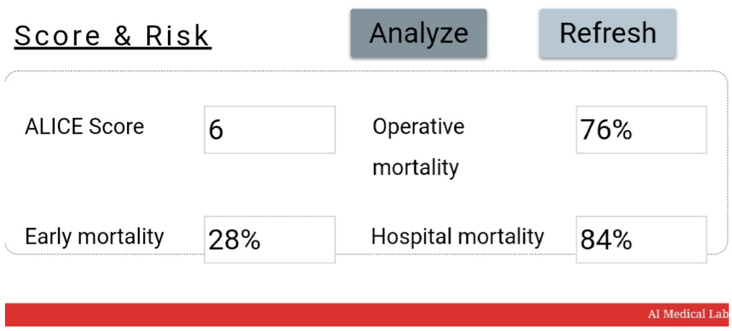
Peer Review File

Article Information: http://dx.doi.org/10.21037/cdt-20-730

Reviewer A

Comment 1: Please integrate the recent publication of Czerny et al. and discuss the results in comparison

Reply 1: Thank you very much for reminding us of the recently published GERAADA score. This new score was established based on a retrospective registry database in German. This new score comprised of 15 variables, including a lot of very specific high-risk events, such as resuscitation, intubation, vasopressors. These specific events usually happened concurrently and would no doubt increase the specificity but at the cost of sensitivity. The accuracy of the final model (AUROC: 0.725) was lower than some previous scores (AUROC:0.74-0.77), let alone our ALICE score (AUROC:0.85). Besides, many items in the new score lack of clear definition, such as coronary mal-perfusion, visceral mal-perfusion and peripheral mal-perfusion. Physicians may have their own interpretation or standards of these parameters, therefore may cause confusion when they decide to use this new score system. In our opinion, it's better to use direct numerical values and well-defined parameters. Third, the new score was developed and validated only with the same registry database. A convincible new score should be externally validated with different database. Our ALICE score was developed with a database in east China and validated with data from another cardiovascular center in west China. Fourth, this new score only focused on postoperative mortality. Our ALICE score demonstrated good accuracy in predicting both pre-operative mortality and post-operative mortality (as shown in the following picture).



(Our website tool: *http://www.aimedicallab.com/tool/alice_en.html*)

Changes in the text: We further discussed this article in our manuscript as advised. We reduced, optimized and integrated the third and fourth paragraphs in discussion section as following: "*In the proposed model, the involvement of the iliac arteries was introduced as a new predictor. TAAD extending to the iliac arteries is the most severe Debakey type I aortic dissection. In this prospective study, involvement of iliac artery was associated with more tearing of renal, mesenteric, and supra-aortic arteries, and could be considered as a parameter of dissection progression. Recently, Czerny et al(23) developed a scoring system to predict the postoperative 30-day mortality based on a*

German Registry database (GERAADA score). In this study, dissection extending to descending or further downstream was also confirmed as a significant risk factor (OR: 1.443, p=0.005). Some reported variables were not adopted in our model. On the one hand, collinearity widely exists among candidate variables (Supplementary Material, Fig. S2). For instance, cTnT improved the accuracy of predictive model, despite correlation with transaminases. On the other hand, very specific high-risk events, such as iatrogenic dissection(5), massive pericardial effusion(22), resuscitation(5, 8, 23), intubation and vasopressors(23), would increase specificity at the cost of lowering sensitivity. Although the GERAADA score(23) used many specific events as risk factors, this score did not generate better performance (AUROC: 0.73), comparing with previous works (AUROC: 0.74-0.77)(7-11)."

