Published in final edited form as:

Epilepsia. 2014 September ; 55(9): 1366–1373. doi:10.1111/epi.12653.

Interrater agreement for Critical Care EEG Terminology

Nicolas Gaspard[*], **Lawrence J. Hirsch**[*], **Suzette M. LaRoche**[†], **Cecil D. Hahn**[‡], and **M. Brandon Westover**[§] **for the Critical Care EEG Monitoring Research Consortium**

[*]Department of Neurology, Comprehensive Epilepsy Center, Yale University School of Medicine, New Haven, Connecticut, U.S.A

[†]Department of Neurology, Emory University, Atlanta, Georgia, U.S.A

[‡]Division of Neurology, The Hospital for Sick Children and University of Toronto, Toronto, Ontario, Canada

[§]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts, U.S.A

Summary

Objective—The interpretation of critical care electroencephalography (EEG) studies is challenging because of the presence of many periodic and rhythmic patterns of uncertain clinical significance. Defining the clinical significance of these patterns requires standardized terminology with high interrater agreement (IRA). We sought to evaluate IRA for the final, published American Clinical Neurophysiology Society (ACNS)–approved version of the critical care EEG terminology (2012 version). Our evaluation included terms not assessed previously and incorporated raters with a broad range of EEG reading experience.

Methods—After reviewing a set of training slides, 49 readers independently completed a Web-based test consisting of 11 identical questions for each of 37 EEG samples (407 questions). Questions assessed whether a pattern was an electrographic seizure; pattern location (main term 1), pattern type (main term 2); and presence and classification of eight other key features ("plus" modifiers, sharpness, absolute and relative amplitude, frequency, number of phases, fluctuation/evolution, and the presence of "triphasic" morphology).

Results—IRA statistics (κ values) were almost perfect (90–100%) for seizures, main terms 1 and 2, the +S modifier (superimposed spikes/sharp waves or sharply contoured rhythmic delta

Address correspondence to Nicolas Gaspard, Comprehensive Epilepsy Center, Yale School of Medicine, PO Box 208018, LLCI-716, New Haven, CT 06520-8018, U.S.A. ; Email: nicolas.gaspard@yale.edu

Conflict of Interest

All authors are members of the CCEMRC. The CCEMRC received infrastructure support from the American Epilepsy Society/Epilepsy Foundation. NG received support from the Epilepsy Foundation. LJH received research support for investigator-initiated studies from UCB-Pharma, Upsher-Smith, and Lundbeck; consultation fees for advising from Lund-beck, Upsher-Smith, GlaxoSmithKline RSC Diagnostics, and NeuroPace; royalties for authoring chapters for UpToDate-Neurology, and for coauthoring the book *Atlas of EEG in Critical Care*, by Hirsch and Brenner, 2010. SML received research support from UCB and royalties from Demos Publishing. CDH received research support from the Canadian Institutes of Health Research, The Hospital for Sick Children Foundation, and the PSI Foundation. MBW received support from the American Brain Foundation and royalties for coauthoring the book *Pocket Neurology*, LWW, 2010. 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.

Supporting Information

Additional Supporting Information may be found in the online version of this article.

activity), sharpness, absolute amplitude, frequency, and number of phases. Agreement was substantial for the +F (superimposed fast activity) and +R (superimposed rhythmic delta activity) modifiers (66% and 67%, respectively), moderate for triphasic morphology (58%), and fair for evolution (21%).

Significance—IRA for most terms in the ACNS critical care EEG terminology is high. These terms are suitable for multicenter research on the clinical significance of critical care EEG patterns.

Keywords

Continuous EEG monitoring; EEG terminology; Periodic patterns; Rhythmic patterns; Interrater agreement; Intensive care; Critical care; PLEDs; GPEDs

Continuous electroencephalography (EEG) monitoring (CEEG) is important for detecting nonconvulsive seizures (NCSz), and rhythmic and periodic patterns associated with increased risk of seizures and poor outcome in critically ill patients.[1–8] Many widely used EEG terms lack a consensus definition and exhibit high interrater variability, including terms for “epileptiform discharges,”[9] “seizures,”[10] and “triphasic waves.”[11] In addition, the distinction between interictal and ictal patterns can be difficult in acutely ill patients.[1]

Standardized terminology is necessary to enable consistent interpretation of intensive care unit (ICU) EEG studies and to investigate the clinical relevance of ICU EEG patterns. Indeed, it is still debated whether the patterns under study are associated with potential for neuronal injury and warrant treatment, or if they are simply an epiphenomenon of encephalopathy or acute brain injury. Progress in resolving these debates cannot be made without a validated terminology for reliably identifying and distinguishing the patterns in question. This terminology must be sufficiently expressive to capture most EEG patterns encountered in acutely ill patients, and easy to learn for electroencephalographers of varied experience.

A single-center study evaluating the interrater agreement (IRA) of the 2005 American Clinical Neurophysiology Society (ACNS) Critical Care EEG terminology[12] found slight to moderate agreement for tested terms, and a subsequent study found improved IRA on a revised version (Table 1).[13,14] However, the study was limited: some descriptors were not tested; questions had variable numbers of choices, and most raters were involved in its development, thereby limiting generalizability of the findings. Additional revisions were incorporated in the 2012 version,[15] and extensive training materials were developed and placed online at no cost (a link to the training material is provided at the end of the article).

