Published in final edited form as:

Epilepsy Behav. 2014 March ; 32: 102–107. doi:10.1016/j.yebeh.2014.01.011.

EEG Interpretation Reliability and Interpreter Confidence: A Large Single Center Study

Arthur C. Grant^{a,b,*}, Samah G. Abdel-Baki^c, Jeremy Weedon^d, Vanessa Arnedo^a, Geetha Chari^a, Ewa Koziorynska^a, Catherine Lushbough^a, Douglas Maus^{a,b}, Tresa McSween^a, Katherine A. Mortati^a, Alexandra Reznikov^a, and Ahmet Omurtag^{c,e}

^aDepartment of Neurology, SUNY Downstate Medical Center, Brooklyn, NY, USA

^bDepartment of Physiology & Pharmacology, SUNY Downstate Medical Center, Brooklyn, NY, USA

^dThe Scientific Computing Center, SUNY Downstate Medical Center, Brooklyn, NY, USA

^cBioSignal Group, Corp. Brooklyn, NY, USA

^eDepartment of Biomedical Engineering, University of Houston, Houston, TX, USA

Abstract

The intrarater and interrater reliability (I&IR) of EEG interpretation has significant implications for the value of EEG as a diagnostic tool. We measured both I&IR of EEG interpretation based on interpretation of complete EEGs into standard diagnostic categories and rater confidence in their interpretations, and investigated sources of variance in EEG interpretations. During two distinct time intervals six board-certified clinical neurophysiologists classified 300 EEGs into one or more of seven diagnostic categories, and assigned a subjective confidence to their interpretations. Each EEG was read by three readers. Each reader interpreted 150 unique studies, and 50 studies twice to generate intrarater data. A generalizability study assessed the contribution of subjects, readers, and the interaction between subjects and readers to interpretation variance. Five of the six readers had a median confidence of 99%, and the upper quartile of confidence values was 100% for all six readers. *Intrarater* Cohen's kappa (κ_c) ranged from 0.33 to 0.73 with an aggregated value of 0.59. κ_c ranged from 0.29 to 0.62 for the 15 reader pairs, with an aggregated Fleiss kappa of 0.44 for interrater agreement. The κ_c were not significantly different across rater pairs (Chi-Square = 17.3, df=14, $p = 0.24$). Variance due to subjects (i.e. EEGs) was 65.3%, to readers was 3.9%, and to the interaction between readers and subjects was 30.8%. Experienced epileptologists have very high confidence in their EEG interpretations and low to moderate I&IR, a common paradox in clinical medicine. A necessary but insufficient condition to improve EEG interpretation accuracy is to increase intrarater and interrater reliability. This goal could be accomplished, for instance, with an automated on-line application integrated into a continuing medical education module that measures and reports EEG I&IR to individual users.

© 2014 Elsevier Inc. All rights reserved.

*To whom correspondence should be addressed at: SUNY Downstate Medical Center, Comprehensive Epilepsy Center, 450 Clarkson Ave., Box 1275, Brooklyn, NY 11203, 718.270.2959 (tel), 718.270.4711 (fax), arthur.grant@downstate.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

interrater reliability; intrarater reliability; confidence; EEG

1. Introduction

There is no gold standard for an EEG's true interpretation. While EEG findings are described with a standardized terminology, the skill of EEG interpretation is learned primarily in a master-apprentice format. A necessary, though insufficient, condition for accurate EEG interpretation is high intrarater and interrater reliability (I&IR) among interpreters. If interpreters do not agree with themselves when re-reading a study, or with each other when reading the same study, then at least some interpretations must be wrong. Over 50 years of research has consistently shown that EEG I&IR, when measured with the kappa statistic, ranges from slight to substantial depending on the specific EEG finding or interpretive category being examined (Table 1). Few research studies have examined EEG I&IR based on interpretation of complete EEGs rather than short segments. Similarly, most studies have measured I&IR of detecting specific findings such as seizures or interictal epileptiform discharges, rather than of interpreting the EEG into standard diagnostic categories (e.g. Table 2). These important limitations in the accuracy of EEG interpretation are not addressed in EEG textbooks, possibly because expert EEG interpreters paradoxically have very high confidence in their interpretations, a surprisingly common circumstance in clinical medicine [1, 2].

