Early qualitative and quantitative amplitude-integrated electroencephalogram and raw electroencephalogram for predicting long-term neurodevelopmental outcomes in extremely preterm infants in the Netherlands: a 10-year cohort study

Xiaowan Wang, Chiara Trabatti*, Lauren Weeke, Jeroen Dudink, Henriette Swanenburg de Veye, Rian M J C Eijsermans, Corine Koopman-Esseboom, Manon J N L Benders, Maria Luisa Tataranno*

Summary

Background Extremely preterm infants (<28 weeks of gestation) are at great risk of long-term neurodevelopmental impairments. Early amplitude-integrated electroencephalogram (aEEG) accompanied by raw EEG traces (aEEG–EEG) has potential for predicting subsequent outcomes in preterm infants. We aimed to determine whether and which qualitative and quantitative aEEG–EEG features obtained within the first postnatal days predict neurodevelopmental outcomes in extremely preterm infants.

Methods This study retrospectively analysed a cohort of extremely preterm infants (born before 28 weeks and 0 days of gestation) who underwent continuous two-channel aEEG–EEG monitoring during their first 3 postnatal days at Wilhelmina Children's Hospital, Utrecht, the Netherlands, between June 1, 2008, and Sept 30, 2018. Only infants who did not have genetic or metabolic diseases or major congenital malformations were eligible for inclusion. Features were extracted from preprocessed aEEG–EEG signals, comprising qualitative parameters grouped in three types (background pattern, sleep–wake cycling, and seizure activity) and quantitative metrics grouped in four categories (spectral content, amplitude, connectivity, and discontinuity). Machine learning-based regression and classification models were used to evaluate the predictive value of the extracted aEEG–EEG features for 13 outcomes, including cognitive, motor, and behavioural problem outcomes, at 2–3 years and 5–7 years. Potential confounders (gestational age at birth, maternal education, illness severity, morphine cumulative dose, the presence of severe brain injury, and the administration of antiseizure, sedative, or anaesthetic medications) were controlled for in all prediction analyses.

Findings 369 infants were included and an extensive set of 339 aEEG–EEG features was extracted, comprising nine qualitative parameters and 330 quantitative metrics. The machine learning-based regression models showed significant but relatively weak predictive performance (ranging from r=0·13 to r=0·23) for nine of 13 outcomes. However, the machine learning-based classifiers exhibited acceptable performance in identifying infants with intellectual impairments from those with optimal outcomes at age 5–7 years, achieving balanced accuracies of 0·77 (95% CI 0·62–0·90; p=0·0020) for full-scale intelligence quotient score and 0·81 (0·65–0·96; p=0·0010) for verbal intelligence quotient score. Both classifiers maintained identical performance when solely using quantitative features, achieving balanced accuracies of 0·77 (95% CI 0·63–0·91; p=0·0030) for full-scale intelligence quotient score and 0·81 (0·65–0·96; p=0·0010) for verbal intelligence quotient score.

Interpretation These findings highlight the potential benefits of using early postnatal aEEG–EEG features to automatically recognise extremely preterm infants with poor outcomes, facilitating the development of an interpretable prognostic tool that aids in decision making and therapy planning.

Funding European Commission Horizon 2020.

Lancet Digit Health 2023; 5: e895–904

Published Online November 6, 2023 https://doi.org/10.1016/S2589-7500(23)00198-X

See Comment page e853

*Contributed equally

Department of Neonatology (X Wang MSc, L Weeke MD, J Dudink MD, R M J C Eijsermans MSc, C Koopman-Esseboom MD, Prof M J N L Benders MD, M L Tataranno MD), **Psychosocial Department** (H Swanenburg de Veye PhD), **and Child Development and Exercise Centre** (R M J C Eijsermans), **Wilhelmina Children's Hospital, and Brain Centre Rudolf Magnus** (J Dudink, Prof M J N L Benders, M L Tataranno), **University Medical Centre Utrecht, Utrecht, Netherlands; Pediatric and Neonatology Unit, Maggiore Hospital, ASST Crema, Crema, Italy** (C Trabatti MD)

Correspondence to: Dr Maria Luisa Tataranno, Department of Neonatology, Wilhelmina Children's Hospital, University Medical Centre Utrecht, 3584 EA Utrecht, Netherlands **m.l.tataranno-2@umcutrecht.nl**

Introduction

Survival rates of extremely preterm infants (born at <28 weeks of gestation) have seen enormous improvements thanks to advanced obstetric and neonatal care over the past three decades.[1–3] Nevertheless, surviving infants still face a high risk of long-term neurodevelopmental impairments, such as cognitive and motor deficits.[4,5]

Adverse outcomes in extremely preterm infants largely arise from brain abnormalities acquired during their premature extrauterine life.[6] To mitigate long-term morbidity in these infants, targeted care and timely

Research in context

Evidence before this study

We searched PubMed from database inception to May 24, 2022, using the terms ("aEEG" OR "two-channel EEG" OR "single-channel EEG") AND ("neurodevelopmental outcomes" OR "outcome") AND "extremely preterm" NOT "review", with no language restrictions. Our search yielded 13 results, of which four original studies were identified using features extracted from amplitude-integrated electroencephalogram (aEEG) accompanied by raw EEG traces (aEEG–EEG) to predict long-term neurodevelopmental outcomes for extremely preterm infants (<28 weeks of gestation). These studies had relatively small sample sizes, ranging from 22 to 65. Of these studies, one study calculated EEG power, whereas the other three used distinct types of qualitative features. No studies were identified using machine learning-based prediction models or combining different types of aEEG–EEG features for outcome prediction in extremely preterm infants.

Added value of this study

In this study, we used a large, homogeneous retrospective cohort of extremely preterm infants, with aEEG–EEG data collected routinely over a 10-year period and applied machine learning-based regression and classification models to evaluate the predictive ability of an extensive set of qualitative and quantitative aEEG–EEG features on multiple long-term outcome measurements. By focusing on routine aEEG–EEG monitoring in the first 3 postnatal days, we had the opportunity to investigate whether it is feasible to detect extremely preterm infants with poor outcomes as early as possible. We found that the classification models showed acceptable performance at identifying infants with intellectual disabilities at early school age (5–7 years). Remarkably, these classifiers maintained the same level of performance when solely using quantitative features. Our findings contribute towards understanding the role of qualitative and quantitative aEEG–EEG features in predicting subsequent neurodevelopmental outcomes for extremely preterm infants.

