Clinical Neurophysiology 126 (2015) 1661–1669

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings

J.J. Halford [a,*], D. Shiau [b], J.A. Desrochers [b], B.J. Kolls [c], B.C. Dean [d], C.G. Waters [d], N.J. Azar [e], K.F. Haas [e], E. Kutluay [a], G.U. Martz [a], S.R. Sinha [c], R.T. Kern [b], K.M. Kelly [f], J.C. Sackellares [g], S.M. LaRoche [h]

[a] Department of Neurology, Medical University of South Carolina, Charleston, SC, USA
[b] Optima Neurosciences Inc., Alachua, FL, USA
[c] Department of Neurology, Duke University Medical Center, Durham, NC, USA
[d] School of Computing, Clemson University, Clemson, SC, USA
[e] Department of Neurology, Vanderbilt University, Nashville, TN, USA
[f] Center for Neuroscience Research, Allegheny Singer Research Institute, Allegheny General Hospital, Pittsburgh, PA, USA
[g] Department of Neurology, Malcolm Randal VA Medical Center, Gainesville, FL, USA
[h] Department of Neurology, Emory University Hospital, Atlanta, GA, USA

ARTICLE INFO

Article history:
Accepted 7 November 2014
Available online 20 November 2014

Keywords:
EEG
Continuous monitoring
Inter-rater agreement
Seizure detection
Periodic discharge
ICU EEG monitoring

HIGHLIGHTS

- Agreement was moderate among eight experts for labeling the location of seizures in 1 h epochs of continuous ICU EEG monitoring recordings.
- Inter-rater agreement for labeling periodic discharges was considerably lower than for labeling seizures.
- Agreement among experts was improved by the use of EEG education modules.

ABSTRACT

Objective: This study investigated inter-rater agreement (IRA) among EEG experts for the identification of electrographic seizures and periodic discharges (PDs) in continuous ICU EEG recordings.

Methods: Eight board-certified EEG experts independently identified seizures and PDs in thirty 1-h EEG segments which were selected from ICU EEG recordings collected from three medical centers. IRA was compared between seizure and PD identifications, as well as among rater groups that have passed an ICU EEG Certification Test, developed by the Critical Care EEG Monitoring Research Consortium (CCEMRC).

Results: Both kappa and event-based IRA statistics showed higher mean values in identification of seizures compared to PDs ($k = 0.58$ vs. 0.38; $p < 0.001$). The group of rater pairs who had both passed the ICU EEG Certification Test had a significantly higher mean IRA in comparison to rater pairs in which neither had passed the test.

Conclusions: IRA among experts is significantly higher for identification of electrographic seizures compared to PDs. Additional instruction, such as the training module and certification test developed by the CCEMRC, could enhance this IRA.

Significance: This study demonstrates more disagreement in the labeling of PDs in comparison to seizures. This may be improved by education about standard EEG nomenclature.

1. Introduction

Evidence from several studies suggests that electrographic seizures occur frequently in critically ill patients due to a variety of insults to the brain (Privitera et al., 1994; Jordan, 1995; DeLorenzo et al., 1998; Vespa et al., 1999, 2003; Towne et al., 2000; Pandian et al., 2004; Claassen et al., 2004, 2007; Jette et al., 2006; Kilbride

* Corresponding author at: 96 Jonathan Lucas St., Suite 301 CSB, MSC606, Charleston, SC 29425, USA. Tel.: +1 843 792 3223; fax: +1 843 792 8626.
E-mail address: halfordj@musc.edu (J.J. Halford).

http://dx.doi.org/10.1016/j.clinph.2014.11.008
1388-2457/Published by Elsevier Ireland Ltd. on behalf of International Federation of Clinical Neurophysiology.

et al., 2009; Oddo et al., 2009). Approximately 90% of these seizures are clinically unrecognized non-convulsive which can only be reliably diagnosed by continuous EEG (cEEG) monitoring (Hirsch, 2010). For many ICU patients, non-convulsive seizures are potentially harmful if the diagnosis and treatment are delayed (Jordan, 1993, 1999a; Waterhouse et al., 1998; Hirsch, 2004a,b; Hirsch and Kull, 2004; Kull and Emerson, 2005; Kaplan, 2006; Jirsch and Hirsch, 2007; Hyllienmark and Amark, 2007; Oddo et al., 2009; Friedman et al., 2009). Therefore, cEEG monitoring has become standard practice in many ICUs and is rapidly spreading in use. Rapid recognition of nonconvulsive seizures and other abnormal EEG patterns such as periodic discharges can have significant impact on the decision making including ordering neuro-imaging, modification of antiepileptic drug regimen and optimization of cerebral perfusion (Jordan, 1999b; Claassen et al., 2000).

