Peer Review File

Article information: https://dx.doi.org/10.21037/atm-21-1564

Reviewer A

The authors propose automated methods of classifying abnormality in and estimating the age of EEG recordings from infants in the NICU. The method is developed on a very large cohort (<1000 infants) of neonatal EEG recordings. The integration of EEG into the NICU is of high importance and automated methods of interpretation are promising solution.

Comments

Comment 1: The authors should include summary statistics of the post-natal age of EEG recording. This may explain why the EEG classifications does not seem to be able to differentiate infants with the range of conditions as outlined in Table 1.

Reply 1: Thanks for the reviewer's comments. Firstly, we apologize for the possible misunderstanding caused to the reviewer. In the original manuscript, Table 1 was only made to show the baseline information of the cohort, without the purpose to investigate the correlation between the EEG report conclusion and disease systems. Thus, we revised the Table 1 by removing the information of clinical systems to Table S2 which may mis-lead the understanding of the main purpose of our study. Besides, according to the reviewer's suggestion, we add the summary statistics of post-natal age in Table 1 and in Table S2, we also checked the tendency of EEG report conclusion in each diagnosed disease systems. The results showed that HIE, Central nervous system infection, Congenial metabolic disease and unexplained convulsions tend to have moderate and severely abnormal EEG report conclusion, while temporary metabolic disorder tend to be normal.

Changes in the text: In the revised version, we revised Table 1 and added Table S2. We also added related description in the first part of the result (1. *Benchmark EEG dataset*).

Comment 2: There is some concern with the use of the EEG classification system of Liu. While its constituents are in line with several neonatal EEG classifications systems such as Murray et al. 2009 (Murray DM, Boylan GB, Ryan CA, Connolly S. Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. Pediatrics. 2009; 124: e459-67.) and Hellstrom Westas et al. 2006 (Hellström-Westas L, Rosén I, De Vries LS, Greisen G. Amplitude-integrated EEG classification and interpretation in preterm and term infants. NeoReviews. 2006; 7: e76-87.). It is concerning that this classification system does not seem able to differentiate infants with a range of conditions as outlined in Table 1.

Reply 2: Thanks for the reviewer's comments. Firstly, Liu's standard has been widely used and promoted in China and Liu's standard in neonatal is generally translated from the descriptions in Ebersole et al. [Reference: Ebersole, John S., and Timothy A. Pedley, eds. *Current practice of clinical electroencephalography*. Lippincott Williams & Wilkins, 2003.], page 205, TABLE 6.3. One of the main purposes of this study was to train an automatic predictive model which could give a report conclusion from the original EEG signal data. We do recognize that there are certain differences between the existing standards. If the report conclusion from the model-developing dataset was based on Liu's standard, the generated model could only applicable to medical institutions that use Liu's standard in clinic. This was one of the major limitations, which we have pointed out in the original manuscript "*One of the major limitations in our system is the criteria used in the clinical report generation to train the prediction model, rendering it could only generate reports that follow Liu's guideline*." Secondly, the reviewer's main concern is that the EEG report conclusion cannot differentiate infants with different clinical system. The clinical system is one of the influence factors that cause different abnormalities in EEG recordings. The tendency occurs but may not be the only decisive factor. Our main purpose is to relate EEG signals with EEG report conclusion, and clinical system is currently not our prediction task. As the reply in the first comment, we have modified Table1 and added Table S2 to remove possible mis-leading and add the tendency analysis.

Changes in the text: Beside the same changes in replying the first comment, we also added some description in the revised discussion part.

Comment 3: The arbitrary split of the data into training and testing sets can be overcome by using K-fold cross-validation at this stage – the authors should use K combinations of training and testing sets to generate the results (before they optimize the GBM).

Reply 3: Thanks for the reviewer's comments. We may not make it clear about the dataset splitting in the original manuscript. We have a total of 1,851 subjects, of which 1,591 are used for model developing, and 260 are used as an independent validation dataset. It may be our wording "testing dataset" brought misunderstandings to the reviewer. We changed it to "validation dataset" and changed the "training dataset" to "model-developing dataset" in the revised version. Among the 1,591 model-developing datasets, we applied the 10-fold cross-validation strategy to choose the best model with hyper-parameters of GBM according to the cross-validation performance (see the *Model generation and cross validation* part of the method section).
Changes in the text: We changed "testing dataset" to "validation dataset" and changed the "training dataset" to "model-developing dataset" in the revised version, including Abstract, Methods, Result, Figure1, Figure2, Table2, Table S4. Also, we modified description in the Method section (*Machine learning* part).

Comment 4: In Figure 2, the authors mention that age prediction algorithms were trained on infants with normal reports, yet the Fig 2a) shows infants with abnormal reports in the training data, please clarify? Did the authors use separate systems for age prediction and EEG grading? If so, does this mean that they also used different training sets for age prediction and EEG grade methods?
Reply 4: Thanks for the reviewer's comments. The first question, we apologize for the possible misunderstanding caused to the reviewer. The prediction model was constructed using only the samples with normal report conclusions but applied in all samples. When drawing the Figure 2A, the prediction results on all samples were displayed, not only for showing the best fit for samples with normal conclusion but also showing the deviation for the samples with abnormal conclusion. The second question, for the two prediction tasks, the data used for model-developing (model-developing dataset) and the independent validation dataset (validation dataset) are consistent. But the model-developing and validation steps are independent to different tasks.
Changes in the text: We have revised the second part of the result (*Auto-Neo-EEG could successfully estimate brain age*).