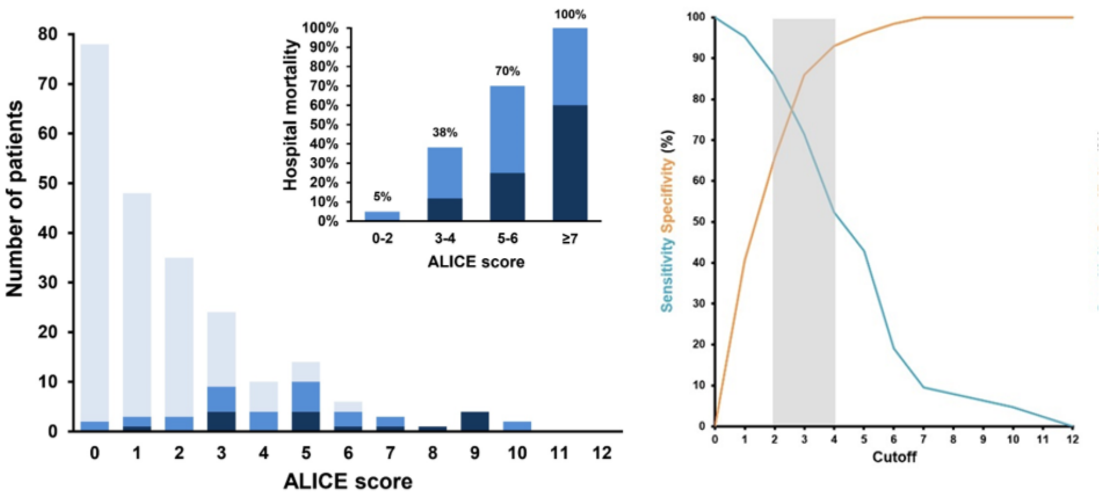
Comment 2: please provide 30 day mortailty, aus usual in surgical publication for better comparison
Reply 2: Thanks for your comments. As we calculated, of the 29 patient died postoperatively, 27 patients died within 30 days after surgery. The post-operative mortality and the post-operative 30-day morality were very close. As you suggested, we added this postoperative 30-day mortality in the manuscript and the **Table 1**.
Changes in the text: We added a row in Table 1: "*30-day mortality 27(13) 20(13) 7(13) 0.823*" and a sentence in the manuscript (p6l20) "*Of these, 27 patients died within 30 days (Table 1).*"

Comment 3: Table 1: what is the meaning of "composite endpoints" ?
Reply 3: Thanks for your comment. Composite endpoint was frequently used in medical researches as an exploratory outcome, especially when the incidence of the primary endpoints was low. In this study, we defined a composite endpoint as hospital mortality or prolonged ICU stay (i.e. >30 days), whichever happed earlier. This means patients who died during hospital stay or stay in hospital for more than one month would be considered to meet the criteria of the "composite endpoints".

Reviewer B

Comment 1: The authors developed and validated a predictive score with much better accuracy. The ALICE score has good specificity and sensitivity at a cut off of 3, which is relative low considering the full range of the score (0 to 12).

Reply 1: Thanks for your question. A relatively low best cutoff is very common. For example, the best cutoff of SOFA score for the diagnosis of sepsis is only 2 compared with a full range of 24. The largest value of a predictive score indicated the most extreme scenario. In this study, a cutoff of 3 is relatively low, considering the full range of 12. However, a higher cutoff would have better specificity at the cost of lower sensitivity (in the gray-zone analysis). Besides, as shown in the Figure 3, there is a big rise of mortality from score of 0-2 to 3-4, suggesting a cutoff of 3 is reasonable. Patients with higher risk score did not destine to death. Therefore, we believe the stratification performance of a predictive model is of greater importance. Our ALICE score has done this particularly well, the mortalities distributed with a stair-step shape (5%, 38%, 70%, and 100% for patients with a score of 0-2, 3-4, 5-6 and ≥7, respectively).

