Original Article

Development and validation of Auto-Neo-electroencephalography (EEG) to estimate brain age and predict report conclusion for electroencephalography monitoring data in neonatal intensive care units

Xinran Dong[1], Yanting Kong[2], Yan Xu[1], Yuanfeng Zhou[3], Xinhua Wang[3], Tiantian Xiao[2,4], Bin Chen[2], Yulan Lu[1], Guoqiang Cheng[2], Wenhao Zhou[1,2]

[1]Center for Molecular Medicine, Pediatric Research Institute, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China; [2]Division of Neonatology, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China; [3]Division of Neurology, Children's Hospital of Fudan University, National Children's Medical Center, Shanghai, China; [4]Department of Neonatology, Chengdu Women's and Children's Central Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China

Contributions: (I) Conception and design: X Dong, W Zhou, G Cheng; (II) Administrative support: W Zhou, G Cheng; (III) Provision of study materials or patients: Y Xu, Y Zhou, X Wang; (IV) Collection and assembly of data: Y Kong, T Xiao, B Chen, Y Lu; (V) Data analysis and interpretation: X Dong, Y Lu; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Guoqiang Cheng, MD, PhD; Wenhao Zhou, MD, PhD. Division of Neonatology, Children's Hospital of Fudan University, National Children's Medical Center, 399 Wanyuan Road, Shanghai 201102, China. Email: gqchengcm@163.com; zhouwenhao@fudan.edu.cn.

Background: Electroencephalography (EEG) monitoring is widely used in neonatal intensive care units (NICUs). However, conventional EEG report generation processes are time-consuming and labor-intensive. Therefore, an automatic, objective, and comprehensive pipeline for brain age estimation and EEG report conclusion prediction is urgently needed to assist clinician's decision-making.

Methods: We recruited patients who underwent EEG monitoring from the NICU at Children's Hospital of Fudan University from Jan. 2016 to Mar. 2018. A total of 1,851 subjects were enrolled, including the patient's conceptional age (CA) and the clinical EEG report conclusion (normal, slightly abnormal, moderately abnormal, or severely abnormal). A total of 1,591 subjects were used to generate predictive models and 260 were used as the validation dataset. We developed Auto-Neo-EEG (an automatic prediction system to assist clinical neonatal EEG report generation), including signal feature extraction, supervised machine learning realized by gradient boosted models, to estimate brain age and predict EEG report conclusion.

Results: The predicted results from the validation dataset were compared with the clinical observations to assess the performance. In the independent validation dataset, the model could achieve accordance 0.904 on estimating brain age for neonates with normal clinical EEG report conclusion, and differences between the predicted and observed brain age were strongly related with EEG report conclusion abnormality. Further, as for the EEG report conclusion prediction, the model could achieve area under the curve (AUC) of 0.984 for severely abnormal situations, and 0.857 for moderately abnormal ones.

Conclusions: The Auto-Neo-EEG has the high accuracy of estimating brain age and EEG report conclusion, which can potentially greatly accelerate the EEG report generation processes assist in clinical decision making.

Keywords: Neonates; electroencephalography monitor; neural signal processing; machine learning model; brain age estimation

Submitted Mar 30, 2021. Accepted for publication Jul 01, 2021.
doi: 10.21037/atm-21-1564
View this article at: https://dx.doi.org/10.21037/atm-21-1564

Introduction

In neonatal intensive care units (NICUs), continuous electroencephalography (EEG) monitoring has been widely applied to the diagnosis of neonatal neurological diseases such as epilepsy, encephalopathy, and central nervous system infection (1,2). Also, the long-term neurological outcome is associated with early-onset EEG changes (3). Though it is easy to perform EEG monitors on neonates lasting for hours or even days at the bedside, the raw EEG signal data are usually very large, which takes experienced neurophysiologists several hours to interpret. Besides, that neonates' brain is developing complicates the evaluation of EEG pattern (4), especially among preterm neonates (5). O'Reilly designed a new EEG signal feature range EEG (rEEG) in 26 newborns with less than 29 weeks of gestational age and found that it is closely related to brain development and maturity (6). Similarly, Stevenson *et al.* constructed a brain age prediction model based on EEG signal features from 65 preterm infants, which could greatly fit actual age and the predicted age difference could be used as a predictor of the neurodevelopmental outcome (5,7,8). These studies are all tested on small data sets, and there is no systematic analysis on how to apply the findings to the clinic.

Developing a machine learning strategy that can quantitatively analyze the EEG signal dataset is crucial, which could make automatically screening the abnormal possible to assist clinicians in EEG report generation and further diagnosis. Therefore, in this study, we aimed to construct an automatic system that could uncover brain age and suggest abnormality from original EEG signals. Here, we collected a large group of EEG datasets from the NICU at Children's Hospital of Fudan University, including 1,851 subjects with signal data and clinical reports. We constructed a system named Auto-Neo-EEG, which consists of an EEG neural signal processing pipeline to extract features from the original signal datasets, machine learning models based on gradient boosted model (gbm) (9) for prediction. The model could achieve great performance on estimating brain age for neonates and figuring out abnormal EEG records, showing great potential in NICU application.

We present the following article in accordance with the STARD reporting checklist (available at https://dx.doi.org/10.21037/atm-21-1564).

Methods

Patients

We retrospectively reviewed patients who underwent EEG monitoring in the NICU in Children's Hospital of Fudan University, from Jan. 2016 to Mar. 2018. A total of 1,851 subjects from 1,692 patients were collected. The study was approved by the ethics committee of the Children's Hospital of Fudan University (No. 2020227), the patients recruited belong to a neonatal project (NCT02544100), and written informed consent was obtained from the guardians of each patient. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013).

Inclusion and exclusion criteria

Shown in *Figure 1A*, the inclusion criteria for subjects were as follows: (I) neonates whose conceptional age (CA) ranged from 29^{+0} to 44^{+6} weeks (203 to 314 days) at the beginning of the recording period; (II) neonates who had video-continuous EEG examinations with a valid observation span over 30 minutes; (III) the first recording was taken if a patient had several recordings within one day. The exclusion criteria were: (I) neonates whose EEG clinical reports were missing or incomplete in some necessary signals.

Dataset acquisition

EEG data were acquired using a Nicolet One machine (sampling frequency: 500 Hz). We followed the International 10–20 system to place the electrodes. The frontal (F3, F4), central (C3, C4), mid-temporal (T3, T4), parietal (P3, P4) scalp electrodes and reference electrode (Cz) was placed (10). We chose the parietal area (P3/P4) instead of the occipital area (O1/O2) because more artifacts detected in the occipital area.

We had three experienced clinicians (Y Zhou, X Wang, Y Xu) in charge of the EEG report generation, who all had attended the uniformly training program and were certified by the Chinese Anti-Epilepsy Association. Y Zhou is the senior clinician with more than 10 years of experience in neonatal EEG reading, and X Wang and Y Xu both have 5 years of experience. For each EEG clinical interpretation, X Wang and Y Xu would perform a double-blind interpretation of the EEG signal and make a conclusion.