In light of the limited scope of prior EEG I&IR studies, and the speculation that EEG experts may have both high confidence in their interpretations and less than "almost perfect" I&IR, we designed a study with the following four goals: 1) measure *intrarater reliability* of established EEG interpretation categories based on interpretation of complete EEGs, 2) measure *interrater reliability* of established EEG interpretation categories based on interpretation of complete EEGs, 3) measure rater confidence in their interpretations, and 4) investigate sources of variance in EEG interpretations. To achieve both these goals and the generalizability of study results we created a large, heterogeneous sample of EEGs for interpretation.

2. Methods

2.1 EEG Recordings

The sample consisted of 300 30-minute EEGs recorded digitally with standard 10 – 20 system electrodes, plus electrodes T1 and T2. The EEGs had been recorded on both in-patients and out-patients, and were selected as illustrated (Fig 1). The sample was designed to include EEGs heterogeneous with respect to patient age (1-year-old) and findings, but with an overrepresentation of abnormal studies, especially those including status epilepticus, seizure(s), and interictal epileptiform discharges. EEG findings were based on the original clinical interpretation, with the realization that different interpretations would likely arise in the research study. The final set of 300 EEGs were de-identified, including removal of technologist comments. Interpreters were aware of intermittent photic stimulation and hyperventilation, when applicable.

2.2 Interpreters

Interpreters consisted of a pool of six epileptologists, three adult and three pediatric, all board-certified in clinical neurophysiology (Table 3). The adult epileptologists had fellowship training at three different centers. Two of the pediatric epileptologists had fellowship training at the same center, and all three had been working together for five

years. Each interpreter spends 20 – 30% of their time interpreting EEG and video-EEG studies. For the purpose of EEG reading assignments, the six interpreters (A – F) were divided into 20 unique groups of 3 readers (ABC, ABD...DEF), with each reader in 10 groups. The readers underwent no specific training or preparation for the study.

2.3 Interpretation Procedure

Readers were aware of only patient age and medications, making interpretation more difficult than in routine clinical practice. They were unaware of indication for the EEG, patient history, and technologist comments. To mimic standard EEG interpretation conditions, readers were free to adjust the montage, voltage sensitivity, time base, and frequency filters. For each study, interpreters assigned a probability (subjective confidence) to each of seven primary diagnostic categories (Table 2). These categories were chosen because they are mutually exclusive and have distinct clinical correlations. The abnormal categories were considered hierarchical, with status epilepticus at the top and slowing at the bottom. In other words, an EEG with a single seizure, epileptiform discharges and slowing would be categorized as “seizure.”

One category had to have a probability higher than any other, and the sum of all probabilities had to equal 1. For example, if the interpreter was 70% certain that the EEG contained epileptiform discharges and independent slowing, 15% certain that it contained only epileptiform discharges, and 15% certain that it contained only slowing, those three categories would be assigned 0.7, 0.15 and 0.15, respectively, and all other categories would be zero. When appropriate for the primary category assigned the highest confidence, interpreters further qualified the abnormality with respect to laterality and distribution. Here we report and analyze only the category assigned the highest probability.

2.4 EEG assignments

There were two interpretation time intervals (T1, T2), each lasting two months and separated by four months. Each reader group interpreted 10 unique EEGs during T1 and 5 unique EEGs during T2. This arrangement resulted in: 1) each reader interpreting 150 unique EEGs, 2) each of the 300 EEGs being interpreted by 3 readers, and 3) each pair of readers interpreting 60 EEGs in common. In addition, each reader re-interpreted 50 studies during T2 that he or she had interpreted during T1. These studies were selected specifically for each reader to include every EEG interpreted at T1 in a rare category (SE, Sz, Ep, Ep + SI, U, Table 2), plus a random sample of EEGs from the other categories to make a total of 50. Readers knew that some T2 studies were repeats, but were unaware of how many or which ones.

2.5 I&IR analysis

Fleiss' kappa (κ_f) was used as a measure of interrater agreement among the six readers [3]. The more traditional Cohen's kappa coefficient (κ_c) was used as a measure of interrater agreement among rater pairs, and was compared to the aggregate κ_f . These analyses were performed using firstly all seven EEG diagnostic categories, then as a secondary analysis using the following three condensed categories with uninterpretable (U) excluded: 1) SE & Sz, 2) Ep+SI & Ep & SI, 3) NI. These three categories separated EEGs into those with ictal patterns, non-ictal abnormalities, and normal. Cohen's kappa was also used as a measure of intrarater agreement for the 50 studies interpreted by each reader during both T1 and T2, and as a measure of interrater agreement among adult and pediatric trained readers interpreting adult and pediatric EEGs. Confidence intervals for Fleiss kappas were calculated by a nonparametric bootstrap method with 1,000 replications.

