Peer Review File

Article information: https://dx.doi.org/10.21037/jmai-22-35

Reviewer A

I congratulate you on the greatness, importance and quality of the work carried out. And I have the following constructive comment: it would be interesting to know if this methodology is equally efficient for other surgical procedures in congenital heart diseases.

Reply to Reviewer A:

Thank you for your comments and we indented to test other surgical procedures in the near future.

Reviewer B

The authors developed a convolutional neural network model that predicts the clinical stability of patients who underwent the Norwood operation. Considering that there's a clinical demand for tools to quantitatively and continuously assess these patients' unstable conditions after the operation, it is important to test the hypothesis that a deep-learning-based model that inputs various modalities of clinical data, including ECG, predicts the clinical stability of patients. However, because of the design of the study, particularly the output of the model, the model developed in the study does not provide sufficient significance to the hypothesis.

In this study, it is stated that the CNN model was designed to discriminate critically ill patients from stable patients. However, the model was actually trained to predict if the input ECG waveform is recorded within 48 hours postoperatively or within 24 hours before transfer. Because the ECG after the Norwood operation can be characterized by many factors, including cardiac arrest during the operation, pericardial effusion or hemorrhage, postoperative pericarditis, and paralytics/sedation, it is assumed that the prediction of the model presented in the study is likely affected more by such parameters rather than patients' stability itself. As a corroboration of this, the change in predicted scores happened later than the change in the clinical observation score in some patients shown in Figure 5. Considering that the clinical observation score was calculated mainly based on interventions already performed on patients except for lactate, the predicted score, which is purely derived from physiological and clinical values, should change preceding the clinical observation score if it reflects the clinical status of patients. The authors need to address this issue by adding data to support their conclusions or re-design the CNN model to predict patients' clinical instability directly.

Response:

Thank you for this comment.

When providing clinical care for these patients in the intensive care setting, it is clear that the initial 48 hours are the most critical, where the majority of hemodynamic

instability and risk of death occurs. Patients gradually progress during their ICU stay with the removal of the breathing tube, initiation of feeds, etc., eventually getting to a clinical state that is stable and ready for transfer. We sought to see if a CNN model can distinguish patients from a critical state to a stable state prior to transfer. As we see clinical progress occur, we suspected this could also be detected by an AI/ML model, progressing from critical (0) to stable (1). We trained the model on 45 Norwood patients as they progressed from critical to stable and then tested it on a separate cohort of 10 Norwood patients. We were very interested in the results for each postoperative day as seen by the model, aiming to develop another marker of clinical stability for the care of these complex patients.

While it is true that the retrospective clinical observation score was calculated based on interventions already performed, and it is logical that the predicted score from the model changes prior to the clinical observation score, that didn't happen in all observations but was observed in several instances. Achieving such precision would require dedicated clinicians to score and label patients' data closely and at short intervals. We anticipate using this algorithm to augment clinicians' decision-making and not to substitute it.

In addition, I would suggest that the authors consider addressing the following issues:

1. Provide patients' characteristics for class 0 and class 1, respectively, to let readers assess the possibility of biases in input data.

Response:

Each patient in the training model was labeled class 0 and class 1. Therefore, demographic table 1 has all characteristics of these patients. To delignate training vs. test patients we will provide a supplemental table separating the 45 training patients and 10 test patients and their characteristics.

We will add a supplemental table 1.a

	Training Cohort (N =45)	Validation Cohort (N=10)	p-Values
Age at Norwood procedure, days	5 ±3.6	6.9 ± 3.1	0.5100
Gestational age at birth, weeks	38 (35.4-41.1)	38 (36-39.3)	0.1663
Postmenstrual Age at Norwood Procedure, weeks	38.7 (36.1-42)	39 (36.7-40.5)	0.2895
Weight at Norwood procedure, kg	3 ± 0.5	2.9 ± 0.5	0.2499
ICU length of stay, days	20.3 (7-85)	11.2 (7-28)	0.0990
Sex			0.9234
Male, n (%)	26 (57.8)	5 (50)	
Female, n (%)	19 (42.2)	5 (50)	

