an error you pointed out in this study. In addition, I am very sorry, but the description error caused by our carelessness must be corrected here. We delete the patient whose hospital stay is less than 48 hours instead of the 3 days. The reason why we delete these patients is not because they are least sick. The main reason is that the patients with ICU hospitalization time less than 48 hours produce a large number of missing values in the data extraction. This is the inherent defect caused by the design of this study.

AKI diagnosis depends on creatinine baseline and trend, urine volume, which are routinely monitored once a day. If the patient's hospital stay is less than 48 hours, a large number of variables are null during the 48-hour data retrospective even if the patient is finally classified as a severe AKI patient. Therefore, we delete patients staying in hospital for less

than 48 hours is to reduce the lack of data. We have added it in limitation so that readers are aware of the shortcomings of our research in Line 265: “As a result of the study design, we deleted patients with ICU hospitalization of less than 48 hours, which may result in the exclusion of most relatively mild patients and may reduce false positives.”

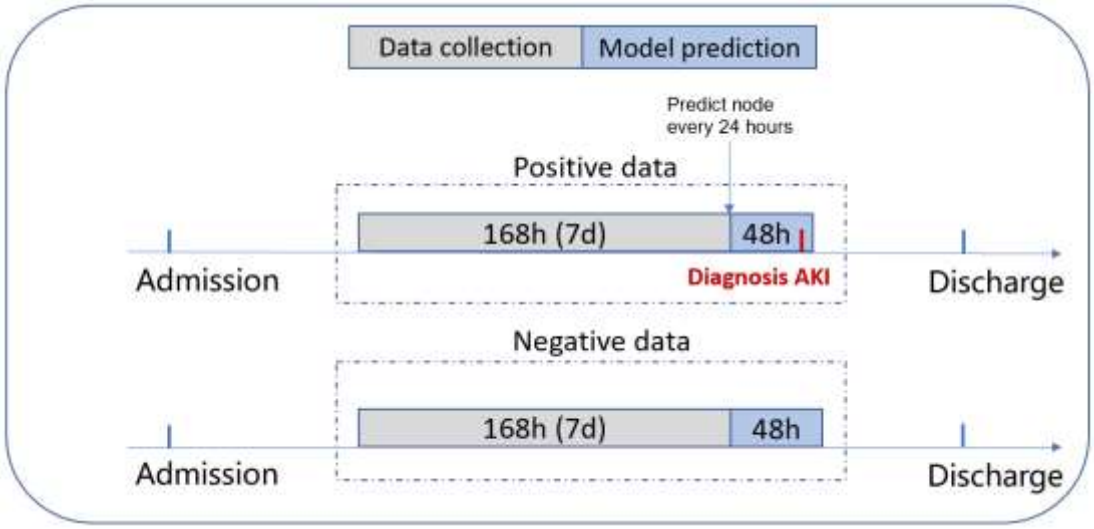


Comment 4- *I understand how the data were chosen to include, but how did the model cope with missing information for some patients? More information is needed here. Inevitably data will be missing for patients as not all patients have all labs recorded.*

Reply: We deleted variables missing more than 50%. Variables missing more than 30% but less than 50% are listed to clinicians who determine the potential correlation between these variables and AKI. We carry out multiple interpolation for these variables but clinicians require to be retained, and the others deleted. Variables missing less than 30% are fill in multiple interpolation. We added to it in Line 119.

Comment 5- *How frequent were predictions? I think the discussion mentions daily - why were more frequent predictions not made? How was data grouped? Did the model have access to data for only previous days (i.e. model that predicts on day t, only sees data up to and including day t-1) or would that model have seen data from day t as well?*

Reply: The frequent of our prediction is once a day. This frequency is mainly limited by a large number of laboratory test results such as creatinine and blood routine once a day, or even once every other day. Therefore, the data required for predictions less than one day are repetitive and unnecessary. Our model predicts on day T, only sees data from Day T-7 to day T.



Comment 6- *The the authors say "Comorbidity included hypertension, diabetes, cardiopathy, liver disease, and malignant tumors" are these just examples, or is this an exhaustive list? If an exhaustive list, why were these chosen and not others that may be risk factors for AKI?*

Reply: Thank you for asking such questions. The comorbidity are all the complications listed in this study. The principles include two points: one is that they may lead to kidney injury possibly, because of hypertensive nephropathy, diabetic nephropathy, cardiorenal syndrome, hepatorenal syndrome and so on. Second, these complications are relatively common. If comorbidity is relatively rare, they will be ignored by machine learning. The selection is also based on previous studies on the risk factors of AKI.

Comment 7*-"unifying the unit and diagnostic codes" can the authors say more about how they did this?*

Reply: There are different units in the laboratory test variables of the three databases, and unifying these units is the initial content of this study, so we did not describe it in detail. Although the variable units are different, it is not very difficult to unify the unit, because among the more than 1000 test variables, there are about 200 common variables that have a great impact on the study, of which 120 variables units are consistent. For the remaining 80 indicators, although there are two or three units, the unit's conversion has an internationally recognized fixed coefficient. we can get a unified unit after we write the coefficient conversion code. The diagnostic code is relatively simple because our diagnosis follows ICD 9 or ICD 10 code. There is not much change between the two groups.