The goal of this study was to assess the IRA of the 2012 ACNS Critical Care EEG Terminology, including terms not evaluated previously and raters with a broad range of expertise with EEG reading and familiarity with the terminology.

Subjects/Materials and Methods

EEG samples

EEG samples (n = 37) were collected prospectively between January 1, 2012, and September 1, 2012. Samples consisted of 12 examples of lateralized periodic discharges (LPDs), 6 of generalized periodic discharges (GPDs), 3 of bilateral independent periodic discharges (BIPDs), 4 of lateralized rhythmic delta activity (LRDA), 4 of generalized rhythmic delta activity (GRDA), 2 of generalized rhythmic spike-and-wave complexes (GSWs), 1 of lateralized rhythmic spike-and-wave complexes (LSWs), and 5 of seizures. Samples were obtained from unselected consecutive cases where the pattern of interest was present. A greater emphasis was placed on periodic discharges, as in our experience they exhibit the largest amount of variability of morphology and allow most components of the terminology to be used.

Ten second to 60 s examples were presented as consecutive 10 s pages in a longitudinal bipolar montage, with the low-pass filter at 70 Hz and the high-pass filter at 1 Hz. None of the samples contained significant amounts of electrical noise, and thus the notch filter was left off. Sensitivity was individually set for each example to optimize visualization. Raters could not change these settings. The duration of each example was chosen to present enough information for all terms to be applied, although it was expected that for some cases, the determination of evolution and fluctuation could not be captured in samples of this brief duration. The durations of nonictal (median 30 s and interquartile range [IQR] 30–40 s) and ictal samples (median 40 s and IQR 30–50 s) were not significantly different ($p = 0.29$; Mann-Whitney U-test).

Completing the IRA assessment was part of a certification test that each rater had to take to participate in studies of the Critical Care EEG Monitoring Research Consortium (CCEMRC). Before participating, the raters were invited to review a set of training slides presenting the most recent version of the Critical Care EEG Terminology.[15] These slides can be obtained on the website of the ACNS (a link is provided at the end of the article). The test was designed with a Web-based survey engine (SurveyMonkey, Palo Alto, CA, U.S.A.). Raters were free to complete the test in multiple sittings, and were allowed to refer to any reference text while taking the test, including the recent version of the terminology and the training slides. For each sample, they were asked to first decide whether the pattern represented an unequivocal electrographic seizure (defined in the ACNS terminology as any pattern composed of generalized spike-and-wave discharges at >3 Hz, or any pattern of evolving discharges reaching a frequency of >4 Hz). In cases where raters opted to classify the pattern as a definite seizure, the rest of the questions were skipped and the raters were brought to the next sample. If they classified the pattern as not a definite seizure, then raters were asked to describe it using 11 descriptors defined in the terminology (Table 1). We derived four additional concepts about the plus (+) modifiers, namely “any +”, and the presence or absence of individual plus modifiers (+F, superimposed fast activity; +R, superimposed rhythmic delta activity; and +S, superimposed spikes/sharp waves or sharply contoured delta activity), yielding a total of 15 concepts.

A gold standard was defined by combining the responses of five raters who were considered to have a high experience with the terminology (SML, CDH, and LJH) and/or designed the study (NG and MBW). It was possible in all cases to identify a majority choice among these five experts, as no ties occurred.

Raters were also asked to report their level of confidence with the nomenclature using a semiquantitative scale (very uncomfortable, uncomfortable, neutral, comfortable, and very comfortable), as well as the number of years of experience reading EEG studies (stratified as 0–2, 2–5, 5–10, 10–15, >15). A total of 49 raters reviewed all 37 EEG samples and answered all of the associated questions. Eighty-two percent (40/49) of readers were specialized in adult neurology; the remaining 18% (9/49) practiced primarily in pediatric neurology. The range of EEG reading experience among raters was broad, as follows: ≤2 years of experience, 45% of raters; 2–5 years, 14%, 5–10 years, 20%, 10–15 years, 8%, ≥15 years, 12%. Further details regarding rater characteristics are provided in the Table S1.

Statistical analysis

IRA was quantified using Gwet's multirater agreement coefficients AC1 (for categorical data) and AC2 (for ordinal data),[16,17], hereafter referred to simply as kappa (κ) statistics. Gwet's κ statistics were recently developed to overcome certain well-known shortcomings of conventional IRA coefficients such as Cohen's and Fleiss' κ statistics; specifically, these latter κ statistics perform badly (exhibit "paradoxes") when raters exhibit a high or low degree of agreement, and when the true prevalence of classes among the cases being rated is nonuniform.[16,18] Further details about κ statistics for measuring IRA are provided as Supplemental Material.