Comment 5: Table 3 is unwielding, any EEG grading algorithm should output a single grade so the results should be presented as a confusion matrix with 4 possible grades (normal, slightly, moderately, severly abnormal). Not pairwise comparisons.

Reply 5: Thanks for the reviewer's comments. Firstly, we apologize for the possible misunderstanding caused to the reviewer. Although we applied a cascade strategy to deal with multi-class prediction, it is true that each sample will have a clear predicted report conclusion level (normal, slightly, moderately, severely abnormal). Therefore, we moved the original Table3 to TableS5 and created new Table2 which contain the confusion matrix with four levels.
Changes in the text: We have moved the original Table3 to TableS5 and created new Table2 as the reviewer suggested.

Comment 6: Why did the authors only choose 24 EEG recordings for the assessment of inter-observer agreement – this seems limited.
Reply 6: Thanks for the reviewer's comments. According to the sample size calculation for Cohen's KAPPA score [reference paper: Bujang, M. A. B. N. "Guidelines of the minimum sample size requirements for Cohen's Kappa. Epidemiol." Biostatistics Public Health 14 (2017): e12267-12261.], if only consider two categories (normal vs abnormal), at least 96 samples are required (power = 90%, alpha = 0.05) to conclude that our good consistency (K=0.9) is significantly higher than moderate consistency (K=0.7). If consider four categories, at least 60 samples are required. Therefore, we added the consistency results up to 96 samples. Since the table is too long, we move it to TableS4.
Changes in the text: We have revised the original Table2 and move to TableS4, and descriptions in the first part of the result (1. *Benchmark EEG dataset*).

Minor comments
Comment 7: Introduction: Stevenson et al used over 60 preterm infants to build their age prediction model. They also compared predicated age difference for the prediction of neurodevelopmental outcome in (see)
Reply 7: Thanks for the reviewer's comments. We checked the paper we cited in the original manuscript (Stevenson, N. J., et al. "Functional maturation in preterm infants measured by serial recording of cortical activity." *Scientific reports* 7.1 (2017): 1-7.). Their initial cohort has 67 infants, with 43 infants had normal neurodevelopmental outcome. After the application of artefact detection, 39 preterm infants with a gestational age less than 28 weeks and normal neurodevelopmental outcome at 12 months of age were used to build the age prediction model. Then, we checked more papers published by the same group (Stevenson, Nathan J., et al. "Automated cot-side tracking of functional brain age in

preterm infants." *Annals of clinical and translational neurology* 7.6 (2020): 891-902.), where 65 preterm infants were used and the difference between the brain age and postmenstrual age was evaluated as a predictor the neurodevelopmental outcome. We really appreciate their work and added this one in our introduction part.
Changes in the text: we have revised the introduction and added one new citation.

Comment 8: Methods: Where did the authors place the reference electrode? In their prediction algorithm did they used a referential or bipolar montage?
Reply 8: Thanks for the reviewer's comments. The Cz was used as the reference electrode. The prediction algorithm used the signal dataset from each montage adjusted to the reference.
Changes in the text: we have revised the method section (3. *Dataset acquisition*).

Comment 9: Methods: please clarify the statements - If the result from the two experts was consistent, they three would directly apply it as the final report. If not, they three would discuss it together to make a final decision.
Reply 9: Thanks for the reviewer's comments. We apologize for the possible misunderstanding caused to the reviewer. The statement is clarified to “If the result from the two experts (X.W and Y.X) was consistent, they would directly apply it as the final report. If not, Y.Z will join and they three would discuss it together to make a final decision.”.
Changes in the text: We have revised the method section (3. *Data acquisition*).

Comment 10: Methods: trance discontinuous should be trace discontinu
Reply 10: Thanks for the reviewer's comments. We have fixed this typo.
Changes in the text: We changed to “Tracé discontinu” in the revised method (*Standard-scheme for manually reported EEG conclusion*).

Comment 11: Methods: when outlining methods of assessing agreement there is nothing in between the brackets ‘Kappa () function’, is

something missing here?
Reply 11: Thanks for the reviewer's comments. We added the brackets to show that Kappa was the function's name in the R package. We have removed the brackets and formatted it to italic in the revised version.
Changes in the text: We changed to "*Kappa*" in the revised method (*Interrater agreement assessment*).