Implications of all the available evidence

The current findings strengthen existing evidence and support the possibility of using early aEEG–EEG to automatically identify extremely preterm neonates at risk of long-term neurodevelopmental impairments. This could enable clinical teams to allocate more medical resources to newborn infants at high risk and implement appropriate and timely supportive care to facilitate optimal brain development. Furthermore, it is helpful in counselling parents about what to expect regarding their child's development and needs, thereby preparing them emotionally and practically.

interventions are vital to protect their vulnerable developing brains.[7] For this reason, there is a heightened interest in developing reliable brain-based markers to predict future outcomes in these infants at the earliest possible stage during their stay in the neonatal intensive care unit (NICU).[8]

Amplitude-integrated electroencephalogram (aEEG), in conjunction with its corresponding raw EEG traces (collectively referred to as aEEG–EEG), has gained increasing popularity worldwide as a useful bedside tool for monitoring the brain function of infants admitted to NICUs.[9,10] As a simplified alternative to conventional multichannel EEG, the aEEG–EEG uses only one or two channels and is easy to set up.[10] It can be started upon a neonate's admission to the NICU, enabling early detection and intervention of brain dysfunction. Moreover, both qualitative and quantitative aEEG–EEG parameters have shown great potential in predicting long-term outcomes for preterm infants.[9,11–18] When compared with neuroimaging markers such as brain MRI, aEEG–EEG shows similar performance in future outcome prediction[19] and allows the identification of preterm infants without any overt brain injury but at risk of poor outcomes.[20]

Despite the ever-growing body of literature exploring the role of aEEG–EEG in predicting outcomes for preterm infants, several crucial aspects require further investigation to enhance our understanding of this topic. One of the primary concerns is the scarcity of research in extremely preterm infants.[13] In existing studies, these infants are often merely incorporated as a subgroup, and the sample size is relatively small. Considering the weekly evolution of aEEG–EEG characteristics during the preterm period,[21] findings from other age groups cannot be directly applied to these infants.

The diversity in procedural and analytical approaches among existing studies adds complexity to drawing consistent conclusions.[22] One notable issue lies in outcome measurement: the assessment time varies across studies, and the use of internationally standardised scales is not guaranteed. Another complication is the inconsistency in the timing of aEEG–EEG data collection. To identify the earliest possible brain-based markers, assessing the predictive value of aEEG–EEG features within the first postnatal days could be of value.

Furthermore, previous research often focuses on a specific aEEG–EEG feature, ignoring the complementarity between different feature types. This oversight could potentially hinder the improvement of outcome prediction performance.[23] With the advent of machine learning prediction algorithms, we can now process an extensive set of input features. Unlike conventional statistical inferences (eg, group difference or in-sample association), which dominated previous research on the relationship between aEEG–EEG and outcomes, the machine learning algorithms enable

making predictions on an individual level, which are sought after for future precision medicine.[24]

In this context, we aimed to examine whether aEEG–EEG features obtained during the first 3 postnatal days could serve as early indicators for neurodevelopmental outcomes at preschool and early school age within a large, homogeneous cohort of extremely preterm infants. A comprehensive set of qualitative and quantitative aEEG–EEG features was extracted and fed into machine learning-based regression and classification models to evaluate their predictive value for future outcomes. The machine learning regression models were used to explore the relationships between aEEG–EEG features and outcomes. The machine learning-based classification models were used to distinguish between infants with optimal and impaired outcomes, aiming to make the findings clinically applicable.

Methods

Study design and population

This retrospective cohort study examined extremely preterm infants (born before 28 weeks and 0 days of gestation) admitted to the NICU of the Wilhelmina Children's Hospital (WKZ), Utrecht, the Netherlands, between June 1, 2008, and Sept 30, 2018. Only infants who received continuous two-channel aEEG–EEG monitoring within the first 3 postnatal days and did not have genetic or metabolic diseases or major congenital malformations were eligible for inclusion. Permission to use the patient data was obtained from the Medical Research Ethics Committee of the University Medical Centre Utrecht (protocol number 20-660-C). Because all data used in this study were collected as part of standard medical care and the analysis was retrospective, written parental consent was not required. All data were pseudonymised and de-identified before analysis.

Procedures

As part of standard care for extremely preterm infants admitted to the NICU of the WKZ, two-channel aEEG–EEG monitoring was started as soon as possible after a neonate's admission and maintained at the bedside for the first 3 postnatal days. The BrainZ monitor (BRM2 or BRM3; Natus Medical, Seattle, WA, USA) was used for the aEEG–EEG recording. Raw EEG signals were recorded from pairs of needle electrodes subcutaneously placed over the frontoparietal cortex (left channel: F3-P3, right channel: F4-P4) at a sampling rate of 256 Hz and were subsequently processed to generate aEEG traces. Additionally, a reference electrode was placed over the vertex (Cz).

All aEEG–EEG traces were first reviewed by experienced clinicians (CT, LW, and MLT) using Analyze Research software (version 2.0, BrainZ Instruments). For each of the 3 days, a so-called best hour period of aEEG–EEG data, characterised by more mature patterns and fewer artifacts, was manually selected within specific time periods: 20–24 h for day 1, 44–48 h for day 2, and 68–72 h for day 3. The best hour of aEEG–EEG data was used for subsequent qualitative background pattern assessment and quantitative feature calculation due to the sensitivity of these features to artifacts (figure 1).

The experienced clinicians (CT, LW, and MLT) further conducted qualitative aEEG–EEG analysis using the Hellström-Westas classification system.[25] To ensure the accuracy and reliability of the qualitative classification, at least two clinicians jointly reviewed and evaluated aEEG–EEG traces during each analysis, while remaining