2.6 Probability of a reader being wrong

The interrater aggregated κ_f and data in Table 4 can be used to calculate a statistic that is more intuitive than the kappa value, namely the probability of a reader being wrong in their assignment of the most likely diagnostic category for a randomly selected EEG. Kappa is a measure of the degree of agreement between a pair of readings, corrected for the probability of agreement expected by chance alone. Specifically, $\text{kappa} = (P - P_e) / (1 - P_e)$, where P is the probability that a randomly selected pair of readings (i.e. from two different *readers* of the same EEG) will agree on a randomly selected category, and P_e is the probability that such agreement will occur by chance. If P_n is the prevalence of the n th category, then P_e equals the sum over n of P_n squared. The P_n are calculated from the right most column of Table 4, and the aggregated κ_f are known. From these data the value P is easily obtained, and thus $1 - P$, which is the probability that a randomly selected pair of readings will *disagree* on a randomly selected category.

2.7 Generalizability Analysis

A generalizability study (G-study) was performed in which propensity of EEG interpretations to be assessed (with highest probability) as either normal or non-normal (including U) was modeled as being made up of three variance components, with the intention of assessing the contribution of each component to total variance. The components were: 1) subjects (identified by the EEG study number), representing agreement across all raters in whether pathology exists or not; 2) raters, representing the notion that some raters will be more generally inclined than others to “find” pathology; 3) the interaction of subjects with raters, representing the notion that some EEG patterns will be assessed consistently across raters, whereas others will not. A generalized mixed linear model was constructed for this purpose, with random factors subject, reader, and subject x reader.

SAS (SAS Institute, Cary NC) Release 9.2 software was used for all analyses.

2.8 Sample size considerations

Based on a projected Fleiss kappa of 0.7 for interrater reliability, it was estimated that 300 EEG studies read by each of 3 raters would yield a 95% confidence interval for kappa of width no more than 0.1 units. Since most studies would not be read twice by the same rater it was accepted that intrarater reliability would necessarily be assessed with less precision.

3. Results

3.1 Frequency of categorical assessments for each rater

The frequency of the seven primary diagnostic EEG categories assigned by each rater is shown in Table 4.

3.2 Subjective confidence in EEG interpretations

Confidence levels for each reader, based on the diagnostic category given the highest probability for each of the 150 EEGs (*i.e.*, excluding re-readings), are shown in Table 5A. Five of the six readers had a median confidence of 99%, and the upper quartile of confidence values was 100% for all six readers. Confidence values for the 50 EEGs interpreted at both T1 and T2 are shown in Table 5B. The median confidence was 100% for four of the six readers at both T1 and T2. The upper quartile of confidence was 100% for all six readers at T1 and T2, except for reader D whose confidence was 99% at T2.

3.3 Intrarater reliability

Intrarater κ_c ranged from 0.33 to 0.73 (Table 6A), with the aggregated κ_c of 0.59 in the high moderate range. Three readers had a κ_c in the substantial range (Table 1). A test of the null hypothesis that the κ_c were equal across raters was significant (Chi-Square = 14.6, df=5, $p = 0.012$), meaning that there was a significant difference among the six κ_c . After excluding reader A, whose κ_c of 0.33 was substantially lower than the rest, the aggregated κ_c was 0.63 (95% CI 0.56 – 0.70), and the remaining five κ_c were not significantly different from each other (Chi-Square = 4.4, df=4, $p = 0.35$).

When EEG interpretations were condensed into the three categories of ictal patterns, non-ictal abnormalities, and normal, the aggregated κ_c was in the substantial range (aggregated $\kappa_c = 0.70$, 95% CI 0.62 – 0.78).

3.4 Interrater reliability

For the 15 reader pairs κ_c ranged from 0.29 to 0.62, with an aggregated κ_f of 0.44 (Table 6B). The κ_c were not significantly different across rater pairs (Chi-Square = 16.2, df=14, $p = 0.30$). With EEG interpretations condensed into the three clinically meaningful categories the κ_c ranged from 0.33 to 0.72, with an aggregated κ_f of 0.55 (Table 6B). Again, the κ_c were not significantly different across rater pairs (Chi-Square = 17.3, df=14, $p = 0.24$). Interrater agreement among the adult and pediatric trained readers interpreting EEGs from adult and pediatric patients is shown in Table 6C. Pediatric epileptologists interpreting pediatric EEGs had the highest κ_c of 0.68, while adult epileptologists interpreting pediatric EEGs had the lowest κ_c of 0.33.

