Published in final edited form as:

J Clin Neurophysiol. 2019 January ; 36(1): 1–8. doi:10.1097/WNP.0000000000000534.

Interrater and Intrarater Agreement in Neonatal Electroencephalogram Background Scoring

Shavonne L. Massey, MD MSCE^{1,2,3}, Haochang Shou, PhD⁴, Robert Clancy, MD^{1,2,3}, Marissa DiGiovine, MD^{1,2,3}, Mark P. Fitzgerald, MD PhD^{1,2,3}, France W. Fung, MD^{1,2,3}, John Farrar, MD PhD^{2,4,5}, and Nicholas S. Abend, MD MSCE^{#1,2,3,4}

¹Department of Pediatrics (Neurology), Children's Hospital of Philadelphia, Philadelphia, PA, USA

²Department of Neurology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

³Department of Pediatrics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

⁴Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

⁵Department of Anesthesia and Critical Care, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

These authors contributed equally to this work.

Introduction:

Many neonates undergo EEG monitoring to identify and manage acute symptomatic seizures. Information about brain function contained in the EEG background data may also help predict neurobehavioral outcomes. For EEG background features to be useful as prognostic indicators, the interpretation of these features must be standardized across electroencephalographers. We aimed to determine the interrater and intrarater agreement among electroencephalographers interpreting neonatal EEG background patterns.

Methods:

Five neonatal electroencephalographers reviewed 5–7.5 minute epochs of EEG from full term neonates who underwent continuous conventional EEG monitoring. The EEG assessment tool used to classify background patterns was based on the American Clinical Neurophysiology Society's guideline for neonatal EEG terminology. Interrater and intrarater agreement were measured using Kappa coefficients.

Corresponding Author: Shavonne L. Massey, MD, Colket Translational Research Building Room 10011, 3501 Civic Center Boulevard, Philadelphia, PA 19104, (p) 215-590-1719, (f) 215-590-1771, masseysl@email.chop.edu.

Conflicts of Interest and Source of Funding:

No declared conflicts of interest.

Results:

Interrater agreement was consistently highest for voltage (binary: substantial, kappa=0.783; categorical: moderate, kappa=0.562), seizure presence (fair-substantial; kappa=0.375, 0.697), continuity (moderate; kappa=0.481), burst voltage (moderate; kappa=0.574), suppressed background presence (moderate-substantial; kappa=0.493, 0.643), delta activity presence (fair-moderate; kappa=0.369, 0.432), theta activity presence (fair-moderate; kappa=0.347, 0.600), graphoelements presence (fair; kappa=0.381), and overall impression (binary: moderate, kappa=0.495; categorical: fair-moderate, kappa=0.347, 0.465).

Agreement was poor or inconsistent for all other patterns. Intrarater agreement was variable, with highest average agreement for voltage (binary: substantial, kappa=0.75; categorical: substantial, kappa=0.714) and highest consistent agreement for continuity (moderate-substantial; kappa=0.43–0.67) and overall impression (moderate-substantial; kappa=0.42–0.68).

Conclusions:

This study demonstrates substantial variability in neonatal EEG background interpretation across electroencephalographers, indicating a need for educational and technological strategies aimed at improving performance.

Introduction

An increasing number of neonates undergo continuous electroencephalographic (EEG) monitoring^{1–3} to identify electrographic seizures and assess background activity, as recommended by recent neonatal EEG monitoring guidelines.⁴ The EEG background data contain a vast amount of information about brain function that may help predict neurobehavioral outcomes. Further, since EEG is a direct measure of cerebral function, EEG may be a superior predictor of outcomes compared to clinical measures, such as examination signs or laboratory values. EEG background patterns have been used to predict and/or have been correlated with multiple outcomes including mortality, neurodevelopment, and epilepsy.^{5–22} Studies have established that normal and severely abnormal EEGs are often associated with favorable and unfavorable outcomes, respectively.^{23,24} Further, the persistence of abnormal EEG features increases the predictive value for unfavorable outcomes.^{19,25} However, there are several problems related to the use of EEG data for neuroprognostication. First, until recently, studies have used varying terminology and definitions.²⁴ In 2013, the American Clinical Neurophysiology Society (ACNS) published standardized neonatal EEG terminology, thereby providing standardized interpretation methods.²⁶ Second, although EEG is an objective measure of cortical and subcortical network activity, EEG interpretation is subjective. Numerous studies have shown that the interpretation of EEG features among electroencephalographers is not standard,^{27–40} but only one prior study evaluated neonatal EEG interpretation.³⁴ To optimize the use of neonatal EEG data as a predictive tool, EEG interpretation must be reliable across electroencephalographers. We aimed to measure the intrarater and interrater agreement of neonatal EEG background patterns by pediatric electroencephalographers using standardized neonatal EEG terminology.

Methods

Study Participants

The study was performed at the Children's Hospital of Philadelphia (CHOP). All electroencephalographers who participated in this study were: (1) board certified in pediatric neurology, (2) formally trained in pediatric electroencephalography, (3) spending at least 20% patient care time with EEG interpretation, and (4) primary English speakers.

EEG tracings were selected from 13 full-term neonates (most of whom were diagnosed with hypoxic ischemic encephalopathy [HIE]) who underwent clinically indicated continuous conventional EEG monitoring between January 1, 2010 and December 31, 2016. Electrodes were placed according to the International 10–20 System modified for neonates. Most neonates were monitored for up to 96 hours, and interpretable EEG samples were selected from time points throughout the recording period. Prior to scoring for this study, EEG tracings were de-identified, and clinical notations were deleted. Participants were only aware that tracings were from full-term neonates.

