FULL-LENGTH ORIGINAL RESEARCH

Interrater agreement in the interpretation of neonatal electroencephalography in hypoxic-ischemic encephalopathy

*Courtney J. Wusthoff, †Joseph Sullivan, †Hannah C. Glass, ‡Renée A. Shellhaas, §Nicholas S. Abend, ¶Taeun Chang, and ¶Tammy N. Tsuchida

Epilepsia, 58(3):429–435, 2017
doi: 10.1111/epi.13661

Dr. Courtney Wusthoff is an assistant professor of neurology and pediatrics at Stanford University

Summary

Objective: Research using neonatal electroencephalography (EEG) has been limited by a lack of a standardized classification system and interpretation terminology. In 2013, the American Clinical Neurophysiology Society (ACNS) published a guideline for standardized terminology and categorization in the description of continuous EEG in neonates. We sought to assess interrater agreement for this neonatal EEG categorization system as applied by a group of pediatric neurophysiologists.

Methods: A total of 60 neonatal EEG studies were collected from three institutions. All EEG segments were from term neonates with hypoxic-ischemic encephalopathy. Three pediatric neurophysiologists independently reviewed each record using the ACNS standardized scoring system. Unweighted kappa values were calculated for interrater agreement of categorical data across multiple observers.

Results: Interrater agreement was very good for identification of seizures ($\kappa = 0.93$, $p < 0.001$), with perfect agreement in 95% of records (57 of 60). Interrater agreement was moderate for classifying records as normal or having any abnormality ($\kappa = 0.49$, $p < 0.001$), with perfect agreement in 78% of records (47 of 60). Interrater agreement was good in classifying EEG backgrounds on a 5-category scale (normal, excessively discontinuous, burst suppression, status epilepticus, or electrocerebral inactivity) ($\kappa = 0.70$, $p < 0.001$), with perfect agreement in 72% of records (43 of 60). Other specific background features had lower agreement, including voltage ($\kappa = 0.41$, $p < 0.001$), variability ($\kappa = 0.35$, $p < 0.001$), symmetry ($\kappa = 0.18$, $p = 0.01$), presence of abnormal sharp waves ($\kappa < 0.20$, $p < 0.05$), and presence of brief rhythmic discharges ($\kappa < 0.20$, $p < 0.05$).

Significance: We found good or very good interrater agreement applying the ACNS system for identification of seizures and classification of EEG background. Other specific EEG features showed limited interrater agreement. Of importance to both clinicians and researchers, our findings support using the ACNS system in identifying seizures and classifying backgrounds of neonatal EEG recordings, but also suggest limited reproducibility for certain other EEG features.

Accepted December 9, 2016; Early View publication February 6, 2017.

*Divisions of Child Neurology and Neonatal and Developmental Medicine, Stanford University, Palo Alto, California, U.S.A.; †Departments of Neurology and Pediatrics, San Francisco Benioff Children's Hospital, University of California, San Francisco, California, U.S.A.; ‡Department of Pediatrics and Communicable Diseases (Division of Pediatric Neurology), University of Michigan, Ann Arbor, Michigan, U.S.A.; §Departments of Neurology and Pediatrics, The Perelman School of Medicine at the University of Pennsylvania and the Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, U.S.A.; and ¶Division of Neurophysiology, Epilepsy and Critical Care, Children's National Health System, Washington, District of Columbia, U.S.A.

Address correspondence to Courtney J. Wusthoff, Division of Child Neurology, Stanford University, 750 Welch Road, Suite 317, Palo Alto, CA 94304, U.S.A. E-mail: wusthoff@stanford.edu

Wiley Periodicals, Inc.

KEY WORDS: Electroencephalography, Hypoxic-ischemic encephalopathy, Neonate, Seizure, Neurocritical care.