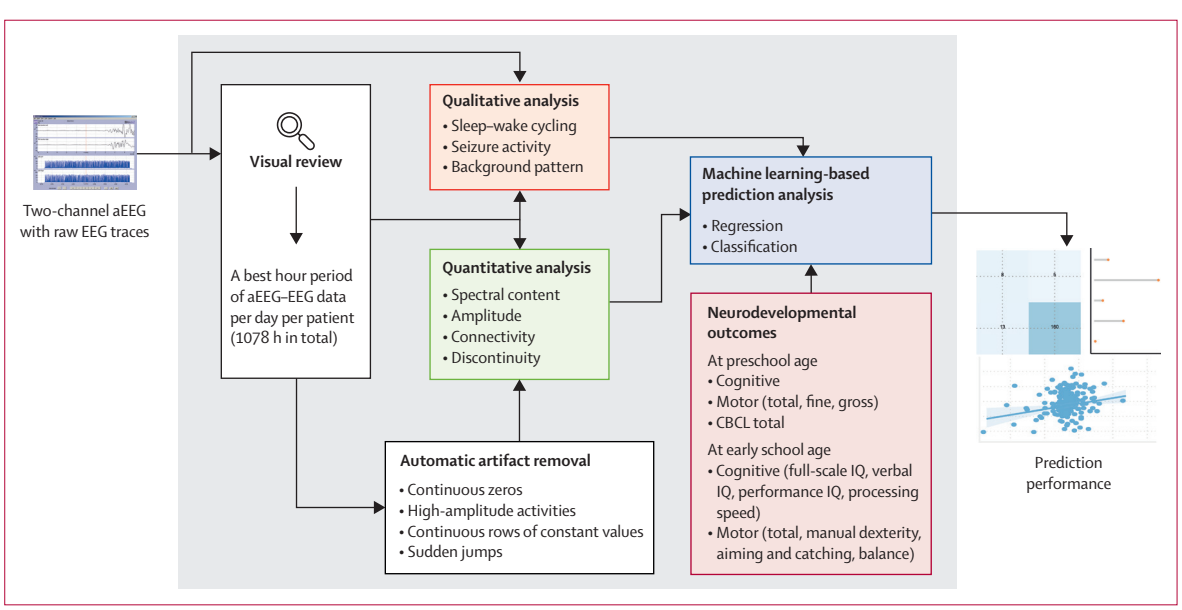

Figure 1: **Overview of the entire analysis pipeline**
aEEG=amplitude-integrated electroencephalogram. CBCL=Child Behavior Checklist. EEG=electroencephalogram. IQ=intelligence quotient.

masked to patients' clinical characteristics and outcomes throughout the process.

The qualitative analysis involved three different feature types: background pattern, sleep–wake cycling, and seizure activity. During background pattern evaluation, the selected best hour of aEEG–EEG data for each corresponding day was segmented into six discrete epochs, each lasting 10 min. From these epochs, the most prevalent category was extracted and used for further analysis. The evaluation of sleep–wake cycling and seizure activity was conducted using complete data for each day.

Each of these features could be further classified as multiple subtypes (appendix p 2). However, not all these subtypes are commonly observed in extremely preterm infants. To focus on the most relevant information and facilitate comprehension, we converted the initial classification results of each feature into a binary form: normal or abnormal (appendix p 2).

The quantitative analysis relied on raw EEG signals and was performed using Matlab scripts (MathWorks, Natick, MA, USA) that were developed in-house, together with a Matlab software package NEURAL.[26]

Before extracting quantitative features, the visually selected periods of raw EEG data were preprocessed for automated artifact removal and noise reduction. Data segments containing obvious artefacts—continuous zeros, high-amplitude activities, continuous rows of constant values, and sudden jumps—were removed. Subsequently, a band-pass filter of 0·5–30 Hz was used to remove low-frequency and high-frequency noises, and a notch filter at 50 Hz was applied to attenuate power line interference.

From the preprocessed EEG data, a set of 110 quantitative features was computed. These features can be grouped into four distinct categories: spectral content (26 features), amplitude (56 features), connectivity (21 features), and discontinuity (seven features). The properties of these features are detailed in appendix (pp 3–5).

Both qualitative and quantitative aEEG–EEG features were calculated for each of the 3 days.

See Online for appendix

	Extremely preterm infants (n=369)
Maternal and demographic characteristics	
Sex*	
Female	164 (44%)
Male	205 (56%)
Gestational age at birth, weeks	26·4 (1·1; 23·9–27·9)
Birthweight†, g	880 (180)
Maternal education level	
No education	1 (<1%)
Primary education	11 (3%)
Some secondary education	37 (10%)
Completed secondary education	80 (22%)
University education	96 (26%)
Missing	144 (39%)
Clinical characteristics during the NICU stay	
Morphine	
Yes	222 (60%)
No	142 (39%)
Missing	5 (1%)
Morphine cumulative dose‡, mg/kg	1·3 (2·3)
The administration of antiseizure, sedative, or anaesthetic medications§	
Yes	140 (38%)
No	229 (62%)
Illness severity	
Severe	183 (50%)
Mild	181 (49%)
Missing	5 (1%)
The presence of severe brain injury	
Yes	118 (32%)
No	246 (67%)
Missing	5 (1%)
Apgar score	
At 1st min after birth¶	5 (3–7)
At 5th min after birth‖	8 (6–8)
At 10th min after birth**	8 (8–9)
Follow-up age	
Age at the time of BSID-III administration, years	2·5 (0·2; 2·2–3·0)
Age at the time of CBCL administration, years	2·5 (0·2; 2·2–3·0)
Age at the time of WPPSI-III administration, years	5·9 (0·2; 5·1–6·8)
Age at the time of MABC-2 administration, years	5·9 (0·2; 5·1–7·0)

Data are n (%), mean (SD; range), mean (SD), or median (IQR). BSID-III=Bayley Scales of Infant and Toddler Development, Third Edition. CBCL=Child Behavior Checklist. MABC-2=Movement Assessment Battery for Children, Second Edition. NICU=neonatal intensive care unit. WPPSI-III=Wechsler Preschool & Primary Scale of Intelligence, Third Edition. *The use of "female" and "male" refers to sex assigned at birth. †One infant with missing data. ‡11 infants with missing data. §The antiseizure, sedative, or anaesthetic medications include phenobarbital, lidocaine, levetiracetam, clonazepam, midazolam, and other potential surgical anaesthetics. ¶Seven infants with missing data. ‖Four infants with missing data. **86 infants with missing data.

Table 1: **Patient characteristics**

Outcomes

Neurodevelopmental outcomes for extremely preterm infants were measured during their regular follow-up visits at the outpatient clinic of the WKZ at preschool age (2–3 years) and early school age (5–7 years) by raters masked to patients' aEEG–EEG characteristics.