3.5 Probability of a reader being wrong

When interpreting EEGs into the seven primary categories, the probability that a randomly selected *pair* of readers will *disagree* on a randomly selected category is about 42%, implying that the probability of *one reader* being wrong is at least 21%. With the EEG diagnostic categories condensed to normal, ictal, and non-ictal abnormalities, the probability of disagreement between a randomly selected *pair* of readers is about 23%, and thus the probability of *one reader* being wrong is at least 11.5%. These data refer only to choice of the diagnostic category with highest probability, and not to how the interpreters would ultimately have rendered a clinical correlation.

3.6 Generalizability Analysis

The G-study revealed that variance due to subjects was 65.3%, (Wald 95% CI 52.4% – 83.8%) and variance due to the readers was only 3.9% (Wald 95% CI 1.4% – 35.5%). Variance due to the interaction between readers and subjects was 30.8% (Wald 95% CI 26.0% – 37.1%).

4. Discussion

This is the first study to measure confidence of EEG readers in their interpretations, measure both intrarater and interrater reliability of EEG interpretation based on interpretation of complete EEGs into standard diagnostic EEG categories, and use a generalizability analysis to assess the contribution of subjects, readers, and the interaction of subjects with readers to the overall variance in EEG interpretations.

We found that experienced epileptologists have remarkably high confidence in their EEG interpretations (Table 5). This exuberance is not surprising, since in clinical practice EEG readers are accustomed to making definitive interpretations. Any equivocation about findings the reader may have is forgotten at the rendering of the EEG report. Although

confidence in EEG interpretation has not been previously examined, high confidence and overconfidence are common psychologies in clinical medicine [1, 2].

Overconfidence occurs when high confidence is not commensurate with high accuracy. Since there is no gold standard for EEG interpretations, true accuracy can not be measured. However, maximum possible accuracy depends on I&IR, since high prevalence of disagreement necessarily implies high prevalence of inaccurate interpretation. The EEG readers in this study were overconfident, since they had both high confidence and less than “almost perfect” I&IR. The absence of a statistically significant difference between the *intrarater* κ_c of all but one reader, and between *interrater* κ_c of the 15 reader pairs suggests that interpretive disagreements did not arise from a broad spectrum of expertise among the readers. We acknowledge that this is not a definitive conclusion, as the set of confidence intervals around the kappa values span a wide range, and statistical significance between kappas may have emerged with a larger data set. Nonetheless, these readers would be considered experts by any reasonable standard.

For these reasons we propose that the aggregated intrarater κ_c of 0.59 and interrater κ_f of 0.44 among these six epileptologists most likely reflects the inherently subjective nature of EEG interpretation, a vexing problem in the absence of a gold standard. The clinical significance of this problem is illustrated by the calculation that, in this dataset, the probability of a reader rendering a wrong (seven category) interpretation on a randomly selected EEG is at least 21%. Even when the interpretive categories are reduced to normal, ictal, and non-ictal abnormalities, the probability of one reader being wrong is at least 11.5%.

The data in Table 6C suggest two specific *possible* influences of epileptologist training and experience on EEG interpretation skill. First, and not surprisingly, pediatric epileptologists had significantly higher interrater agreement interpreting pediatric EEGs than did adult epileptologists. Second, epileptologists with fellowship training at the same institution, and / or who work together at the same institution for many years (as was true of the three pediatric epileptologists) have higher interrater agreement than those who don't.

Prior studies of EEG intrarater agreement are sparse. Fifty years ago Little and Raffel examined intrarater agreement of a single experienced electroencephalographer interpreting 100 EEGs into a nine point scale ranging from definitely normal to “borderline” to definitely abnormal [4]. The second interpretation was made days to months after the first. Rating deviation was zero for 33 EEGs, ± 1 for 52 EEGs, ± 2 for 12 EEGs, and ± 3 for 3 EEGs [4]. Gerber and colleagues [5] assessed intra- and interrater agreement among board-certified clinical neurophysiologists using a proposed standardized terminology for rhythmic and periodic EEG patterns encountered in critically ill patients [6]. Intrarater agreement of five readers interpreting 58 10-second EEG samples from 11 critically ill adults ranged from slight to substantial (κ between 0.10 and 0.76) depending on the EEG finding evaluated [5]. Combined with our data, these studies support the concept that EEG interpretation is a complex process with an inherently subjective component.