Key Points

- We assessed interrater agreement for the American Clinical Neurophysiology Society (ACNS) standardized neonatal EEG categorization system
- We found good or very good interrater agreement for identification of seizures and categorization of neonatal EEG background
- We found limited interrater agreement for other background features and abnormal sharp waves
- Although identification of seizures and background are reproducible in neonatal EEG, identification of other features may be less reliable

Electroencephalography (EEG) is an important tool for assessing brain function in neonates, for whom clinical observation alone is often insufficient to accurately evaluate cerebral function. EEG is used for two main purposes in neonates. First, continuous video-EEG monitoring is the optimal modality for the identification and diagnosis of neonatal seizures, because >80% of neonatal seizures are subclinical and there is poor observer agreement in correctly identifying seizures based on clinical semiology alone.[1–4] Second, EEG is used to assess the severity of brain dysfunction in neonates with known or suspected brain injury; abnormal EEG background features may serve as indicators of unfavorable long-term prognosis.[5–7]

Despite these widely accepted uses of EEG, until recently, there was no unified approach to the interpretation of neonatal EEG. A recent study of neonatal EEG recordings at a single institution reported good interobserver agreement in identification of seizures using multichannel EEG, but did not attempt to assess agreement for other EEG features.[8] Individual studies typically developed idiosyncratic EEG scoring methods, and have relied on small numbers of patients to demonstrate an association between these scores and either concurrent neurologic injury or subsequent outcomes. The lack of standardized terminology for the description of neonatal EEG has been a barrier to understanding the clinical significance of specific EEG findings, and to conducting multicenter research using neonatal EEG.

In 2013, the American Clinical Neurophysiology Society (ACNS) published a guideline for standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates.[9] The terminology, developed by consensus among an ad hoc committee of experts, was intended to parallel terminology used to describe continuous EEG monitoring in children and adults, and to improve reproducibility of research efforts by providing a common standard for EEG interpretation in neonates. The aim of the current study was to assess the interrater agreement for the ACNS neonatal EEG terminology among three pediatric neurophysiologists.

Methods

Raters and EEG records

Three centers (Children's Hospital of Philadelphia [CHOP], Children's National Health System [CNHS], and the University of California, San Francisco Benioff Children's Hospital [UCSF]) each contributed 20 EEG studies, for a total of 60 neonatal 3-h EEG records. Each center had previously collected these EEG recordings as part of single-center research studies between 2007 and 2011. At each center, all term neonates receiving therapeutic hypothermia for hypoxic-ischemic encephalopathy underwent continuous video-EEG recording for at least 72 h as part of clinical care. For this study, consecutively recorded EEG studies were selected from term neonates <24 h old who received therapeutic hypothermia according to published criteria.[10,11] The first 3 h of each EEG recording was used for the present study. Institutional review board or Committee on Child Health Research approval was obtained at each center. A waiver of informed consent was granted.

Each EEG was recorded according to the International 10–20 system modified for neonates using each institution's clinical EEG software. At a minimum, EEG recordings all included the following channels: Fp1-T3, T3-O1, Fp2-T4, T4-O2, Fp1-C3, C3-O1, Fp2-C4, C4-O2, T3-C3, C3-Cz, Cz-C4, and C4-T4, along with electrocardiography (ECG) for artifact detection. Annotations were removed, the EEG recordings were anonymized, and the tracings were reformatted for re-interpretation by the investigators using Persyst Insight software (Persyst, San Diego, CA, U.S.A.). During independent review of the study, EEG raters could adjust the montage, paper speed, and voltage settings in any way preferred.

The 60 EEG records were reviewed independently by three pediatric neurophysiologists (CW, JS, and TT) who are board certified in child neurology, fellowship-trained in clinical neurophysiology, and have at least 5 years of neonatal EEG clinical and research experience. The reviewers were blinded to the patients' clinical details.