We divided the 15 concepts for each case into categorical and ordinal assessment types as follows: *ordinal*: sharpness, absolute amplitude, relative amplitude, frequency, number of phases, and evolution; and *categorical*: definite seizure, main term 1, main term 2, "plus" modifier, and triphasic morphology. Agreements on categorical assessments were considered to be all-or-none. Partial agreement for ordinal data was scored using a conventional quadratic penalty function, adjusted for the number of possible choices.[17] The following qualitative divisions are used to categorize κ values falling into different ranges: slight agreement 0.01–0.20; fair agreement 0.20–0.40; moderate agreement 0.40–0.60; substantial agreement 0.60–0.80; and almost perfect agreement 0.80–1.00. Precision of the overall group estimates for IRA was quantified by 95% confidence intervals (Cis) for the estimated κ values calculated by the Jackknife method.[17,19]

Overall individual performance scores were calculated as average percentage of correct answers, averaged over all 407 questions (11 questions for each of 37 cases), where "correct" answers were defined to be the majority answer given by the five-member expert panel described earlier. Answers were considered either correct or incorrect, without allowance for partial credit (in contrast with calculations below for IRA, in which partial credit was possible for questions about ordinal variables). The application of some of the descriptors required categorical classifications (e.g., main term 1: lateralized [L] vs. generalized [G] vs. bilateral independent [BI] vs. multifocal [M]), whereas the others required ordinal (rank-ordered) classifications (e.g., sharpness: spiky >sharp >sharply

contoured >blunt). “Corrected” overall scores were also calculated by replacing the raw score with an average in which the score for each concept was weighted by either the observed percent agreement or by the estimated degree of IRA (κ). This correction was performed to mitigate the negative bias imparted by poorly defined concepts, that is, to deemphasize the contribution to each individual’s overall performance score of questions for which there was low IRA.

Sources of disagreement were investigated by plotting confusion matrices,[20] showing for each of eight nonbinary concepts the percentage of each available answer chosen by respondents versus the correct choice, as determined by expert consensus (SML, CDH, LJH, NG, and MBW).

Possible effects of experience and comfort with the terminology were investigated by linear regression analysis applied to plots of the scores of individual readers (relative to expert panel consensus answers) versus years of experience reading EEG recordings or degree of comfort.

Statistical calculations and figure creation were performed using the MATLAB Statistics Toolbox and custom software developed in MATLAB (The MathWorks, Natick, MA, U.S.A.).

Standard protocol approvals, registrations, and patient consents

This study was part of a larger multicenter initiative, the Critical Care EEG Monitoring Research Consortium, which was approved by the institutional review boards of all participating institutions. Informed consent was not required because the “subjects” were the investigators.

Results

Raters’ characteristics are detailed in Table S1. Individual scores on each of 15 tested concepts are displayed as a heat map in Figure 1A. A “raw” overall score was assigned to each respondent by averaging the percentages over the 15 response categories (Fig. 2B, orange bars). “As expected, agreement-corrected overall scores were slightly higher than raw scores. The median κ-corrected overall score was 80.2 (IQR 73.1–87.3)%.

The percentages of observed agreement and estimated chance-corrected level of IRA (κ value) with 95% CIs for all 15 concepts are shown in Table 1. Agreement was “almost perfect” for seizures, main terms 1 and 2, the +S modifier, sharpness, absolute amplitude, frequency, and number of phases; “substantial” for the +F and +R modifiers; “moderate” for triphasic morphology; and “fair” for evolution and for agreement on precise combinations of plus (+) modifiers (i.e., choosing between F, R, S, FR, FS, and no +). “Poor” agreement was found only when assessing for the presence of any plus (+) modifier (κ = 19.2%). Relationships between the observed percentage-agreement values and the estimated IRA κ values are displayed graphically in Figure 2.

Of interest, a “wisdom of the crowd” phenomenon was evident, in that there was a high degree of agreement between the expert consensus answers and the majority answer among

all 49 respondents: The majority answer differed from the expert consensus for only 13 (3%) of 407 questions. These questions all addressed terms other than main terms 1 and 2 and 10 of 13 of these addressed terms with a low IRA.

We next investigated the nature of terminology disagreements for nonbinary concepts using confusion matrices (Fig. 3). The minority of incorrect responses was generally “close” to the correct response (near the diagonal line), reflecting the difficulty of choosing between reasonable alternatives in borderline cases, but nevertheless providing evidence of at least partial agreement. For example, when the correct choice for main term 1 was “BI,” the two significant minorities chose either G or M; and when the correct sharpness modifier was “spiky,” most incorrect responses fell into the neighboring category, “sharp,” whereas very few responses fell into “sharply contoured,” and none fell into “blunt.”

Linear regression between EEG reading experience and individual performances showed nonsignificant results (small slopes indistinguishable from zero) in all but one category (absolute amplitude) (Fig. 4). Similar results were obtained with self-reported comfort (Fig. S2).

Discussion

In this IRA study of the 2012 ACNS Critical Care EEG Terminology, we found substantial to almost perfect agreement ($\kappa > 0.6$) between raters for most of the investigated terms, with the exception of triphasic morphology and evolution, which showed moderate and fair agreement, respectively. For terms that were previously tested, we found a level of IRA that was comparable with or superior to previous smaller scale studies, with the exception of main term 2, which showed a mild decrease. A direct comparison with previous studies should, be approached with caution, however, as they differed from the present study in several respects, including the selection and number of EEG samples, the number and experience of raters, and the methods of statistical analysis. Finally, we also found that performance relative to expert consensus was similar among EEG readers with different levels of expertise, suggesting that the terminology can be learned quickly by young, motivated, and well-trained electroencephalographers, even those without many years of EEG reading experience.