Comment 12: Methods: Can the authors provide more detail on the artefact removal process – were all recordings adjusted using ICA analysis or were high artefact EEG recordings ignored, or just channels or segments of the recording? What were 'improper' correlations between ICA components and the EEG, was it high correlations? Was this process performed using manual annotation or automatically?
Reply 12: Thanks for the reviewer's comments. We have added the detailed artefact removal process in the revised version. ICA (independent component analysis) was performed at the last step after improper channel removed. ICA was used to decompose the EEG signal into independent components, and each component was compared to the EOG channel to identify any improper correlations. The effects of the rejected components were removed from the original data. This process was performed automatically.
Changes in the text: We added detailed description in the revised method (1. *EEG signal pre-processing and feature extraction from EEG dataset*).

Comment 13: Methods: the multi-class setup of the algorithm is a 'one vs all' approach and typically you would assign the class using the maximum posterior probability, rather than a ranked approach.
Reply 13: Thanks for the reviewer's comments. Our task to predict EEG report conclusion was a multi-label classification issue. Typically, it was an ordered multi-label classification issue. Thus, we applied a cascade strategy to transfer it into three binary-classification issue. This design was out of the following considerations: (1) finding out the severely abnormal and moderately abnormal EEG recordings were more important to clinical diagnosis and the corresponding binary-classification model could be directly applied; (2) the results showed that currently prediction model to distinguish slightly abnormal was worse than others, and if consider multi-label classification, this level may influence the model generation. (3) some predictive features were important in distinguish severely abnormal but may not useful in distinguish other levels.

Therefore, we did not adopt the one-vs-all approach in routine multi-class problems.

Besides, as reviewer suggested, we also tried to directly generate the multi-label classification model using GBM (by the same strategy of data-splitting, feature selection, cross-validation) and the confusion matrix was shown below:

Strategy	**Dataset**	**Predicted label**	**Original label**				**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**	**Accuracy**
			normal	Slightly abnormal	Moderately abnormal	Severely abnormal							
Predicted by original EEG signals and actual CA	Model-developing dataset	Normal	751	185	31	10	751	520	226	94	88.88%	69.71%	79.89%
		Slightly Abnormal	90	397	19	2	397	896	111	187	67.98%	88.98%	81.27%
		Moderately Abnormal	0	1	47	0	47	1492	1	51	47.96%	99.93%	96.73%
		Severely Abnormal	4	1	1	52	52	1521	6	12	81.25%	99.61%	98.87%
	Validation dataset	Normal	120	46	1	0	120	66	47	27	81.63%	58.41%	71.54%
		Slightly Abnormal	26	44	7	1	44	136	34	46	48.89%	80%	69.23%
		Moderately	0	0	2	4	2	239	4	15	11.76%	98.35%	92.69%

		Abnormal											
		Severely Abnormal	1	0	7	1	1	246	8	5	16.67%	96.85%	95%

Compared with the results in revised Table2, this model was better in predicting normal and slightly abnormal but significantly worse in predicting severely abnormal and moderately abnormal recordings, especially the sensitivity value.

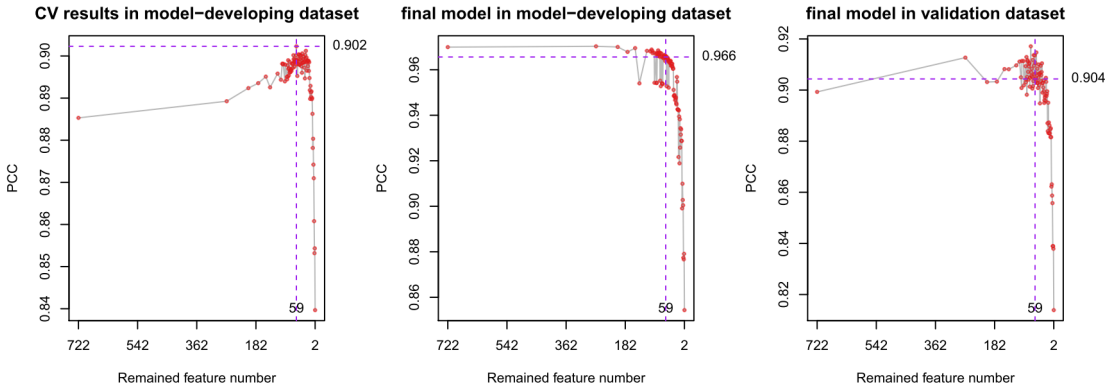
Moreover, as our task was an ordered multi-label classification issue. We checked current available models, such as ordered logistic regression (**https://stats.idre.ucla.edu/r/dae/ordinal-logistic-regression/**). However, these models require some additional assumptions that the relationship between each pair of outcome groups is the same, which could not meet the data condition of EEG report conclusion.

Changes in the text: We added the discussion of model strategy selection in the revised discussion part.