EEG interpretation and data collection

Each EEG was interpreted according to the 2013 ACNS terminology and classification system for neonatal EEG studies[9] and results were recorded on a standardized form. Seizures were defined as abnormal, rhythmic, and evolving patterns lasting longer than 10 s. Raters recorded the total number of individual seizures in the recording. The number of seizures was categorized as follows: 0 seizures, 1–10 seizures, 11–20 seizures, and >20 seizures. EEG background was classified according to a number of standard features. Continuity was described as normal (interburst intervals ≤6 s in duration), excessively discontinuous (interburst intervals >6 s), or burst suppression (severely suppressed interburst intervals (<5 μV pp), with bursts showing no normal patterns with an overall unreactive tracing). Symmetry was described as present or absent, with absent symmetry defined as a >2:1 difference in voltages or a clear discrepancy in frequencies or graphoelements between hemispheres. Synchrony was described as present or absent, with normal synchrony defined as hemispheric bursts occurring within 1.5 s of each other. Voltage was defined as normal (25–50 μV peak-to-peak amplitude while awake or in active sleep), borderline low (≥10 μV but <25 μV), or abnormally low (<10 μV). Variability was described as present (having spontaneous EEG changes in any domain or responses to internal stimuli), absent, or unable to assess. Sharp waves were assessed as present or absent, including presence of physiologic negative sharp waves, pathologic negative sharp waves, normal positive sharp waves, abnormal positive sharp waves, and brief rhythmic discharges (BRDs). Periodic and rhythmic activity were identified as periodic discharges, rhythmic delta activity, and nonseizure spike and slow wave activity. Finally, records were classified as either normal or abnormal, with abnormal defined as an EEG containing any seizures or any interictal abnormality.

EEG scoring data were collected and managed using REDCap secure electronic data capture tools hosted at Stanford University.[12]

Analysis

Interrater agreement was evaluated as both kappa statistic (κ) and simple percent agreement. Statistical analysis was performed using SAS Enterprise Guide 5.1 (Cary, NC, U.S.A.). Kappa values for categorical data were calculated using the MAGREE macro to apply the method described in Fleiss.[13,14] This is a generalization of the Cohen kappa statistic to the measurement of agreement among multiple raters, wherein each rater rates each item only once. For multiple categories, all nonagreements were weighted equally. The significance of kappa values was determined by corresponding p-values, with $p < 0.05$ considered significant. Kappa values of <0.2 were considered poor, 0.2–0.4 fair, 0.41–0.6 moderate, 0.61–0.8 good, and >0.8 very good.[15]

Results

Seizure identification

The three pediatric neurophysiologists independently reviewed the 60 EEG recordings for a total of 180 responses (Tables 1 and 2). The frequency of "no seizure" responses was 54% (97 of 180 responses). There was very good agreement in identifying seizures as present or absent: $\kappa = 0.93$, $p < 0.001$, and all three raters agreed in 95% of records (57 of 60 EEG studies).

Raters also quantified the number of seizures in each recording. Responses indicated a range of 0–28 individual seizures per EEG. Among records with any seizures, the median number of seizures identified was 8 (interquartile range [IQR] 3.5–13). Using seizure quantity categories, overall interrater agreement was good ($\kappa = 0.72$, $p < 0.001$), with all raters agreeing on seizure burden category in 80% of records. Raters had good agreement identifying records with fewer than 10 seizures ($\kappa = 0.72$, $p < 0.001$). However, there was only fair agreement in categorizing records as having 11–20 seizures ($\kappa = 0.35$,

Table 1. Summary of neonatal EEG features independently identified by three neurophysiologists

EEG feature	Frequency[a]
Seizures	
None	54% (97/180)
1–10	32% (57/180)
11–20	11% (19/180)
>20	4% (7/180)
Background continuity	
Normal	53% (85/159)
Excessively discontinuous	36% (58/159)
Burst suppression	10% (16/159)
Background symmetric	97% (163/168)
Background synchronous	96% (161/168)
Background voltage	
Normal	73% (123/168)
Borderline	15% (25/168)
Abnormal	11% (18/168)
Background variability present	60% (100/168)
Physiologic negative sharps present	60% (101/168)
Pathologic negative sharps	51% (85/168)
Positive sharps	
Normal	4% (7/168)
Pathologic	15% (25/168)
Absent	81% (136/168)
Brief rhythmic discharges present	19% (32/168)
Any abnormality	83% (149/180)

[a]Frequency totals are among all responses for interpretable records; varying denominators reflect number of records interpretable for each EEG feature. For seizure quantification and identification of any abnormality, there were three independent reviews of 60 EEG recordings for a total of 180 responses. For background continuity, seven EEG studies were considered by at least one reviewer to have background uninterpretable due to status epilepticus (n = 4) or electrocerebral inactivity (n = 3), for a total of 53 EEG recordings and 159 responses. For all other features, 56 records were interpretable (after removing 4 with status epilepticus), for a total of 168 responses.