The strengths of our study lie in the large number of EEG readers and in the number and variety of samples that were assessed, which ensured a more accurate estimation of agreement between raters. In addition, in contrast to the previous study, the design of the present IRA study was planned to allow the calculation of κ statistics for all concepts. The wide range of rater experience (from 1 to almost 30 years) supports the generalizability of our findings to EEG readers of all levels who are involved in interpreting critical care EEG studies. Finally, cases were consecutively selected on the basis of the presence of a pattern that would fall under the tested terminology and were not “cherry picked” to be exemplary examples of each pattern. We thus believe that they offer a reasonable representation of the variety of cases that are encountered in routine practice, including difficult cases, thereby further strengthening the generalizability of our results.

Although this was not the primary aim of this study, we found excellent agreement for the distinction between ictal and interictal epochs. This surprisingly high degree of agreement differs from what had been reported previously in critically ill patients.[10] Although the focus of the terminology is not to standardize interpretation of electrographic seizures, it does provide simple rules to define electrographic seizures for *operational research purposes* (i.e., not intended for clinical use). Our results indicate that these rules can be applied with a high level of consistency. Further research using a larger variety of ICU EEG seizure examples is needed to confirm and extend the findings of the present study regarding interrater agreement for electrographic seizures.

We found a low IRA for the "any +" concept, which may seem paradoxical in light of the much better agreement for each individual plus (+) modifier. However, in this case, performance was degraded by poor agreement, on which of the 37 cases simultaneously *lacked all three* individual plus modifiers (+F, +R, +S). In other words, agreeing on whether any given individual plus feature is absent from an EEG is relatively easier than agreeing on whether *all* plus features are absent, since a perceived detection of any one of these three features automatically destroys agreement with a reader who classifies the EEG as lacking any "plus" features.

This study was subject to a number of limitations. First, despite the large number of samples, some modifiers were not tested (+FS, very low amplitude, frequency >3 Hz, multifocal distribution), most of which are uncommonly encountered.

Second, the fact that the EEG samples used in our study were short duration (10–60 s) might have affected IRA, particularly for features defined in terms of temporal evolution. Viewing only a limited sample of a pattern might artificially enhance agreement by giving the impression that it is more regular (or recurrent?) than it actually is. It is possible that the excellent agreement for seizures seen in this study is partly attributable to such an effect. In other cases, a more extended sample might enhance agreement, allowing the temporal characteristics of a pattern to appear more obvious, whereas a briefer sample left room for greater uncertainty. In particular, the relatively poor IRA seen in the present study for the concept of evolving versus fluctuating versus static could be partly due to the short duration of EEG epochs used for this study. Alternatively, the poor agreement for this concept could be due to the complexity of the current definitions or that these concepts address features that are too subtle and/or too complex to be compactly codified into words. Further work with longer duration samples is needed to clarify IRA for electrographic seizures and evolution.

Third, the use of the term "triphasic morphology" exhibited only a moderate agreement between raters. This is in agreement with a previous study that found that the presence of triphasic waves was an important cause for divergence between two readers scoring EEGs of comatose patients.[11] A dedicated IRA study to determine which aspects of triphasic wave morphology can be applied most consistently (number of phases, polarity of the most prominent phase, duration of each phase) could help improve this definition. It is also possible that, despite its historical significance, the concept of triphasic wave morphology is

no longer clinically useful, and may ultimately need to be discarded or replaced with another concept for which there would be wider agreement.

Overall, these results largely validate the ACNS Standardized Critical Care EEG Terminology and indicate that the test can serve as a reasonable method for certifying collaborators for research purposes. Readers who wish to take the certification test can obtain the manuscript describing the latest version of the terminology[15] and the training slides from the website of the American Clinical Neurophysiology Society (http://www.acns.org/research/critical-care-eeg-monitoring-research-consortium-ccemrc/education). The link to the certification test can be obtained by contacting the authors (NG, MBW).)

Conclusions

We assessed the agreement between 49 raters using the most recent ACNS-endorsed version of the ACNS Critical Care EEG Terminology. Of the 15 tested concepts, most showed substantial to near perfect agreement, with the exception of triphasic morphology and evolution, which showed moderate and fair agreement, respectively. The current terminology was applied consistently among raters with variable years of experience interpreting EEG studies, but with specific training in ICU EEG. These results indicate that the terminology is defined sufficiently to enable meaningful investigation of the included terms and possibly for standardization of clinical EEG reports.

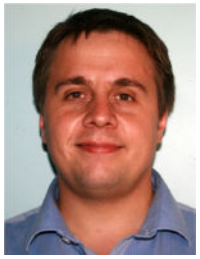
Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to acknowledge all the electroencephalographers who participated in the study. We also wish to thank members of the CCEMRC for helpful discussion.

Biography



Nicolas Gaspard is a Postdoctoral Research Associate at the Yale Comprehensive Epilepsy Center.