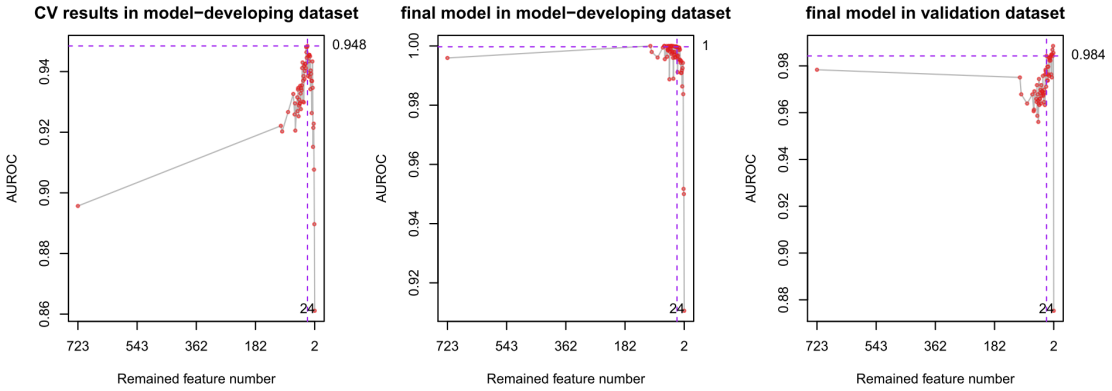
Comment 14: Methods: the process of feature selection is not clear. The authors appear to use a filter approach where filters with only high correlation with age or brain damage were included in the final feature set, and in subsets of features with high correlation with each other, the single feature that had the highest correlation with age or brain damage was selected. Have the authors considered a more systematic approach using wrappers such as backward selection?

Reply 14: Thanks for the reviewer's comments. According to the reviewer's great suggestion, we have re-done the feature selection steps by the backward selection. For each iteration, gbm model was generated and features with the minimum importance value were removed. The remained features with the best performance for the cross-validation (CV) results in the model-developing dataset were used.

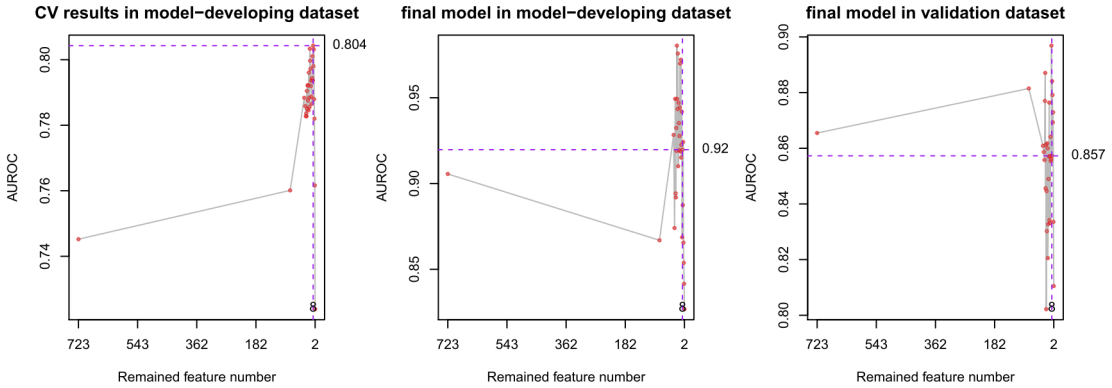
In the brain age estimation task, the performance value (PCC value) at each iteration step was shown below and 59 features were remained.



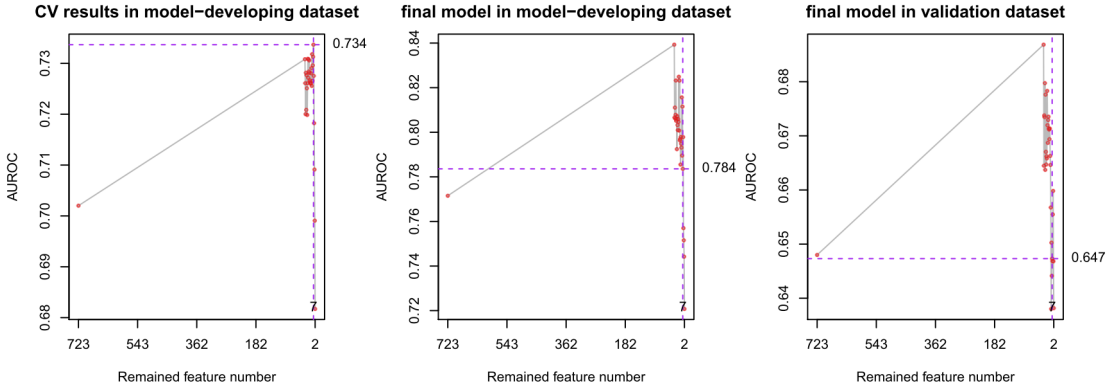
In the severely abnormal prediction task, the performance value (AUROC value) at each iteration step was shown below and 24 features were remained.