Table 2. Summary of interrater agreement for specific EEG features

EEG feature	Perfect agreement, %[a]	κ	p–Value
Seizures (present/absent)	95	0.93	<0.001
Quantification of seizures (by categories of 0, 1–10, 11–20, and over 20)	80	0.72	<0.001
Background categorization (normal, excessively discontinuous, burst suppression, status epilepticus, or electrocerebral inactivity)	72	0.70	<0.001
Background continuity (normal, excessively discontinuous/burst suppression)	81	0.78	<0.001
Background symmetry (present/absent)	93	0.18	0.01
Background synchrony (present/absent)	89	0.11	0.09
Background voltage (normal, borderline, abnormal)	63	0.41	<0.001
Background variability (present/absent)	54	0.35	<0.001
Physiologic negative sharps (present/absent)	29	0.069	0.18
Pathologic negative sharps (present/absent)	30	0.17	0.02
Positive sharps (normal/pathologic/absent)	59	0.13	0.02
Brief rhythmic discharges (present/absent)	63	0.19	0.007
Any abnormality (present/absent)	78	0.49	<0.001

[a]Percent of records for which all three raters agreed.

$p < 0.001$), and moderate agreement in identifying EEG recordings with >20 seizures ($\kappa = 0.41$, $p < 0.001$).

Background features

The overall interrater agreement for background classification using a 5-category scale of normal, excessively discontinuous, burst suppression, status epilepticus, or electrocerebral inactivity was good ($\kappa = 0.70$, $p < 0.001$). There was agreement across all raters in 72% of records using this 5-category scale (43 of 60). After removing those records for which at least one rater indicated they were unable to assess the first hour due to either status epilepticus (four records) or due to electrocerebral inactivity (three records), 53 EEG studies were categorized as having normal continuity, excessive discontinuity, or burst suppression by each of the three raters, for a total of 159 responses. Interrater agreement remained good for degree of continuity ($\kappa = 0.78$, $p < 0.001$), with agreement across all raters in 81% of EEG recordings (43 of 53).

EEG background was further described by symmetry and synchrony. Of the 56 records interpretable for these features (after removing the four EEG studies that were uninterpretable due to status epilepticus), only four records were categorized by any rater as asymmetric, equivalent to 3% of total responses (5 of 168 responses). There was a high percentage agreement for this feature of the EEG background, but given the small number of records with the finding, the kappa remained low ($\kappa = 0.18$, $p = 0.01$). Similarly, of these 56 records, only 6 were categorized as asynchronous by at least one rater (4%, 7 of 168 responses), and in this case the kappa value did not reach statistical significance ($\kappa = 0.11$, $p = 0.09$).

EEG background amplitude for these 56 records was categorized by voltage as normal, borderline, or abnormal. There was moderate interrater agreement ($\kappa = 0.41$, $p < 0.001$) with total agreement across all raters in 63% of EEG studies (35 of 56).

EEG background was categorized as having variability present or absent. There was fair interrater agreement ($\kappa = 0.35$, $p < 0.001$).

Sharps and brief rhythmic discharges

Interrater agreement for presence of physiologic negative sharp waves was poor, but did not achieve statistical significance ($\kappa = 0.069$, $p = 0.18$). Interrater agreement for presence of pathologic negative sharp waves was also poor, and did reach statistical significance ($\kappa = 0.17$, $p = 0.02$).

Positive sharp waves could be identified as present and normal; present and pathologic; or absent. For this feature, there was again poor interrater agreement ($\kappa = 0.13$, $p = 0.02$).