Comment 8-*"We deleted more than 50% of the missing values and replaced the remaining missing values with multiple interpolations" What do the authors mean by deleting >50% of the missing values? Why were these deleted?*

Reply: Sorry, there may be some deviation in my statement here. We deleted variables missing more than 50%. Variables missing more than 30% but less than 50% are listed to clinicians who determine the potential correlation between these variables and AKI. We carry out multiple interpolation for these variables but clinicians require to be retained, and the others deleted. Variables missing less than 30% are fill in multiple interpolation.

Comment 9-*"There were more negative data than positive data, so we randomly sampled the negative datasets and constructed a new data subset with a sampling ratio of positive and negative data of 1:5."*

Why was this needed? Shouldn't the models be trained on representative data?

Reply: Severe AKI patients is relatively small, account for about 10% of all patients. Secondly, patients with non-severe AKI produce a large number of negative data, while patients with severe AKI have more than 1 to 30 negative data before diagnosis of severe AKI, so the ratio of positive data to negative data is about 300:1. The model had high specificity but poor sensitivity if we included all the negative data.

And as you said, we refer to a large number of similar studies in order to ensure that the data of the training model is representative and finally determine to form datasets for model

construction by randomly selecting some of the negative data and positive data with the final proportion of 1:5.

Comment 10*- Did the authors check if any patients represented in the internal validation dataset were also included in the prospective validation? What would the results have been if these were excluded?*

Reply: Our prospective validation data will not be included in the internal retrospective validation set because the internal validation had be completed by the end of 2019, while the prospective validation includes patients in 2020. According to your suggestion, we specially checked the admission time of patients in our internal and prospective data set, and no mixture was found.

Comment 11*-"We sampled the 20% predicted data every month and deleted samples with more than 50% missing values to ensure data correctness. The criteria to terminate prediction were A. a positive diagnosis; B. Transfer out of ICU or death with negative diagnosis."*

This is confusing. It is not clear to me why examples were deleted, and why discharge or a positive diagnosis would result in deletion. This should be much better described, as it seems like it reduces the clinical applicability of the results:

Why was it necessary to delete patients? What was this criterion based on? What would the results have been without this deletion? It's it particularly important to predict accurately

on patients who died, as those deaths may be preventable? And if the authors mean a positive diagnosis for AKI, isn't that what this study cares about? Why were only 20% sampled? How did the authors arrive at this percentage?

Reply: I am sorry that there may be irregularities in my statement, which may be misunderstood by you. We extract 20% of the data for verification to evaluate whether there are data deletions in the samples of prospective studies and errors in the crawling algorithm. Sometimes, the prospective prediction algorithm cannot update the data synchronously because of the update of the hospital database, which may affect the predicted results. During the one-year prospective prediction, database crashes two times resulted in 87 patients failing to make predictions in time, which were found by verification. Although these patients still have predictive results, we believe that the data are incomplete and we delete these patients to ensure the accuracy of study. We think this proportion is very small and does not affect our research.

Secondly, the prediction model automatically grabs the data to predict the probability of severe AKI in the next 48 hours. **The frequency of our prediction is once a day, and the end of prediction is not to delete the patient.** If the patient has already met the AKI diagnosis, it is an idle work to predict the occurrence of AKI within 48 hours. If a patient leaves ICU or dies, our data will be missing and cannot be accurately predicted, so the system will end the prediction by default. It is not mean that we delete this part of the data when the patient dies or transferred to hospital. We changed our statement: "When a patient has the following conditions, AKI prediction system will end the patient's prospective prediction" in Line 146.

Comment 12- *In the discussion, the authors mention this about a previous study: "A total of 55.8% of severe AKI patients were predicted within the first 48 hours, although each accurate prediction was accompanied by two mispredictions". But I can't seem to find the same number for the present study from the authors. How many false positives were predicted for every true positive? This figure can help understand how clinically useful the model is.*

Reply: Each correct prediction is followed by two wrong predictions you mentioned above, which derived from a literature we cited. In our analysis, we included this information in sensitivity, specificity, PPV and NPV, and if we list true and false positives prediction, the table would be very complex because the study included six machine learning methods and eight statistical analyses. We use the PPV to show true positive and false positive, because PPV= true positive / (true positive + false positive). According to our statistics, the lowest PPV is 0.49 and the highest is 0.74, that is, the worst model is that one true positive prediction is accompanied by one false positive prediction, and the best model is that three true positive predictions are accompanied by one false positive prediction.

Comment 13-*"The model code can be obtained by email if readers need it, but we cannot guarantee that all the code will be provided" Is the code available or not? This is a very unusual code availability statement. If the code is not available, why not? If it is, please make it publicly available. Can the code be provided for review?*

Reply: Part of our code involves patents, so we can't make all of it public. We have uploaded the public part as a supplementary file including “AKI diagnosis code”, “model train code”, “data partition code”.

Comment 14- *The abstract says 358 positive results predicted and 344 patients diagnosed with Severe AKI, but doesn’t say what the crossover is in the abstract. Better to instead say sensitivity, specificity, PPV, NPV, and AUROC. The accuracy is also not needed.*

Reply: According to your suggestions, we have carefully revised the abstract: “358 positive results were predicted, and 344 patients were diagnosed with severe AKI, with the best sensitivity of 0.72, the specificity of 0.80, and the AUROC of 0.84.” in Line 40.