At preschool age, the Bayley Scales of Infant and Toddler Development, Third Edition (BSID-III), were administered to evaluate cognitive and motor functioning.[27] We used four index scores provided by the BSID-III—cognitive composite score, total motor composite score, and fine and gross motor subscaled scores. The Child Behavior Checklist (CBCL) was used to assess behavioural or emotional problems, yielding a total problem score.[28]

At early school age, the Wechsler Preschool & Primary Scale of Intelligence, Third Edition (WPPSI-III), was used to assess cognitive abilities by providing a full-scale intellectual quotient (IQ), and composite scores for verbal IQ, performance IQ, and processing speed.[29] The

Movement Assessment Battery for Children, Second Edition (MABC-2), was used to evaluate motor abilities by generating an overall estimate of the child's motor performance and three subscaled scores: manual dexterity, aiming and catching, and balance.[30]

The administration of these outcome measurements was carried out using their Dutch versions. All outcome measurements are standardised, norm-referenced tests. Lower scores on the BSID-III, WPPSI-III, and MABC-2 indicate poorer functioning, whereas higher scores on the CBCL indicate more behavioural or emotional problems. For classification purposes, each outcome was converted into a binary form (optimal *vs* impaired) using thresholds set at –2 SD or +2 SD from the normative mean value. For measurements with mean values of 100 (SD 15), including BSID-III composite scores and WPPSI-III scores, a score of 70 or less was categorised as impaired. More specifically, an IQ score of 70 or less on the WPPSI-III at age 5–7 years indicates intellectual disability. For measurements with mean values of 10 (SD 3), including BSID-III subscaled scores and MABC-2 scores, a score of 4 or less was categorised as impaired. For the CBCL total problem score with a mean value of 50 (SD 10), a score of 70 or higher was categorised as impaired.

Statistical analysis

The relationship between aEEG–EEG and outcomes can be affected by a range of patient-related factors. To accurately determine the inherent predictive value of aEEG–EEG for future outcomes, several crucial confounders were considered in subsequent prediction analysis. These confounders were gestational age at birth, maternal education, illness severity, morphine cumulative dose, the presence of severe brain injury, and the administration of antiseizure, sedative, or anaesthetic medications (appendix p 6). A Spearman's correlation coefficient (r_s) matrix was created to describe the relationship between each pair of the aEEG–EEG features, confounders, and outcomes. Before being fed into a prediction model, each aEEG–EEG feature was adjusted for confounders using ordinary least-squares regression. Missing values were imputed with the median (for continuous variables) or most frequent values (for categorical variables).

For each outcome measurement, a support vector regression model and a histogram-based gradient-boosting classification model were built. Each prediction (regression or classification) model was trained through a nested 3-fold cross-validation procedure, in which a grid search was used to find the optimal hyperparameters. For classification models specifically, the folds were stratified to maintain the proportion of samples for each class. Moreover, we only built classification models for outcomes with a minimum of nine samples per class. This requirement ensured that each inner fold contained at least two samples, enabling a robust hyperparameter optimisation process.

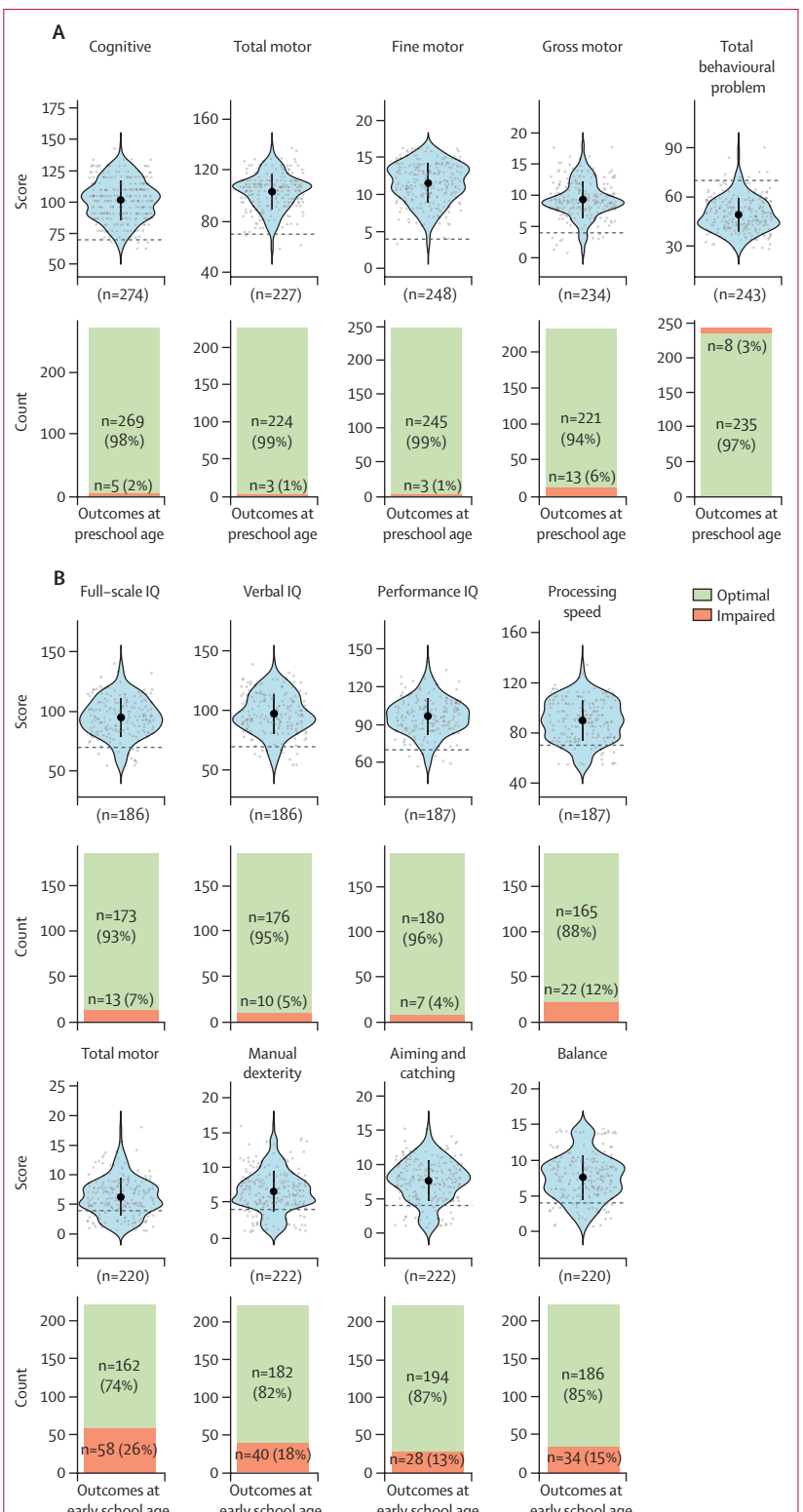

Figure 2: **Distributions of neurodevelopmental outcomes at preschool age (A) and early school age (B)**
On each violin plot, the central black dot denotes the mean, the vertical line through the dot represents the SD, and the horizontal dash line indicates the threshold of –2 SD or +2 SD from the normative mean value. The number of infants who had outcome measurements is indicated below each violin plot. IQ=intelligence quotient.