Norwood			0.7570
BTs, n (%)	32 (71.1)	6 (60)	
RV-PA, n (%)	13 (28.9)	4 (40)	
Anatomy			0.9252
HLHS, n (%)	38 (84.4)	8 (80)	-
Unbalanced AVSD, n (%)	3 (6.7)	1 (10)	-
Others	4 (8.9)	1 (10)	-
HLHS subtype			0.8925
MS/AS, n (%)	14 (36.8)	3 (37.5)	-
MS/AA, n (%)	11 (28.9)	3 (37.5)	-
MA/AA, n (%)	11 (28.9)	2 (25)	-
MA/AS, n (%)	2 (5.3)	0	-
Required ECMO immediate postoperatively	4 (8.9)	1 (10)	1.0
Disposition from ICU			
Transferred stepdown, n (%)	39 (86.7)	9 (90)	0.6650
Mortality, n (%)	8 (17.8)	1 (10)	1.0

Table 1.a. Patient Demographics and Characteristics. Means (25-75 percentiles) are reported for continuous variables. Frequencies (percentage) are reported for categorical variables and T-test on age, gestational age, weight, and length of stay. A Chi-square test contingency is used on the rest of the rows for the significance test. RVPAS = right ventricle to pulmonary artery shunt; BTS = Blalock-Taussig shunt; kg = kilograms; HLHS = hypoplastic left heart syndrome; AVSD = atrioventricular septal defect; MS = mitral stenosis; AS = aortic stenosis; AA = aortic atresia; MA = mitral atresia; ECMO = extracorporeal membrane oxygenation; ICU = intensive care unit

2. While the authors mentioned the lack of interpretability in CNN models, visualization of the model, such as grad-CAM, needs to be performed because there's a concern for significant biases in the data.

Response: Thank you for this comment. Grad-CAM is not applicable to ECG signals since grad-CAM was designed for 2-D signals

3. What is the structure of the unified CNN model for all three ECG leads?
Response: Please refer to (Fig 2)

4. Which data were used for the development of the model for Group 4? If data from 10 patients in the test group were used, the AUC of 0.98 in Group 4 has possibly been overestimated compared to the other groups.
Response:

Data from the 45 patients included in the training model were used in all groups. The 10 test patients' data wasn't entered in the development of the model.

5. Page 4, Line 78: "While much effort has focused solely on prevention of significant clinical events, very little data exists on detecting changes (both positive and negative) in a patient's clinical status." This seems to explain the study's originality, but it seems to require a more concrete description.

Response:

We substituted the sentence "While much effort has focused solely on prevention of significant clinical events, very little data exists on detecting changes (both positive and negative) in a patient's clinical status'' with the following:

"For high risk patient populations, a great deal of effort is focused solely on the prevention of significant clinical deterioration events. There is very limited ability to detect or quantify subtle changes in clinical status (both positive and negative) in a patient's clinical status over time."

Changes made to page 4- lines 78-81.

6. Page 13, Line 265: "The results of the present study achieved an AUC-ROC of 0.98 that focused on discriminating between clinical wellness and clinical instability, rather than focusing on detecting only deterioration events." Also, this discussion should be addressed more. It doesn't seem appropriate to compare the AUC value of 0.98, which is for detecting data recorded within 48 hours after surgery, to the AUC values in the previous studies that focused on prediction of patients' prognosis.

Response:

The AUC value of 0.98 was not only reflecting data from the first 48 hours post operative. It included date of the last 24 hours prior to transfer to stepdown.

7. Page 12, Line 244: "Closer correlation was observed in the first two days postoperatively and the final day before transfer to the floor." Provide objective results for this statement.

Response:

On review, we believe that this statement does not reflect how the results are interpreted, so we will remove this statement.

8. Page 14, Line 277-281. What was the reason why the patient remained in the ICU despite the high clinical score? Because the CSS was not used prospectively for the decision-making of this patient, it is not reasonable to assume that the patient was not transferred to the floor because of "clinical uncertainty," which clinicians did not recognize.