Interrater agreement regarding the presence of brief rhythmic discharges was poor ($\kappa = 0.19$, $p = 0.007$).

Rhythmic activity

EEG recordings were categorized by presence or absence of periodic discharges, rhythmic delta activity, and spike and slow wave discharges. However, each of these findings was relatively rare. Periodic discharges were identified in 12% of responses, rhythmic delta activity in 11%, and spike and slow wave discharges in 1.7%. Thus there was not statistical power to calculate kappa values for these features ($p > 0.05$ in all cases).

Any abnormality

The data from the preceding features were compiled to categorize EEG studies as either normal or abnormal. If there was any abnormality in continuity, voltage, symmetry, or synchrony, or if pathologic sharp waves, brief rhythmic discharges, or seizures were identified, the EEG was classified as abnormal. Using this system, EEG studies were deemed normal in only 17% of responses. There was moderate interrater agreement ($\kappa = 0.49$, $p < 0.001$).

Discussion

In this interrater study, three expert pediatric neurophysiologists evaluated 60 EEG studies from neonates who were

treated with therapeutic hypothermia for neonatal encephalopathy. We identified varying degrees of agreement for specific EEG features as defined in the ACNS guideline on neonatal terminology and classification.

The best interrater agreement was in identifying the presence or absence of electrographic seizures. The ACNS terminology is specific in defining a seizure as an abnormal, evolving, rhythmic pattern lasting longer than 10 s in duration.[9] Although the requirement of 10 s duration is somewhat arbitrary, this historical convention facilitates identification of seizures as distinct from other, briefer changes in the EEG background. In this study, the neurophysiologists were able to apply that definition with near-perfect agreement, as reflected in the κ value of 0.93. This suggests that seizure identification using the ACNS guideline is reliable, even when performed independently by different neurophysiologists. This finding supports the recent work by Stevenson and colleagues, which suggested that interobserver agreement is good for identification of any neonatal seizures within an EEG.[8] This result is certainly encouraging for clinical practice, given the need to accurately and reliably determine whether a given patient has seizures in order to guide medical management. This also supports the methodology of research studies in which identification of electrographic neonatal seizures plays a role.

Quantification of seizures by quantity bins of 10 also had good interrater agreement. We recognize this is a coarse measure of seizure quantity. There is much controversy regarding how to optimally define seizure burden—specifically, as to whether seizure counts are adequate, or whether other methods incorporating seizure durations and temporal extent are superior. Although this study found good agreement using arbitrarily defined and relatively broad seizure count categories, it is unknown whether these specific categories have clinical or physiologic relevance. Further work is needed to determine whether simple seizure counts are a clinically relevant and practical measure of seizure burden. Although more detailed methods of seizure quantification may be more clinically relevant and precise, they are also more labor intensive than the method employed here, and could be subject to lower agreement between observers. Additional work is needed to determine interrater agreement using alternate approaches to seizure quantification.

Our results also suggest that categorization of neonatal background continuity by the ACNS system is reproducible. The lower interrater agreement for other specific background features (symmetry, voltage, and variability) suggests that these may be more subjective aspects of visual analysis, or that their definitions might benefit from revision in future guidelines, in order to improve their reproducibility.

Previous research has proposed correlations between specific numbers of negative or positive sharp waves and various pathologies.[5,16–19] Herein we found unexpectedly low interrater agreement for these EEG features, even though they were categorized simply as present or absent, and not quantified. Pathologic sharp waves have been proposed as biomarkers of cerebral dysfunction. Our data suggest that additional work is needed to determine if a better definition might improve interrater agreement, or if other features might serve as more reliable and reproducible biomarkers.

Finally, application of a combination of the features to categorize EEG recordings as normal or abnormal resulted in only moderate interrater agreement. This suggests that alternate definitions or global categories might be more useful.