Interrater reliability has been examined with EEGs recorded from adult and pediatric outpatients [4, 7–10], patients undergoing pre-surgical video-EEG monitoring [11–13], critically ill patients [5, 14], and other patient populations [15–21]. Most of these studies have measured the agreement on detecting specific EEG patterns, such as interictal epileptiform discharges [8, 16, 17], seizures [11–13, 15, 21], or patterns specific to critically ill patients [5, 14]. Kappa values in these studies (when reported) ranged from – 0.2 to 0.83, with the highest values associated with detecting the presence or absence in the EEG of a single finding, such as interictal epileptiform discharges [9].

Theoretically, all variability in assessment of pathology should be due to subjects, i.e. the EEG recordings. The G-study revealed that subjects accounted for nearly two-thirds of the variance in assessing an EEG as normal or abnormal. The small variance of 3.9% due to readers alone indicates that they did not vary much in their basic proclivity to find pathology. The substantial variance of 30.8% due to the interaction between subjects and readers implies that variability of EEG interpretation depends to a substantial extent on the nature of the EEG findings being interpreted. In other words, some EEG patterns are more difficult to interpret than others, which results in greater interpretive variability.

This study has several weaknesses. The EEG readers were at one academic center, and thus do not necessarily represent the diversity of skill and knowledge of the expert EEG reading community. However, the readers were trained at five institutions and were board certified in clinical neurophysiology. They were also equally divided between adult and pediatric neurologists. Interpreting the EEG studies was more difficult than in usual practice because the readers did not know the indication for the EEG, patient history, or technologist comments, which were all omitted to prevent introduction of bias. Had the readers known this additional information the intrarater and interrater kappa values would likely have been somewhat higher. Indeed, in a debriefing after study completion the interpreters commented that there were a few instances when they could not confidently determine whether diffuse theta activity represented normal drowsiness or abnormal slowing in a mildly encephalopathic patient. We did not address the *clinical significance* of interpretive disagreements, and some disagreements are clearly more significant than others. However, by design each of the seven primary diagnostic EEG categories has a unique clinical correlation. Finally, the study design did not allow for determination of why readers disagreed on the findings in any particular EEG.

5. Conclusions

This comprehensive study of EEG I&IR found that a group of six fellowship trained, board-certified epileptologists have at least a 21% chance of being wrong when interpreting an EEG into one of seven clinically meaningful diagnostic categories. Thus, professional board certification does not necessarily guarantee an accurate interpretation. Our data also revealed that variability in EEG interpretation is due primarily to how readers interpret specific (as yet undefined) EEG patterns, rather than differences in expertise or inclination to find pathology in an EEG recording. These findings suggest that when an EEG interpretation is inconsistent with other clinical data, it may be worthwhile to have the EEG interpreted by a second independent reader, or to obtain a second EEG study. Our results also support the proposal of Gerber et al. to increase EEG interobserver agreement through an interactive training module that promotes consistent use of EEG descriptive terminology at multiple centers [5]. In particular, we suggest that EEG I&IR could be increased, for example, with an automated, on-line, interactive process integrated into a continuing medical education module that continuously measures and reports I&IR of EEG interpretations.

Acknowledgments

Supported by NIH 1RC3NS070658 to Bio-Signal Group Corp. with a subcontract to SUNY Downstate Medical Center. The authors thank Madeleine Coleman for assistance with manuscript preparation.

References

1. Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med.* 2008; 121:S2–23. [PubMed: 18440350]
2. Croskerry P, Norman G. Overconfidence in clinical decision making. *Am J Med.* 2008; 121:S24–9. [PubMed: 18440351]