Response:

Since the clinical observation score was conducted retrospectively, the score could be high, but the decision to transfer the patient to stepdown remained the clinician's decision at that time, which was difficult to capture from the retrospective scoring. Clinician scoring could have easily not recognized the reason behind the uncertainty. As clinicians, we are faced with many patients who, on chart review, they seem ''fine,'' but during exam or observation of subtle changes as well as the "gut feeling of the clinician,'' we decide to observe longer in the ICU.

9. Page 14, Line 284: "The goal, therefore, of adapting this AI model to …." The logical connection of this sentence to the preceding is unclear.

Response:

Thank you for pointing that out. We decided to change it to start a new paragraph. Page 14, line 285, as follows "The goal, therefore, of adapting this AI model to a real-time assessment at the bedside would be to indicate wellness as a prompt to progress clinical care, as well as indicate deterioration as a prompt to investigate possible changes to clinical care that are needed. This represents a novel application compared to the existing clinical applications of AI.''

10. Figure 5: To show the clinical course of the patients in detail, consider adding plots for the timings of clinical events, including extubation, discontinuation of inotropes and sedatives, chest closure, and other interventions.

Response:

Thank you for this suggestion. We a modified figure 5 to show the correlation between the clinical stability score, retrospective clinical observation score and clinical events demonstrated as plots throughout the ICU stay.

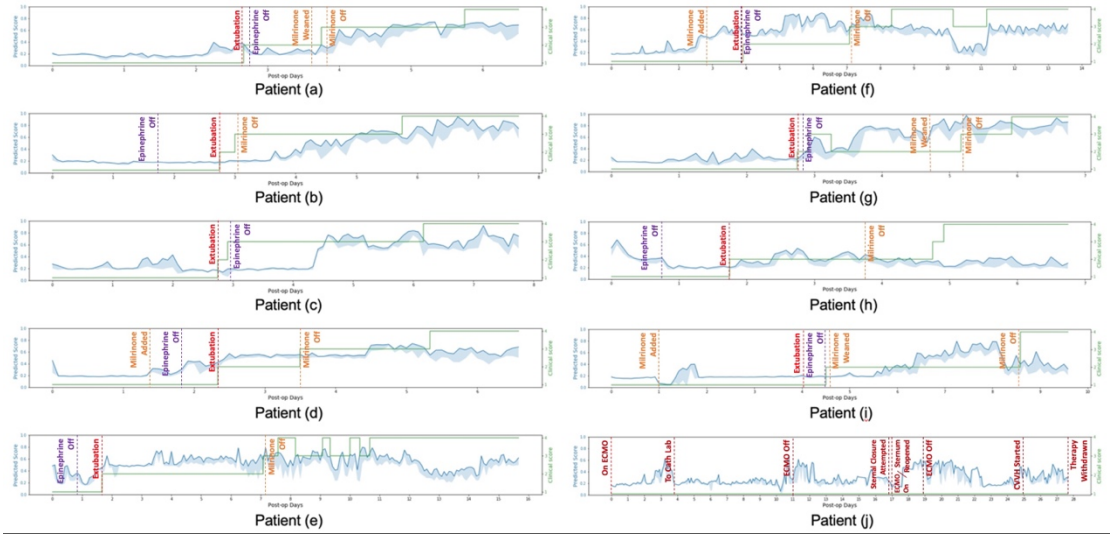

Figure 5. Performance of the final model for each patient in the test cohort. The blue line in each pane represents the linear transformation of the final model (per 2 hours), with the shaded area representing the 95% confidence interval. The clinical stability

score (0-1) is on the left y-axis as the Predicted Score. The green line represents the retrospective clinical observation score (0-4) and is on the right y-axis as Clinical Score. Clinical events are shown as plots throughout the ICU stay; red= extubation, purple=discontinuation of epinephrine infusion, and yellow= milrinone infusion (addition and discontinuation). Patient J uniquely shows other clinical events (ECMO initiation and decannulation, attempted chest closure, and CVVH initiation)

11. Table 1: While the type of Stage I palliation (RVPAS/BTS) is clinically important in this cohort, this information doesn't seem to provide much significance for the study because the patients were not separated based on it. Consider presenting a similar table for the training and test groups.
The objectives presented in the study are of significant importance and interest. I hope the authors find these suggestions valuable.