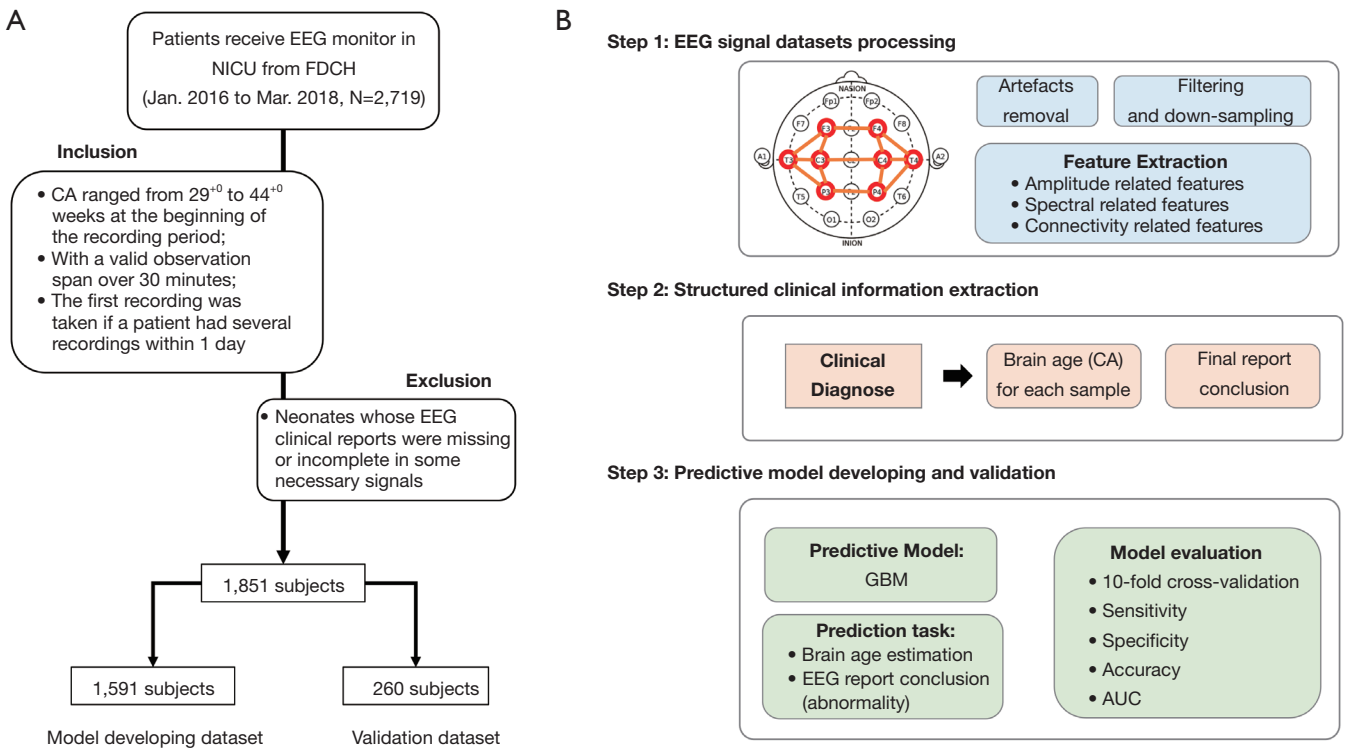

Figure 1 The design of this study. (A) Flowchart for patient recruiting; (B) the design of Auto-Neo-EEG. NICU, neonatal intensive care unit; FDCH, Children's Hospital of Fudan University; EEG, electroencephalography; CA, conceptional age.

If the result from the two experts (X Wang and Y Xu) was consistent, they would directly apply it as the final report. If not, Y Zhou will join and they three would discuss it together to make a final decision.

Standard-scheme for manually reported EEG conclusion

The clinicians followed the guideline—a handbook for clinical EEG reading (11)—to give a clinical report. The levels of EEG abnormalities are defined as slightly abnormal, moderately abnormal, and severely abnormal, and the details are as follows:

Slightly abnormal

(I) Compared with the actual CA, the background activity is mature but slightly delayed, and the trace alternant (TA) or Tracé discontinu (TD) are slightly discontinuous;
(II) The waveform or rhythm compatible with CA is slightly lacking, or the immature waveform compatible with CA disappears;
(III) Focal electrical attenuation;
(IV) A small number of focal or multifocal discharges on a normal or mildly abnormal background.

Moderately abnormal

(I) Compared with the actual CA, the background activity is moderately discontinuous [for neonates whose CA is less than 30 weeks, the interval between bursts is over 30 s; for neonates over 30 weeks (CA), the interval is between 20 and 60 s];
(II) The waveform or rhythm compatible with CA is lacking, or the immature waveform compatible with CA obviously disappears;
(III) The hemisphere continues to be asymmetrical and/or out of sync, but not exceeding 50% of the entire record;
(IV) Continuous universal voltage reduction, the background activity below 25 μV in all states;
(V) Single-disciplinary or other forms of electrical attack without severe background abnormalities;
(VI) Other abnormality whose level is between slightly and moderately abnormal but does not belong to the above descriptions of moderately abnormal,

such as multiple discharge waveform.

Severely abnormal

(I) Compared with actual CA, the background activity is significantly discontinuous (for example, the burst interval exceeds 60 s);
(II) Focal or one-sided periodic discharge exists;
(III) Over-synchronization and or asymmetry between hemispheres, making up more than 50% of the entire record;
(IV) Sharp waves frequently appear in the Rolandic or midline area;
(V) Severely low voltage (below 5 μV in all states);
(VI) Burst suppression;
(VII) Equipotential.

Interrater agreement assessment

Here, we randomly selected 96 subjects for detailed interrater agreement assessment by two experts (X Wang, Y Xu). Cohen's Kappa was used to investigate interrater agreement level for binary level and weighted Kappa was used for ordinal report level. The test was realized by using *Kappa* function from R package vcd (version 1.4-7).

Establishment of Auto-Neo-EEG pipeline

In total, 1,851 subjects were recruited for analysis, of which 1,591 subjects were used in a retrospective study for prediction model generation, while the remaining 260 were included in the validation dataset. The model-developing dataset was used to generate models, including the optimization of model hyper-parameters, while the validation dataset, independent from the model training procedure, was used to evaluate prediction performance.

As shown in *Figure 1B*, the Auto-Neo-EEG pipeline consists of:

EEG signal pre-processing and feature extraction from EEG dataset

For each EEG recording, the original signal dataset was pre-processed, including adjusted to the reference electrode, artifact removal, filtering, and down-sampling. The artifact removal steps were as follows: (I) improper electrode placement: Improper electrode placement can lead to the removal of channels with low correlation coefficients with all the other channels (the threshold was set at 0.15). (II) Electrode coupling: electrode coupling mainly allowed the identification of channels with relatively low power compared with all channels from the same hemisphere. (III) Continuous zeros or constant values: continuous zeros or constant values may result from procedures such as testing the electrode impedance, and these values should be removed. (IV) High amplitude & (V) sudden jump removal: the high amplitude and sudden jump removal steps mainly focused on segments with abnormally high values or value changes possibly caused by movement. The thresholds were 1,500 μV for the absolute amplitude and 200 μV for the amplitude difference, which are also suitable for preterm infants. In the first two steps, a whole channel was removed, and the values of the other channels changed to "NA" (not available) for the segments that failed the filtering procedure. (VI) Too many "NA" values: all "NA" values across all channels were examined and segments with an excessive number of "NA" values were removed. (VII) Improper correlations between independent component analysis (ICA) and electrooculography (EOG): the last step was performed if the original dataset contained an EOG channel. 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. The first five steps were reported by O'Toole *et al.* (12).

Then, a notch filter was performed at 50, 100, 150, 200, and 250 Hz; and a finite-impulse response (FIR) filter was applied in a range between 0.3 to 50 Hz. Finally, the dataset was down-sampled from 500 to 100 Hz.

A total of 722 signal features were extracted to reflect the amplitude, rEEG, spectral density, and connectivity-related aspects. Firstly, the original signal dataset could be decomposed into four frequency bands, i.e., 0.5–4, 4–8, 8–13, and 13–30 Hz, and the number of signal channels was eight. For each frequency band and in each channel, six amplitude features (amplitude_total_power, amplitude_SD, amplitude_skew, amplitude_kurtosis, amplitude_env_mean, amplitude_env_sd), eight rEEG features (rEEG_mean, rEEG_median, rEEG_lower_margin, rEEG_upper_margin, rEEG_width, rEEG_SD, rEEG_CV, rEEG_asymmetry), five spectral features (spectral_power, spectral_relative_power, spectral_flatness, spectral_entropy, spectral_diff) were extracted, thus the first number of signal features was 608 [(6+8+5)*4 (frequency band)*8 (channel)]. Secondly, two spectral related features (spectral_edge_frequency, FD) were extracted in each of eight channels: 2*8=16. For the above aspects, the average figure for all channels was also

calculated as new features, which resulted in (6+8+5)*4+2=78 more features. Besides, five connectivity related features (connectivity_BSI, connectivity_corr, connectivity_coh_mean, connectivity_coh_max, connectivitiy_coh_freqmax) were extracted for each of four frequency bands (5*4=20). In total, 722 (608+16+78+20) neural signalling features were obtained for each sample.

Figure S1 shows the flowchart of the signal processing procedure. We referred to the research by O'Toole *et al.* (12) to perform some of the signal processing steps. The processing steps were developed based on the Python3.6 environment with MNE (13).

EEG signal finding extraction from clinical reports

In clinical reports, CA and report conclusions were extracted. CA was normalized to days and EEG report conclusions were labelled as one of the four ordinal categories (normal, slightly abnormal, moderately abnormal, and severely abnormal).

Prediction model outcome declaration and evaluation criteria

We had two tasks: CA estimation and report conclusion (abnormal severity) prediction, and the outcomes were declared below:

CA estimation; the observed CA was normalized into days and the machine learning model could directly output the predicted days. We built the prediction model merely using the samples with a normal conclusion from the model-developing dataset. The comparison between the predicted CA and the observed CA would be measured by Pearson correlation coefficient (PCC) with significance and 95% confidence interval (CI).