Subsequently, a leave-one-subject-out cross-validation procedure was conducted to evaluate the out-of-sample predictive performance of each model with the best hyperparameters. The performance of regression models was measured by the Pearson's correlation coefficient (r) and mean square error (MSE) between actual and predicted outcome scores. r ranges from –1 to 1, where 1 indicates a perfect match, 0 implies a chance prediction, and a negative value suggests an even worse performance. The MSE ranges from 0 to infinity, where 0 represents a perfect fit, whereas increasing values mean decreasing performance. The performance of classification models was measured by balanced accuracy and the Fβ score. Both metrics range from 0 (completely wrong) to 1 (completely correct), with higher values being better. A balanced accuracy of 0·5 is equivalent to random guessing. The Fβ score is a robust metric that considers both precision and recall by calculating a weighted harmonic mean of the two. We used a β value of 10 to assign more weight to recall than to precision.

The significance of these performance metrics was evaluated using permutation tests, in which outcome scores were randomly shuffled 1000 times per test. The p value is calculated as (n+1) divided by 1001, where n represents the number of times that randomised values yield results superior to the original result from the non-permuted data. The significance level was set at p values of less than 0·05. For a regression model, significance in both r and MSE indicated better-than-chance performance, and for a classification model significance in both balanced accuracy and Fβ score indicated better-than-chance performance. We further estimated 95% CIs for their performance metrics based on 1000 bootstrap resamples. Finally, we used Shapley additive explanations (SHAP) to evaluate the contributions of each feature category and features from each day to outcome predictions, with higher SHAP values indicating greater contributions.

The descriptive statistical analysis and data visualisation were implemented using Python (version 3.10.9) and R (version 4.2.3) in RStudio (version 2023.03.0+386). The machine learning-based prediction analyses were implemented using scikit-learn (version 1.2.1) within Python (version 3.10.9).

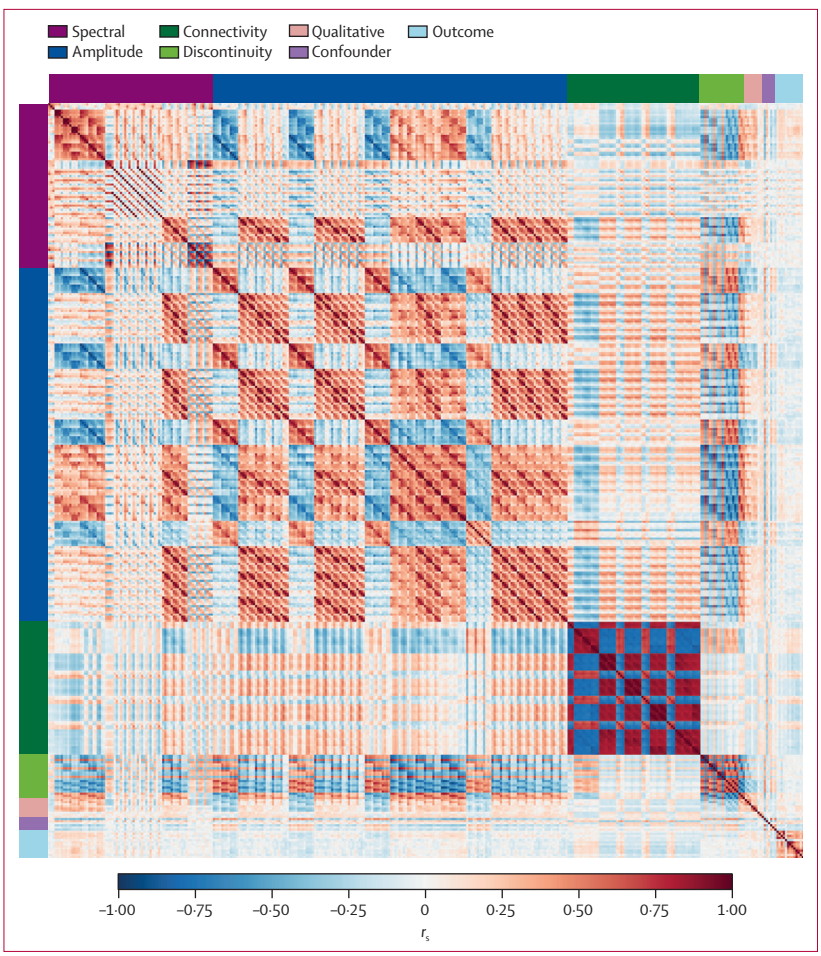

Figure 3: **Spearman's correlation coefficient matrix of aEEG–EEG features, confounders, and outcomes**
Different types of aEEG–EEG features and outcomes are represented in different colours according to the colour bars located at the top and left side of the matrix. Within the matrix, red refers to positive and blue to negative Spearman's correlation coefficient (r_s) values. The intensity of the colour gradient corresponds to the strength of the correlation, with darker shades indicating stronger correlations and lighter shades indicating weaker correlations. aEEG=amplitude-integrated electroencephalogram. EEG=electroencephalogram.

Role of the funding source
The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

Results

A total of 601 extremely preterm infants were screened. Of these infants, 108 were excluded due to the absence of aEEG–EEG data or the presence of congenital birth defects, and 124 were excluded due to insufficient, non-valid, or poor-quality aEEG–EEG data. Ultimately, 369 infants met the eligibility criteria and were included in this study. These infants' demographics and clinical characteristics during their NICU stay are summarised in table 1.

343 (93%) of 369 enrolled infants had aEEG–EEG data available for all three time periods (ie, 20–24 h, 44–48 h, and 68–72 h), whereas the remaining infants had data available for one (three [1%]) or two (23 [6%]) of the time periods. In total, 1078 h of aEEG–EEG data with relatively good quality were obtained and used for feature extraction (figure 1). A total of 339 aEEG–EEG features—nine qualitative and 330 quantitative features—were obtained for each infant. The distributions of these features over the three periods are shown in the appendix (pp 7–8).

Of the 369 infants, a minimum of 227 (62%; total motor) and a maximum of 274 (74%; cognitive) had outcomes measured at preschool age and a minimum of 186 (50%; full-scale IQ) and a maximum of 222 (60%; aiming and catching) at early school age (figure 2). At preschool age, a range of three (1%) to 13 (6%) infants

exhibited impaired outcomes, and at early school age a range of seven (4%) to 58 (26%) infants exhibited impaired outcomes. The age details at the time of outcome measurements are shown in table 1.