Study Protocol

There were two EEG scoring sessions for each participant. In the first session, participants scored 25 EEG tracings lasting 5–7.5 minutes in random order. Interpretation was performed using a standardized data collection tool based on the ACNS terminology for neonatal EEG²⁶ that assessed behavioral state, continuity, voltage, frequency, symmetry, synchrony, dysmaturity, graphoelements, EEG transients, artifact, seizures, and overall impression. While there are published scales for overall impression, participants provided a gestalt impression since there is no formal ACNS definition.^{41,42} Given the brief duration of the EEG tracings, variability and reactivity were not adequately assessed. Additionally, participants used a visual analog scale (range 0–100) to report their confidence in scoring of each feature. The second scoring session occurred two months later. Participants received the same 25 EEG tracings plus an additional 5 naive tracings, presented in random order and scored using the same data collection tool. One new field, 'initial impression', was presented to reviewers first to capture their overall impression of the EEG background prior to completing the detailed analysis of individual features. Although study participants were aware of the ACNS terminology, they were not provided with any criteria or definitions embedded in the scoring tool.

The data collection responses were entered directly by participants into REDCap (Research Electronic Data Capture), an electronic data capture tool that provides a secure, web-based system for data capture.⁴³

Several supplemental variables not included in the ACNS terminology but felt to provide alternative ways of presenting the data were generated post-hoc. 'Normal continuity' separated normal continuity (continuous and normal discontinuous) and abnormal discontinuity. 'Best frequency' identified the fastest reported background activity during the EEG epoch, except for beta activity, as this may represent exogenous influences such as medication effect rather than true cerebral function.

Statistical Analysis

Statistical analysis was performed with Stata (version IC, 14). Frequency counts for the presence of each response for the EEG features were recorded. Categorical features were also recalibrated to binary scales when possible and assessed in both manners. Features were assessed using percent agreement for both intrarater and interrater agreement and non-weighted Kappa coefficient. Given the exploratory nature of the study, seeking to identify variables for consideration in future analysis, a conservative p-value of <0.10 was considered statistically significant, and Fleiss kappa values were recorded.⁴⁴ Interrater agreement values were considered reliable if the Fleiss kappa p-value was significant ($p<0.10$) in both assessment periods (see Supplement Table A for EEG feature categories). Interrater agreement values were considered to have probable reliability if the Fleiss kappa p-value was significant in only one of the assessment periods. Interrater agreement values were considered unreliable if the Fleiss kappa p-value was not significant in either assessment period. Intrarater agreement was assessed individually and as an average across raters. Individual agreement values were considered reliable if the Fleiss kappa p-value was significant ($p<0.10$) across all (5) or most (4) reviewers. Agreement was reassessed with statistical significance cutoff values of 0.05 and 0.20 to evaluate the effect of increasing and decreasing the p-value on interrater kappa. The kappa agreement scale was interpreted as poor (<0.1), slight (0.1–0.2), fair (0.21–0.4), moderate (0.41–0.6), substantial (0.61–0.8), almost perfect (0.81–1), or perfect (1).

Results

Study Participants and Neonatal EEG Tracings

Five pediatric electroencephalographers participated in the study. Table 1 provides participant demographics. Table 2 provides demographics for the 13 full-term neonates used to derive the EEG tracings. The primary diagnostic etiologies included HIE for 12 (92%) and congenital cardiac disease status post-surgical repair for 1 (8%) neonate.

Interrater Agreement

Table 3 provides the interrater and intrarater kappa agreement values for all variables, grouped by their reliability across both assessment periods. Supplement Table A provides all statistical results, including the interrater kappa and percent agreement values of all measured variables.

Variables with Stable Reliability:

Some variables had significant p-values for kappa measurements of interrater agreement in both assessments, and thus those variables were considered reliably assessed with the EEG tracings used in the study. Interrater agreement was substantial for binary measurement of voltage (assessment 1, 2), suppression presence (assessment 1), and seizure presence (assessment 2). Interrater agreement was moderate for continuity (assessment 1, 2), normal continuity (assessment 1, 2), voltage (assessment 1, 2), burst voltage (assessment 1, 2), binary overall impression (assessment 1, 2), awake state (assessment 1), delta activity presence (assessment 2), theta activity presence (assessment 2), suppression presence

(assessment 2), and overall impression (assessment 1). Interrater agreement was fair for initial impression (assessment 2), binary initial impression (assessment 2), indeterminate state (assessment 1, 2), graphoelements presence (assessment 1, 2), delta activity presence (assessment 1), theta activity presence (assessment 1), seizures presence (assessment 1), and overall impression (assessment 2). Interrater agreement was slight for awake state assessment 2).

Variables with Probable Reliability:

Interrater agreement was moderate for transient type (assessment 1) and binary transient type (assessment 1). Interrater agreement was fair for sleeping state (assessment 2), active sleep (assessment 2), and EEG transients presence (assessment 2). Interrater agreement was slight for discontinuity type (assessment 2) and binary discontinuity type (assessment 2).