Report conclusion (abnormal severity) prediction; since it was a multi-class prediction issue, we adopted the cascade strategy, i.e., we transformed the classes into four bi-classification questions based on abnormal severity: severely abnormal *vs.* the rest, moderately abnormal *vs.* normal/slightly abnormal, slightly abnormal *vs.* normal, and abnormal *vs.* normal. The final predicted result would be severely abnormal if the first strategy (severely abnormal *vs.* the rest) was severely abnormal. The final predicted result would be moderately abnormal if the first strategy (severely abnormal *vs.* the rest) was the rest and the second strategy (moderately abnormal *vs.* normal/slightly abnormal) was moderately abnormal. The final predicted result would be slightly abnormal if the first two strategies were the rest and normal/slightly abnormal but the third strategy (slightly abnormal *vs.* normal) was slightly abnormal. If the result from the third strategy was normal, the sample would be classified as normal. Combining the above three binary classification models, we could give a unique final label for each sample. The fourth strategy was set as a reference for clinician judgment (if the prediction system user does not want to classify the abnormality with three different levels), and this would not be used in the report conclusion judgment in our study. We calculated area under the curve (AUC) under receiver operating characteristic (ROC) curve with 95% CI, sensitivity, specificity, and accuracy in the model-developing dataset and an independent validation dataset to evaluate the model's performance.

Machine learning steps

Dataset splitting

We split the dataset according to the sample collection time. 1,591 subjects before 2018 were used as the model-developing dataset and the remaining 260 subjects after 2018 were treated as the independent validation dataset.

Machine learning model selection and results evaluation

The classification and regression prediction tasks were all based on the gbm. As described above, PCC was used to evaluate the performance of regression prediction tasks and AUC under ROC curve with sensitivity, specificity, and accuracy at the optimal threshold was used to evaluate the performance of binary classification tasks. For the report-conclusion prediction with four levels, the confusion matrix was generated with sensitivity, specificity, and accuracy to evaluate the performance.

Feature selection

A backward selection procedure was applied. For each iteration, gbm model was generated and features with the minimum importance value were removed. The remained features from the model with the best performance for the cross-validation (CV) results in the model-developing dataset were used. The feature selection was performed separately for different prediction tasks.

Model generation and CV

In the model-developing dataset, we applied a 10-fold CV strategy for training the prediction model. All predictions were realized by gbm and within each CV, different value combinations for each of the parameters would be iterated, i.e., interaction.depth, n.trees, shrinkage, and n.minobsinnode. The final model was the one with the best performance (highest PCC value or AUC value) for the CV results.

For the final used model, there would be "importance

Table 1 Summary statistics for the 1,851 EEG subjects

Clinical features	Model-developing dataset (N=1,591)	Validation dataset (N=260)	Overall (N=1,851)
Gender, n (%)			
Female	671 (42.2)	114 (43.8)	785 (42.4)
Male	920 (57.8)	146 (56.2)	1,066 (57.6)
CA (days)			
Mean (SD)	269.0 (22.9)	266.0 (23.8)	268.0 (23.0)
Median [min, max]	272 [204, 314]	268 [211, 314]	271 [204, 314]
Post-natal age (days)			
Mean (SD)	17.9 (19.3)	17.3 (16.8)	17.8 (19.0)
Median [min, max]	11.0 [0, 142]	12.0 [0, 81.0]	11.0 [0, 142]
Monitoring time (h)			
Mean (SD)	2.72 (1.77)	2.75 (1.52)	2.72 (1.73)
Median [min, max]	2.25 [0.556, 24.8]	2.31 [0.586, 15.3]	2.26 [0.556, 24.8]
EEG report conclusion level, n (%)			
Normal	845 (53.1)	147 (56.5)	992 (53.6)
Slightly abnormal	584 (36.7)	90.0 (34.6)	674 (36.4)
Moderately abnormal	98.0 (6.2)	17.0 (6.5)	115 (6.2)
Severely abnormal	64.0 (4.0)	6.00 (2.3)	70.0 (3.8)

EEG, electroencephalography; CA, conceptional age; SD, standard deviation.

values" to show the signal feature importance with a higher value to indicate more contribution in the prediction model. For binary classification, the "Youden's J statistic" was employed to get the optimal cut-off. The selected final model with the optimal threshold was applied in both the model-developing and independent validation dataset to estimate the performance.

Code implementation

All scripts for prediction and visualization were written in R version 3.6, with packages caret (https://cran.r-project.org/package=caret) and gbm (https://cran.r-project.org/package=gbm).

Statistical analysis

We applied a two-tailed Student's *t*-test for comparison of continuous variables between two groups and an F-test for overall significance in linear regression. Fisher's Exact test was used for the enrichment testing for 2*2 categorical data. Statistical significance was defined as $P<0.05$ and false discovery rate (FDR) correction was used for multiple tests. Statistical test in the model generation procedure is described above. All statistical analyses were performed using R version 3.6.

Results

Benchmark EEG dataset

A total of 1,851 subjects of video-continuous EEG recordings were recruited (*Figure 1A*), with the corresponding clinical reports and disease diagnoses collected. The basic statistics were summarized in *Table 1*. The detailed clinical reports about 30 recording findings could result in a four-level report conclusion judged by experienced clinicians (Table S1 and Method section). The number of samples with four categories of EEG report conclusions in each corrected gestational age (CA) week was shown in Figure S2. The average monitoring time was 2.7 hours. The corresponding 1,692 patients (some patients had multiple recordings) were finally diagnosed (with the help of EEG recordings and other clinical diagnoses) and classified into 11 disease systems. Hypoxic ischemic

encephalopathy (HIE), central nervous system infection, congenital metabolic disease and unexplained convulsions tend to have moderate and severely abnormal EEG report conclusion, while temporary metabolic disorder tends to be normal (Table S2, P value <0.05). In the following sections, we would build an intelligent system—Auto-Neo-EEG (the detailed process is shown in *Figure 1B* and Figure S1)—to systematically investigate the correlation between EEG signals and CA in neonates, and based on that we generate a predictive model to facilitate EEG report generation.

For the Interrater agreement assessment, 96 randomly selected subjects were shown in Table S3. Only nine subjects were inconsistent for report conclusion level between two experts (patients 5, 8, 48, 58, 63, 66, 92, 93, 96) with 90.63% agreement percentage. The weighted Kappa value was 0.923 if consider the ordinal report level and the Cohen's Kappa value was 0.913 if merge three abnormal levels into one. Generally, the good interrater agreement level for outcome ensures the quality of the benchmark dataset used in this study.

Auto-Neo-EEG could successfully estimate brain age

In total, 59 signal features passed the signal filtering step were used to fit the prediction model which achieved high accordance with PCC 0.966 (95% CI: 0.961–0.970, P value <2e−16, F-test) in identifying the CA for 845 datasets with normal reported conclusion level (red points in *Figure 2A*), but the difference occurs for samples with not-normal conclusion levels (orange, green and purple points in *Figure 2A*). When estimating the deviation—CA_diff (difference between the predicted and real CAs), it was significantly less than zero in all three abnormal groups, and the differences increased with the severity of the abnormality (*Figure 2B*, all P value <1e−6, Student's *t*-test, with median values −4.27, −8.36, and −14.90). The severely abnormal samples were significantly enriched in groups with CA difference smaller than 14 days, followed by moderately abnormal (*Figure 2C*, P value <0.05). Samples with CA difference larger than 14 days were also enriched in severely abnormal (*Figure 2C*, P value <0.05). The model also achieved great performance in the validation dataset with accordance of PCC 0.904 (95% CI: 0.870–0.930, P value <2e−16, F-test, red points in *Figure 2D*) for the samples with normal conclusion level. Similarly, the CA_diff decreases with the severity of abnormality (*Figure 2E*, all P value <0.05, Student's *t*-test, with median values −3.46, −13.71, and −15.12). Further, we divided all samples into five groups according to the difference between the predicted CA and actual CA {(−Inf, −14], (−14, −7], (−7,7], (7,14], (14, Inf]} and calculated the odds ratio for each EEG report level with statistical testing. Similarly, in the validation dataset, the severely abnormal samples were significantly enriched in groups with CA difference smaller than 14 days, followed by moderately abnormal (*Figure 2F*, P value <0.05). We tried to directly apply the absolute CA difference as the predictor to the four EEG reporting levels by using "Youden's J statistic" to get the optimal threshold, the result showed that the absolute CA difference was a good marker for the severely abnormal prediction (Table S4, accuracy >80%). Therefore, we could apply Auto-Neo-EEG to quantitative estimation for EEG maturity, and the deviation to the real CA highlights the severity of the abnormality.