With the lack of reliable automated detection, a major constraint in recognizing nonconvulsive seizures is that there is complete reliance on visual analysis of the raw EEG recordings by clinical neurophysiologists. Furthermore, like interictal epileptiform discharges, although there are published criteria of EEG signal characteristics for recognizing electrographic seizures and PDs, it is not rare that inter-rater agreement (IRA) among EEG experts can be very poor, especially for cases with equivocal patterns or with more complex and abnormal background activities (Ronner et al., 2009). Since IRA reliability in recognizing these critical EEG events has significant implications for the value of EEG as a diagnostic tool, it is important to design studies that are similar to the clinical practice to assess IRA.

In spite of its importance, there has only been one published study that quantitatively assessed IRA in recognizing electrographic seizures in critically ill patients (Ronner et al., 2009), and two which assessed IRA in recognizing PDs (Gerber et al., 2008; Mani et al., 2012). Ronner et al. sampled discontinuous EEG epochs (three 10 s epochs for each of the 30 EEG recordings), put them into screenshots in PowerPoint slides, and the experts were asked to determine whether a seizure was present. Although this study was able to provide an assessment of inter-rater variability, the EEG review the experts performed in this study is very different from typical daily clinical practice because of three reasons. First, the experts were provided with short duration EEG samples, which limited an adequate informed review since determination of an electrographic seizure often requires analysis of both the ictal pattern and the background activity preceding it. Secondly, this study did not allow the experts to adjust typical EEG visualization settings (such as montage, signal sensitivity, filters, etc.) during the EEG review. Third, there was only a measurement of the presence or absence of patterns but not their durations (i.e., onset and offset), which is suboptimal since the length of an EEG pattern, especially seizures, is clinically important. In the study by Gerber et al., a conventional digital EEG review station was used to categorize the EEG epochs with standardized terminology. As in Ronner et al.'s study, most EEG epochs were 10 s in duration, selected from 11 ICU patients who all had the diagnosis of subarachnoid hemorrhage. In Mani et al.'s study, a large number of clinical neurophysiologists (16 experts) labeled a selection of EEG epochs of periodic and rhythmic EEG activity based on the American Clinical Neurophysiology Society (ACNS) nomenclature. A high IRA was found for labeling the location and pattern type but other diagnostic criteria for the presence of fast activity and sharp/spike wave activity showed a low IRA. Similarly, only a small number (<15) of EEG samples with short duration (10 s) were used for this study and the reviews were based on screen-captured images that the experts could not adjust visualization settings. As in Ronner et al.'s study, the experts in these two studies labeled only the presence or absence of patterns but not their durations (i.e., onset and offset).

The purpose of the present study was to statistically evaluate IRA among EEG experts in identifying electrographic seizure and PDs from continuous EEG recordings collected from prospective ICU patients using a review system which replicates a typical digital EEG review station. Specifically, we aimed to investigate (1) the degree of agreement in identifying seizures versus PDs, (2) the agreement with respect to the duration of seizure and PD events, and (3) the effect of studying and passing the ACNS CCEMRC ICU EEG quiz on expert agreement levels for identifying and marking the duration of these events.