Comment 15- *Severe AKI in 9% of patients - this seems low, given that the prospective data collection occurred during 2020, when one would expect ICUs to have a large number of COVID patients, up to 50% of which develop AKI in hospital, and perhaps more in ICU. Why was the incidence much lower?*

Reply: Although novel coronavirus wreaked havoc in 2020, it was mainly concentrated in Wuhan, China. Zhejiang Province, where our center is located, has confirmed a total of 1439 cases so far in 2020, and there has been no case of the epidemic getting out of control. In addition, our center is not the special hospitals for COVID-19, so we have not received any

COVID-19 patients in 2020. Generally speaking, COVID-19 did not have a significant impact on our research.

Comment 16*- In the discussion authors suggest their visualization graphic is a benefit over a previous work that compared performance with clinicians. Can the authors say more about the visualization tool? How was it developed? Is there any evidence it improves predictive performance of humans?*

Reply: Thank you for your interest in our visualization tool. This visualization program has patent restrictions and is an important part of unpublished papers, so I can't provide you with more details at this time. If you are interested, we can share it with you after our paper is published.

Comment 17*- In Figure S3, why is sepsis also being flagged? Does a similar model also flag sepsis and other scores that was not included in this manuscript?*

Reply: Thank you for this keen observation. As mentioned above, our visual graphics integrate real-time and dynamic APACHE II scores, SOFA scores, Sepsis prediction, and AKI prediction. This system can complete the prospective prediction content and present it to clinicians timely. Our AKI prediction model is an important part of it so I cannot be completely separated in the presentation process.

Comment 18*-"Marine F" in the discussion should be "Flechet, M et al". Marine is a first name.*

Reply: Thank you for pointing out our mistake, and we have revised it.

Reviewer #2:

Comment 1*- We need details on how the investigators coded the disease codes / comorbidities. These were standardised as ICD10 codes, but there are multiple permutations of diagnosis codes. For instance, chronic kidney disease may be called chronic renal… or renal impairment….etc. How did the investigators recode these diagnoses for purpose of data analysis, and how were these transformed in the real-time data feed for prospective validation? What do the authors mean by "unifying the diagnostic/diagnosis codes"?*

Reply: Thank you for your question. We need to clarify here. We bring the diagnosis into the analysis in the early stage. We match it according to the ICD10 data set firstly, so as to extract the specific diagnosis. Because these diagnoses are standardized, then, we use Countvectorizer to extract features and classify them. Finally, we give the diagnosis list to clinicians for manual screening. However, the results are not satisfactory because the words are too scattered to form specific characteristics and is not included in the final prediction model. Therefore, the final model includes no diagnostic variables. In the prospective study, we did not have a problem with the diagnosis.

We didn't found a very appropriate way to analyze the diagnostic variables, considering that there may be too many diagnostic features to dilute the variables importance. We are looking for the method of cluster analysis to analyze the diagnosis, but the research is in progress.

In addition, the unified code we are talking about refers to that we convert ICD-9 code used by some patients into unified ICD-10 code through data conversion.

Comment 2- *Data coding – I refer to the use of Pearson correlation coefficient to analyse characteristics before selecting variables with high correlation. We need details on how this was actually performed. What's the threshold to include or exclude features? Please pardon me but I don't often see such interim steps when it comes to machine learning techniques*.

Reply: We admit that there are expression problems about variables selection in the writing process. We use a method similar to the forward incremental method in the multivariate logic regression

model, that is, the combination of embedded feature selection and forward addition for feature selection. First of all, all variables are trained in the model, then list by variables importance. variables are added to the model one by one according to the variable importance. a variable is retained if it causes the AUC growth to be greater than 0.01, otherwise delete it. We are very sorry and have corrected the error in the article in Line 113.

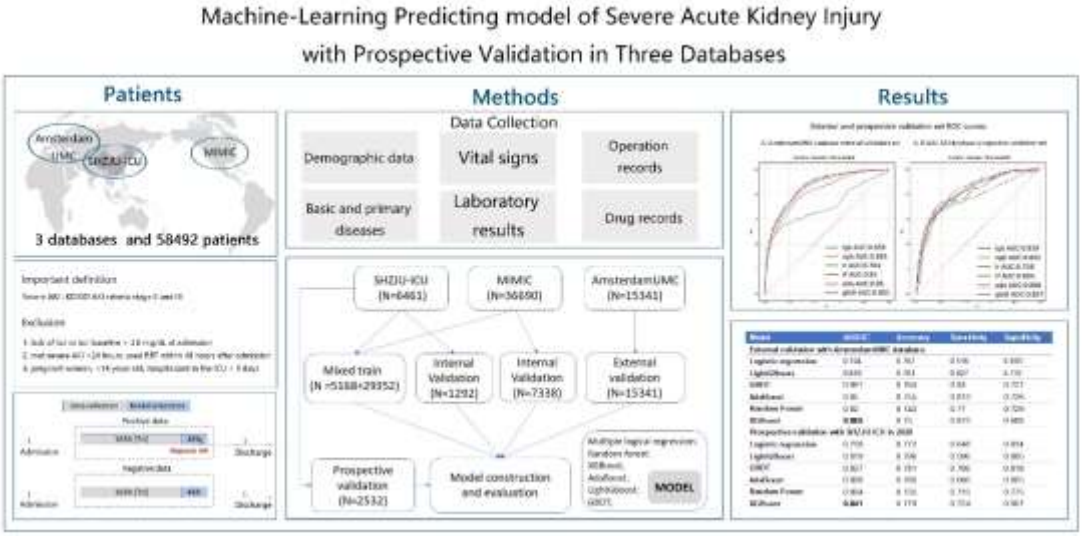
Comment 3- *Missing values – this is a common issue in all machine learning studies. I don't quite understand what the authors mean when they say "we deleted more than 50% of missing values"; what happened to "remaining missing values"? Do they mean they handled missing values in different ways – some by imputation and some by null? Please clarify*.

