APM ANNALS OF PALLIATIVE MEDICINE
AN OPEN ACCESS JOURNAL FOR HIGH-QUALITY RESEARCH IN PALLIATIVE MEDICINE

Peer review file

Article Information: http://dx.doi.org/10.21037/apm-20-1183

Review Comments A:

The present paper reports the result of review and meta-analysis regarding different delirium risk assessment tools for Intensive Care Unit. This represents an important issue, due to relevance of delirium for prognosis and care burden of patients admitted to ICUs. Authors conclude that PRE-DELIRIC and E-PRE-DELIRIC are recommended to assess delirium risk of ICU patients, in spite of heterogeneity of results among studies.

I suggest that the Authors take into account the following point

Comment 1- A review regarding PRE-DELIRIC only was recently published (Ho MH, Chen KH, Montayre J, et al. Diagnostic test accuracy meta-analysis of PRE-DELIRIC (PREdiction of DELIRium in ICu patients): A delirium prediction model in intensive care practice. Intensive Crit Care Nurs. 2020;57:102784), and results are partially different from the present one In fact in the present paper PRE-DELIRIC shows a cumulative AUC of 0.844 (95% CI=0.793-0.896), while in the previous paper AUC is 0.78 (95% CI 0.74–0.81), which is similar to the value reported in the present paper for E-PRE-DELIRIC. Authors should discuss the reason of these differences, which seem to underline even more the heterogeneity of available data.

Reply1: "Thank you very much for your efforts in reviewing our manuscript and providing helpful comments. We have now compared our review with the Diagnostic test accuracy meta-analysis of PRE-DELIRIC (PREdiction of DELIRium in ICu patients). The reason for these differences is that we have included different original studies. We analyze the performance of PRE-DELIRIC without adding the recalibrated model, whose accuracy is not as good as the former one. Plus, we have included two Chinese studies whose AUC is nearly 0.93, which attributes to better performance.

Changes in the text: we have modified our text as advised (see Page10, line 230-236)

Comment 2- Regarding discriminative ability, Authors report on AUC, but give no data regarding sensitivity and specificity, which would be of interest to judge on clinical utility of the tools. These data should be added, if available for all studies.

Reply2: We thank you for pointing out this weakness of the present study. We regard sensitivity and specificity as good measures to express model discriminative abilities. But the probability for delirium development was divided into groups, like very low, low, moderate, and high risk of delirium. Sensitivity and specificity were calculated for these different groups. Based on the available data, we cannot derive sensitivity and specificity for all studies. In the revised manuscript, we have discussed these limitations.

Changes in the text: we added limitation. (see Page 12, line 293-298)

Comment 3- Moreover they state in the introduction that "the premise of effective prevention was to evaluate the risk of delirium, and to give different targeted interventions according to the risk degree of delirium" (lines 52-54), and again later "take corresponding preventive and nursing measures according to risks" (lines 58-59). These statements seems largely unsupported by data and should be presented as research hypotheses. In fact no data are available, as far as I know, that delirium strategies would be more effective in subjects at higher risk, and actually the opposite may be true. Moreover, the ABCDEF bundle, that Authors correctly cite as delirium prevention strategy, has been proven effective in improving a vast array of outcomes, including mortality (Pun BT, Caring for Critically Ill Patients with the ABCDEF Bundle: Results of the ICU Liberation Collaborative in Over 15,000 Adults, Crit Care Med, 2018). Due to the citied reasons, and the heterogeneity of meta-analysis results, with sensitivity estimates of PRE-DELIRIC which seem to be around 75% according to available literature, I would state clearly in the conclusion that an universal application of preventive intervention, irrespectively of delirium risk score, would be advisable at the moment. Conversely, the statement "that careful consideration should be given" (line 260) is very ambiguous and should be modified and clarified. Moreover, I would clarify the statement "At the same time, high-quality RCT research is needed to explore the economic and social benefits of the clinical application of the model, thus providing research basis for the construction of a more accurate prediction model of delirium in intensive care units (lines 253-255). In fact RCTs should include delirium risk score as inclusion/stratification criteria, if a risk cut-off has to be chosen to deliver delirium preventive intervention.

Reply3:

Reply①Thank you very much for your helpful comments. Indeed, the ABCDEF bundle should be a universal application of the preventive intervention. In the revised manuscript, we have reorganized the "Introduction" section.

Changes in the text: we have modified our text as advised (see Page 3, line 57-58)

Reply②the statement "that careful consideration should be given"

Changes in the text: we have modified our text as advised (see Page13, line321)

Reply③Thank you for providing helpful comments on lines 253-255. We have revised it.