2. Method

2.1. Subject population and test EEG dataset

EEG recordings used in this study were collected from critically ill patients 18 years of age or older who were admitted to the Medical University of South Carolina (Charleston, SC), Emory University Hospital (Atlanta, GA), or Duke University Medical Center (Durham, NC). Collection of EEG data was approved by each institution's Investigational Review Board, as well as the Western Investigational Review Board (WIRB). The only criterion for including a subject's EEG recording in the study is that there should be at least one seizure noted in the daily clinical EEG report. Therefore, the ICU patients included in the study had a variety of etiologies, background EEGs, and ictal EEG patterns. Based on the seizure occurrences described in the clinical reports, a total of 30 1-h EEG segments were randomly sampled from 20 subjects' long-term EEG recordings that contained seizures. Although selected patients had at least one seizure documented in their report, the randomly sampled 1 h EEG segments did not necessarily include seizures. All recordings were made using the International 10–20 recording system recorded on XLTEK or Nihon-Kohden equipment and were acquired at a sampling rate of 200 or 256 Hz.

2.2. Expert raters

Eight board-certified academic EEG experts, who review and interpret ICU EEGs for their clinical duties, were recruited to participate in the study. At the time of the study, four experts had taken and passed the ICU EEG Certification Test developed by CCEMRC and based on the 2012 version of the ACNS Standardized Critical Care EEG Terminology (Hirsch et al., 2013), which mainly focused on periodic discharges and rhythmic delta activities. The task was to independently mark the onset and offset of electrographic seizures and PDs in each of the 1-h EEG segments. Experts were not asked to differentiate between different locations of PDs (i.e., generalized vs. lateralized) but were asked to mark PDs of any location or spatial distribution. Three expert raters were involved with the collection of some test EEG segments and therefore were not asked to review those EEG segments. As a result, five experts reviewed and marked events on all 30 EEG segments, whereas the other three experts only reviewed and marked events on 20 EEG segments. Therefore, each test EEG segment was independently reviewed and marked by seven EEG experts.

2.3. Review process

A web-based EEG review and scoring system, *EEGnet* (Halford et al., 2011, 2013), was used by the experts to perform the tasks. All 30 sampled EEG segments were uploaded onto the *EEGnet* server for access. A password-protected user account was created for each of the experts and the files were assigned to each individual rater by the study administrator. Once a rater logged into the account, he/she would see a list of the test EEG segments with a

"Progress" column, which indicated whether the review of an individual segment had been completed.

After the expert selected an EEG file to review, the first 10 s of the EEG displayed on the screen, and the expert was able to change the EEG display settings, including sensitivity, channel montage, low-pass, high-pass, and notch filter parameters, as necessary at any time during the review. Once the rater identified a seizure or a PD, he/she marked a line (red for seizures and blue for PDs) at the onset time of the event and another line for the offset time, which completed the marking of an event with a gray area between two vertical lines. An example of event marking in EEGnet is demonstrated in Fig. 1.

2.4. Inter-rater agreement analysis

2.4.1. Kappa statistic

After collecting all of the marking results from all of the expert raters, Cohen's κ (kappa) statistic was calculated for each pair of raters in order to better observe the distribution of IRA. It is calculated as:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where Pr(a) is the relative observed agreement between raters, and Pr(e) is the hypothetical probability of chance agreement, using the observed responses to calculate the probabilities of each observer randomly assigned to each category. Kappa has been described as the ideal statistic to quantify agreement for dichotomous variables. Magnitude guidelines in the literature suggested that: values <0 as indicating no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement (Landis and Koch, 1977).

Implicit in the kappa is the assumption that the rated items, subjects, or targets are independent. However, identification of "transient events" during a serially observed process such as seizures in EEG data contains responses that are highly correlated among the neighboring responses, which violates the independence assumption of kappa. Therefore, in this study, we applied a Monte-Carlo-based permutation technique to produce an empirical distribution of kappa in the presence of dependence (Norman and Scott, 2007). The main purpose of this technique is to calculate expected agreement due to chance (i.e., Pr(e)) between two raters. To achieve this, we first generated two sequences (one for seizure events and the other for PDs) comprised of binary responses from each rater's markings. Each binary response represents the marking in each second – i.e., 1 if the second is within an event marking and 0, otherwise. Secondly, for each binary sequence, 10,000 random permutations of runs of 1 s and 0 s were sampled, and the pairs of permuted sequences were cross-tabulated to create an agreement table. Repetition of this permutation process provided a sample from all possible random agreements of all possible pairs of sequences. The R statistics and development system was used to perform the simulations.