Reply: I'm very sorry, but this sentence is ambiguous. We deleted variables missing more than 50%. Variables missing more than 30% but less than 50% are listed to clinicians who determine the potential correlation between these variables and AKI. We carry out multiple interpolation for these variables which clinicians require to be retained, and the others deleted. Variables missing less than 30% are fill in multiple interpolation. Generally speaking, the method we use to deal with missing values is only multiple interpolation.

Comment 4- 2 training datasets were merged into 1 (which 2? The SHZIU and MIMIC?).

Reply: I'm sorry for the trouble caused by it, but the mixed training set is SHZJU and MIMIC, and we have made changes in Line 125: "The SHZJU and MIMIC databases training sets were mixed

into a new training set".



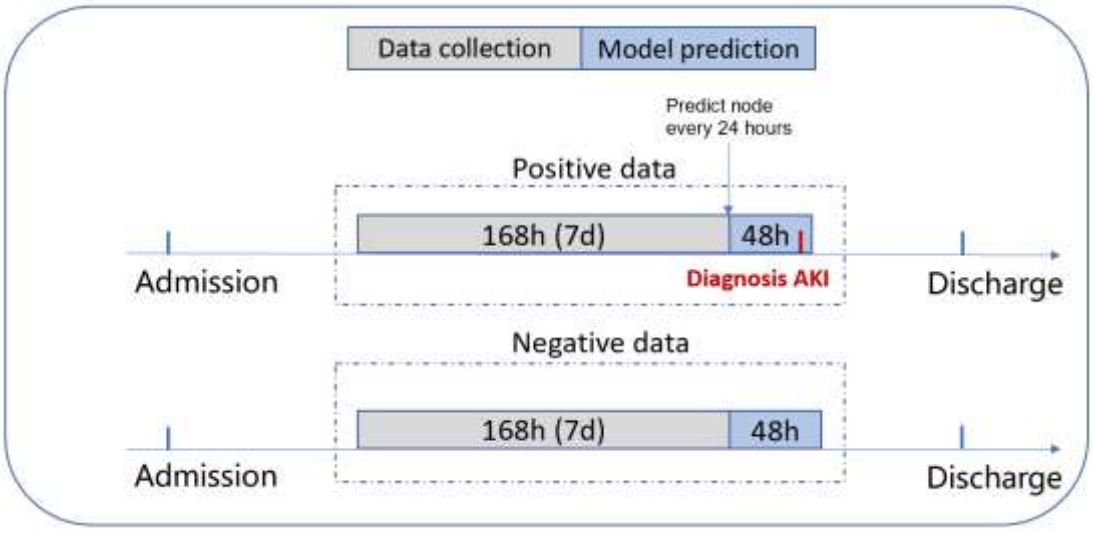
Comment 5*- Random sampling of negative dataset – the authors decided to trim the non-AKI patients randomly to reduce the imbalance of the case to control numbers. I think this is a reasonable strategy when it comes to retrospective analysis, but could the authors share in discussion, how would that be processed when it comes to real-time data feed and prospective validation?*

Reply: The research team discussed this topic at the beginning of the study. The focus of our discussion is whether the randomized selection of control group data is right, but the answer is no. It is not suitable for real scenarios and not consistent with the prospective analysis. Although acute kidney injury is very common, the proportion is still around 10%, that is, 90% of patients do not occur. It is very common in clinical practice that the negative data are much higher than the positive data. However, artificial intelligence can still learn enough features to build an appropriate model if negative data is large enough. This is why the study included MIMIC and AmsterdamUMC databases.

However, it is impossible to know whether the current data is positive or negative in the prospective study, so we do not have any restrictions in practice. Data screening is not allowed for prospective research. One of the purposes of the prospective study is to confirm that the model constructed by our method can cope with clinical scenarios. To be fair, there are many AKI prediction models constructed by artificial intelligence at present, and one of the highlights of our study is that it has been verified by prospective research for one year.