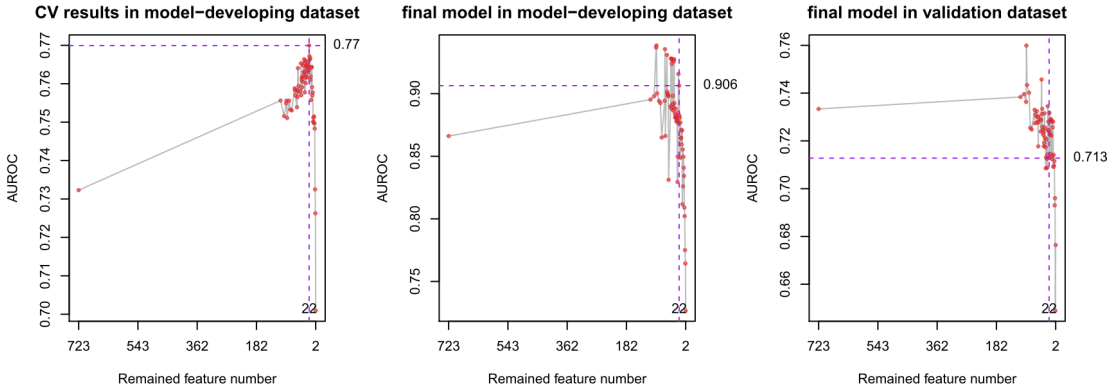
In the moderately abnormal prediction task, the performance value (AUROC value) at each iteration step was shown below and 8 features were remained.



In the slightly abnormal prediction task, the performance value (AUROC value) at each iteration step was shown below and 7 features were remained.



In the abnormal prediction task, the performance value (AUROC value) at each iteration step was shown below and 22 features were remained.



Changes in the text: We added re-written the method section (feature selection). Due to features used in the final model changed, the related results (figure2, table2， figure3， figure4, Table S5) were changed. And the result description in the Abstract, Result also changed.

Comment 15: Discussion: Awal et al. not Lai et al – see reference (18)
Reply 15: Thanks for the reviewer's comments. We have fixed this typo.
Changes in the text: We changed to “Awal et al” in the revised discussion.

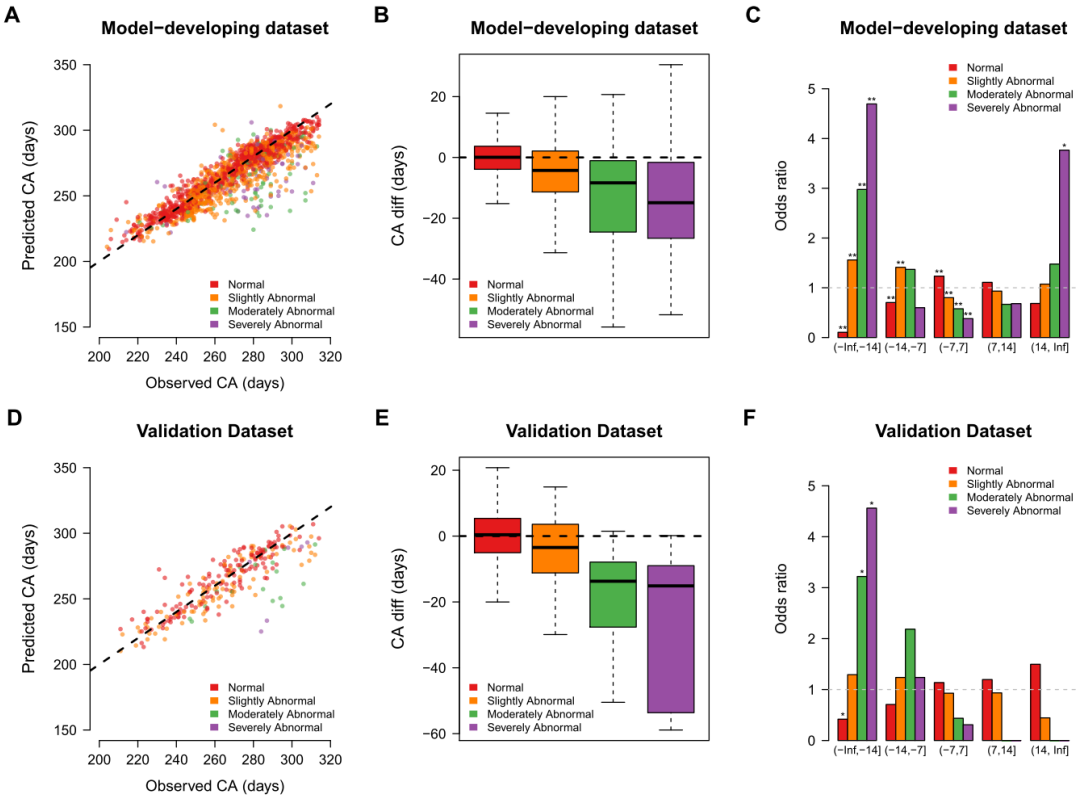
Reviewer B

The authors are to be congratulated on this study of a very large and valuable dataset of neonatal eeg and their attempts to automate reporting of this complex data using a neural network.

There is indeed a great need for automated reporting of neonatal EEG, as the authors point out expert neurophysiologists are just not available 24/7 in real time to interpret the data.

In order to be ready for publication this manuscript should be substantially revised and reorganised in my opinion.
I would recommend a different emphasis in the interpretation of the results.