Auto-Neo-EEG could achieve high performance to classify EEG abnormality

Next, we developed models to predict four levels of EEG report conclusions directly from the original EEG signals. As we applied the cascade strategy with four binary comparisons to deal with the multi-class classification issues, the original pairwise performance was shown in *Figure 3* and Table S5 with the final confusion matrix shown in *Table 2*. The best prediction accuracy was achieved in the severely abnormal category (AUC 1 for the model-developing dataset with 95% CI: 0.999–1.000, 0.984 the validation dataset with 95% CI: 0.970–0.999), followed by the moderately abnormal category (AUC 0.919 for the model-developing dataset with 95% CI: 0.885–0.955 and 0.857 for the validation dataset with 95% CI: 0.741–0.973). Distinguishing of these two levels could achieve great accuracy and specificity (all higher than 85%), but relatively low sensitivity was observed in moderately abnormal, where several samples were over-estimated to the severely abnormal level. Differentiating the normal and slightly abnormal conditions tended to be more challenging than that in other conditions, which may be due to slight abnormalities in clinical judgments that are relatively more subjective (*Figure 3C,3G*, Table S5). Besides, the prediction model separating abnormal (combining three abnormal levels) and normal subjects only achieved ordinary performance (*Figure 3D,3H*, Table S5). The features that contributed the most to identify the severely abnormal conditions (with highest importance value) were rEEG (similar to aEEG) lower median and asymmetry, whereas to identifying the slightly abnormal conditions, the observed

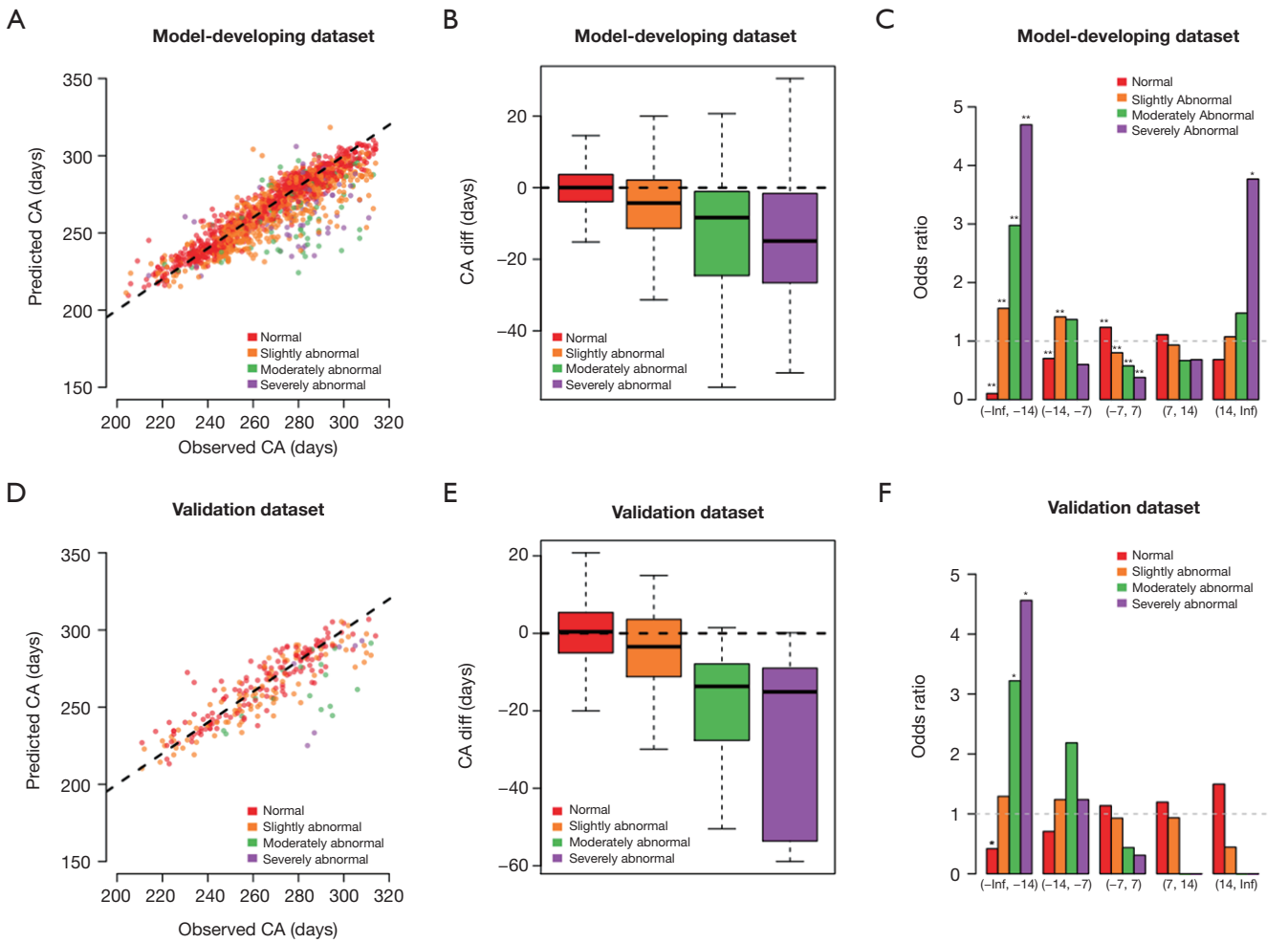

Figure 2 The performance of Auto-Neo-EEG in CA identification. (A) Scatter plot of the predicted CA *vs.* observed CA in the model-developing dataset. Dots were colored according to the corresponding EEG report conclusion label. (B) Boxplot of the CA difference (CA diff) between the predicted CA and the observed CA in the four EEG report conclusions in the model-developing dataset. (C) Barplot for the odds ratio of subjects with each EEG report conclusion label within each interval of CA difference in the model-developing dataset. (D) Scatter plot of the predicted CA *vs.* observed CA in the validation dataset. (E) Boxplot of the CA difference between the predicted CA and the observed CA in the four EEG report conclusions in the validation dataset. (F) Barplot for the odds ratio of subjects with each EEG report conclusion label within each interval of CA difference in the validation dataset. *, the P value significance compared to background is smaller than 0.05; **, the P value significance compared to background is smaller than 0.01. EEG, electroencephalography; CA, conceptional age.

CA and spectral difference contributed the most (*Figure 4*). Generally, Auto-Neo-EEG could successfully predict EEG report conclusion abnormality.

Discussion

Continuous EEG as a non-invasive tool of brain function monitoring in the NICU is recommended by the American Clinical Neurophysiology Society (ACNS) (14,15). In acute neonatal encephalopathy, EEG can provide useful information on brain function. Severe background activity abnormalities reflect serious brain damage and are prognostic predictors of long-term adverse outcomes (1). Awal et al reviewed 52 studies and showed that burst suppression and low voltage could accurately predict the neurologic sequelae of newborns with HIE (16). Moreover, few biomarkers have been used to assess brain maturation in neonates (17). To data, magnetic resonance imaging