Comment 6- *We need more details about the selection of features. In this study, from my understanding, the prediction window (i.e. final feature included seems to be within 48 hours of the AKI-defining creatinine, but was there a time-gap between the feature to be included and the AKI-defining creatinine?) If there was no time-gap, that would be a major limitation as the features selected were too proximate to the event to be predicted, making it less useful for risk prediction. In addition, what was the time period from which features were selected for prediction? 1 day? 2 day? 3 days? Please kindly add these details. (it seems to me the feature window was 7 days, since Figure S2 illustrates most features to be 7D max or min or std). The authors should comment on these aspects in discussion, as risk prediction would serve us best if a reasonable lead time is created to try prevent a complication of care.*

Reply: Thank you for your question. First of all, the time for the AKI diagnosis by urine volume and creatinine is relatively fixed in the process of retrospective study, because creatinine was tested routine once a day or every other day in ICU, and often before fasting in the morning. Our prediction node is to collect data every 24 hours from the patient's admission to ICU, and collect the variables in the past 7 days at this time point to form a piece of data. If there is no diagnosis within the next 48 hours, the data is considered to be negative, otherwise, it is positive. Our node selection is in line with the doctor's logic of thinking. Our diagnostic nodes and predictive feature selection nodes have a clear time sequence. It will not be staggered together.



Comment 7- *Analysis – "non-normally distributed data was log-transformed and deleted if it was still non-normal in distribution"? Really? But the authors said they used non-parametric tests? Please clarify.*

Reply: I am very sorry and the expression here is wrong. We choose non-parametric tests for variables with non-normal distribution.

Comment 8- *It is interesting to note that in the Amsterdam database, many of the comorbidities such as cancer, cirrhosis, cardiomyopathy, diabetes, hypertension, were missing. Could the authors comment how did that affect the validation of the severe AKI prediction algorithm in that database? If we refer to Table S1 – it seems that "comorbidities" were all removed in the final model? Please clarify this in the methodology and feature selection*.

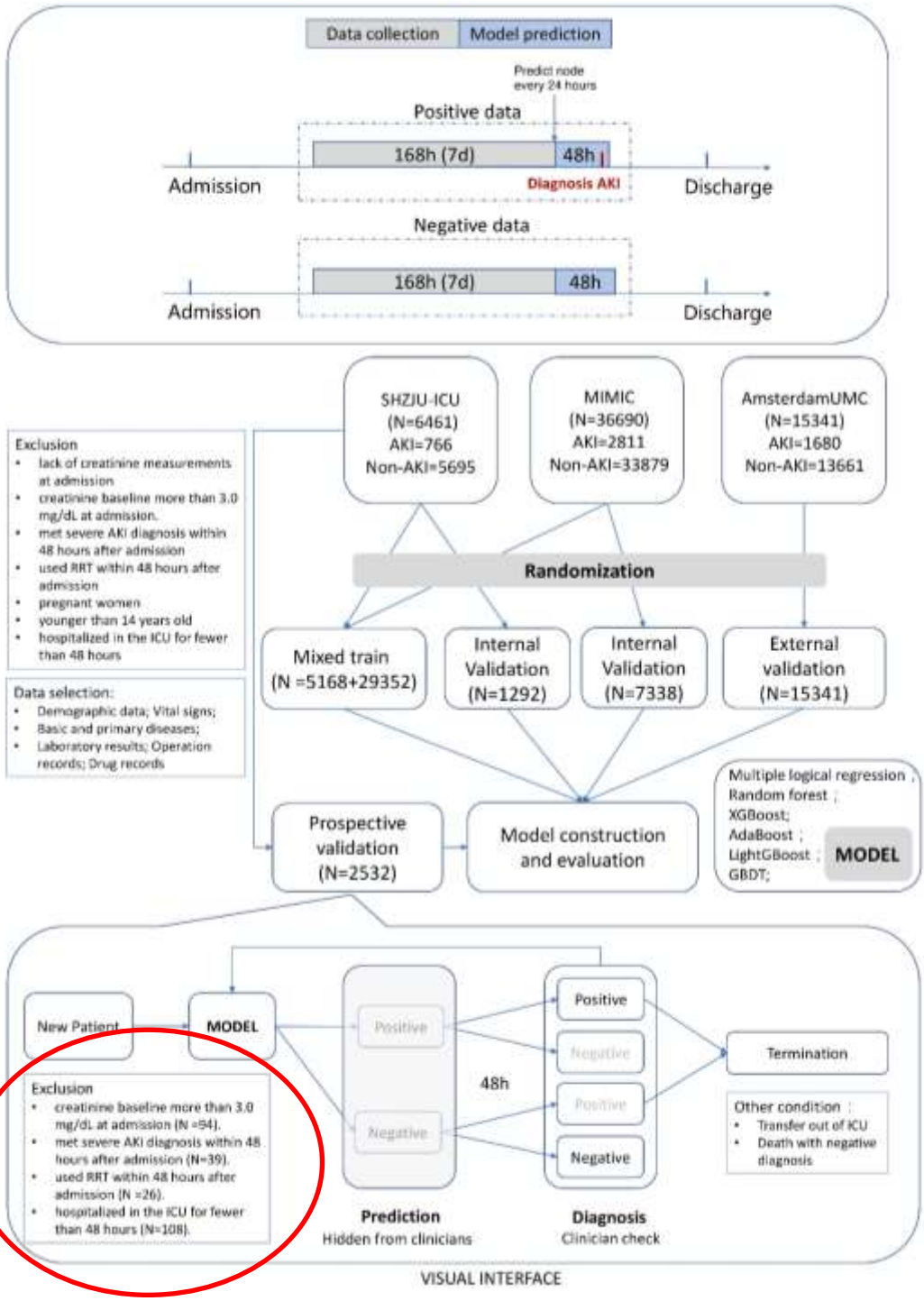
Reply: The impact is small, because the diagnostic variables are not ultimately included in the model construction. Most of the variables included in the model are laboratory test variables, which may be the reason why we have achieved good results in using AmsterdamUMC database. Although the distribution of such variables is not so satisfactory, it is to be expected. The laboratory test variables have a large amount of data and changes, and is easy to obtain which are more likely to reflect the changes of the AKI.