Changes in the text: we have modified our text as advised (see Page 13, line 310-315)

Comment 4- Regarding calibration, the Authors only report how many studies performed the assessment of external validation and the instrument used ("In terms of external validation, nine articles reported calibration, of which three (16, 21, 33) carried out Hosmer-Lemeshow test and six articles(17-19, 25, 30, 31) reported the calibration plot or belt."), but give no information regarding the results of such assessment across different studies. Although a formal meta-analysis might not be performed, I feel that a review of these data should be reported together with discriminative ability.

Reply4: Thank you for pointing out this issue. In table 3, we give the P-value of the H-L test. If $P > 0.05$, calibration was good.

Changes in the text: we have modified our text as advised (see Page 8, line 199)

Comment 5- Results of ref 31 are discussed extensively (and this is justified, due to the use of both DELIRIC and PREDELIRIC in a large sample of subjects). Yet I feel that the synthesis of results provided at lines 208-210 is inaccurate. In fact according to ref 31 E-PREDELIRIC should be used at admission and PREDELIRIC should be completed after 24 hour if delirium did not occur, to improve the detection of low risk cases. Moreover in ref 31 a suboptimal sensitivity for a screening instrument is reported, especially to deny preventive measures. (sensitivity and specificity: 60 and 65%, respectively, for the E-PRE-DELIRIC model and 69 and 66% for the PRE-DELIRIC model).

Reply5: Thanks for your supportive and constructive comments. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page11, line257-260)

Comment 6- English language is generally well understandable, but not fluent and with some mistakes. Therefore a thorough language revision has to be performed

Reply6: We have checked the writing of the manuscript carefully and made a thorough revision via an assisting language checker.

Minor suggestions include the following

Comment 7- Abstract (line 21): change "determine the current delirium risk prediction model in intensive care unit and evaluate its performance" into "compare the performance of available delirium risk prediction models for intensive care units"

Reply7: Modified accordingly

Changes in the text: we have modified our text as advised (see Page1, line21-23)

Comment 8- "The E-PRE-DELIRIC model and/or PRE-DELIRIC MODELS ARE recommended". Similarly, in several statements across the paper the past tense should be changed with a present tense. Take care of singular and plural too.

Reply8: Thank you for spotting these mistakes. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page2, line37-38)

Comment 9- Lines 137-139: a statistical reference should be given for sample size standards

Reply9: Thank you for your helpful comments. A statistical reference is according to "PROBAST" Criteria. In PROBAST question "Were there a reasonable number of participants with the outcome?". Answer Yes/probably yes: For model development studies, if the number *of participants with the outcome relative to the number of candidate predictor parameters is ≥20 (EPV ≥20). For EPVs between 10 and 20, the item should be rated as either probably yes or probably no, depending on the outcome frequency, overall model performance, and distribution of the predictors in the model. So, we only report studies that met the sample size standard (events per variable, EPV≥ 20). But for studies of EPVs between 10 and 20, we regard the item as probably yes in our review. To remove ambiguity, we change the sentence into "Five studies met the sample size standard (events per variable, EPV > 10) in model development studies".*

Changes in the text: we have modified our text as advised (see Page7, line155-156)

Comment 10- Line 139: Continuous predictors handled "unreasonably". According to what??

Reply10: Thank you for your question. According to "PROBAST" Criteria, for model development studies, it is unreasonable to change the collected continuous variables into categorical variables. But if categorical predictor groups are defined using a prespecified method. It is appropriate. For model validation studies, it is unreasonable if continuous predictors are included using different definitions or transformations, or categorical variables are categorized using different cut points, as compared with the development study.

Changes in the text: we have modified our text as advised (see Page7, line159-160)

Comment 11- Lines 175-177. Rephrase and explain that three delirium prediction models were externally validated in at least two studies

Reply11: Thank you for your helpful comments. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page8, line200; Page9, line201-203)

Comment 12- Lines 190-191 and 194-195: The explanation of discrimination and calibration might be better include among Methods.

Reply12: Thank you for your helpful comments. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page5, line114-119)

Comment 13- Line 192: explain that AUC for E-PRE-DELIRIC was <0.7 only in ref 31 (in fact the summary AUC score was >0.7). Can an explanation of this heterogeneity be attempted according to between-study comparison? IN fact ref 31, that is extensively discussed, represents a large study and provides a discriminative accuracy estimate lower than other studies.