Variables without Reliability:

Some variables did not have significant p-values for kappa measurements of interrater agreement in either assessment, and thus they were not reliably assessed with the EEG tracings used in the study. These included burst abnormality type, interburst voltage, interburst abnormality type, quiet sleep, transitional sleep, indeterminate sleep, alpha activity presence, beta activity presence, best frequency achieved, symmetry, synchrony, dysmaturity presence, and artifact presence.

Statistical Significance Cut-offs for Interrater Agreement:

The kappa values were also assessed using p-values of 0.05 and 0.20 for significance to assess for any major effect on overall outcome and conclusions. If the p-value decreased to 0.05, the interrater agreement of transient type was no longer significant in assessment 1. If the p-value increased to 0.20, additional variables were included in the significant range. Multiple impressions (slight) and alpha activity presence (poor) become significant in assessment 1. Interburst voltage (slight), interburst abnormality type (slight), and synchrony (poor) become significant in assessment 2.

Intrarater Agreement

The reliability of intrarater agreement assessment was variable with most variables achieving significance for some but not all reviewers. Additionally, within each variable, the range of intrarater agreement was highly variable. Variables with a significant p-value for all reviewers' intrarater agreement assessment included awake state (poor, moderate, and near perfect), continuity (moderate and substantial), and overall impression (moderate and substantial). Variables with a significant p-value for most reviewers' intrarater agreement assessment included quiet sleep (poor, moderate, and substantial) and normal continuity (slight and substantial). Statistical significance of intrarater agreement for individual raters ranged from 34–64%. Table 3 provides average and range of intrarater agreement for all variables and Supplement Table A provides individual intrarater agreement data for all variables.

Confidence in Scoring

See Supplement Table A. Assessment confidence was highest (median score >80) for the presence of seizures and type of EEG transients. Assessment confidence was high (median score 71–80) for continuity, discontinuity type, voltage, burst voltage, interburst voltage, interburst abnormality type, symmetry, synchrony, and presence of artifact. Assessment confidence was intermediate (median score <60–70) for assessment of behavioral state, burst abnormality type, EEG transients presence, dysmaturity presence, and overall impression. Assessment confidence was low (median score <60) for sleep type differentiation. There was a wide spread of confidence scores for each measured variable with interquartile ranges of 18–41 points. There was no correlation between confidence and interrater agreement for EEG variable scoring in assessment 1 (spearman's correlation coefficient $\rho = 0.101$; p -value = 0.7) nor between confidence and degree of intrarater agreement (spearman's correlation coefficient $\rho = -0.22$; p -value = 0.41).

Discussion

This study demonstrates variable agreement in the interpretation of neonatal EEG background features within and between pediatric electroencephalographers. None of the assessed features had perfect or near-perfect interrater agreement. Interrater agreement was highest for voltage (moderate), seizure presence (fair-substantial), continuity (moderate), burst voltage (moderate), suppressed background (moderate-substantial), delta activity (fair-moderate), theta activity (fair-moderate), and overall impression (moderate). Assessment of voltage as a binary variable increased interrater agreement by one agreement category (moderate to substantial). All other conversions of categorical to binary variables did not increase interrater agreement category.

Our data are generally consistent with the only other study to assess interrater agreement of neonatal EEG background features in term neonates. Wusthoff et al. assessed the interrater agreement among three pediatric electroencephalographers who reviewed EEG tracings from 60 term neonates with HIE using a standardized scoring system based on the ACNS terminology.³⁴ Agreement was very good for identification of seizures ($\kappa=0.93$); good for continuity ($\kappa=0.78$) and overall background ($\kappa=0.70$); moderate for classifying the record as normal versus abnormal ($\kappa=0.49$) and background voltage ($\kappa=0.41$); fair for the presence of background variability ($\kappa=0.35$); and poor for symmetry ($\kappa=0.18$), synchrony ($\kappa=0.11$), physiologic negative sharps ($\kappa=0.07$), pathologic negative sharps ($\kappa=0.17$), and positive sharps ($\kappa=0.13$).

While interrater agreement in seizure identification was not a primary goal of the current study, our data are somewhat consistent with previous work. In the current study, interrater agreement for the presence of seizures varied from fair ($\kappa=0.375$) to substantial ($\kappa=0.697$) between the two scoring sessions. Wusthoff et al reported very good ($\kappa=0.93$) agreement for the identification of the presence of seizures and good ($\kappa=0.72$) agreement for the quantification of seizure burden among three pediatric electroencephalographers at separate institutions.³⁴ Stevenson et al reported the interrater agreement of neonatal seizure detection by three neonatal EEG experts who reviewed EEG from 70 full-term neonates at risk for encephalopathy.²⁸ Agreement for temporal assessment

of seizures was high ($\kappa=0.83$), but agreement decreased when a neonate's seizures were brief in duration (<30 seconds) and rarer during the recording. Notably, in our study all seizures were brief duration and no record contained more than one seizure, potentially reducing our agreement for seizure detection.

Intrarater agreement has not been assessed previously for neonatal EEG interpretation. This study demonstrates that intrarater agreement is extremely variable in neonatal EEG background assessment. Intrarater agreement was consistently highest for continuity (moderate and substantial agreement), normal continuity (slight and substantial agreement), overall impression (moderate and substantial agreement), awake state (poor, moderate, and near perfect agreement), and quiet sleep (poor, moderate, and substantial agreement). Intrarater agreement varied (spanned four or more agreement categories) across reviewers for most variables, typically ranging from poor to perfect or near-perfect. Notably, there was a correlation in which variables with high interrater reliability also have high intrarater agreement, indicating that these variables have the most consistent interpretation.