Comment 9- *The risk prediction threshold was said to be 0.6 (in methodology) but specified to be 0.423 in the results (line 177). Please clarify.*

Reply: This is a writing error; our rule is 0.423!

Comment 10- *Line 194: "we deleted 267 patients among 94 patients with creatinine baseline more than 3.0 mg/dL" Please clarify. These details are best described in an enhanced patient flow diagram please.*

Reply: Thank you for your positive response to our work and the kind advice. We have added this information to Figure1.



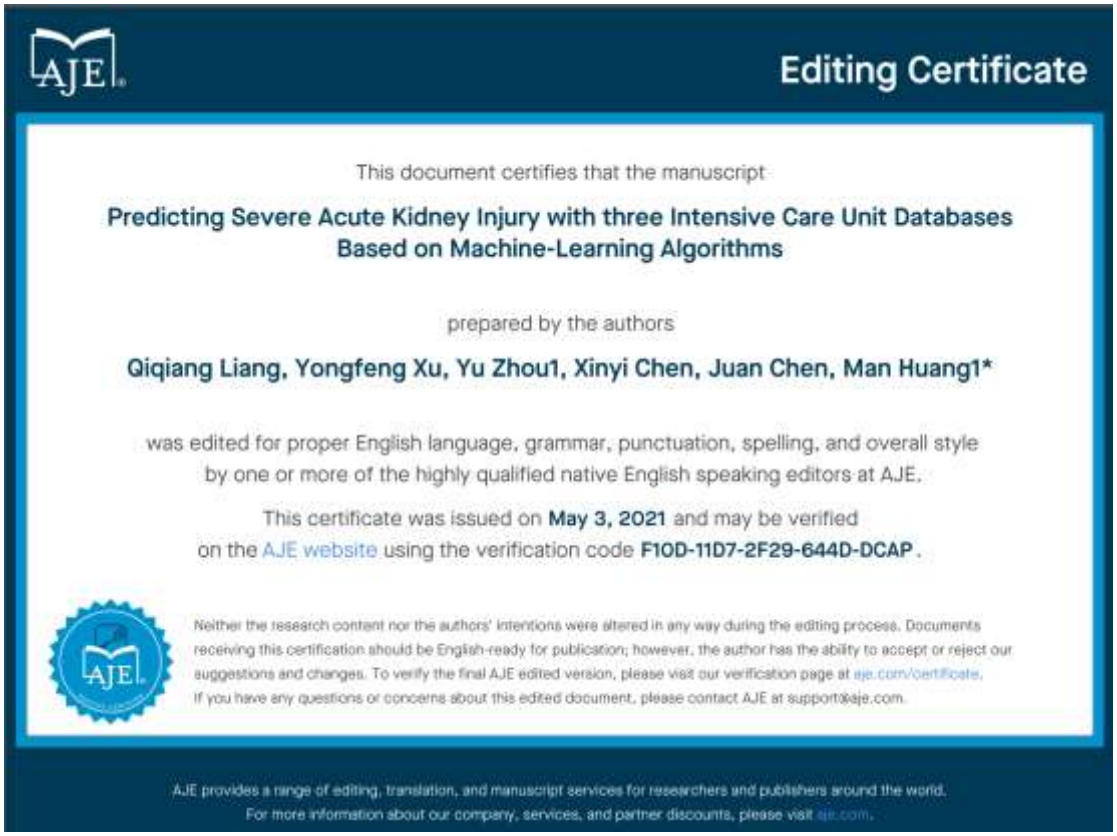
Comment 11- *The authors should comment on the sensitivity and positive predictive value. Usually, such analysis scores really well in accuracy (due to an imbalanced outcome for which non-*

AKI are usually much higher than AKI). However, the interest is in predicting AKI (versus predicting non-AKI). Therefore, the sensitivity and positive predictive value help us decide on the clinical performance and validity with regard to false alerts and recall. These should be included in the discussion.

Reply: Thanks for your valuable suggestions. We add an explanation in Line 280: “The diagnostic performance of severe AKI is good with the sensitivity as high as 0.85 in model construction and external validation. however, the sensitivity decreases to 0.72 in the prospective validation, and the overall PPV effect is general. Our model seems to be superior to diagnostic non-AKI patients rather than AKI because of the proportion of positive data that we include. A large number of negative data will increase the specificity and reduce the sensitivity. In the retrospective study, we reduced the proportion of negative data by randomization but retain all date in prospective phase with the sensitivity decreases. We believe that such results are still acceptable and need to be viewed by the reader as a whole.”