References

1. Chong DJ, Hirsch LJ. Which EEG patterns warrant treatment in the critically ill? Reviewing the evidence for treatment of periodic epileptiform discharges and related patterns. J Clin Neurophysiol. 2005; 22:79–91. [PubMed: 15805807]
2. Brophy GM, Bell R, Claassen J, et al. Guidelines for the evaluation and management of status epilepticus. Neurocrit Care. 2012; 17:3–23. [PubMed: 22528274]
3. Orta DSJ, Chiappa KH, Quiroz AZ, et al. Prognostic implications of periodic epileptiform discharges. Arch Neurol. 2009; 66:985–991. [PubMed: 19667220]
4. Claassen J, Taccone FS, Horn P, et al. Recommendations on the use of EEG monitoring in critically ill patients: consensus statement from the neurointensive care section of the ESICM. Intensive Care Med. 2013; 39:1337–1351. [PubMed: 23653183]
5. Abend NS, Dlugos DJ, Hahn CD, et al. Use of EEG monitoring and management of non-convulsive seizures in critically ill patients: a survey of neurologists. Neurocrit Care. 2010; 12:382–389. [PubMed: 20198513]
6. Hughes JR. Periodic lateralized epileptiform discharges: do they represent an ictal pattern requiring treatment? Epilepsy Behav. 2010; 18:162–165. [PubMed: 20554251]
7. Foreman B, Claassen J, Abou Khaled K, et al. Generalized periodic discharges in the critically ill: a case–control study of 200 patients. Neurology. 2012; 79:1951–1960. [PubMed: 23035068]
8. Gaspard N, Manganas L, Rampal N, et al. Similarity of lateralized rhythmic delta activity to periodic lateralized epileptiform discharges in critically ill patients. JAMA Neurol. 2013; 70:1288–1295. [PubMed: 23921464]
9. Gilbert DL, Sethuraman G, Kotagal U, et al. Meta-analysis of EEG test performance shows wide variation among studies. Neurology. 2003; 60:564–570. [PubMed: 12601093]
10. Ronner HE, Ponten SC, Stam CJ, et al. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. Seizure. 2009; 18:257–263. [PubMed: 19046902]
11. Young GB, McLachlan RS, Kreeft JH, et al. An electroencephalographic classification for coma. Can J Neurol Sci. 1997; 24:320–325. [PubMed: 9398979]
12. Hirsch LJ, Brenner RP, Drislane FW, 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; 22:128–135. [PubMed: 15805813]
13. Gerber PA, Chapman KE, Chung SS, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. J Clin Neurophysiol. 2008; 25:241–249. [PubMed: 18791475]
14. Mani R, Arif H, Hirsch LJ, et al. Interrater reliability of ICU EEG research terminology. J Clin Neurophysiol. 2012; 29:203–212. [PubMed: 22659712]
15. Hirsch LJ, Laroche SM, Gaspard N, et al. American Clinical Neurophysiology Society's Standardized Critical Care EEG Terminology: 2012 version. J Clin Neurophysiol. 2013; 30:1–27. [PubMed: 23377439]
16. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008; 61:29–48. [PubMed: 18482474]
17. Gwet, KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. Gaithersburg, MD: Advanced Analytics, LLC; 2010.
18. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol. 1990; 43:543–549. [PubMed: 2348207]
19. Tukey J. Bias and confidence in not-quite large samples. Ann Math Stat. 1958; 29:614.
20. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ. 1997; 62:77–89.