3. Fleiss, J. Statistical methods for rates and proportions. New York: John Wiley; 1981.
4. Little SC, Raffel SC. Intra-rater reliability of EEG interpretations. *J Nerv Ment Dis.* 1962; 135:77–81. [PubMed: 14465903]
5. Gerber PA, Chapman KE, Chung SS, Drees C, Maganti RK, Ng YT, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. *J Clin Neurophysiol.* 2008; 25:241–9. [PubMed: 18791475]
6. Hirsch LJ, Brenner RP, Drislane FW, So E, Kaplan PW, Jordan KG, et al. The ACNS subcommittee on research terminology for continuous EEG monitoring: proposed standardized terminology for rhythmic and periodic EEG patterns encountered in critically ill patients. *J Clin Neurophysiol.* 2005; 22:128–35. [PubMed: 15805813]
7. Azuma H, Hori S, Nakanishi M, Fujimoto S, Ichikawa N, Furukawa TA. An intervention to improve the interrater reliability of clinical EEG interpretations. *Psychiatry Clin Neurosci.* 2003; 57:485–9. [PubMed: 12950702]
8. Halford JJ, Pressly WB, Benbadis SR, Tatum WO, Turner RP, Arain A, et al. Web-based collection of expert opinion on routine scalp EEG: software development and interrater reliability. *J Clin Neurophysiol.* 2011; 28:178–84. [PubMed: 21399515]
9. Stroink H, Schimsheimer RJ, de Weerd AW, Geerts AT, Arts WF, Peeters EA, 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]
10. Woody RH. Inter-judge reliability in clinical electroencephalography. *J Clin Psychol.* 1968; 24:251–6. [PubMed: 5650559]
11. Spencer SS, Williamson PD, Bridgers SL, Mattson RH, Cicchetti DV, Spencer DD. Reliability and accuracy of localization by scalp ictal EEG. *Neurology.* 1985; 35:1567–75. [PubMed: 4058746]
12. Walczak TS, Radtke RA, Lewis DV. Accuracy and interobserver reliability of scalp ictal EEG. *Neurology.* 1992; 42:2279–85. [PubMed: 1461379]
13. Wilson SB, Scheuer ML, Plummer C, Young B, Pacia S. Seizure detection: correlation of human experts. *Clin Neurophysiol.* 2003; 114:2156–64. [PubMed: 14580614]
14. Abend NS, Gutierrez-Colina A, Zhao H, Guo R, Marsh E, Clancy RR, et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. *J Clin Neurophysiol.* 2011; 28:15–9. [PubMed: 21221016]
15. Benbadis SR, LaFrance WC Jr, Papandonatos GD, Korabathina K, Lin K, Kraemer HC. Interrater reliability of EEG-video monitoring. *Neurology.* 2009; 73:843–6. [PubMed: 19752450]
16. Spatt J, Pelzl G, Mamoli B. Reliability of automatic and visual analysis of interictal spikes in lateralising an epileptic focus during video-EEG monitoring. *Electroencephalogr Clin Neurophysiol.* 1997; 103:421–5. [PubMed: 9368485]
17. Barkmeier DT, Shah AK, Flanagan D, Atkinson MD, Agarwal R, Fuerst DR, et al. High inter-reviewer variability of spike detection on intracranial EEG addressed by an automated multi-channel algorithm. *Clin Neurophysiol.* 2012; 123:1088–95. [PubMed: 22033028]
18. 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]
19. Palmu K, Wikstrom S, Hippelainen E, Boylan G, Hellstrom-Westas L, Vanhatalo S. Detection of ‘EEG bursts’ in the early preterm EEG: visual vs. automated detection. *Clin Neurophysiol.* 2010; 121:1015–22. [PubMed: 20395172]
20. Piccinelli P, Viri M, Zucca C, Borgatti R, Romeo A, Giordano L, et al. Inter-rater reliability of the EEG reading in patients with childhood idiopathic epilepsy. *Epilepsy Res.* 2005; 66:195–8. [PubMed: 16118044]
21. Ronner HE, Ponten SC, Stam CJ, Uitdehaag BM. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. *Seizure.* 2009; 18:257–63. [PubMed: 19046902]
22. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33:159–74. [PubMed: 843571]

Highlights

- Intrarater and interrater reliability of EEG interpretation is fair to moderate
- Confidence in the accuracy of interpretations is extremely high
- The probability that an expert EEG interpretation is wrong is at least 21%.
- An on-line interactive process to increase I&IR of EEG interpretations is feasible

Figure 1.
Flow chart illustrating selection process for the dataset of 300 EEG studies.

Table 1Strength of Agreement Using the Kappa (κ) Statistic [22]

κ Value	Strength of Agreement
< 0	Poor
0 – 0.20	Slight
0.21– 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Table 2

EEG diagnostic categories

category	abbreviation
status epilepticus	SE
seizure	Sz
epileptiform discharges with independent slowing	Ep + Sl
epileptiform discharges without independent slowing	Ep
slowing only	Sl
normal	NI
uninterpretable	U

Studies could be uninterpretable due to artifacts, technical quality, or both.