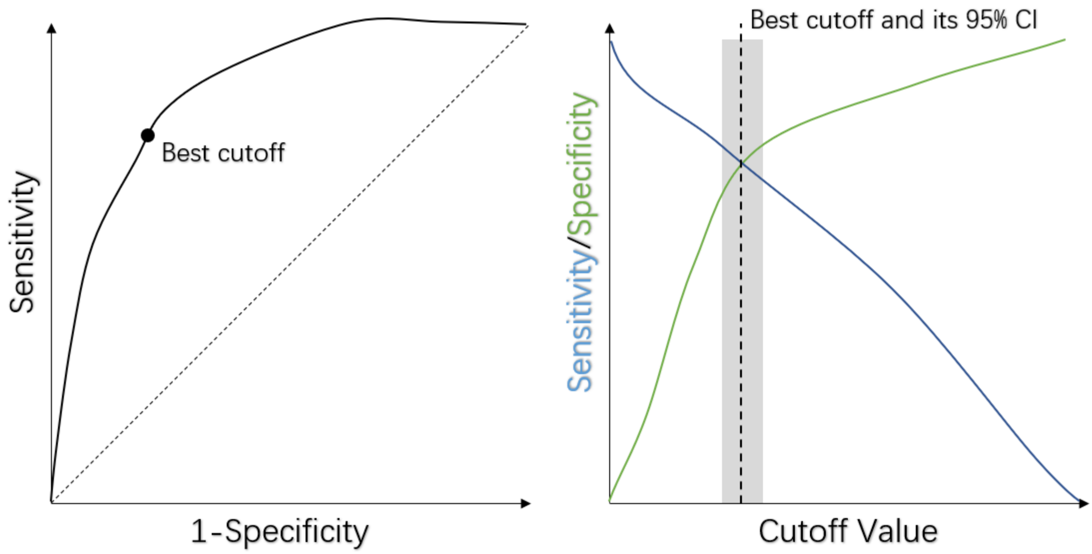


Comment 2: Preoperative acidosis is also a parameter of critical ill state, why did not the authors consider the preoperative pH as a candidate variable.

Reply 2: Thanks for your excellent question. Severe acidosis is a very specific parameter of circulatory failure and critical illness. However, we did not use pH for the following reasons: (1) pH of blood gas reflects the metabolic and respiratory status of a patient. A metabolic acidosis, however, has a corresponding respiratory compensation. Due to pain, discomfort, fear and anxiety, patients were likely to have an increased drive of breathing, therefore lead to lower PaCO2 and higher pH. (2) By contrast, the lactate is rarely affected by breath. We often observed normal pH in many hyperlactatemia patients. For example, a 46-female aTAAD patients (ID 1236330) admitted to our ICU with a lactate of 8.3 mmol/L and a BE of -10.4 mmol/L while a pH of 7.32 under a respiratory compensation ($PaCO_2$ 28 mmHg), therefore died before surgery unfortunately. (3) The pH was not contained in previously published scores, therefore, was not collected in our dataset.

Comment 3: Table 4 is confusing to me. I don't understand grey zones. Please provided some instructions about this method.

Reply 3: Thanks for your good question. Let me introduce the grey zone analysis to you. For a predictive score or a quantified parameter, we can calculate the sensitivity and specificity at each degree to draw a ROC curve. For a predictive parameter, the sensitivity decreased while specificity increased with the cutoff value. The best cutoff value was determined by Youden's Index (Sensitivity+Specificity-1). We can estimate the 95% CI of AUROC, sensitivity, specificity, PPV and NPV at the best cutoff. Here raises a question: how can we estimate the variability of the best threshold? The grey zone analysis approach was proposed to answer the question. Briefly, the grey-zone is the 95% CI of the best threshold calculated by bootstrap resampling method. The range of grey-zone suggested how many patients would risk misclassification with the given cutoff value. This grey zone has become a popular and necessary part for estimation of a predictive parameter in recent years[1-3].



[1] Georges D, de Courson H, Lanchon R, Sesay M, Nouette-Gaulain K, Biais M. End-expiratory occlusion maneuver to predict fluid responsiveness in the intensive care unit: an echocardiographic study. Crit Care 2018;22:32.

[2] Xu LY, Tu GW, Cang J, Hou JY, Yu Y, Luo Z et al. End-expiratory occlusion test predicts fluid responsiveness in cardiac surgical patients in the operating theatre. Ann Transl Med 2019;7:315.

[3] Yonis H, Bitker L, Aublanc M, Perinel Ragey S, Riad Z, Lissonde F et al. Change in cardiac output during Trendelenburg maneuver is a reliable predictor of fluid responsiveness in patients with acute respiratory distress syndrome in the prone position under protective ventilation. Crit Care 2017;21:295.