Response:
Thank you for this suggestion. We added supplemental table 1.a

	Training Cohort (N =45)	Validation Cohort (N=10)	p-Values
Age at Norwood procedure, days	5 ±3.6	6.9 ± 3.1	0.5100
Gestational age at birth, weeks	38 (35.4-41.1)	38 (36-39.3)	0.1663
Postmenstrual Age at Norwood Procedure, weeks	38.7 (36.1-42)	39 (36.7-40.5)	0.2895
Weight at Norwood procedure, kg	3 ± 0.5	2.9 ± 0.5	0.2499
ICU length of stay, days	20.3 (7-85)	11.2 (7-28)	0.0990
Sex			0.9234
Male, n (%)	26 (57.8)	5 (50)	
Female, n (%)	19 (42.2)	5 (50)	
Norwood			0.7570
BTs, n (%)	32 (71.1)	6 (60)	
RV-PA, n (%)	13 (28.9)	4 (40)	
Anatomy			0.9252
HLHS, n (%)	38 (84.4)	8 (80)	-
Unbalanced AVSD, n (%)	3 (6.7)	1 (10)	-
Others	4 (8.9)	1 (10)	-
HLHS subtype			0.8925
MS/AS, n (%)	14 (36.8)	3 (37.5)	-
MS/AA, n (%)	11 (28.9)	3 (37.5)	-
MA/AA, n (%)	11 (28.9)	2 (25)	-
MA/AS, n (%)	2 (5.3)	0	-
Required ECMO immediate	4 (8.9)	1 (10)	1.0

postoperatively			
Disposition from ICU			
Transferred stepdown, n (%)	39 (86.7)	9 (90)	0.6650
Mortality, n (%)	8 (17.8)	1 (10)	1.0

Table 1.a. Patient Demographics and Characteristics. Means (25-75 percentiles) are reported for continuous variables. Frequencies (percentage) are reported for categorical variables and T-test on age, gestational age, weight, and length of stay. A Chi-square test contingency is used on the rest of the rows for the significance test. RVPAS = right ventricle to pulmonary artery shunt; BTS = Blalock-Taussig shunt; kg = kilograms; HLHS = hypoplastic left heart syndrome; AVSD = atrioventricular septal defect; MS = mitral stenosis; AS = aortic stenosis; AA = aortic atresia; MA = mitral atresia; ECMO = extracorporeal membrane oxygenation; ICU = intensive care unit

Reviewer C

The authors are to be commended for trying to use ML to evaluate critically ill infants post the Norwood procedure

I have some comments that I feel should be addressed
1) I like the pragmatic method of labelling that the authors used. However, it was not obvious from the text if the data between day 2 and the last ICU data - please clarify.
Response:
The data from the first two days post operatively and the one day prior to transfer to step down were used to train the model. When testing the model on the unique10 patients, we applied the model on all post operative days until transfer to step down.

2) I found the description of preparation of ECG data slightly confusing - how was data combined when there is >50% missingness on lead.
Response:
If there is any missingness on one lead, then that lead will be discounted for the duration, and the data from the other two leads will be used to produce the score. The model didn't produce any score if all leads were missing for the same duration, but that didn't occur in our test data set.

3) The comparison of ML models (CNN vs RF/LR) is not a fair comparison. Why not use a neural network for prediction of clinical state from Lab data and Vital sign
Response:
The data containing labs and vital signs is not sufficient to train neural network which can lead to the network to be underfitted

4) It is unclear how lab and vital sign data was combined for the final model (or ho 3

lead data was combined. The best way would be concatenate the direct data to the flattened data between the CNN and DENSE network

Response:
The way we combined all modalities is by averaging the predictions from different models. Please refer to page 11- lines 230-232

5) It is slightly odd that in the test data the truth is between 1-4 while in training the choice is binary. I understand this partly equates to the probability of stability but why not use the more granular measure during training and reformulate as a regression problem.

Response:
That means we need to manually label all 45 training patients throughout their ICU stay by multiple clinicians which is currently not feasible

This is an interesting paper - I have some concerns about whether the comparisons are completely fair.