Taken as a whole, these findings have implications both for research and clinical practice. For researchers, our findings suggest caution must be used in selecting features for neonatal EEG analysis. Some features (background continuity, categorization, and seizures) had good interrater agreement and might serve as reliable features to include in future studies. However, given the very poor interrater agreement for features such as negative and positive sharps, it may be that these are not reliable markers to study, or that studies including these markers should include a demonstration of reproducibility of these findings.

Clinically, our results suggest a potential hierarchy of importance when applying EEG findings to guide clinical intervention and prognosis. Because there was good interrater agreement for identifying presence or absence of neonatal seizures, and in categorizing background continuity, these might be among more important EEG features to include in clinical decision making. Likewise, overall categorization of EEG background had good reliability, and thus is reasonable to include in clinical use. Coarse quantification of seizures had good reliability, suggesting that this might be a meaningful measure for clinical use, although other metrics of seizure burden warrant further study and comparison. In contrast, we found surprisingly low interrater agreement regarding presence or absence of sharp waves, for both negative and positive sharp waves. Thus, although in individual cases the presence of abnormal sharp waves may still be clinically relevant, caution should be used when placing great importance on the finding (or absence) of sharps. To be clear, given this study's limitations, we do not suggest that identification of sharp waves is always meaningless. Rather, we point out there may be variation between providers in whether and how sharps are identified on a given EEG. Awareness of this potential variation may be a consideration for the clinician applying EEG findings to clinical care for any particular patient.

A major strength of this study was the inclusion of a large number of conventional neonatal EEG recordings from a relatively homogeneous subject population. Including 60 prolonged recordings allowed sufficient statistical power for analysis of most EEG features. Similarly, the population included (term neonates undergoing therapeutic hypothermia for hypoxic ischemic encephalopathy) frequently

receives continuous EEG monitoring clinically and for research; these findings are relevant to a commonly studied population. Furthermore, the raters for this project were all experienced pediatric neurophysiologists with expertise in neonatal EEG; all three were among the authors of the ACNS guideline. Although the use of expert readers might lead to higher agreement and kappa values than with the less-experienced raters, our results showing low interrater reliability was not likely due to lack of experience or familiarity with the definitions applied here.

The limitations of this study are important to recognize and address through future research. First, this study examined interrater agreement across three specific raters, and may not be generalizable to the work of all other neurophysiologists. These results require replication by other groups. In addition, this study examined EEG studies from only the first 3 h of recording in term neonates receiving therapeutic hypothermia. Our findings may not be applicable to other patient populations, or in EEG recorded at later points during the clinical course. In particular, EEG features in healthy term and preterm neonates may be different, and interrater agreement of preterm EEG features warrants separate investigation. It may be that the duration of EEG used here affected findings, and that analysis of a full 24 h of recording would have led to different results. However, given a "routine" neonatal bedside EEG is 1 h in duration, the time samples used were relevant to clinical practice. Finally, this study utilized kappa statistics as a measure of interrater agreement. The kappa statistic is commonly used in studies of interrater agreement because, in contrast to reporting just percent agreement, the kappa statistic takes into account the likelihood of multiple raters agreeing by chance. Related to this, kappa statistics are heavily influenced by the prevalence of findings in the sample studied. That is, even if percent agreement is similar across multiple samples, varying incidence of specific findings between samples could create very different kappa values.[20] Even if there is a high percentage of agreement for a finding that is uncommon in the sample, the kappa may remain low. Although this may seem counterintuitive, it reflects the way in which the analysis accounts for the likelihood of agreement by chance alone. In our study, this feature of the kappa statistic may have resulted in a low kappa value for relatively rare findings despite high percentage of agreement, such as for normal positive sharps or asynchrony (Tables 1 and 2). At the same time, the p-values in Table 2 indicate that for almost all features, the kappa result we found was unlikely due to chance regardless of how common the feature was in the sample. Only background synchrony and physiologic negative sharps had a kappa with $p > 0.05$, suggesting that a larger sample size might be helpful to examine these features in particular.