Confidence assessments demonstrated a wide range of average values across electroencephalographers. Overall, the confidence scores suggest that electroencephalographers are ambivalent about analysis of specific features, especially when asked to recall and utilize specific criterion defined by ACNS. Further, there was no correlation between confidence scores and the degree of interrater or intrarater agreement. This finding suggests that electroencephalographers' perceived strengths and weaknesses in neonatal EEG interpretation did not impact the degree to which they consistently assessed EEG features with themselves or with their peers.

Though the overall impression of the EEG background was reliably assessed in this study, the interpretation of individual EEG variables on which the overall impression was based demonstrated significant interrater and intrarater variability. If EEG data are used at the bedside to direct care and prognostication, and as a potential biomarker for neonatal brain injury in research, then it is imperative to ensure that neonatal EEG interpretation is standardized. There are potential strategies to improve agreement which warrant future study. Educational interventions and tools which simplify EEG scoring may improve scoring reliability. Hussain et al evaluated interrater agreement for scoring of hypsarrhythmia patterns and showed that interrater reliability was high for deciphering between a normal and abnormal EEG ($\kappa=0.89$) but variable when evaluating specific hypsarrhythmia features (κ range 0.08–0.68).³² The authors then introduced a simplified hypsarrhythmia grading system, and after a brief tutorial on use of this grading system, interrater reliability for interpreting hypsarrhythmia increased to κ of 0.87 (compared with κ of 0.09).³¹ It is feasible that similar tools which incorporate the specific definitions for neonatal EEG background terminology could improve interrater agreement, and groups have begun work on such strategies.^{45,46} Once neonatal EEG assessments are reliable, we can test the true predictive value of individual neonatal EEG for both acute and chronic neonatal outcomes. Knowing the most informative features and patterns can allow for redirection of educational efforts for neonatal EEG readers to focus on correctly and reliably scoring those EEG features that offer the greatest impact to the neonate's care.

There are several limitations to this study. First, the study was performed at a single institution with a homogenous pool of electroencephalographers in terms of training, potentially resulting in exaggerated levels of interrater agreement. Second, the limited number and duration of EEG clips provided for assessment during each session limited the number of examples provided for each background feature. A large discrepancy in the presence of a feature resulted in lower kappa values than might have been demonstrated if there had been a more balanced presentation of features. This is evidenced by the wide discrepancy between kappa and percent agreement and has been previously reported⁴⁷. Having more representative samples for EEG features with highly skewed presence (multiple impressions, sleep type, interburst voltage, interburst abnormality type, symmetry, synchrony, dysmaturity, and seizures) might have increased the kappa values for those features. Tracing length also directly affected the assessments since we could not assess variability and reactivity. Third, assessment of a tracing from a neonate's EEG without access to the entire record is not equivalent with EEG interpretation in real world situations. With access to clinical information and the whole EEG record, an electroencephalographer can review and re-review unclear features and patterns in the context of an entire EEG record to form their overall assessments. Fourth, the Hawthorne Effect may have also slightly influenced the study since participants knew that they were in a study and that their responses would be highly scrutinized, perhaps altering their typical EEG interpretation methodology. Their EEG assessments in the study, therefore, may not be entirely reflective of their neonatal EEG assessments in daily clinical care. Fifth, HIE was an over-represented diagnosis in this study. However, HIE is commonly encountered in neonatal neurocritical care and many NICUs primarily use EEG to monitor these neonates.^{48,49} Lastly, while conceivable confounders were measured and presented in Table 1, there is always the potential for unmeasured confounders to affect agreement.

Given that the current gold standard of EEG interpretation is visual analysis by an electroencephalographer, it is disconcerting that there is not high reliability in interpretation for all variables. However, the major neonatal EEG background categories (continuity, voltage, overall impression) were reliably measured in this study, all with moderate intrarater and interrater agreement. As the use of neonatal EEG increases, it is increasingly important to develop strategies to optimize interpretation for the remaining variables. Enhanced educational courses and increased exposure to neonatal EEGs will likely improve familiarity with neonatal EEG terminology. Use of standardized EEG interpretation templates that incorporate the ACNS terminology may increase interpretation agreement.⁴⁶ Yet, even with optimization through standardization and educational strategies, EEG interpretation through the eyes of a human reader will always incorporate some degree of subjectivity. Thus, this study highlights the potential role for accurate automation of neonatal EEG interpretation. With increased utilization of EEG due to heightened recognition of the significant sequelae of neonatal brain injury, there is an increased demand on medical centers to provide neonatal EEG interpretation, a skill requiring advanced training. Automated interpretation technology could assuage this increased need. Promising automated programs and quantitative EEG software remove subjectivity, but these approaches require more investigation.^{27,50,51}

Interrater and intrarater variability exists in the interpretation of visually-assessed neonatal EEG features, demonstrating an increased need for improved educational and automation strategies addressing neonatal EEG interpretation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

None

This work was funded by NIH T32-NS061779 and K02-NS096058.