Comment 12*- Grammar may need correction in many areas of the script. E.g. Abstract “many diagnosis models lack of external….”; “the prediction model of severe AKI exhibits promises as a clinical application….”. The abstract does not read well, and may I suggest reviewing the grammar with appropriate rectifications please. Other segments – for example; Introduction – “decrease in urine volume lags the onset of AKI”, “common defect in these studies”; methodology – “patients who used RRT”; data collection – “basic and primary diseases (what were “basic” diseases?)”, “comorbidity included hypertension, diabetes….”; line 202 – “more details see in supplementation….”. Would suggest use online tools such as Grammarly application to revise the text.*

Reply: Thank you very much for your advice on the grammar of this study. our paper has been improved by using Grammarly and has improved the quality of English through AJE.



Editing Certificate

This document certifies that the manuscript

Predicting Severe Acute Kidney Injury with three Intensive Care Unit Databases Based on Machine-Learning Algorithms

prepared by the authors

Qiqiang Liang, Yongfeng Xu, Yu Zhou1, Xinyi Chen, Juan Chen, Man Huang1*

was edited for proper English language, grammar, punctuation, spelling, and overall style by one or more of the highly qualified native English speaking editors at AJE.

This certificate was issued on **May 3, 2021** and may be verified on the AJE website using the verification code **F10D-11D7-2F29-644D-DCAP**.

Neither the research content nor the authors' intentions were altered in any way during the editing process. Documents receiving this certification should be English-ready for publication; however, the author has the ability to accept or reject our suggestions and changes. To verify the final AJE edited version, please visit our verification page at aje.com/certificate. If you have any questions or concerns about this edited document, please contact AJE at support@aje.com.

AJE provides a range of editing, translation, and manuscript services for researchers and publishers around the world. For more information about our company, services, and partner discounts, please visit aje.com.

Comment 13*- Abstract – best to provide improved clarity, with regard to the prediction window of the model; was the prediction window a minimum of 48 hours prior to severe AKI or was it within 48 hours of a severe AKI?*

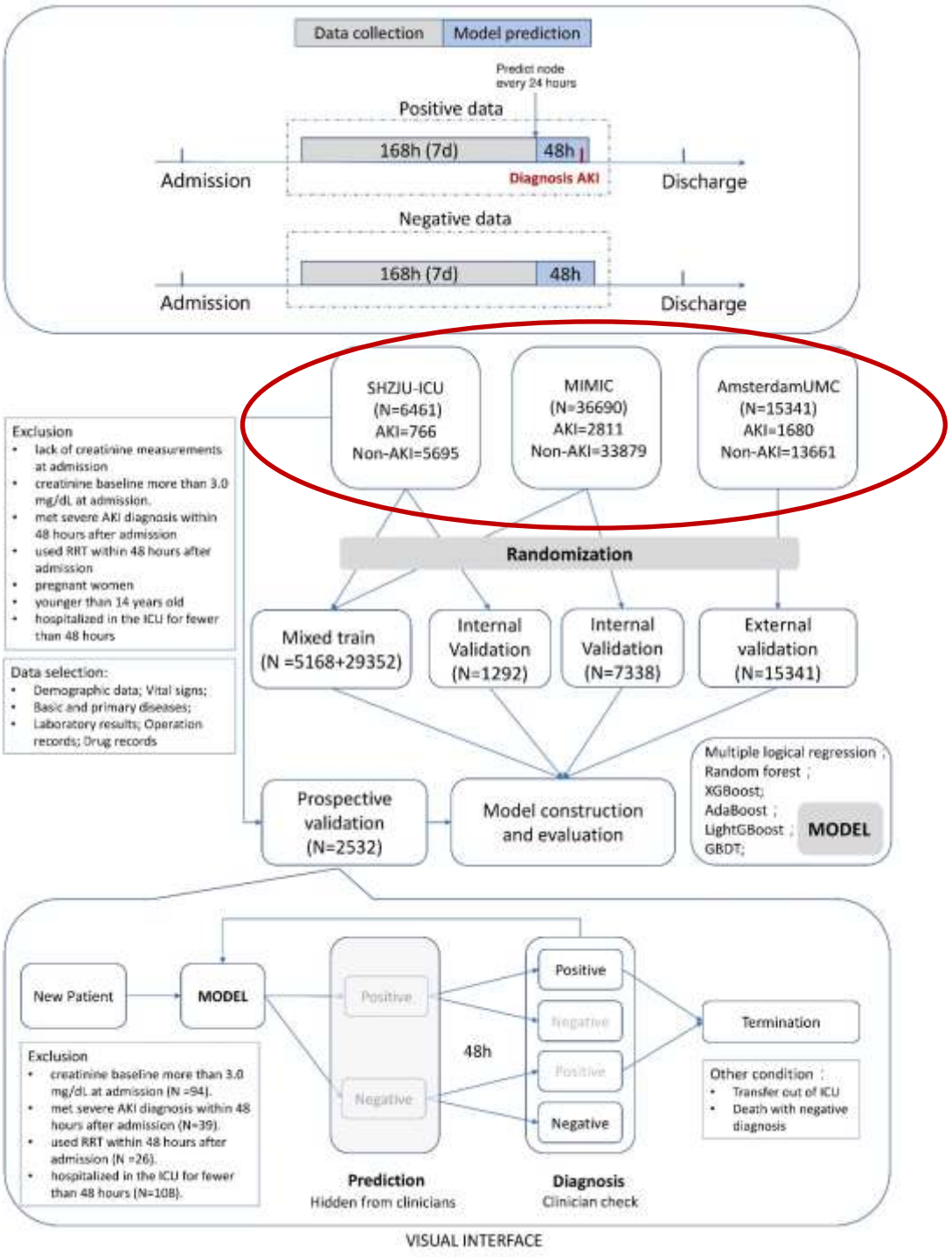
Reply: The purpose of this study is to predict whether severe AKI will occur within the next 48 hours at the current predicted time.