Our results support the use of the ACNS guideline on neonatal EEG terminology and classification as a standard for identifying seizures and describing neonatal EEG background classifications. The results support current clinical practice: there is reliability in neurophysiologist interpretation of EEG regarding presence or absence of neonatal seizures and degree of continuity—two clinically important prognostic features of the EEG. For the researcher, these two elements may be most valid to include in EEG analysis in future studies. At the same time, a number of other EEG features had only poor or fair agreement. Presence or absence of various types of sharp waves had particularly low interrater agreement. This suggests that visual analysis is insufficient for identification of sharp waves for research purposes. Finally, each of the findings in this project should be considered by the ACNS in future guideline revisions.

Acknowledgments

Stanford's REDCap database was made available through the Stanford Center for Clinical Informatics grant support (Stanford CTSA award number UL1 RR025744 from NIH/NCRR). The authors thank Dr. Catherine Clark for her assistance with data management for this study.

Disclosure

None of the authors has any conflict of interest to disclose. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

References

1. Clancy RR, Legido A, Lewis D. Occult neonatal seizures. *Epilepsia* 1988;29:256–261.
2. Murray DM, Boylan GB, Ali I, et al. Defining the gap between electrographic seizure burden, clinical expression and staff recognition of neonatal seizures. *Arch Dis Child Fetal Neonatal Ed* 2008;93:F187–F191.
3. Malone A, Ryan CA, Fitzgerald A, et al. Interobserver agreement in neonatal seizure identification. *Epilepsia* 2009;50:2097–2101.
4. 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;28:611–617.
5. Novotny EJ Jr, Tharp BR, Coen RW, et al. Positive rolandic sharp waves in the EEG of the premature infant. *Neurology* 1987;37:1481–1486.
6. Holmes GL, Lombroso CT. Prognostic value of background patterns in the neonatal EEG. *J Clin Neurophysiol* 1993;10:323–352.
7. Almubarak S, Wong PKH. Long-term clinical outcome of neonatal EEG findings. *J Clin Neurophysiol* 2011;28:185–189.
8. Stevenson NJ, Clancy RR, Vanhatalo S, et al. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Ann Clin Transl Neurol* 2015;2:1002–1011.
9. Tsuchida TN, Wusthoff CJ, Shellhaas RA, et al. American clinical neurophysiology society standardized EEG terminology and categorization for the description of continuous EEG monitoring in neonates: report of the American Clinical Neurophysiology Society critical care monitoring committee. *J Clin Neurophysiol* 2013;30:161–173.
10. Shankaran S, Laptook AR, Ehrenkranz RA, et al. Whole-body hypothermia for neonates with hypoxic-ischemic encephalopathy. *N Engl J Med* 2005;353:1574–1584.
11. Bonifacio SL, Glass HC, Vanderpluym J, et al. Perinatal events and early magnetic resonance imaging in therapeutic hypothermia. *J Pediatr* 2011;158:360–365.
12. Harris PA, Taylor R, Thielke R, et al. Research electronic data capture (REDCap) – a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–381.

13. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. *Psychol Bull* 1979;86:974–977.
14. Fleiss JL. *Statistical methods for rates and proportions*. 3rd Ed. New York, NY: John Wiley & Sons, Inc.; 2003.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159–174.
16. Biagioni E, Boldrini A, Bottone U, et al. Prognostic value of abnormal EEG transients in preterm and full-term neonates. *Electroencephalogr Clin Neurophysiol* 1996;99:1–9.
17. Rowe JC, Holmes GL, Hafford J, et al. Prognostic value of the electroencephalogram in term and preterm infants following neonatal seizures. *Electroencephalogr Clin Neurophysiol* 1985;60:183–196.
18. Scher MS, Bova JM, Dokianakis SG, et al. Positive temporal sharp waves on EEG recordings of healthy neonates: a benign pattern of dysmaturity in pre-term infants at post-conceptional term ages. *Electroencephalogr Clin Neurophysiol* 1994;90:173–178.
19. Clancy RR, Tharp BR. Positive rolandic sharp waves in the electroencephalograms of premature neonates with intraventricular hemorrhage. *Electroencephalogr Clin Neurophysiol* 1984;57:395–404.
20. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–549.