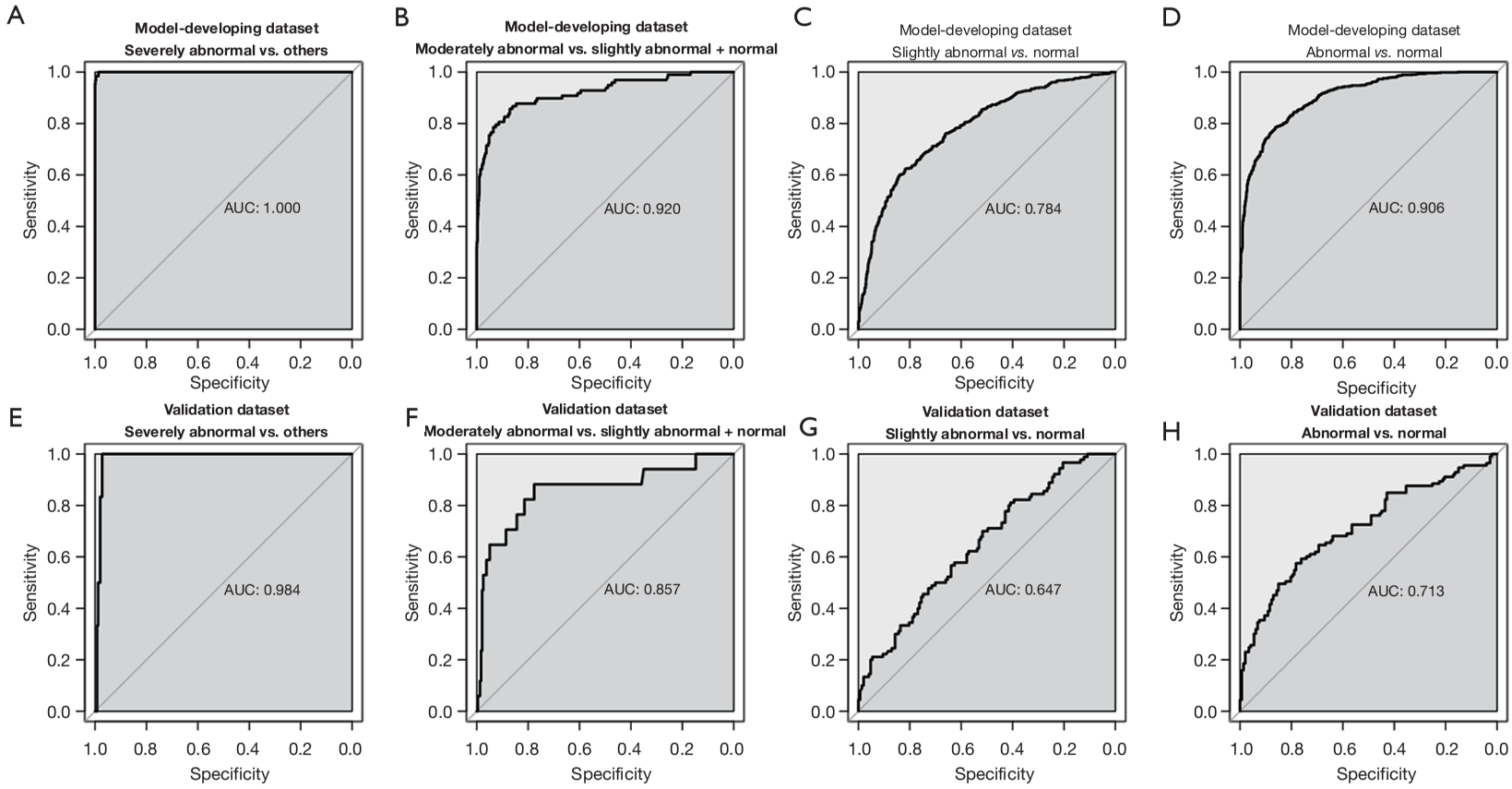

Figure 3 ROC curves of the conclusion label prediction in the model-developing and validation dataset in four binary comparison prediction strategies. (A-D) ROC curves in the mode-developing dataset. The value in the middle shows the AUC value under ROC curve. (E-H) ROC curves in the validation dataset. (A,E) Binary comparison to distinguish severely abnormal and other levels. (B,F) Binary comparison to distinguish moderately abnormal and slightly abnormal and normal levels. (C,G) Binary comparison to distinguish slightly abnormal and normal level. (D,H) Binary comparison to distinguish abnormal (severely, moderately and slightly abnormal) and normal. ROC, receiver operating characteristic; AUC, area under the curve.

Table 2 The performance of Auto-Neo-EEG in predicting final conclusions

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	619	200	11	0	619	535	211	226	73.25%	71.72%	72.53%
		Slightly abnormal	135	290	3	0	290	869	138	294	49.66%	86.3%	72.85%
		Moderately abnormal	84	90	75	0	75	1,319	174	23	76.53%	88.35%	87.62%
		Severely abnormal	7	4	9	64	64	1,507	20	0	100.00%	98.69%	98.74%
	Validation dataset	Normal	90	44	1	0	90	68	45	57	61.22%	60.18%	60.77%
		Slightly abnormal	39	36	4	0	36	127	43	54	40.00%	74.71%	62.69%
		Moderately abnormal	16	10	4	0	4	217	26	13	23.53%	89.30%	85.00%
		Severely abnormal	2	0	8	6	6	244	10	0	100.00%	96.06%	96.15%

EEG, electroencephalography; CA, conceptional age; TP, true positive; TN, true negative; FP, false positive; FN, false negative.

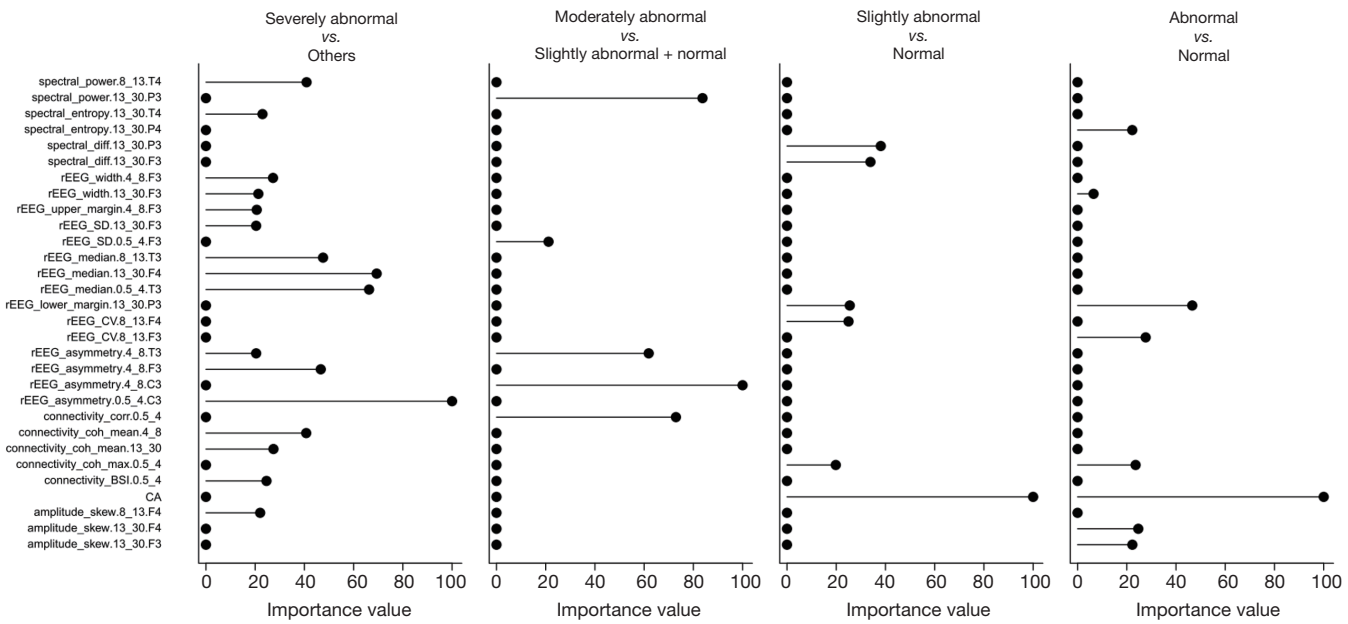

Figure 4 The features' importance value in the EEG report conclusion label predictions. The four sub-figures were the four pairwise binary comparisons, where each bar represents the importance values of the features for the corresponding prediction purposes. Features with a maximum importance value higher than 20 are shown. EEG, electroencephalography; rEEG, range EEG.

(MRI) is the only clinical tool for assessing the brain integrity in neonates (18). Recently, one study showed that preterm EEG can be used to estimate functional brain age. Therefore, continuous EEG monitoring is critical in NICU. However, many difficulties still exist in the clinics. For example, neonatal EEG interpretation requires a specialized neurophysiologist who, however, definitely not be available to interpret EEG signals 24/7 (19). To address the gap in EEG monitoring and interpretation, real-time and automatic equipment to interpret EEG data is needed.

Here, the Auto-Neo-EEG, which focuses on automatic brain age estimation and classifying the EEG abnormality greatly accelerates the generation of EEG reports for neonates. Our findings suggested that Auto-Neo-EEG could predict brain maturity and classify the EEG abnormality with high accuracy. This system has multiple advantages. First, it is generated according to a large-scale dataset. Children's Hospital of Fudan University has launched EEG monitoring in NICU for 7 years and has assisted over 5,000 patients in disease diagnosis so far. We have six Nicolet One machines with the specially assigned technologist to perform EEG signal recording, which has guaranteed the effectiveness in the data collection procedure. We found clear correlation patterns between CA and signal patterns and uncovered the relationship between CA difference and abnormal EEG signals quantitatively, which could directly be applied to automatic maturity abnormality screening. We also tried to directly apply the CA difference to predict the EEG report conclusion (Table S4), which showed that the performance to distinguish severely abnormal and moderately abnormal was much better than normal and slightly abnormal but still worse than the results shown in *Table 2*, indicating that the difference for CA was a good marker for the abnormality prediction but not the only affected signal features. Besides, it has been previously reported that EEG interpretation lacks consistency (20), and in our dataset, the interrater agreement for the report conclusion could reach a perfect level.