A 358×358 Spearman's correlation coefficient matrix is shown in figure 3, illustrating the pairwise relationships between the 339 aEEG–EEG features, six confounders, and 13 outcomes. The correlations within the same category of aEEG–EEG features seemed to be stronger than the inter-category correlations. The correlations between individual aEEG–EEG features and outcomes were relatively weak (absolute values <0·29). Additionally, the confounders exhibited obvious correlations with both aEEG–EEG features and outcomes, indicating their potential effects on the relationships between aEEG–EEG and outcomes.

In machine learning-based regression analyses, nine of 13 outcome measurements were significantly predicted by aEEG–EEG features (table 2; appendix p 9). However, the prediction performance of the significant models was relatively low (ranging from r=0·13 to r=0·23).

Eight outcomes had sufficient infants per class (optimal and impaired) for robust hyperparameter optimisation in classification analysis (figure 2). Among them, gross motor and aiming and catching scores were not significantly predicted by aEEG–EEG features using regression models (table 2). Therefore, the two outcomes were excluded from further classification analysis, resulting in the development of six classification models (table 3). The classifiers for full-scale and verbal IQ scores achieved significant performance (table 3; figure 4A). The optimal hyperparameters for the two classifiers are detailed in the appendix (p 10).

The highest contribution to the full-scale IQ classifier, indicated by mean absolute SHAP values, came from amplitude features, followed by discontinuity, spectral, connectivity, and qualitative features (figure 4B). Additionally, features from day 2 contributed most to the full-scale IQ classifier, followed by day 3 and day 1 (figure 4C). For the verbal IQ classifier, only discontinuity features from day 3 played a part in making accurate predictions.

We further trained classifiers based solely on quantitative features to predict full-scale and verbal IQ scores, which achieved identical performance as the complete feature set: for full-scale IQ, balanced accuracy was 0·77 (95% CI 0·63–0·91, p=0·0030) and the Fβ

	r	Permutation p value for r	Mean square error	Permutation p value for mean square error
Outcomes at preschool age				
BSID-III				
Cognitive composite score	0·14	0·013	38·72	0·013
Total motor composite score	0·10	0·056	57·32	0·058
Fine motor scaled score	0·13	0·014	19·95	0·010
Gross motor scaled score	0·08	0·13	1·08	0·055
CBCL				
Total behavioural problem score	0·13	0·023	7·55	0·022
Outcomes at early school age				
WPPSI-III				
Full-scale IQ score	0·22	0·0050	1·09	0·013
Verbal IQ score	0·23	0·0030	1·14	0·040
Performance IQ score	0·19	0·0080	1·17	0·057
Processing speed score	0·18	0·017	1·11	0·019
MABC-2				
Total motor score	0·20	0·0010	4·75	0·0010
Manual dexterity score	0·21	0·0020	2·27	0·0030
Aiming and catching score	0·11	0·055	2·81	0·064
Balance score	0·17	0·0060	6·40	0·0060

BSID-III=Bayley Scales of Infant and Toddler Development, Third Edition. CBCL=Child Behavior Checklist. IQ=intelligence quotient. MABC-2=Movement Assessment Battery for Children, Second Edition. WPPSI-III=Wechsler Preschool & Primary Scale of Intelligence, Third Edition.

Table 2: **Prediction performance of machine learning-based regression models**

	Balanced accuracy (95% CI)	Permutation p value for balanced accuracy	Fβ score (95% CI)	Permutation p value for Fβ score	Precision (95% CI)	Recall (95% CI)
WPPSI-III						
Full-scale IQ score	0·77 (0·62–0·90)	0·0020	0·61 (0·33–0·88)	0·0020	0·38 (0·17–0·61)	0·62 (0·33–0·89)
Verbal IQ score	0·81 (0·65–0·96)	0·0010	0·69 (0·37–0·98)	0·0010	0·35 (0·14–0·56)	0·70 (0·38–1·00)
Processing speed score	0·55 (0·47–0·63)	0·095	0·18 (0·04–0·35)	0·089	0·22 (0·05–0·43)	0·18 (0·04–0·35)
MABC-2						
Total motor score	0·51 (0·46–0·57)	0·31	0·21 (0·10–0·32)	0·31	0·29 (0·16–0·44)	0·21 (0·10–0·32)
Manual dexterity score	0·54 (0·48–0·61)	0·11	0·20 (0·08–0·33)	0·12	0·28 (0·11–0·44)	0·20 (0·08–0·33)
Balance score	0·52 (0·46–0·59)	0·27	0·15 (0·04–0·27)	0·24	0·20 (0·05–0·37)	0·15 (0·04–0·27)

The Fβ score was calculated with a β value of 10. IQ=intelligence quotient. MABC-2=Movement Assessment Battery for Children, Second Edition. WPPSI-III=Wechsler Preschool and Primary Scale of Intelligence, Third Edition.

Table 3: **Prediction performance of machine learning-based classification models**