Comment 1: Clinically the most useful single classification into 2 groups is to separate normal from abnormal (any severity). There is also considerable value in being able to reliably separate normal and slightly abnormal from more abnormal. These groupings allows the reviewing clinician to then focus resources for continued review of the cEEG to the subset of neonates with abnormal EEG's. This is important since the abnormality of the EEG highlights a greatly increased risk of seizures and need for continued frequent review, which is labor intensive (1)
It should be more prominently acknowledged that the automated reporting is not yet sufficiently sensitive and specific for these separation tasks. In testing the automated reporting achieved 36.28% sensitivity, 93.20% specificity and 68.46% accuracy in separating normal from abnormal of any severity. The authors should not be disheartened by this, it is a difficult problem.
The finding that discrepancy between automated estimation of corrected gestational age and actual corrected gestational age is a strong predictor of an overall finding of abnormality is the most intriguing result to me as a clinician and potentially the most useful clinically. It is difficult and time consuming to calculate interburst intervals at the bedside and to come up with a CGA suggested by the EEG. This appears to be a task that the automated reporter is naturally well suited to.
This result should be explored further or reported in more detail. For example I would be interested in odds ratios ie if the predicted CA = actual CA what is the odds this was an overall normal EEG compared with an EEG where the gap between predicted CA and actual CA was x, vs 2x vs 3x.
Reply 1: Thanks for the reviewer's comments and we really appreciate reviewer's recognition of our work. According to the suggestion, we have divided all samples into five groups according to the difference between predicted CA and actual CA (<-14, -14~-7, -7~7, 7~14 and >14) and calculated the odds ratio for each EEG report level with statistical testing. Shown below (revised Figure2):



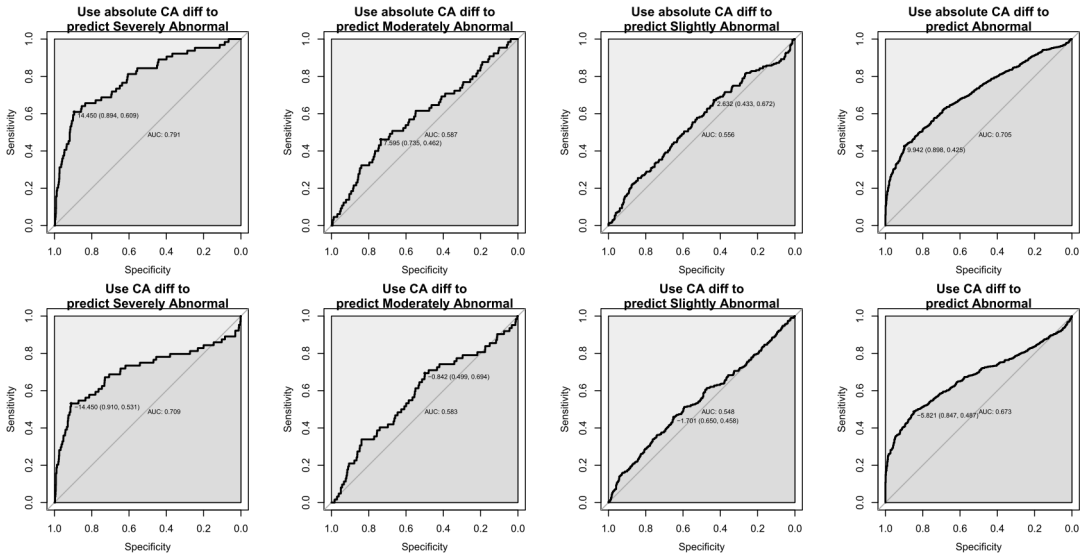
The severely abnormal samples were significantly enriched in groups with CA difference smaller than 14 days, followed by moderately abnormal. Samples with CA difference larger than 14 days were enriched in severely abnormal only in model-developing dataset.

Changes in the text: We re-generated Figure2 and modified the second part of the result section (*Auto-Neo-EEG could successfully estimate brain age*).

Comment 2: Alternatively, what is the sensitivity , and specificity of this feature alone to detect an overall abnormal EEG as interpreted by human experts.

(Of note the accuracy of the auto-neo-eeg to predict CA is not in itself particularly valuable, since we have this clinical data. What is promising/ potentially clinically useful is that the difference between predicted CA and actual CA is such a useful feature in predicting whether the overall EEG is abnormal.)

Reply 2: Thanks for the reviewer's comments. We tried as reviewer suggested by using CA diff (difference between predicted CA and actual CA) and absolute CA diff (absolute difference between predicted CA and actual CA) as the direct predictor to predict EEG report conclusions. the “Youden’s J statistic” was employed to get the optimal cut-off. The results in the model-generation dataset were shown below:



Generally, the absolute CA diff shows better than CA diff and the threshold was finally decided as:

Normal: (0,2.9]

Slightly abnormal: (2.9,7.6]
Moderately abnormal: (7.6,14.5]
Severely abnormal: (14.5, Inf]

And the statistics by using this classification model were as follows:

Strategy	**Dataset**	**Predicted label**	**Original label**				**TP**	**TN**	**FP**	**FN**	**Sensitivity**	**Specificity**	**Accuracy**
			normal	Slightly abnormal	Moderately abnormal	Severely abnormal							
Predicted by absolute CA difference	Model-developing dataset	Normal (0,2.9]	352	129	19	5	352	593	153	493	41.66%	79.49%	59.4%
		Slightly Abnormal (2.9,7.6]	321	177	16	12	177	658	349	407	30.31%	65.34%	52.48%
		Moderately Abnormal (7.6,14.5]	152	172	27	8	27	1161	332	71	27.55%	77.76%	74.67%
		Severely Abnormal (14.5,Inf]	20	106	36	39	39	1365	162	25	60.94%	89.39%	88.25%
	Validation dataset	Normal (0,2.9]	42	21	2	1	42	89	24	105	28.57%	78.76%	50.38%

		Slightly Abnormal (2.9,7.6]	51	25	2	0	25	117	53	65	27.78%	68.82%	54.62%
		Moderately Abnormal (7.6,14.5]	35	25	5	1	5	182	61	12	29.41%	74.9%	71.92%
		Severely Abnormal (14.5,Inf]	19	19	8	4	4	208	46	2	66.67%	81.89%	81.54%

Compared with results in the revised Table2, the performance was much worse in normal and slightly abnormal as the difference for CA was not that large to distinguish these two groups. The performance was relatively better in severely abnormal and moderately abnormal but still worse than using original EEG signals, showing that the difference for CA was a good marker for the abnormality prediction but was not the only affected signal features.

Changes in the text: We have added the results in Table S5 and added the description in the revised result and discussion part.

Comment 3: Figure S4: The importance value of features in the final conclusion predictions, is a very interesting figure and could perhaps be promoted into the main article.

Reply 3: Thanks for the reviewer's comments. We have moved Figure S4 to Figure 4. As the feature selection process has changed according to the Reviewer A's suggestion, the importance value of features have updated.

Changes in the text: We have moved Figure S4 to Figure 4.

Additional recommendations and comments:

Comment 4: It would be good to improve the overall readability and linguistic correctness of the text by engaging an editor
Reply 4: Thanks for the reviewer's comments. We have asked an editor to help us improve the readability and linguistic correctness of the revised version.
Changes in the text: We have modified the language of the revised version.

Comment 5: Figure 2C is mislabelled training, whereas I think it presents testing set data. It might be better to use preterm/ term and postterm as x axis labels for this graph rather than immature , normal mature, over mature
Reply 5: Thanks for the reviewer's comments. We have re-generated the Figure2 by removing the concept of immature, normal mature and over mature. Instead, we directly marked the difference interval (instead of using preterm/term/postterm). And we labelled the model-developing dataset and validation dataset.
Changes in the text: We have modified Figure 2 in the revised version.

Comment 6: Was the process of artefact elimination an automated process, if not how was it done, and could it be automated?
Reply 6: Thanks for the reviewer's comments. We have added the detailed artefact removal process in the revised version. This process was performed automatically.
Changes in the text: We added detailed description in the revised method (1. *EEG signal pre-processing and feature extraction from EEG dataset*).

Comment 7: In the discussion you appear to equate delayed brain maturity with EEG discrepancy between apparent and actual CGA. This is erroneous. I would recommend you eliminate the sentence Studies has indicated that delayed brain maturation is associated with developmental disorders such as attention-deficit/hyperactivity disorder (ADHD)(18).
1. Macdonald-Laurs E, Sharpe C, Nespeca M, Rismanchi N, Gold JJ, Kuperman R, et al. Does the first hour of continuous electroencephalography predict neonatal seizures? Arch Dis Child Fetal Neonatal Ed. 2021;106(2):162-7.

Reply 7: Thanks for the reviewer's comments. We have removed this citation in the revised version.
Changes in the text: We have removed this citation in the revised discussion.

Reviewer C
Comment 1.Dates of the study do not match in the abstract and manuscript.
Reply 1: Thanks for the reviewer's comments. We have fixed this typo.
Changes in the text: We changed to "Jan. 2016 to Mar.2018" in the revised abstract.

Comment 2. There is no aim or hypothesis in the Introduction.
Reply 2: Thanks for the reviewer's comments. We have added the aim in the revised version.
Changes in the text: We have added the aim in the revised Introduction.

Comment 3. For EEG electrode placement, was the International 10-20 system modified for neonates used? There is no mention of occipital electrodes used.
Reply 3: Thanks for the reviewer's comments. We followed the International 10-20 system to place the EEG electrode. We chose the parietal area (P3/P4) instead of occipital area (O1/O2) because of more artefact detected in occipital area.
Changes in the text: We have revised the method section (3. *Data acquisition*).