References

1. Glass HC, Kan J, Bonifacio SL, Ferriero DM. Neonatal seizures: treatment practices among term and preterm infants. *Pediatr Neurol* 2012;46:111–5. [PubMed: 22264706]
2. Boylan G, Burgoyne L, Moore C, O'Flaherty B, Rennie J. An international survey of EEG use in the neonatal intensive care unit. *Acta Paediatr* 2010;99:1150–5. [PubMed: 20353503]
3. Filan PM, Inder TE, Anderson PJ, Doyle LW, Hunt RW. Monitoring the neonatal brain: a survey of current practice among Australian and New Zealand neonatologists. *J Paediatr Child Health* 2007;43:557–9. [PubMed: 17635686]
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;28:611–7. [PubMed: 22146359]
5. Holmes G, Rowe J, Hafford J, Schmidt R, Testa M, Zimmerman A. Prognostic value of the electroencephalogram in neonatal asphyxia. *Electroencephalogr Clin Neurophysiol* 1982;53:60–72. [PubMed: 6173201]
6. Khan RL, Nunes ML, Garcias da Silva LF, da Costa JC. Predictive value of sequential electroencephalogram (EEG) in neonates with seizures and its relation to neurological outcome. *J Child Neurol* 2008;23:144–50. [PubMed: 18160554]
7. Lamblin MD, Walls Esquivel E, Andre M. The electroencephalogram of the full-term newborn: review of normal features and hypoxic-ischemic encephalopathy patterns. *Neurophysiol Clin* 2013;43:267–87. [PubMed: 24314754]
8. Laroia N, Guillet R, Burchfiel J, McBride MC. EEG background as predictor of electrographic seizures in high-risk neonates. *Epilepsia* 1998;39:545–51. [PubMed: 9596208]
9. Mariani E, Scelsa B, Pogliani L, Introvini P, Lista G. Prognostic value of electroencephalograms in asphyxiated newborns treated with hypothermia. *Pediatr Neurol* 2008;39:317–24. [PubMed: 18940554]
10. Monod N, Pajot N, Guidasci S. The neonatal EEG: statistical studies and prognostic value in full-term and pre-term babies. *Electroencephalogr Clin Neurophysiol* 1972;32:529–44. [PubMed: 4112304]
11. Murray DM, Boylan GB, Ryan CA, Connolly S. Early EEG findings in hypoxic-ischemic encephalopathy predict outcomes at 2 years. *Pediatrics* 2009;124:e459–67. [PubMed: 19706569]
12. Murray DM, O'Connor CM, Ryan CA, Korotchikova I, Boylan GB. Early EEG Grade and Outcome at 5 Years After Mild Neonatal Hypoxic Ischemic Encephalopathy. *Pediatrics* 2016.
13. Murray DM, Ryan CA, Boylan GB, Fitzgerald AP, Connolly S. Prediction of seizures in asphyxiated neonates: correlation with continuous video-electroencephalographic monitoring. *Pediatrics* 2006;118:41–6. [PubMed: 16818547]
14. Nagarajan L, Palumbo L, Ghosh S. Neurodevelopmental outcomes in neonates with seizures: a numerical score of background encephalography to help prognosticate. *J Child Neurol* 2010;25:961–8. [PubMed: 20223749]

15. Pressler RM, Boylan GB, Morton M, Binnie CD, Rennie JM. Early serial EEG in hypoxic ischaemic encephalopathy. *Clin Neurophysiol* 2001;112:31–7. [PubMed: 11137658]
16. 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;60:183–96. [PubMed: 2578926]
17. Selton D, Andre M. Prognosis of hypoxic-ischaemic encephalopathy in full-term newborns--value of neonatal electroencephalography. *Neuropediatrics* 1997;28:276–80. [PubMed: 9413008]
18. Sinclair DB, Campbell M, Byrne P, Prasertsom W, Robertson CM. EEG and long-term outcome of term infants with neonatal hypoxic-ischemic encephalopathy. *Clin Neurophysiol* 1999;110:655–9. [PubMed: 10378734]
19. Takeuchi T, Watanabe K. The EEG evolution and neurological prognosis of neonates with perinatal hypoxia [corrected]. *Brain Dev* 1989;11:115–20. [PubMed: 2712234]
20. Takenouchi T, Rubens EO, Yap VL, Ross G, Engel M, Perlman JM. Delayed onset of sleep-wake cycling with favorable outcome in hypothermic-treated neonates with encephalopathy. *J Pediatr* 2011;159:232–7. [PubMed: 21353680]
21. Watanabe K, Miyazaki S, Hara K, Hakamada S. Behavioral state cycles, background EEGs and prognosis of newborns with perinatal hypoxia. *Electroencephalogr Clin Neurophysiol* 1980;49:618–25. [PubMed: 6158441]
22. Nash KB, Bonifacio SL, Glass HC, et al. Video-EEG monitoring in newborns with hypoxic-ischemic encephalopathy treated with hypothermia. *Neurology* 2011;76:556–62. [PubMed: 21300971]
23. Tsuchida TN. EEG background patterns and prognostication of neonatal encephalopathy in the era of hypothermia. *J Clin Neurophysiol* 2013;30:122–5. [PubMed: 23545762]
24. Awal MA, Lai MM, Azemi G, Boashash B, Colditz PB. EEG background features that predict outcome in term neonates with hypoxic ischaemic encephalopathy: A structured review. *Clin Neurophysiol* 2016;127:285–96. [PubMed: 26105684]
25. Holmes GL, Lombroso CT. Prognostic value of background patterns in the neonatal EEG. *J Clin Neurophysiol* 1993;10:323–52. [PubMed: 8408599]
26. 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;30:161–73. [PubMed: 23545767]
27. Murphy K, Stevenson NJ, Goulding RM, et al. Automated analysis of multi-channel EEG in preterm infants. *Clin Neurophysiol* 2015;126:1692–702. [PubMed: 25538005]
28. Stevenson NJ, Clancy RR, Vanhatalo S, Rosen I, Rennie JM, Boylan GB. Interobserver agreement for neonatal seizure detection using multichannel EEG. *Ann Clin Transl Neurol* 2015;2:1002–11. [PubMed: 26734654]
29. Abend NS, Gutierrez-Colina A, Zhao H, et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. *J Clin Neurophysiol* 2011;28:15–9. [PubMed: 21221016]
30. Piccinelli P, Viri M, Zucca C, et al. Inter-rater reliability of the EEG reading in patients with childhood idiopathic epilepsy. *Epilepsy Res* 2005;66:195–8. [PubMed: 16118044]
31. Mytinger JR, Hussain SA, Islam MP, et al. Improving the inter-rater agreement of hypsarrhythmia using a simplified EEG grading scale for children with infantile spasms. *Epilepsy Res* 2015;116:93–8. [PubMed: 26280806]
32. Hussain SA, Kwong G, Millichap JJ, et al. Hypsarrhythmia assessment exhibits poor interrater reliability: a threat to clinical trial validity. *Epilepsia* 2015;56:77–81. [PubMed: 25385396]
33. Stroink H, Schimsheimer RJ, de Weerd AW, et al. Interobserver reliability of visual interpretation of electroencephalograms in children with newly diagnosed seizures. *Dev Med Child Neurol* 2006;48:374–7. [PubMed: 16608546]
34. 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. [PubMed: 28166364]