Comment 14*- Abstract – was the model derived from the SHZJU and MIMIC databases with external validation in Amsterdam database, or was it derived from all 3 databases? Please clarify.*

Reply: The training set of the model is derived from MIMIC and SHZJU, and the AmsterdamUMC database is only involved in validation. We have made a clarification in the article.

Comment 15- *Please enhance the patient flow diagram to help us understand (in the derivation cohort at least – how many AKI were selected and how many non-AKIs were selected from the 3 databases). (There is patient flow diagram provided but the exclusion box needs to be more informative, as the authors had deleted many non-AKI patients in the analysis).*

Reply: According to your suggestion, we modified Fig. 1 in the revised manuscript as follows.



Comment 16- *TABLE on model performance – "specificity" is spelt wrongly.*

Reply: Thank you for pointing out our mistake, which has been revised.

VERSION 2 – REVIEW

REVIEWER	Joseph R Ledsam Google Inc
REVIEW RETURNED	11-Oct-2021

GENERAL COMMENTS	The authors have addressed most of my concerns. There is still some information needed to accept however. Comment 1: on AKI 1 I thank the authors for responding and presenting the performance on all AKI. However, my concern has not been addressed. I believe that there is a risk the model is mostly relying on the fact that AKI 1 commonly progresses to AKI 2 or 3, and I wanted the authors to provide evidence that this is not the case. For the review (not necessarily for the main manuscript) can the authors provide the following? - Performance on all AKI (already shown) - Performance on several AKI only (already shown) - Performance on cases of AKI 2+3 that are not proceeded by AKI 1 (this is not currently shown but would help understand what the performance is in the group that is most important to predict) Comment 2 - understood Comment 3 - understood Comment 4 - understood Comment 5 - creatinine being measured only once a day does not preclude more frequent measurements than daily. Previous studies have measured more frequently than daily. This should not block publication, but it is important to note as a limitation. Comment 6 - I don't think it's true that ML will ignore less common co-morbidities. Of course if there is only one or two examples in the dataset then it will be difficult, but there are many ways to address this. Rare autoimmune conditions with renal implications can be grouped as a single group either by clinical group, or using the parent ICD codes for groups of similar diseases. The same is true of cardiovascular diseases that may have been excluded. I would expect the authors to have provided justification for including only these, either by having a threshold for how common the condition is, or by having undertaken a more thorough review. Comment 7 - understood Comment 8 - understood Comment 9 - Reading the manuscript, it is unclear if the 1:5

	enrichment was performed only in the SHZJU-ICU database, or in all the datasets used. If no enrichment was used in MIMIC / AmsterdamUMC, the authors can discount my comment below. However, if the same enrichment was used in all datasets, I have some concerns. "The model had high specificity but poor sensitivity if we included all the negative data." - Isn't this what you would expect in clinical practice then? Doesn't the use of an enriched dataset provide an unrealistic evaluation of model performance? Enriched datasets are typically used when conducting reader studies for practical reasons, but best practice should be to report performance both on a clinically representative dataset, as well as the enriched dataset. I think the authors should report both numbers at the very least, otherwise the study gives an unrealistic impression of expected model performance. The way this sampling has done also appears to have been just dropping large periods of the clinical record for negative patients that would otherwise have produced many false positives. This matters partly because of the TF and FP ratio - and thank you for providing the ratios for the best and worst models in response to comment 12. On the TP and FP ratios, are these the results on the test set that was enriched from 1:300 (positive to negative ratio) to 1:5? So we would expect roughly 60 times the number of false positives in clinical practice - so the TP to FP ratio for the worst model is actually 1 TP for 60 FP, and for the best model 1 TP for 20 FP. I think it's important to report this number in a paper where the authors claim the model can be applied in clinical practice. Comment 10 - understood Comment 11 - understood Comment 12 - Thank you for providing the TP to FP ratios for the best and worse models. See response to comment 9 above for additional comments. Comment 13 - many thanks for uploading the data Comment 14 - many thanks for reporting these. Comment 15 - understood Comment 16 - the authors should not make claims about the visualization tool if they are not backed up by evidence presented in this paper, or in prior work. This claim should be removed. The existence of future evidence to support the use of the visualization tool can be mentioned in limitations. Comment 17 - the authors should find a way to mention this to avoid confusion. Comment 18 - many thanks

REVIEWER	Horng Ruey Chua National University Hospital, Internal Medicine, Nephrology
REVIEW RETURNED	17-Oct-2021

GENERAL COMMENTS	Thank you the revised manuscript. It is now clearer. The diagrams

	are very helpful. They have answered my queries satisfactorily. I do not have further comments to add.