Reply13: Thanks for your helpful comments. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page9, line219-222)

Comment 14- Lines 220-221 ("In addition, due to the limited information presented in the research report, the uncertainty of the model was increased"): please clarify.

Reply14: Thank you for your question. I mean, "Data loss will increase the model's uncertainty," which may make readers confused. I deleted it.

Changes in the text: I deleted the sentence.

Review Comments B:

I am pleased to read this report of a systematic review describing the accuracy of various models used to predict ICU delirium. The authors found the PRE-DELIRIC and E-PRE-DELIRIC models to be the most widely validated models, and both performed well. I agree that early prediction of patients at risk for delirium is important, although the authors could make this argument clearer in their Introduction. I have the following comments about this manuscript:

MAJOR CONCERNS:

Comment 1. In multiple places, the authors report the "incidence" of delirium in case-control studies. Please note that the incidence of a disease cannot be obtained from a case-control study. In a case-control design, the ratio of diseased to non-diseased subjects is determined by the investigators when they are designing the study.

Reply 1: Thank you for pointing out this basic error. We have deleted data on the "incidence" of delirium.

Changes in the text: We have deleted data on the "incidence" of delirium.

Comment 2. Please indicate whether each included study reported model development and/or model validation.

Reply 2: Thank you for your helpful comments. In this review, 14 studies reported model development, and 19 studies reported model validation. In table 3, I've published the statistical method of model development and model validation.

Changes in the text: we have modified our text as advised (see Page 7, line 173-174)

Comment 3. I question some of the criteria the authors used in their risk of bias assessment. For example, requiring > 20 events per variable (page 5, lines 137-138) is a far stricter rule than I have ever heard or seen enforced. Additionally, categorization of a continuous predictor is not an unreasonable analytic step. (Although I agree that keeping the predictor continuous is generally preferable to avoid information loss, I think the PROBAST question about appropriate handling of variables refers to whether any actions were taken that invalidate statistical assumptions of the model.) Please elaborate (perhaps in a supplement) about what other criteria were used in the risk of bias assessment, so the reader can decide if they disagree with any of the other criteria besides those listed here.

Reply 3: Thank you for your question and for providing helpful comments. In PROBAST question, "Were there a reasonable number of participants with the outcome?". Answer Yes/probably yes: For model development studies, if the number of participants with the outcome relative to the number of candidate predictor parameters is ≥20 (EPV ≥20). For EPVs between 10 and 20, the item should be rated as either probably yes or probably no, depending on the outcome frequency, overall model performance, and distribution of the predictors in the model. So, we only report studies that met the sample size standard (events per variable, EPV≥ 20). But for studies of EPVs between 10 and 20, we regard the item as probably yes in our review. To remove ambiguity, we change the sentence into "Five studies met the sample size *standard (events per variable, EPV > 10) in model development studies".*

According to "PROBAST" Criteria, for model development studies, it is unreasonable to change the collected continuous variables into categorical variables, but if it is designed as a categorical variable when collecting data, it is out of the scope of discussion. For model validation studies, it is unreasonable if continuous predictors are included using different definitions or transformations, or categorical variables are categorized using different cut points, as compared with the development study.

Changes in the text: we have modified our text as advised (see Page 5, line 135-139). We have uploaded an appendix to make the criteria clearer (see appendix 1).

Comment 4. Given the focus on PRE-DELIRIC and E-PRE-DELIRIC, the authors should spend some time introducing these models to the reader. What predictor variables are included in each model? At what time point is each model intended to be implemented? In what population was each model initially developed? When should a user choose one model over the other?

Reply 4: Thank you for your helpful comments. We have modified the corresponding text.

Changes in the text: we have modified our text as advised (see Page 10, line 237-243 247-250; Page 11, line 251-253 257-260).

MINOR CONCERNS:

Comment 5. When describing the negative consequences of ICU delirium, another important element to mention is the distress it causes for family members who witness their loved one in this altered state.

Reply 5: Thank you for your helpful comments. We have added the distress it causes for family members in the revised manuscript.

Changes in the text: we have modified our text as advised (see Page 3, line 53-54)

Comment 6. Although I conceptually understand the authors' argument that the ABCDEF bundle might be applied only in higher risk patients, most elements of the bundle are system-level policies or workflows that are either applied to an ICU or not. Is there any precedent in the literature for applying the bundle only to selected patients within an ICU?