Motivated by clinical requirements, we integrated the diagnostic information of patients, clinical reports, and original EEG signal datasets to build a prediction system to assist in generating neonatal EEG reports. We have made considerable progress in the following aspects: (I) we have established a standardized platform for data collection and structured clinical report generation. (II) The Auto-Neo-EEG is an automated, objective, 24/7, and standardized interpretation of bedside EEG monitors, which could effectively assist in clinical report generation. (III) The model to identify CA could provide a quantitative estimation of brain age, which could hardly be judged using traditional strategy. The difference between predicted and observed CA could strongly indicate abnormality as lots of factors such as asphyxia, multi-system malformation will delay brain development. This model could directly be applied without much manual review. In addition, the choice of predictive models: we chose a gradient boosting machine (gbm) as the predictive model in our study and the EEG report conclusion prediction task convert the original multi-class classification issue into three binary classification issues. We also tried some popular algorithms, such as lasso regression for CA prediction and random forest and support vector machine (SVM) for bi-classification prediction, and the performance of those algorithms were worse than that of gbm. We also tried to generate a model to directly predict the four report levels, but the results were worse than the current model, especially in the prediction of severely and moderately abnormal. In addition, we tried deep learning frameworks EEGLearn (21), which had a similar performance, but additional computing resources were required.

Limitations

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, which is generally followed the description of table 6.3 from Ebersole *et al.* (8). Besides, some technical limitations are as follows. (I) Local conclusion prediction: we tried to locate abnormal signal features in specific periods, but the result was not ideal (the false positive rate is high), partially due to lack of accurate label, and more diverse data collections and accurate manual annotations are required in future improvement. (I) Signal finding prediction: we are trying to learn from prediction algorithms, such as an empirical wavelet transformation applied by Bhattacharyya *et al.* (22), sharp and wave calculations applied by Chang *et al.* (23) , different types of entropies mentioned by Arunkumar *et al.* (24), and key-point based local binary patterns proposed by Tiwari *et al.* (25), to further broaden the signal description prediction in the clinical report. (III) Impact of medications: we do not consider the potential impact of medications on the EEG signals in this study, which we will systematically design and discuss the effect in our future work. (V) Clinical application: currently our system still needs clinical experts

to review before a clinical report is done. During this process, the original signal data visualization and a report review system for clinical manipulation are required. Thus, we are developing a server that will include the EEG signal browser, a management system for patient information, and a report review system. The predictive model can also be updated with more data collected during its further clinical application.

Conclusions

In conclusion, the present study shows that Auto-Neo-EEG can successfully estimate brain age and predict signal abnormalities, which could benefit many clinicians in performing neonatal EEG studies.

Acknowledgments

We thank all doctors and nurses in NICU for their patient care and data collection.
Funding: This work was funded by the project supported by Shanghai Municipal Science and Technology Major Project (2018SHZDZX05) and National Key Research and Development Project of China (2018YFC0116903). These two funding had helped in the data collection procedure by paying the nurses who operated the machine and helped the data analysis procedure in purchasing the computer clusters for large-scale data storage and computing. Intelligent Medical Research Project of Shanghai Health and Family Planning Commission (2018ZHYL0225). This funding has helped in the data collection procedure by paying the clinician and researchers to collect and clean the original EEG dataset and clinical report, to perform the clinical report decomposing and label revision.

Footnote

Reporting Checklist: The authors have completed the STARD reporting checklist. Available at https://dx.doi.org/10.21037/atm-21-1564

Data Sharing Statement: Available at https://dx.doi.org/10.21037/atm-21-1564

Peer Review File: Available at https://dx.doi.org/10.21037/atm-21-1564

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://dx.doi.org/10.21037/atm-21-1564). Xinran Dong and Yulan Lu report that funding (2018YFC0116903 and 2018ZHYL0225) was received to support this manuscript. Yanting Kong and Guoqiang Cheng report funding (2018ZHYL0225) was received to support this manuscript. Wenhao Zhou report funding (2018SHZDZX05 and 2018YFC0116903) was received to support this manuscript. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the ethics committee of the Children's Hospital of Fudan University (No. 2020227), the patients recruited belong to a neonatal project (NCT02544100), and written informed consent was obtained from the guardians of each patient.

References

1. Massey SL, Jensen FE, Abend NS. Electroencephalographic monitoring for seizure identification and prognosis in term neonates. Semin Fetal Neonatal Med 2018;23:168-74.
2. McCoy B, Hahn CD. Continuous EEG monitoring in the neonatal intensive care unit. J Clin Neurophysiol 2013;30:106-14.
3. Périvier M, Rozé JC, Gascoin G, et al. Neonatal EEG and neurodevelopmental outcome in preterm infants born before 32 weeks. Arch Dis Child Fetal Neonatal Ed 2016;101:F253-9.
4. Mathieson SR, Stevenson NJ, Low E, et al. Validation of an automated seizure detection algorithm for term neonates. Clin Neurophysiol 2016;127:156-68.
5. Stevenson NJ, Oberdorfer L, Tataranno ML, et al.

Automated cot-side tracking of functional brain age in preterm infants. Ann Clin Transl Neurol 2020;7:891-902.
6. O'Reilly D, Navakatikyan MA, Filip M, et al. Peak-to-peak amplitude in neonatal brain monitoring of premature infants. Clin Neurophysiol 2012;123:2139-53.
7. Stevenson NJ, Oberdorfer L, Koolen N, et al. Functional maturation in preterm infants measured by serial recording of cortical activity. Sci Rep 2017;7:12969.
8. Ebersole JS, Pedley TA. Current practice of clinical electroencephalography. 3rd edition. Philadelphia: Lippincott Williams & Wilkins, 2003.
9. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29:1189-232.
10. Homan RW, Herman J, Purdy P. Cerebral location of international 10-20 system electrode placement. Electroencephalogr Clin Neurophysiol 1987;66:376-82.
11. Liu X. Clinical Electroencephalography (version II). Beijing: People's Medical Publishing House, 2019.
12. O'Toole JM, Boylan GB. NEURAL: quantitative features for newborn EEG using Matlab. arXiv:1704.05694 [physics.med-ph] 2017.
13. Gramfort A, Luessi M, Larson E, et al. MNE software for processing MEG and EEG data. Neuroimage 2014;86:446-60.
14. Shellhaas RA, Chang T, Tsuchida T, et al. The American Clinical Neurophysiology Society's Guideline on Continuous Electroencephalography Monitoring in Neonates. J Clin Neurophysiol 2011;28:611-7.
15. Nagarajan L, Palumbo L, Ghosh S. Classification of clinical semiology in epileptic seizures in neonates. Eur J Paediatr Neurol 2012;16:118-25.
16. Awal MA, Lai MM, Azemi G, et al. EEG background features that predict outcome in term neonates with hypoxic ischaemic encephalopathy: A structured review. Clin Neurophysiol 2016;127:285-96.
17. Somerville LH. Searching for Signatures of Brain Maturity: What Are We Searching For? Neuron 2016;92:1164-7.
18. Pittet MP, Vasung L, Huppi PS, et al. Newborns and preterm infants at term equivalent age: A semi-quantitative assessment of cerebral maturity. Neuroimage Clin 2019;24:102014.
19. Ntonfo GM, Ferrari G, Raheli R, et al. Low-complexity image processing for real-time detection of neonatal clonic seizures. IEEE Trans Inf Technol Biomed 2012;16:375-82.
20. Wusthoff CJ, Sullivan J, Glass HC, et al. Interrater agreement in the interpretation of neonatal electroencephalography in hypoxic-ischemic encephalopathy. Epilepsia 2017;58:429-35.
21. Bashivan P, Rish I, Yeasin M, et al. Learning Representations from EEG with Deep Recurrent-Convolutional Neural Networks. arXiv:1511.06448 [cs.LG] 2015.
22. Bhattacharyya A, Pachori RB. A Multivariate Approach for Patient-Specific EEG Seizure Detection Using Empirical Wavelet Transform. IEEE Trans Biomed Eng 2017;64:2003-15.
23. Chang WD, Cha HS, Lee C, et al. Automatic Identification of Interictal Epileptiform Discharges in Secondary Generalized Epilepsy. Comput Math Methods Med 2016;2016:8701973.
24. Arunkumar N, Ramkumar K, Venkatraman V, et al. Classification of focal and non focal EEG using entropies. Pattern Recognit Lett 2017;94:112-7.
25. Tiwari AK, Pachori RB, Kanhangad V, et al. Automated Diagnosis of Epilepsy Using Key-Point-Based Local Binary Pattern of EEG Signals. IEEE J Biomed Health Inform 2017;21:888-96.