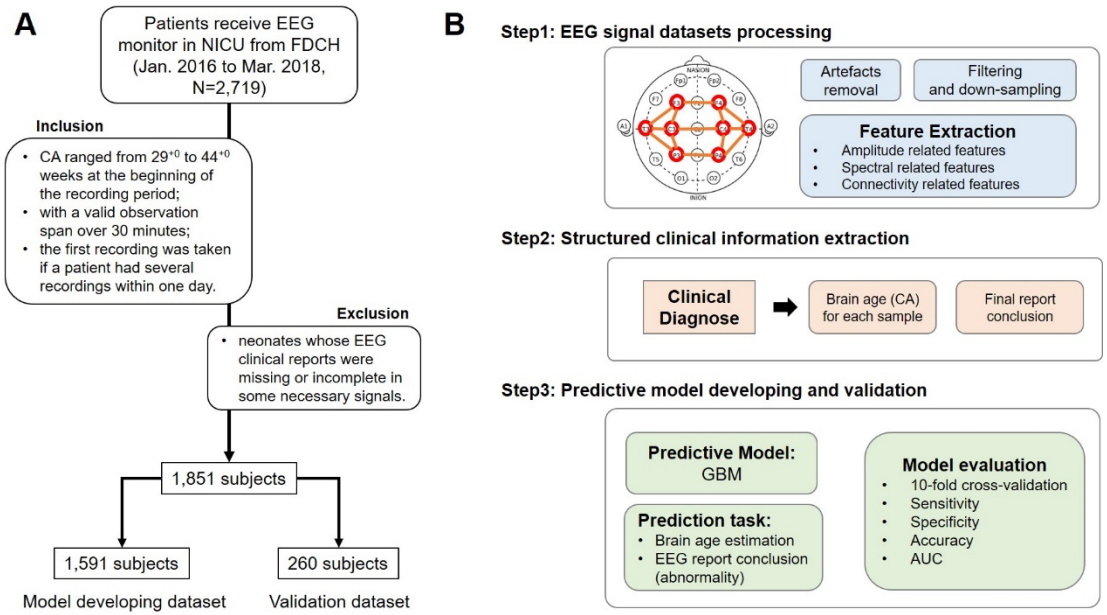
Comment 4. The authors note a 'handbook for clinical EEG reading' as the guideline that clinicians used for scoring EEGs. Can they provide references where this has been used as an acceptable standard neonatal EEG scoring system in prior clinical or research data/publications?
Reply 4: Thanks for the reviewer's comments. Firstly, Liu's standard has been widely used and promoted in China and Liu's standard in neonatal is similar with the descriptions in Table 6.3 from Ebersole et al. [Reference: Ebersole, John S., and Timothy A. Pedley, eds. *Current practice of clinical electroencephalography*. Lippincott Williams & Wilkins, 2003.]. Secondly, our group has published lots of research manuscripts based

on EEG report conclusion which followed Liu's guideline, such as:
Yang, Lin, Yanting Kong, Xinran Dong, Liyuan Hu, Yifeng Lin, Xiang Chen, Qi Ni et al. "Clinical and genetic spectrum of a large cohort of children with epilepsy in China." *Genetics in Medicine* 21, no. 3 (2019): 564-571.
Kong, Yanting, Kai Yan, Liyuan Hu, Mingbang Wang, Xinran Dong, Yulan Lu, Bingbing Wu, Huijun Wang, Lin Yang, and Wenhao Zhou. "Data on mutations and clinical features in SCN1A or Scn2a gene." *Data in brief* 22 (2019): 492-501.
Yang, Lin, Xiang Chen, Xu Liu, Xinran Dong, Chang Ye, Dongli Deng, Yulan Lu, Yifeng Lin, and Wenhao Zhou. "Clinical features and underlying genetic causes in neonatal encephalopathy: A large cohort study." *Clinical Genetics* 98, no. 4 (2020): 365-373.
Changes in the text: We have revised the discussion section.

Comment 5. Did the authors note which medications patients received and was there a way to account for the potential impact of medications (such as sedatives – benzodiazepines, barbituates) on specific background features (voltage, continuity) as this could potentially impact both the conceptual age interpretation and EEG background severity category?
Reply 5: Thanks for the reviewer's comments. This is a great comment and we totally agree that medications will have influence on specific background features, and the following conceptual age interpretation and severity prediction. Generally, the sedative drugs can increase the fast waves, and prolong the IBI time. And the narcotic drugs may also cause burst suppression. In clinic, the clinicians will consider this information together with EEG report conclusion, to make further diagnosis. However, it needs time and labor to systematically design the study which could quantitively measure the effect of medications, and we decide not to discuss the medication issue in depth in this study. This issue is one of our future direction.
Changes in the text: We have modified the discussion part in the revised version.

Comment 6. Regarding figures, I found Figure 1B difficult to follow – a simplified version may be preferable, or consider re-organizing the flow/direction of the boxes – the current flow does not make sense. Figure 1A is missing the total number of patients in the top box from which potential samples were available before applying the inclusion and exclusion criteria.

Reply 6: Thanks for the reviewer's comments. We have modified the Figure 1. The total number is 2,719 subjects. We have re-organized Figure 1B:



Changes in the text: We have modified the Figure 1 in the revised version.

Comment 7: Figure S3 seems to contain some of the most important conclusions from the performance of the algorithm for the 4 EEG background severity groups – I would consider making this a main figure, rather than a supplementary figure.

Reply 7: Thanks for the reviewer's comments. We have moved Figure S3 to Figure 3.

Changes in the text: We have moved Figure S3 to Figure 3.