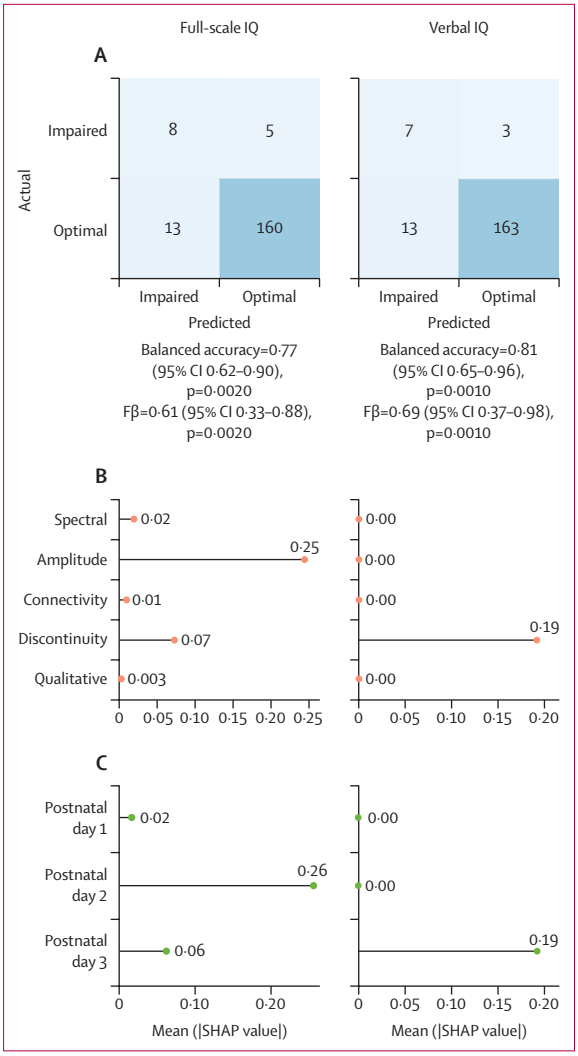

Figure 4: **Classification performance for full-scale IQ and verbal IQ**
Confusion matrix of predicted and actual scores (optimal and impaired; A). The confusion matrix displays the count values of correct and incorrect predictions, visually represented using a gradient. Lighter shades indicate a smaller count, whereas darker shades indicate a larger count. Individual contributions of each feature category (B) and features from each day (C) are represented using SHAP values. Higher absolute SHAP values indicate more contributions to the classification models. IQ=intelligence quotient. SHAP=Shapley additive explanations.

score was 0·61 (95% CI 0·33–0·88, p=0·0030); and for verbal IQ, balanced accuracy was 0·81 (0·65–0·96, p=0·0010) and the Fβ score was 0·69 (0·37–0·98, p=0·0010).

Discussion

This study examined the predictive potential of qualitative and quantitative features extracted from two-channel aEEG–EEG within the first 3 postnatal days for long-term neurodevelopmental outcomes in a large cohort of extremely preterm neonates. Our results revealed that the aEEG–EEG features had significant but relatively weak predictive power for nine of 13 outcomes when using machine learning-based regression models. Nonetheless, machine learning-based classifiers exhibited good performance in distinguishing between infants with intellectual disabilities and those with optimal outcomes at early school age, reaching a balanced accuracy of 0·77 for full-scale IQ and 0·81 for verbal IQ.

Our findings establish the clinical utility of aEEG–EEG features in the first few postnatal days. Different from previous work with small sample sizes,[14–17] the current study enrolled a relatively large cohort of 369 extremely preterm infants, allowing for more robust evidence. By extracting a comprehensive set of features, we provided, for the first time, a fuller picture of aEEG–EEG feature distribution in the extremely preterm population and captured the relationships between these features. Furthermore, the large aEEG–EEG feature set allowed us to develop machine learning-based models for performing individual-level outcome predictions, paving the way for precision medicine.[8,23]

The results that classifiers of full-scale and verbal IQ scores using only quantitative features performed equivalently to models using both qualitative and quantitative features carry substantial clinical implications. Quantitative aEEG–EEG features, compared with qualitative ones, have benefits such as objectivity, convenience, and rapid availability.[13,23] Therefore, our findings open up opportunities to develop a fully automated bedside tool that provides prognosis information, aiding clinicians in decision making and resource allocation. For example, based on predictions from the tool, clinicians can identify neonates at risk of neurodevelopmental impairments and initiate interventions aimed at safeguarding the developing brain at an early stage, ultimately leading to improved outcomes.

The contributions of features to the classification models were ranked using SHAP values. Here, it is worth mentioning that a SHAP value only measures the importance of a given feature to the developed model rather than its real-world importance. Given that predictions from the developed classifiers can be incorrect, SHAP values might not always accurately represent reality. Therefore, even though qualitative features had limited contribution to the IQ classifiers in this study, we should not dismiss their potential value in guiding the development of quantitative metrics and other clinical scenarios.

It is important to note that an extremely preterm infant's future outcomes do not hinge entirely on their first few days of life. To enhance the chances of positive outcomes for these infants, applying outcome predictions at multiple timepoints could be beneficial. For example, longitudinal outcome prediction analyses can be carried out in the NICU, which allows a progressive adaptation of an infant's care plan. Moreover, as an infant matures, a wider range of clinical, neuromonitoring, and neuroimaging data, such as MRI, cerebral ultrasound, and near-infrared spectroscopy, become available. Integrating aEEG–EEG

with these modalities can assist in generating more robust predictions. Further research is needed to explore the feasibility of these proposed concepts.

The current work is limited by its monocentric nature. Although we used a relatively large cohort (n=369) together with a cross-validation procedure to improve statistical power, it would be helpful for future studies to use independent samples to validate our findings. Another concern relates to the challenge of identifying all potential confounders affecting the EEG–outcome relationship, especially post-NICU factors. In addition, although the two-channel aEEG–EEG is user friendly, it provides limited information about brain function, which could limit the outcome prediction performance. By contrast, conventional multichannel EEG offers more information, but its application to extremely preterm neonates can be challenging due to their tiny heads and fragile skin. Thus, future research should consider use of novel techniques, such as dry electrodes, to investigate the possibility of applying multichannel EEG during the first few postnatal days.

To summarise, we showed the value of two-channel aEEG–EEG features, gathered during the first 3 postnatal days, in predicting future neurodevelopmental outcomes of a large group of extremely preterm infants. Using machine learning-based models, we were, for the first time, able to combine and compare the performance of a comprehensive set of qualitative and quantitative aEEG–EEG features in outcome predictions. The quantitative characteristics showed superior predictive power than the qualitative parameters. Our findings provide the possibility of creating an automated tool for long-term disability prognosis, which can guide early personalised treatment in the NICU.

Contributors

MLT had the idea and conceived of the study. XW, CT, and MLT contributed to study design. CT and XW contributed to data collection. CT, LW, and MLT contributed to qualitative analysis. XW contributed to quantitative analysis, statistical analysis, and data visualisation. XW, CT, and MLT directly accessed and verified the data presented. HSdV, RMJCE, and CK-E provided their expertise in child development, and JD and MJNLB in preterm brain injury. XW, CT, and MLT wrote the first draft of the manuscript. LW, JD, HSdV, RMJCE, CK-E, MJNLB, and MLT critically reviewed and revised the manuscript. XW finished the editing of the manuscript. All co-authors approved the final version of the manuscript and had final responsibility for the decision to submit for publication.

Declaration of interests

We declare no competing interests.

Data sharing

Individual participant data that underlie the results reported in this Article, after de-identification, can be made available as part of further research collaborations. Interested parties should contact the corresponding author. Codes for data analysis are available on request from XW (x.wang-5@umcutrecht.nl), subject to authors' approval. Any data sharing will be subject to meeting the Privacy Regulations of University Medical Centre Utrecht, the General Data Protection Regulation, and the General Data Protection Regulation Implementation Act.