35. Mani R, Arif H, Hirsch LJ, Gerard EE, LaRoche SM. Interrater reliability of ICU EEG research terminology. *J Clin Neurophysiol* 2012;29:203–12. [PubMed: 22659712]
36. Halford JJ, Shiau D, Desrochers JA, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clin Neurophysiol* 2015;126:1661–9. [PubMed: 25481336]
37. Foreman B, Mahulikar A, Tadi P, et al. Generalized periodic discharges and ‘triphasic waves’: A blinded evaluation of inter-rater agreement and clinical significance. *Clin Neurophysiol* 2016;127:1073–80. [PubMed: 26294138]
38. Gaspard N, Hirsch LJ, LaRoche SM, Hahn CD, Westover MB, Critical Care EEGMRC. Interrater agreement for Critical Care EEG Terminology. *Epilepsia* 2014;55:1366–73. [PubMed: 24888711]
39. Gerber PA, Chapman KE, Chung SS, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. *J Clin Neurophysiol* 2008;25:241–9. [PubMed: 18791475]
40. Walczak TS, Radtke RA, Lewis DV. Accuracy and interobserver reliability of scalp ictal EEG. *Neurology* 1992;42:2279–85. [PubMed: 1461379]
41. Shellhaas RA, Gallagher PR, Clancy RR. Assessment of neonatal electroencephalography (EEG) background by conventional and two amplitude-integrated EEG classification systems. *J Pediatr* 2008;153:369–74. [PubMed: 18534239]
42. Clancy RR, Dlugos DJ, Nordli DR, Jr.. Normal Pediatric EEG: Neonates and Children In: Ebersole JS, ed. *Current Practice of Clinical Electroencephalography*. 4th ed. Philadelphia, PA: Wolters Kluwer Health; 2014:125–212.
43. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG Research electronic data capture (REDCap) - A metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–81. [PubMed: 18929686]
44. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;33:363–74. [PubMed: 884196]
45. Beniczky S, Aurlen H, Brogger JC, et al. Standardized computer-based organized reporting of EEG: SCORE. *Epilepsia* 2013;54:1112–24. [PubMed: 23506075]
46. Witzman SM MS, Kessler SK, Dlugos DJ, Abend NS Epic for Standardized Pediatric and Neonatal ICU Reporting. American Epilepsy Society Annual Meeting 2017 Washington, DC, USA: Children’s Hospital of Philadelphia; 2017.
47. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. *J Clin Epidemiol* 1990;43:543–9. [PubMed: 2348207]
48. Glass HC, Bonifacio SL, Peloquin S, et al. Neurocritical care for neonates. *Neurocrit Care* 2010;12:421–9. [PubMed: 20066514]
49. Shah NA, Van Meurs KP, Davis AS. Amplitude-integrated electroencephalography: a survey of practices in the United States. *Am J Perinatol* 2015;32:755–60. [PubMed: 25519200]
50. Matic V, Cherian PJ, Koolen N, et al. Objective differentiation of neonatal EEG background grades using detrended fluctuation analysis. *Front Hum Neurosci* 2015;9:189. [PubMed: 25954174]
51. Lees D, Clancy RR Automated Assessment of Neurological Status in Neonates 2003.

Table 1:

Demographics of Study Participants (N=5 electroencephalographers)

Variable	Summary Statistic
Sex, Female	2 (40%)
Completion of Epilepsy/EEG Fellowship	4 (80%)
Epilepsy/EEG Fellowship at Children's Hospital of Philadelphia	2 (50%)
ACGME Approved Epilepsy/EEG Fellowship	2 (50%)
Pediatric Specific Epilepsy/EEG Fellowship	3 (75%)
Years Interpreting EEGs	1 (20%)
1 year	2 (40%)
2–5 years	2 (40%)
6–10 years	1 (20%)
>26 years	1 (20%)
Years Interpreting Neonatal EEGs	Median: 4 years (IQR 2, 10)

EEG, electroencephalogram; IQR, interquartile range.