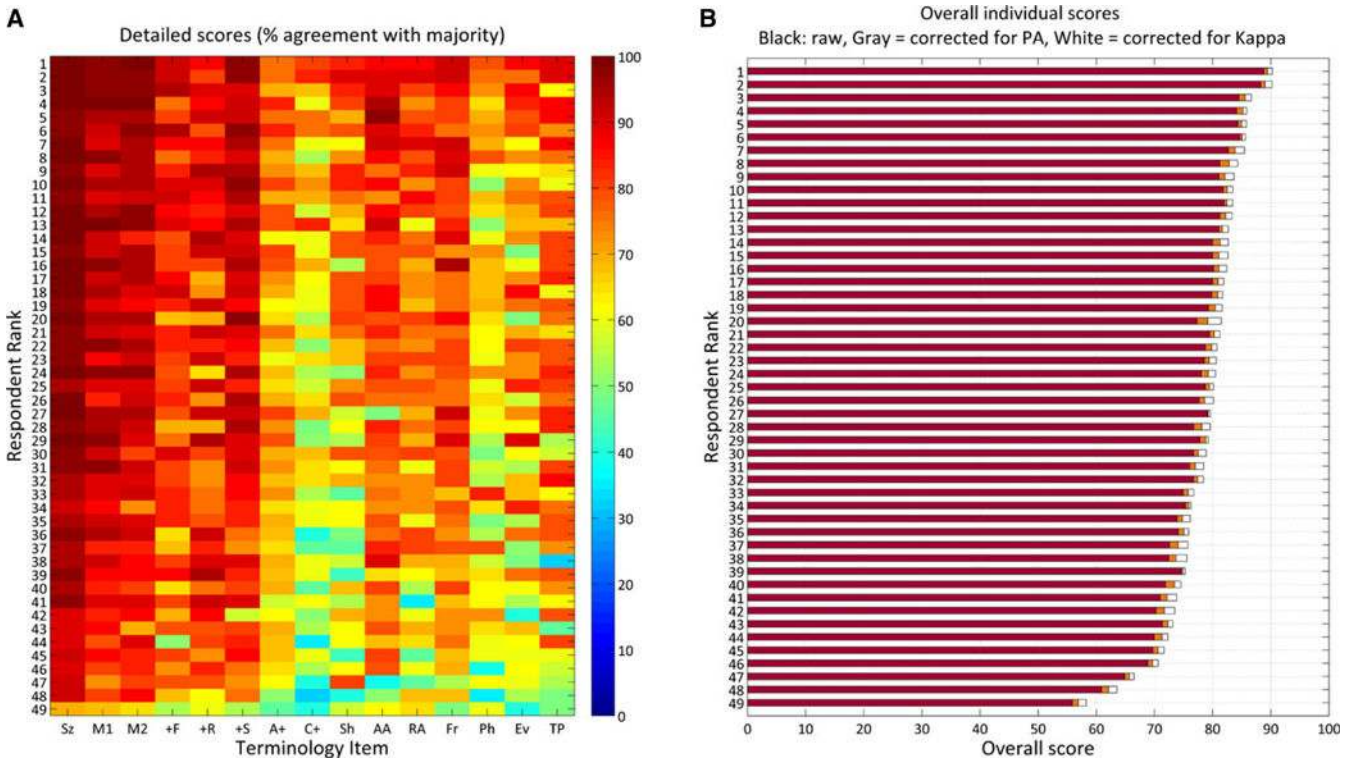


Figure 1.

(**A**) Detailed individual scores. The heatmap shows scores, coded by color according to the colorbar on the right, for all 49 individuals and for each of 15 different ACNS Critical Care EEG terminology concepts. Scores for each concept are calculated as % of answers in agreement with answers determined by consensus of an expert panel. Respondents are rank-ordered according to overall score. (**B**). Overall individual scores. Scores from (**A**) are summarized as a single score obtained in three different ways: By averaging the scores over all 15 concepts, giving equal weight to each (red bars); by averaging with terms weighted by the percentage agreement (orange bars); and by averaging with terms weighted by the IRA value (κ) for each concept (white bars). Sz, seizure; M1, main term 1; M2, main term 2; +F, superimposed fast activity; +R, superimposed rhythmic delta activity; +S, superimposed sharp waves or spikes, or sharply contoured activity; A+, any “plus” modifier (and of +F, +R, or +S vs. No+); C+, specific combination of “plus” modifiers (F, R, S, FR, FS, or No +); Sh, sharpness; AA, absolute amplitude; RA, relative amplitude; Fr, frequency; Ph, number of phases; Ev, evolution; TP, triphasic morphology.

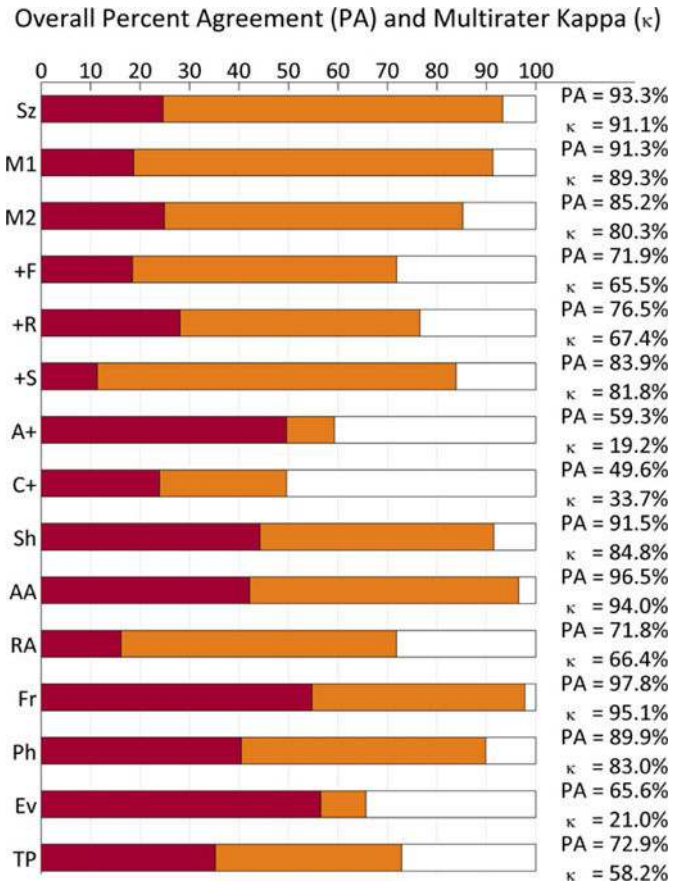


Figure 2.

Kappa (κ) values in relation to percent agreement. Horizontal bars show the percent agreement (PA, red bars + orange bars), relative to the maximal possible percent agreement (100%, ends of white bars). The lengths of red bars show the percent agreement estimated to be attributable to chance, PC, used in estimating the degree of IRA, κ. The total percentage of possible agreement beyond each red bar, 100-PC (orange bar + white bar) is considered the potential degree of agreement achievable "beyond chance," whereas the beyond-chance agreement actually achieved is PA–PC (the length of the orange bars). Mathematically, the chance-corrected IRA, κ, is the percentage of this possible beyond-chance agreement that is actually achieved, that is, $\kappa = \frac{PA - PC}{100 - PC}$. Graphically, this mathematical definition of κ is represented as the fraction of the distance between 100% and the end of the red bar that is taken up by the orange bar. Abbreviations are the same as in Figure 1.