Cite this article as: Dong X, Kong Y, Xu Y, Zhou Y, Wang X, Xiao T, Chen B, Lu Y, Cheng G, Zhou W. Development and validation of Auto-Neo-electroencephalography (EEG) to estimate brain age and predict report conclusion for electroencephalography monitoring data in neonatal intensive care units. Ann Transl Med 2021;9(16):1290. doi: 10.21037/atm-21-1564

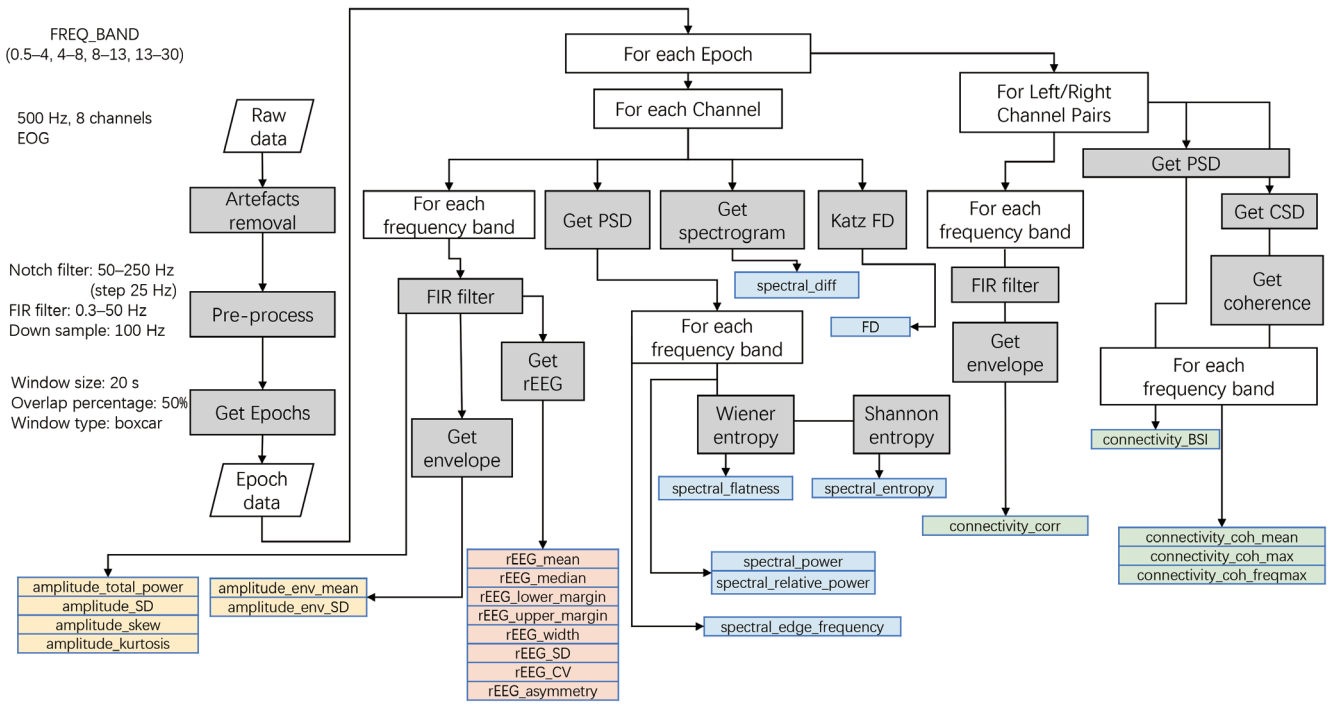

Figure S1 Flowchart of the EEG signal feature extraction procedure for Auto-Neo-EEG. The boxes in yellow, orange, blue and green represent features related to amplitude, rEEG, spectral density and connectivity, respectively. EEG, electroencephalography; FIR, finite-impulse response; PSD, power spectral density; FD, fractal dimension; CSD, cross-PSD; rEEG, range EEG.

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

Table S1 Thirty findings for each EEG recording clinical report

Finding class	General description for the findings
Sleep-wake cycling	Abnormal sleep-wake cycling
	Sleep cycling can be divided into AS and QS period
	Sleep cycling cannot be divided into AS and QS state
Background	Tracé discontinu pattern in sleep state
	Tracé alternant pattern in sleep state
	Continuous pattern in sleep state
	Tracé discontinu pattern in awake state
	Tracé alternant pattern in awake state
	Continuous pattern in awake state
	Burst suppression
	Abnormal symmetry and synchrony
	Hemisphere asymmetry/asynchronous ≤50%
	Hemisphere asymmetry/asynchronous >50%
	Borderline low voltage
	Abnormally low voltage
	Dysmaturity
Seizures	No obvious discharge
	Seizure
Waves	Spike waves rhythmic discharges
	Sharp waves rhythmic discharges
	Low amplitude fast wave rhythmic discharges
	Sleep state sharp wave
	Sleep state sharp-slow wave complex
	Sleep state spike wave
	Sleep state spike-slow wave complex
	Awake state abnormal wave
	Awake state sharp wave
	Awake state sharp-slow wave complex
	Awake state spike wave
	Awake state spike-slow wave complex

EEG, electroencephalography; AS, active sleep; QS, quiet sleep.

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

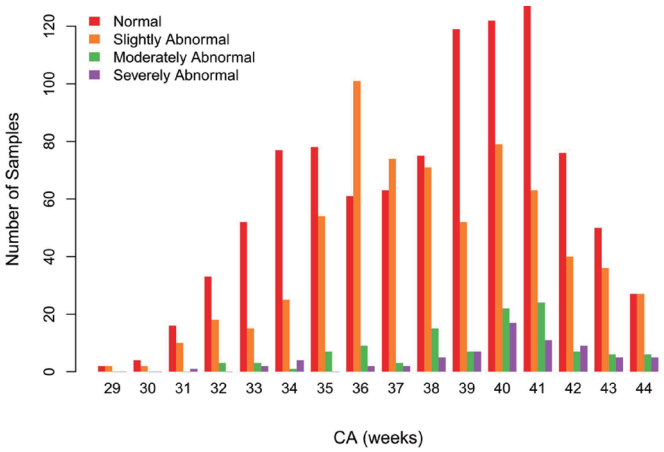

Figure S2 Detailed CA distribution for all samples in the neonatal EEG dataset. Sample numbers in each CA week are divided into four groups according to the EEG report conclusion label. EEG, electroencephalography; CA, conceptional age.

Table S2 Relationship between clinical systems and EEG report conclusion

Disease Classification	Normal (N=992)	Slightly abnormal (N=674)	Moderately abnormal (N=115)	Severely abnormal (N=70)
HIE	71 (50%, OR =0.93, P=0.7)	40 (28%, OR =0.77, P=0.2)	19 (13%, OR =2.2, P=0.006)	12 (8%, OR =2.2, P=0.02)
Cardiopulmonary disease	62 (56%, OR =1, P=0.8)	44 (40%, OR =1.1, P=0.6)	5 (5%, OR =0.73, P=0.7)	0 (0%, OR =0, P=0.03)
Central nervous system infection	43 (41%, OR =0.76, P=0.1)	41 (39%, OR =1.1, P=0.8)	13 (12%, OR =2, P=0.05)	9 (8%, OR =2.2, P=0.04)
Intracranial hemorrhage	81 (63%, OR =1.2, P=0.3)	32 (25%, OR =0.68, P=0.06)	11 (9%, OR =1.4, P=0.4)	5 (4%, OR =1, P=0.8)
Congenial metabolic disease	7 (24%, OR =0.45, P=0.05)	11 (38%, OR =1, P=0.9)	3 (10%, OR =1.7, P=0.4)	8 (28%, OR =7.3, P=7e-05)
Temporary metabolic disorder	131 (72%, OR =1.3, P=0.02)	44 (24%, OR =0.66, P=0.02)	5 (3%, OR =0.44, P=0.09)	2 (1%, OR =0.29, P=0.09)
Perinatal stroke	5 (38%, OR =0.72, P=0.6)	6 (46%, OR =1.3, P=0.6)	1 (8%, OR =1.2, P=0.6)	1 (8%, OR =2, P=0.4)
Premature	482 (53%, OR =0.99, P=0.9)	372 (41%, OR =1.1, P=0.1)	35 (4%, OR =0.62, P=0.02)	22 (2%, OR =0.64, P=0.07)
Genetic factors/syndrome	32 (50%, OR =0.93, P=0.8)	21 (33%, OR =0.9, P=0.8)	8 (12%, OR =2, P=0.08)	3 (5%, OR =1.2, P=0.7)
Unexplained convulsions	31 (35%, OR =0.65, P=0.05)	38 (43%, OR =1.2, P=0.4)	14 (16%, OR =2.5, P=0.005)	6 (7%, OR =1.8, P=0.2)
Other	47 (63%, OR =1.2, P=0.4)	25 (33%, OR =0.92, P=0.8)	1 (1%, OR =0.21, P=0.1)	2 (3%, OR =0.71, P=1)

Each box contains the number of patients, percentage of patients in this clinical system, OR compared to background and P value (P) by Fisher's exact test. EEG, electroencephalography; HIE, hypoxic ischemic encephalopathy; OR, odds ratio.