Table 3

Reader Training.

Reader	Adult/Pediatric	Years Since Fellowship
A	Pediatric	6
B	Adult	6
C	Pediatric	10
D	Adult	3
E	Adult	6
F	Pediatric	9

Table 4

Frequency of Categorical Assessment for Each Reader (A – F)

	A	B	C	D	E	F	Total
SE	8	1	1	3	0	1	14
Sz	3	7	2	5	2	1	20
Ep+SI	33	63	26	35	25	16	198
Ep	8	6	15	12	16	13	70
SI	59	38	54	40	38	65	294
NI	27	35	50	52	67	47	278
U	12	0	2	3	2	7	26
Total	150	150	150	150	150	150	900

Each of the six readers interpreted 150 of the 300 EEGs. No two readers interpreted the same set of 150 studies.

Table 5A

Confidence of Readers in Interpretation of 150 EEG studies

Reader	Min	10 th %ile	Lower Quartile	Median	Upper Quartile
A	40%	60%	70%	85%	100%
B	40%	63%	85%	100%	100%
C	60%	75%	100%	100%	100%
D	55%	70%	90%	99%	100%
E	60%	100%	100%	100%	100%
F	60%	78%	90%	100%	100%

Readers assigned a probability to each of seven EEG diagnostic categories. The data in this table are derived from the EEG category assigned the highest probability.

Confidence of Readers in Interpretation of 50 EEG studies used for analysis of intra-rater agreement

Table 5B

Reader	Time Interval	Min	10 th %ile	Lower Quartile	Median	Upper Quartile
A	T1	40	60	70	80	100
	T2	45	55	60	80	100
B	T1	40	60	85	100	100
	T2	60	70	90	100	100
C	T1	75	100	100	100	100
	T2	60	80	100	100	100
D	T1	60	80	90	100	100
	T2	51	64	80	95	99
E	T1	60	100	100	100	100
	T2	70	100	100	100	100
F	T1	60	73	85	100	100
	T2	60	80	100	100	100

Table 6A

Intrarater reliability based on two interpretations of 50 EEGs

Reader	κ_c	95% CI
A	0.33	0.16 0.50
B	0.50	0.33 0.68
C	0.58	0.42 0.74
D	0.67	0.51 0.82
E	0.73	0.58 0.88
F	0.64	0.48 0.80
Aggregate	0.59	0.52 0.65

CI confidence interval

Table 6B

Interrater reliability by reader pair, for the seven primary EEG diagnostic categories and three condensed categories.

	7 EEG categories		3 EEG categories	
Reader pair	κ_c	95% CI	κ_c	95% CI
AB	0.43	0.27 0.59	0.66	0.43 0.89
AC	0.52	0.35 0.68	0.66	0.46 0.86
AD	0.37	0.23 0.52	0.35	0.13 0.57
AE	0.37	0.23 0.52	0.49	0.30 0.68
AF	0.50	0.34 0.66	0.51	0.26 0.76
BC	0.48	0.33 0.64	0.67	0.48 0.86
BD	0.41	0.26 0.56	0.63	0.43 0.82
BE	0.37	0.21 0.53	0.46	0.25 0.66
BF	0.29	0.14 0.45	0.47	0.24 0.70
CD	0.49	0.33 0.64	0.59	0.39 0.79
CE	0.56	0.39 0.72	0.53	0.33 0.74
CF	0.62	0.46 0.77	0.72	0.53 0.91
DE	0.48	0.33 0.64	0.52	0.31 0.73
DF	0.35	0.18 0.52	0.33	0.09 0.57
EF	0.42	0.25 0.59	0.40	0.16 0.64
κ_f Aggregate	0.44	0.40 0.48	0.55	0.49 0.60

Each reader pair interpreted 60 EEGs in common. CI confidence interval.

Table 6C

Interrater reliability of the 3 adult and 3 pediatric epileptologists interpreting EEGs from adult and pediatric (15 years-old) patients, reported as κ_C (95% CI) summarized across rater pairs.

Interpreter Training ↓	Adult EEG	Pediatric EEG
Adult	0.41 (0.31, 0.52) n = 133	0.33 (0.17, 0.48) n = 47
Pediatric	0.47 (0.35, 0.58) n = 145	0.68 (0.50, 0.86) n = 35