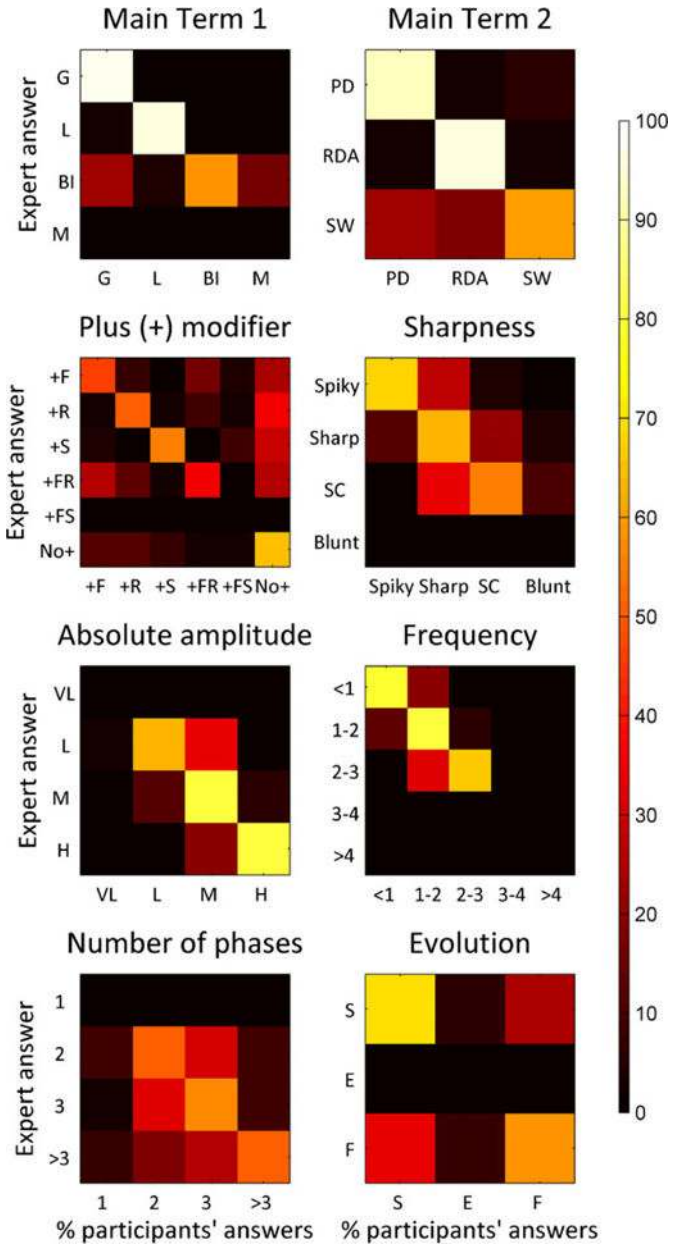


Figure 3.

Confusion matrices. Color-coded confusion matrices for 10 ACNS standardized critical care EEG terminology items are displayed as heat map (scale bar on the right). The available choices for each concept are shown along the vertical and horizontal axes. Heat map intensities indicate the percentage of respondents who chose each option along the horizontal axis when the correct response (determined by panel-of-experts consensus) was the one on the vertical axis. The percentage of respondents choosing each available option is displayed on a color scale along the horizontal (column) axes. A perfect result would produce a diagonal white line from upper left to lower right, with all squares not on or adjacent to that line being black. The majority of respondents selected the correct term on average in all cases shown. It is notable that the minority of incorrect responses was

generally “close” to the correct response (near the diagonal line). Abbreviations are as in Figure 1.