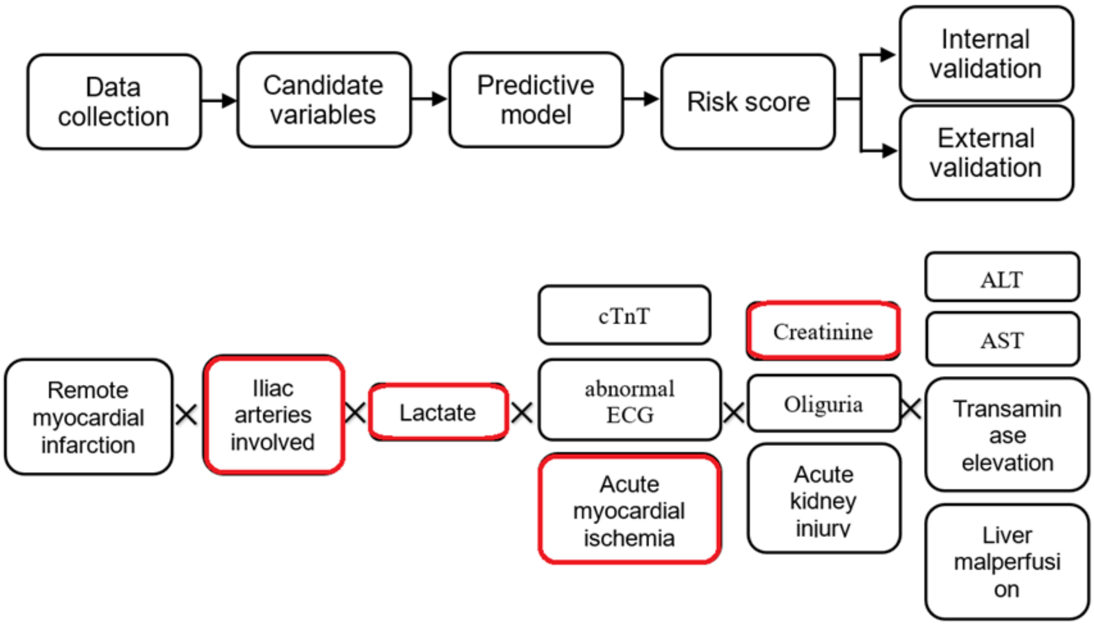
Comment 4: The authors split the data with a ratio of 7:3 rather than 1:1 for development and validation of the new score. Was there a precedent for using such a ratio?

Reply 4: Thanks very much for your comment. Generally, the training set required more population to derive a model. A ratio of 7:3 to split the dataset could guarantee an adequate sample for model derivation and leave enough population for internal validation. In the chapter "logistic regression" of book "Mastering Machine Learning with R", the ratio 7:3 was recommended to split data into train and test sets for building and examining a logistic model. Besides, according to a recently published review[1], it also mentioned: "It should be noted here that we can also randomly divide the data set into a training set and an internal validation set according to a 7:3 ratio".
Both internal and external validation is necessary. According to the TRIPOD statement, a reliable prediction model should be examined internally and externally[2]. The internal validity reflects the reproducibility of the model, while external validity reflects the generalizability of the model and needs to be validated with data sets not from the study itself, which are temporally and geographically independent or completely independent[1]. Internal and external validation of the model are necessary steps to assess the stability and applicability of the model[1]. For example, Duan, et al[3]. developed a risk score for noninvasive failure. This score was also validated internally and externally. Besides, the internal validation ensured the usability of the model in our own center.
The idea of merging these two datasets as a bi-center cohort for external validation of previous scores is creative. We did not do this for the following reason: The Zhongshan Hospital dataset was prospectively collected while the Xijin Hospital dataset was retrospectively collected. Therefore, there were a lot of missing variables in the Xijin Hospital dataset, making it difficult to validate all these published risk scores.

[1] Zhou ZR, Wang WW, Li Y, Jin KR, Wang XY, Wang ZW et al. In-depth mining of clinical data: the construction of clinical prediction model with R. Ann Transl Med 2019;7:796.
[2] Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 2015;162:W1-73.
[3] Duan J, Wang S, Liu P, Han X, Tian Y, Gao F et al. Early prediction of noninvasive ventilation failure in COPD patients: derivation, internal validation, and external validation of a simple risk score. Ann Intensive Care 2019;9:108.

Comment 5: Can the authors extent and provide a bit more information on the bivariate analysis?
Reply 5: Thanks very much for your comment. Bivariate analysis, which explored the relationship between dependent variable (the outcome) and independent variable, played a very important role in choose candidate variables. In this step, patients were divided into two parts: died and survived. Then, all the variables were compared between these two parts. Variables with $p<0.05$ were considered as candidate risk factors for the predictive models. Given the collinearity among some kind of variables, we put these variables in the regression with different combinations ($1\times1\times1\times3\times3\times4=36$)

and choose the combination with the lowest AIC as the final model.



Comment 6: "Xijin hospital" should be "Xijing hospital" in line 27 page 5.
Reply 6: Thanks for your comment. We carefully checked the manuscript to correct spelling mistakes.
Changes in the text: We corrected the spelling mistake as you pointed out.