Acknowledgments

This work was funded by the European Commission (grant agreement number EU H2020 MSCA-ITN-2018-#813483, INtegrating Functional Assessment measures for Neonatal Safeguard). We thank all parents and infants involved in this study. We thank all WKZ neonatal clinicians and nurses who implemented aEEG–EEG in their daily work. We also thank all those who have given valuable advice and comments to this study.

References

1 Costeloe KL, Hennessy EM, Haider S, Stacey F, Marlow N, Draper ES. Short term outcomes after extreme preterm birth in England: comparison of two birth cohorts in 1995 and 2006 (the EPICure studies). *BMJ* 2012; **345:** e7976.

2 Glass HC, Costarino AT, Stayer SA, Brett CM, Cladis F, Davis PJ. Outcomes for extremely premature infants. *Anesth Analg* 2015; **120:** 1337–51.

3 Stoll BJ, Hansen NI, Bell EF, et al. Trends in care practices, morbidity, and mortality of extremely preterm neonates, 1993–2012. *JAMA* 2015; **314:** 1039–51.

4 Rogers EE, Hintz SR. Early neurodevelopmental outcomes of extremely preterm infants. *Semin Perinatol* 2016; **40:** 497–509.

5 Twilhaar ES, Wade RM, de Kieviet JF, van Goudoever JB, van Elburg RM, Oosterlaan J. Cognitive outcomes of children born extremely or very preterm since the 1990s and associated risk factors: a meta-analysis and meta-regression. *JAMA Pediatr* 2018; **172:** 361–67.

6 Volpe JJ. Dysmaturation of premature brain: importance, cellular mechanisms, and potential interventions. *Pediatr Neurol* 2019; **95:** 42–66.

7 Bonifacio SL, Van Meurs K. Neonatal neurocritical care: providing brain-focused care for all at risk neonates. *Semin Pediatr Neurol* 2019; **32:** 100774.

8 Tataranno ML, Vijlbrief DC, Dudink J, Benders MJNL. Precision medicine in neonates: a tailored approach to neonatal brain injury. *Front Pediatr* 2021; **9:** 634092.

9 El-Dib M, Abend NS, Austin T, et al. Neuromonitoring in neonatal critical care part II: extremely premature infants and critically ill neonates. *Pediatr Res* 2023; **94:** 55–63.

10 Toet MC, Lemmers PMA. Brain monitoring in neonates. *Early Hum Dev* 2009; **85:** 77–84.

11 Iyer KK, Roberts JA, Hellström-Westas L, et al. Cortical burst dynamics predict clinical outcome early in extremely preterm infants. *Brain* 2015; **138:** 2206–18.

12 Klebermass K, Olischar M, Waldhoer T, Fuiko R, Pollak A, Weninger M. Amplitude-integrated EEG pattern predicts further outcome in preterm infants. *Pediatr Res* 2011; **70:** 102–08.

13 van 't Westende C, Geraedts VJ, van Ramesdonk T, et al. Neonatal quantitative electroencephalography and long-term outcomes: a systematic review. *Dev Med Child Neurol* 2022; **64:** 413–20.

14 Nordvik T, Schumacher EM, Larsson PG, Pripp AH, Løhaugen GC, Stiris T. Early spectral EEG in preterm infants correlates with neurocognitive outcomes in late childhood. *Pediatr Res* 2022; **92:** 1132–39.

15 Richardson J, Goshen S, Meledin I, Golan A, Goldstein E, Shany E. Predictive value of early amplitude integrated EEG in extremely premature infants. *J Child Neurol* 2020; **35:** 737–43.

16 Shibasaki J, Toyoshima K, Kishigami M. Blood pressure and aEEG in the 96h after birth and correlations with neurodevelopmental outcome in extremely preterm infants. *Early Hum Dev* 2016; **101:** 79–84.

17 Welch C, Helderman J, Williamson E, O'Shea TM. Brain wave maturation and neurodevelopmental outcome in extremely low gestational age neonates. *J Perinatol* 2013; **33:** 867–71.

18 Wikström S, Pupp IH, Rosén I, et al. Early single-channel aEEG/EEG predicts outcome in very preterm infants. *Acta Paediatr* 2012; **101:** 719–26.

19 Leen AT, Ouwehand S, Olsthoorn M, et al. Electroencephalography and brain magnetic resonance imaging in asphyxia comparing cooled and non-cooled infants. *Eur J Paediatr Neurol* 2019; **23:** 181–90.

20 Cainelli E, Vedovelli L, Wigley ILCM, Bisiacchi PS, Suppiej A. Neonatal spectral EEG is prognostic of cognitive abilities at school age in premature infants without overt brain damage. *Eur J Pediatr* 2021; **180:** 909–18.

21 Bourel-Ponchel E, Gueden S, Hasaerts D, et al. Normal EEG during the neonatal period: maturational aspects from premature to full-term newborns. *Neurophysiol Clin* 2021; **51:** 61–88.

22 Fogtmann EP, Plomgaard AM, Greisen G, Gluud C. Prognostic accuracy of electroencephalograms in preterm infants: a systematic review. *Pediatrics* 2017; **139:** e20161951.

23 O'Toole JM, Boylan GB. Quantitative preterm EEG analysis: the need for caution in using modern data science techniques. *Front Pediatr* 2019; **7:** 174.

24 van Boven MR, Henke CE, Leemhuis AG, et al. Machine learning prediction models for neurodevelopmental outcome after preterm birth: a scoping review and new machine learning evaluation framework. *Pediatrics* 2022; **150:** e2021056052.

25 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.

26 Toole JMO, Boylan GB. NEURAL: quantitative features for newborn EEG using Matlab. *arXiv* 2017; published online April 19. https://doi.org/10.48550/arXiv.1704.05694 (preprint).

27 Bayley N. Bayley Scales of Infant and Toddler Development, Third Edition. San Antonio, TX: Pearson, 2006.

28 Achenbach TM, Ruffle TM. The Child Behavior Checklist and related forms for assessing behavioral/emotional problems and competencies. *Pediatr Rev* 2000; **21:** 265–71.

29 Wechsler D. Wechsler Preschool & Primary Scale Of Intelligence, 3rd Edition. San Antonio, TX: The Psychological Corporation San Antonio, 2002.

30 Henderson SE, Sugden D, Barnett AL. Movement Assessment Battery for Children-2. Washington, DC: APA PsycTests, 1992.