Reply 6: Thank you for pointing this out. Indeed, the ABCDEF bundle should be a universal application of the preventive intervention. In the revised manuscript, we have reorganized the "Introduction" section.

Changes in the text: we have modified our text as advised (see Page 3, line 57-58)

Comment 7. For each of the included studies, please indicate if the patient population included medical, surgical, or mixed ICU patients.

Reply 7: Thank you for your helpful comments. We have modified patient population in table 1

Changes in the text: we have modified our text as advised (see table 1)

Comment 8. Page 4, line 91 – The included studies use multivariable regression (which have more than one independent variable and one dependent variable), not multivariate regression (which have more than one dependent variable).

Reply 8: Thank you for pointing this out. We have revised it.

Changes in the text: we have modified our text as advised (see Page 5, line 101)

Comment 9. Page 5, lines 124-125 – What do the numbers "223-2299" and "25-2178" represent? Why is the "sample size" different from the number of subjects? What do the terms "sample size for model development" and "model verification sample size" mean in this context?

Reply 9: Thank you for pointing this out. In the review, model development and validation use different participants, so we classified the number of subjects into "sample size for model development" and "model verification sample size," which may lead to confusion. We have revised it in table 1 and the text.

Changes in the text: we have modified our text as advised (see Page 6, line 143)

Comment 10. Page 6, lines 151-152 – Please clarify how one of the included papers could have used a "multiple linear regression" to predict ICU delirium, which should be a categorical variable.

Reply 10: Thank you for pointing this basic error out. We used the term "multiple linear regression" incorrectly. Indeed, in the original article, it means "the linear predictors and the intercept in a logistic regression model," and we misunderstood. In the revised manuscript, we corrected them.

Changes in the text: we have modified our text as advised (see Page 7, line 174-175)

Comment 11. Page 6, line 156 – What do the authors mean by "stratified randomization" in this context? Usually, stratified randomization is used to refer to a randomized clinical trial where the randomization process is carried out separately in two different groups (such as separate randomization of male patients and female patients).

Reply 11: Thank you for your helpful comments. We mean "random split validation" is used for internal validation. We have revised it in the manuscript.

Changes in the text: we have modified our text as advised (see Page 8, line 178-179)

Comment 12. Page 7, line 193 – AUC does not provide information about "power." It provides information about calibration.

Reply 12: Thank you for pointing this out. We have revised it.

Changes in the text: we have modified our text as advised (see Page 9, line 219)

Comment 13. Page 8, line 235 to page 9, line 242 – The discussion of the TRIPOD guidelines is a bit misfocused. Failure to follow TRIPOD guidelines may prevent readers from assessing the quality of a prognostic model, but it does not directly provide evidence of poor quality. A more relevant discussion point would be if the authors could comment on whether they had trouble assessing the quality of any included studies due to missing information. Then comment whether the source papers would have contained the needed information if the TRIPOD guidelines had been followed.

Reply 13: Thank you for your helpful comments. We have revised it.

Changes in the text: we have modified our text as advised (see Page 12, line 290-292)

Comment 14. The authors spent a lot of time talking about the ABCDEF bundle in the introduction. If that was indeed their motivation for seeking a delirium prediction model, then the ABCDEF bundle should be discussed again at some point in the Discussion. Will you conduct an interventional study applying the ABCDEF bundle selectively to patients with high risk according to the PRE-DELIRIC model?

Reply 14：Thank you for your helpful comments. Indeed, the ABCDEF bundle should be a universal application of the preventive intervention. In the revised manuscript, we have reorganized the "Introduction" section and "Discussion" section.

Changes in the text: we have modified our text as advised (see Page 3, line 62-64; Page 13, line 310-315)

Comment 15. In figure 1, please explain what "risk prediction for ICU delirium and no final model" means.

Reply15.I mean risk prediction for ICU delirium without model development and validation, and I have revised it.

Changes in the text: we have modified our text as advised (see figure 1)

Comment 16. Table 4 – Why is the EPV included in this table? It is reasonable to consider EPV if a model has not been validated or has only undergone internal validation. In that case, high EPV could signal a high risk for overfitting, meaning the model will not perform well in an external dataset. But once external validation has been performed, the model performance in the validation dataset is a much better indicator of whether the model was overfit to the development data.

Reply16. We are very grateful for identifying the potential weakness of including EPV in the table. We have now removed this in table 4.

Changes in the text: we have modified our text as advised (see table 4)

Comment 17. In addition, this manuscript would benefit from language editing services to improve both grammar and diction.

And sorry for the language problem, we had improved the language using a language checker.