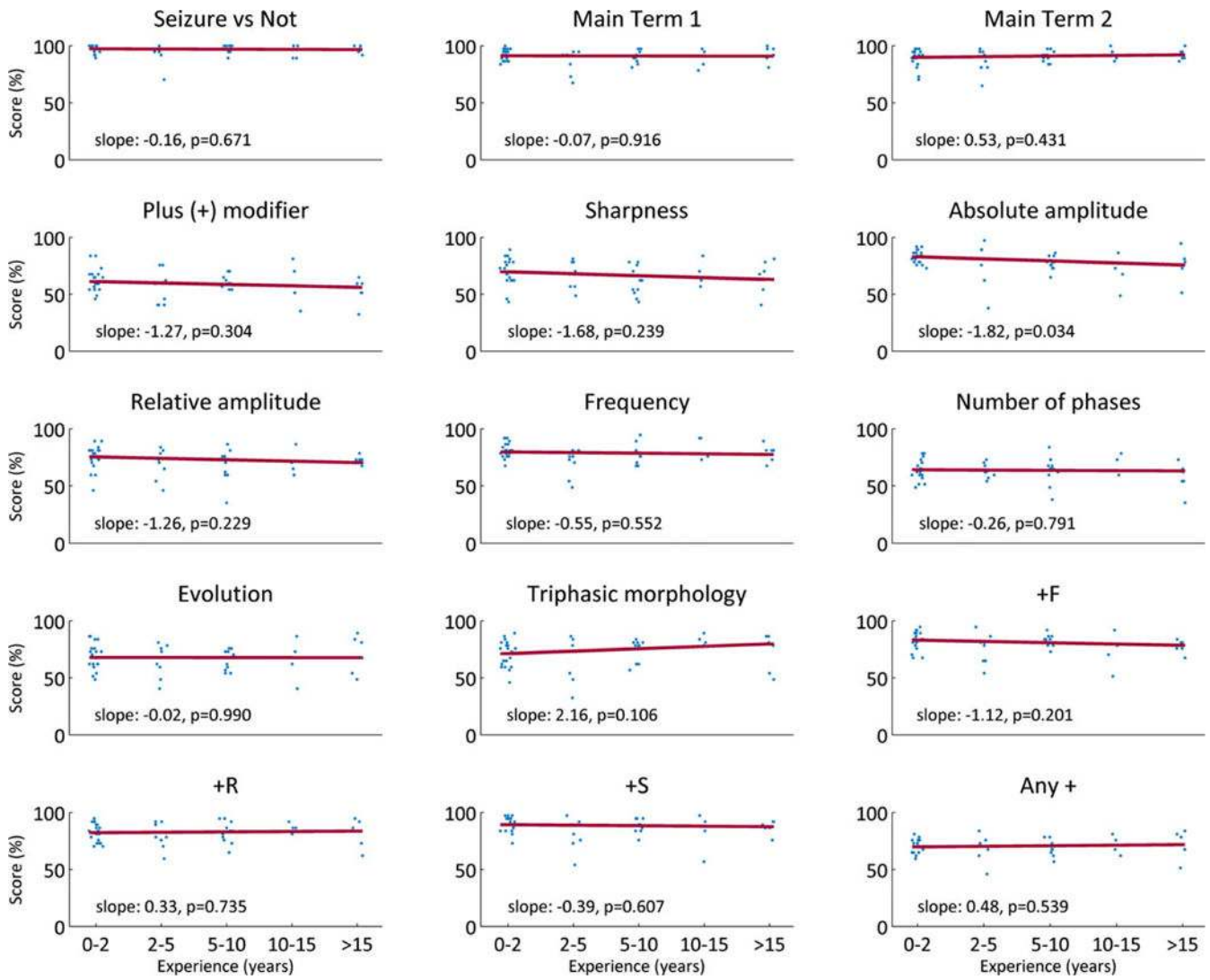


Figure 4.

Effects of EEG experience on individual performances. Scores (% correct, relative to expert panel consensus answers) were plotted for all raters within five groups of self-reported EEG reading experience: 0–2, 2–5, 5–10, 10–15, and >15 years. Linear regression lines (red dashed lines, of the form y = slope x + b), were fitted to each set data, with x values equal to 1, 2, 3, 4, or 5 for the different levels of experience. The slopes and associated p-values are shown on each graph.

Table 1

Percent agreement and kappa (κ) values for tested terms and comparison to prior studies

				Gerber et al. (2008)[13]	Mani et al. (2012)[14]	
Terminology item	**Choices**	**Agreement (%)**	**Kappa (κ), % (95% CI)**	**Kappa, κ (%)**	**Agreement (%)**	**Kappa, κ (%)**
Seizure	Yes, no	93.3	91.1 (90.6–91.6)	NA	NA	NA
Main term 1	G, L, BI, M	91.3	89.3 (89.1–89.6)	−/49	96	87
Main term 2	PD, RDA, SW	85.2	80.3 (79.4–81.2)	−39	98	92
Plus (+) modifier	F, R, S, FR, FS, No +	49.6	33.7 (32.4–35.1)	−/12	NA	NA
Any “+”	Yes, no	59.3	19.2 (17.5–20.9)	NA	82	25
+Fast activity (F)	Yes, no, not applicable	71.9	65.5 (64.4–66.7)	NA	83	54
+Rhythmic activity (R)	Yes, no, not applicable	76.5	67.4 (66.5–68.3)	NA	88	62
+Spike or sharply contoured (S)	Yes, no, not applicable	83.9	81.8 (81.2–82.5)	NA	82	16
Sharpness	Spiky, sharp, sharply contoured, blunt, not applicable	91.5	84.8 (84.3–85.2)	NA	NA	NA
Absolute amplitude	Very low, low, medium, high	96.5	94.0 (93.8–94.2)	NA	93 (jointly assessed)	NA
Relative amplitude	<2, >2, not applicable	71.8	66.4 (65.3–67.4)	NA		
Frequency	<1, 1–2, 2–3, 3–4, >4 Hz	97.8	95.1 (94.9–95.2)	−/34	80	NA
Phases	1, 2, 3, >3, not applicable	89.9	83.0 (82.6–83.4)	NA	NA	NA
Evolution	Static, evolving, fluctuating	65.6	21.0 (19.7–22.2)	NA	NA	NA
Triphasic morphology	Yes, no, not applicable	72.9	58.2 (56.1–60.2)	NA	NA	NA