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

Table S3 Interrater agreement for report conclusion between two experts for 96 EEG recording subjects

Patient	CA	Gender	Outcome by expert 1	Outcome by expert 2
1	42+4	Male	Severely abnormal	Severely abnormal
2	40+3	Male	Normal	Normal
3	34+2	Male	Slightly abnormal	Slightly abnormal
4	43+4	Female	Severely abnormal	Severely abnormal
5*	41+3	Male	Severely abnormal	Moderately abnormal
6	39+6	Male	Normal	Normal
7	40+2	Male	Normal	Normal
8*	39+1	Female	Slightly abnormal	Normal
9	41+5	Female	Slightly abnormal	Slightly abnormal
10	35+2	Male	Moderately abnormal	Moderately abnormal
11	41+6	Male	Normal	Normal
12	40	Male	Normal	Normal
13	44+4	Male	Moderately abnormal	Moderately abnormal
14	40+1	Female	Moderately abnormal	Moderately abnormal
15	42	Female	Moderately abnormal	Moderately abnormal
16	41	Female	Severely abnormal	Severely abnormal
17	39+5	Female	Severely abnormal	Severely abnormal
18	39	Male	Normal	Normal
19	40+4	Male	Normal	Normal
20	37+3	Female	Slightly abnormal	Slightly abnormal
21	37+2	Male	Moderately abnormal	Moderately abnormal
22	41+5	Male	Normal	Normal
23	41+1	Female	Normal	Normal
24	44	Male	Severely abnormal	Severely abnormal
25	34+3	Male	Normal	Normal
26	41+1	Male	Normal	Normal
27	42+5	Female	Slightly abnormal	Slightly abnormal
28	40+5	Male	Normal	Normal
29	34+4	Female	Normal	Normal
30	37+3	Female	Normal	Normal
31	33+2	Female	Normal	Normal
32	38+4	Female	Slightly abnormal	Slightly abnormal
33	36+2	Female	Slightly abnormal	Slightly abnormal
34	35+2	Male	Normal	Normal
35	37	Male	Slightly abnormal	Slightly abnormal
36	39	Male	Normal	Normal
37	34+3	Male	Normal	Normal
38	39+6	Female	Slightly abnormal	Slightly abnormal
39	41+1	Male	Slightly abnormal	Slightly abnormal
40	36+3	Female	Normal	Normal
41	41+3	Male	Normal	Normal
42	35+3	Female	Normal	Normal
43	38+2	Female	Slightly abnormal	Slightly abnormal
44	32+1	Male	Normal	Normal
45	33+4	Female	Slightly abnormal	Slightly abnormal
46	42+1	Male	Normal	Normal
47	39	Male	Normal	Normal
48*	37+5	Female	Slightly abnormal	Normal
49	32+3	Male	Slightly abnormal	Slightly abnormal
50	39	Female	Slightly abnormal	Slightly abnormal
51	31+4	Male	Slightly abnormal	Slightly abnormal
52	34+5	Male	Slightly abnormal	Slightly abnormal
53	37+2	Female	Slightly abnormal	Slightly abnormal
54	38+1	Female	Slightly abnormal	Slightly abnormal
55	38+3	Male	Moderately abnormal	Moderately abnormal
56	33+3	Female	Slightly abnormal	Slightly abnormal
57	35+3	Male	Moderately abnormal	Moderately abnormal
58*	37+3	Female	Normal	Slightly abnormal
59	35+1	Male	Normal	Normal
60	34+3	Male	Normal	Normal
61	36+5	Male	Normal	Normal
62	39+6	Male	Normal	Normal
63*	36+4	Male	Normal	Slightly abnormal
64	37+6	Male	Normal	Normal
65	34+4	Male	Normal	Normal
66*	40+1	Male	Slightly abnormal	Moderately abnormal
67	41+3	Male	Normal	Normal
68	33+1	Male	Slightly abnormal	Slightly abnormal
69	38+4	Male	Normal	Normal
70	39+1	Female	Slightly abnormal	Slightly abnormal
71	33+6	Male	Normal	Normal
72	35+2	Female	Normal	Normal
73	42+1	Female	Moderately abnormal	Moderately abnormal
74	41+4	Male	Normal	Normal
75	34+6	Female	Normal	Normal
76	35	Male	Normal	Normal
77	40+6	Female	Moderately abnormal	Moderately abnormal
78	43+4	Male	Moderately abnormal	Moderately abnormal
79	40+3	Male	Severely abnormal	Severely abnormal
80	40+2	Male	Severely abnormal	Severely abnormal
81	39+5	Female	Moderately abnormal	Moderately abnormal
82	39+6	Male	Severely abnormal	Severely abnormal
83	40+1	Male	Moderately abnormal	Moderately abnormal
84	41+4	Male	Severely abnormal	Severely abnormal
85	40+3	Female	Severely abnormal	Severely abnormal
86	35+5	Male	Severely abnormal	Severely abnormal
87	42+2	Male	Moderately abnormal	Moderately abnormal
88	38+1	Female	Moderately abnormal	Moderately abnormal
89	40+1	Male	Moderately abnormal	Moderately abnormal
90	40+2	Female	Moderately abnormal	Moderately abnormal
91	39+2	Female	Severely abnormal	Severely abnormal
92*	40+1	Male	Moderately abnormal	Severely abnormal
93*	38+3	Female	Moderately abnormal	Severely abnormal
94	41+5	Male	Severely abnormal	Severely abnormal
95	43+2	Male	Severely abnormal	Severely abnormal
96*	40+4	Male	Moderately abnormal	Slightly abnormal

*, subjects inconsistent for report conclusion level between two experts. EEG, electroencephalography.

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

Table S4 The performance of Auto-Neo-EEG in predicting report conclusions by absolute CA difference

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%

EEG, electroencephalography; CA, conceptional age; TP, true positive, TN, true negative, FP, false positive, FN, false negative.

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

Table S5 The performance of Auto-Neo-EEG in predicting report conclusions in each pair-wise comparison

Prediction strategy	Dataset	Prediction Label	Original label				Sensitivity	Specificity	Accuracy	AUC (95% CI)
			Normal	Slightly abnormal	Moderately abnormal	Severely abnormal				
Severely abnormal *vs.* others	Model-developing dataset	F	1,507			0	100%	98.69%	98.74%	1.000 (0.999–1.000)
		T	20			64				
	Validation dataset	F	244			0	100%	96.06%	96.15%	0.984 (0.970–0.999)
		T	10			6				
Moderately abnormal *vs.* slightly abnormal + normal	Model-developing dataset	F	1,245		14	–	85.71%	87.12%	87.03%	0.919 (0.885–0.955)
		T	184		84	–				
	Validation dataset	F	210		6	–	64.71%	88.61%	87.01%	0.857 (0.741–0.973)
		T	27		11	–				
Slightly abnormal *vs.* normal	Model-developing dataset	F	692	221	–	–	62.16%	81.89%	73.83%	0.784 (0.759–0.808)
		T	153	363	–	–				
	Validation dataset	F	107	48	–	–	46.67%	72.79%	62.87%	0.647 (0.576–0.718)
		T	40	42	–	–				
Abnormal *vs.* normal	Model-developing dataset	F	728		160		78.55%	86.15%	82.59%	0.906 (0.892-0.921)
		T	117		586					
	Validation dataset	F	109		45		60.18%	74.15%	68.08%	0.713 (0.648-0.777)
		T	38		68					
Total number	Model-developing dataset		845	584	98	64	–	–	–	–
	Validation dataset		147	90	17	6	–	–	–	–

EEG, electroencephalography; AUC, area under the curve; CI, confidence interval; T, true, F, false.

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