Table 2:

Demographics of Neonates (N=13)

Variable	Summary Statistics N (%) or Median (IQR)
Gestational Age	39 weeks (37.7, 40)
Sex, Male	9 (69%)
Birth Weight	3.54 kg (2.84, 3.79)
Mode of Delivery	
Cesarean-section, urgent	7 (54%)
Vaginal, induced	3 (23%)
Vaginal, spontaneous	1 (7.7%)
Vaginal, assisted	1 (7.7%)
Cesarean-section, non-urgent	1 (7.7%)
Apgar, 1-minute	1 (0, 3)
Apgar, 5-minute	3 (0, 5)
Apgar, 10-minute	4 (2, 6)
Initial pH	
All	7.07 (6.8, 7.2)
Cord	6.88 (6.8, 6.9)
Arterial Blood Gas	7.07 (6.8, 7.2)
Venous Blood Gas	7.21 (7.2, 7.3)
Not specified	7.07 (6.5, 7.3)
Initial Base Deficit	
All	-12.2 (-20, -8.5)
Cord	-15.5 (-19.5, -10)
Arterial Blood Gas	-12.4 (-27, -8)
Venous Blood Gas	-8.4 (-9, -7.8)
Not specified	-12 (-31, -10)
Abnormal Initial Examination	12 (92.3%)
Intubated	12 (92.3%)
Primary Etiology	
Hypoxic Ischemic Encephalopathy	12 (92.3%)
Congenital Cardiac Disease s/p Repair	1 (7.7%)
Secondary Etiology	
Venous Thrombosis	1 (50%)
Congenital Cardiac Disease s/p Repair	1 (50%)

Variable	Summary Statistics N (%) or Median (IQR)	
Therapeutic Hypothermia	12 (100% of HIE)	
Location of EEG Monitoring		
Neonatal Intensive Care Unit	11 (85%)	
Cardiac Intensive Care Unit	2 (15%)	
EEG Indication		
Encephalopathy	9 (69%)	
Therapeutic Hypothermia	2 (15%)	
Evaluate Seizure Burden	1 (7.7%)	
Differential Diagnosis	1 (7.7%)	
Length of EEG Monitoring	5 days (4, 5)	
Seizures		
EEG or Clinical	9 (69%)	
Clinical-only	3 (33%)	
Prescribed Antiseizure Medications		
Phenobarbital	10 (77%)	
Lorazepam	1 (7.7%)	
Levetiracetam	1 (7.7%)	
Midazolam	1 (7.7%)	
>1 Antiseizure Medication	3 (23%)	
Prescribed Sedatives & Paralytics		
Morphine	9 (69%)	
Midazolam	8 (62%)	
Fentanyl	7 (54%)	
Vecuronium	4 (31%)	
Lorazepam	2 (15%)	
Ketamine	1 (7.7%)	
Head Ultrasound Completed	12 (92.3%)	1 Vasculopathy
Day of Life	1 (1, 1)	2 Echogenicity
Abnormal	7 (58%)	3 Cerebral Edema
		1 Slit-like ventricles
MRI Completed	12 (92.3%)	4 HIE
Day of Life	6 (5, 6.5)	5 Hemorrhage
Abnormal	11 (92%)	1 Periventricular leukomalacia
		1 T2 Signal Change
Death Prior to Hospital Discharge	2 (15%)	
Withdrawal of Technological Support	2 (100%)	

EEG, electroencephalogram; HIE, hypoxic-ischemic encephalopathy; IQR, interquartile range.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3:

Interrater and Intrarater Agreement for All Variables

EEG Feature		Interrater Agreement				Intrarater Agreement	
		Assessment 1		Assessment 2			
		Kappa Value	Kappa Agreement Category	Kappa Value	Kappa Agreement Category	Kappa Value Average	Average Kappa Agreement Category
		<i>p-value</i>		<i>p-value</i>		<i>Range</i>	
Stable Reliably Assessed Variables							
Initial Impression	Categorical	n/a	n/a	0.368 <i>0.000</i>	Fair	n/a	n/a
	Binary	n/a	n/a	0.370 <i>0.003</i>	Fair	n/a	n/a
Behavioral State	Awake	0.415 <i>0.001</i>	Moderate	0.139 <i>0.069</i>	Slight	0.447 <i>0.00–0.84</i>	Moderate
	Indeterminate	0.226 <i>0.025</i>	Fair	0.250 <i>0.018</i>	Fair	0.426 <i>0.07–0.90</i>	Moderate
Continuity	Continuity	0.481 <i>0.000</i>	Moderate	0.410 <i>0.002</i>	Moderate	0.558 <i>0.43–0.67</i>	Moderate
	Normal Continuity	0.456 <i>0.000</i>	Moderate	0.463 <i>0.000</i>	Moderate	0.575 <i>0.19–0.75</i>	Moderate
Voltage	Voltage, continuous portion	0.562 <i>0.002</i>	Moderate	0.427 <i>0.038</i>	Moderate	0.714 <i>0.00–1.00</i>	Substantial
	Voltage, continuous portion – Binary	0.783 <i>0.000</i>	Substantial	0.776 <i>0.000</i>	Substantial	0.750 <i>0.00–1.00</i>	Substantial
	Burst Voltage	0.459 <i>0.025</i>	Moderate	0.574 <i>0.010</i>	Moderate	0.430 <i>0.16–1.00</i>	Moderate
Frequency	Delta	0.369 <i>0.037</i>	Fair	0.432 <i>0.001</i>	Moderate	0.492 <i>–0.13–1.00</i>	Moderate
	Theta	0.347 <i>0.007</i>	Fair	0.600 <i>0.000</i>	Moderate	0.459 <i>0.39–0.63</i>	Moderate
	Suppressed	0.643 <i>0.000</i>	Substantial	0.493 <i>0.001</i>	Moderate	0.562 <i>0.19–0.87</i>	Moderate
Graphoelements		0.360 <i>0.001</i>	Fair	0.381 <i>0.001</i>	Fair	0.599 <i>0.07–1.00</i>	Moderate
Seizures		0.375 <i>0.000</i>	Fair	0.697 <i>0.014</i>	Substantial	0.459 <i>0.00–1.00</i>	Moderate
Overall Impression	Categorical	0.465 <i>0.000</i>	Moderate	0.347 <i>0.000</i>	Fair	0.481 <i>0.42–0.68</i>	Moderate
	Binary	0.495 <i>0.000</i>	Moderate	0.421 <i>0.001</i>	Moderate	0.514 <i>0.24–0.82</i>	Moderate
Probable Reliably Assessed Variables							
Behavioral State	Asleep	0.053 <i>0.382</i>	Poor	0.223 <i>0.018</i>	Fair	0.332 <i>0.00–0.60</i>	Fair
	Active Sleep	–0.016 <i>0.168</i>	Poor	0.282 <i>0.010</i>	Fair	–0.022 <i>–0.09–0.00</i>	Poor
Continuity	Discontinuity Type	0.006 <i>0.920</i>	Poor	0.112 <i>0.021</i>	Slight	0.150 <i>0.00–0.60</i>	Slight
	Discontinuity Type – Binary	0.089 <i>0.576</i>	Poor	0.183 <i>0.022</i>	Slight	0.000 <i>0.00–0.00</i>	Poor
EEG Transients	Transients	0.187 <i>0.156</i>	Slight	0.247 <i>0.019</i>	Fair	0.403 <i>0.21–0.56</i>	Fair
	Transient Type	0.421 <i>0.059</i>	Moderate	0.114 <i>0.148</i>	Slight	0.167 <i>–0.50–0.60</i>	Slight
	Transient Type – Binary	0.524 <i>0.011</i>	Moderate	0.111 <i>0.222</i>	Slight	0.143 <i>–0.50–0.50</i>	Slight
Unreliably Assessed Variables							
Multiple Impressions		0.124 <i>0.150</i>	Slight	–0.018 <i>0.718</i>	Poor	0.186 <i>–0.14–1.00</i>	Slight

EEG Feature		Interrater Agreement				Intrarater Agreement	
		Assessment 1		Assessment 2			
		Kappa Value	Kappa Agreement Category	Kappa Value	Kappa Agreement Category	Kappa Value Average	Average Kappa Agreement Category
		<i>p-value</i>		<i>p-value</i>		<i>Range</i>	
Behavioral State	Quiet Sleep	0.032 <i>0.660</i>	Poor	0.069 <i>0.349</i>	Poor	0.237 <i>-0.11-0.65</i>	Fair
	Transitional Sleep	n/a	n/a	-0.008	Poor	0.000 <i>n/a</i>	Poor
	Indeterminate Sleep	0.003 <i>0.948</i>	Poor	0.065 <i>0.246</i>	Poor	0.412 <i>0.00-0.78</i>	Moderate
Voltage	Burst abnormality type	0.249 <i>0.551</i>	Fair	-0.086 <i>0.758</i>	Poor	0.200 <i>0.00-0.40</i>	Slight
	Interburst voltage	0.091 <i>0.820</i>	Poor	0.143 <i>0.148</i>	Slight	0.100 <i>0.00-0.20</i>	Slight
	Interburst abnormality type	0.012 <i>0.959</i>	Poor	0.121 <i>0.180</i>	Slight	0.143 <i>-0.33-0.54</i>	Slight
Frequency	Alpha	-0.068 <i>0.172</i>	Poor	-0.034 <i>0.549</i>	Poor	0.152 <i>-0.06-0.51</i>	Slight
	Beta	-0.044 <i>0.281</i>	Poor	-0.030 <i>0.460</i>	Poor	0.158 <i>0.00-0.52</i>	Slight
	Best frequency achieved	0.118 <i>0.216</i>	Slight	0.039 <i>0.589</i>	Poor	0.304 <i>0.11-0.44</i>	Fair
Symmetry		0.063 <i>0.375</i>	Poor	0.063 <i>0.375</i>	Poor	0.231 <i>0.00-0.47</i>	Fair
Synchrony		0.038 <i>0.621</i>	Poor	0.078 <i>0.176</i>	Poor	0.193 <i>0.00-0.47</i>	Slight
Dysmaturity		0.026 <i>0.688</i>	Poor	0.032 <i>0.500</i>	Poor	0.269 <i>-0.04-0.42</i>	Fair
Artifact		0.063 <i>0.396</i>	Poor	0.056 <i>0.540</i>	Poor	0.238 <i>-0.18-0.43</i>	Fair