2.4.2. Positive event agreement and event duration agreement

The kappa statistic gives an overall assessment of the expert agreement; however, its calculation is based on the binary response for each of the arbitrarily pre-determined time epochs and thus may lose some information regarding the characteristics of the identified events. Therefore, in this study, we calculated an additional IRA statistic that combines two event-related agreement statistics: (1) the fraction of event agreement (FEA), and (2) the fraction of event duration agreement (FEDA). One may interpret FEA as the "inter-rater sensitivity", i.e., if an event is marked by Rater A, how likely it will also be marked by Rater B, and vice versa. The FEA is calculated between a pair of raters as the number of events agreed (with a minimum overlap of 1 s) by both raters divided by the sum of agreed and disagreed events. It is worth noting that, if Rater B marked two events that both overlapped with the same event marked by Rater A, only one "agreed" event was included in the FEA calculation. Fig. 2 gives a schematic example of the FEA calculation. In this example, Rater A marked a total of

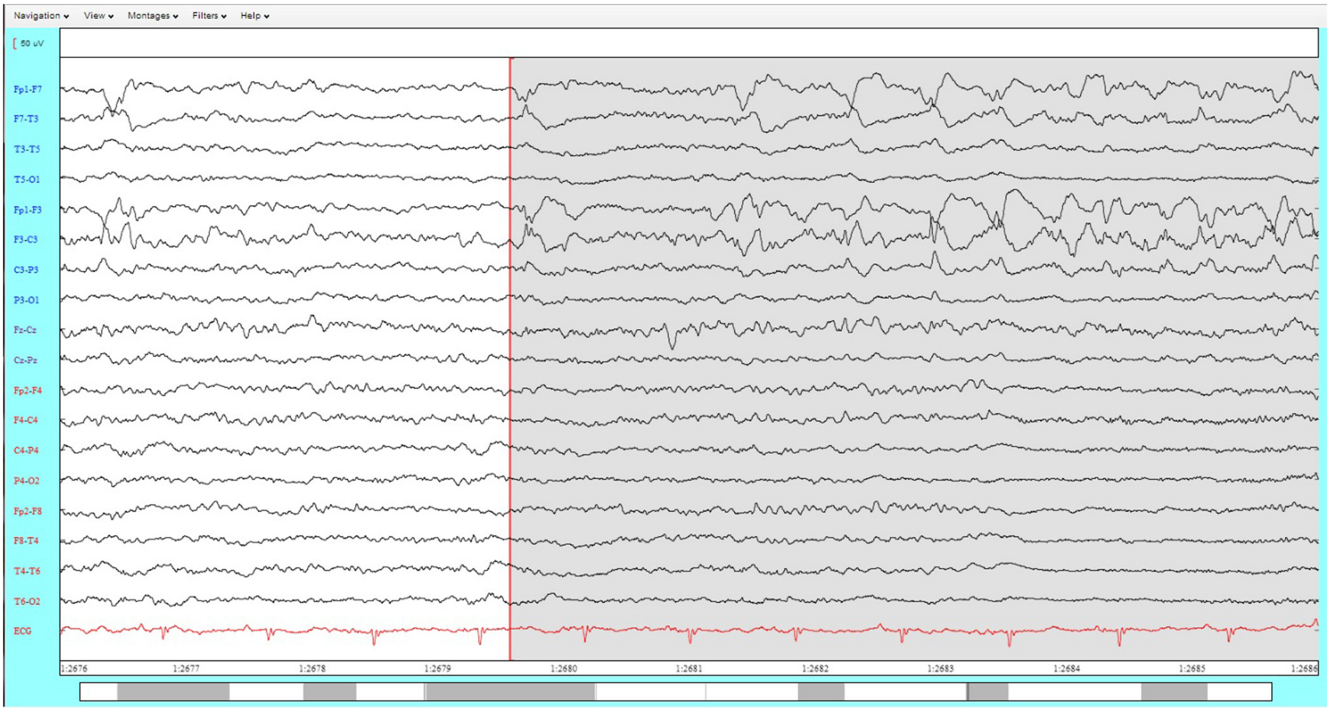

Fig. 1. Marking of a seizure event – this 10-s EEG page contains the first 7 s of a seizure (onset marked by a red vertical line).

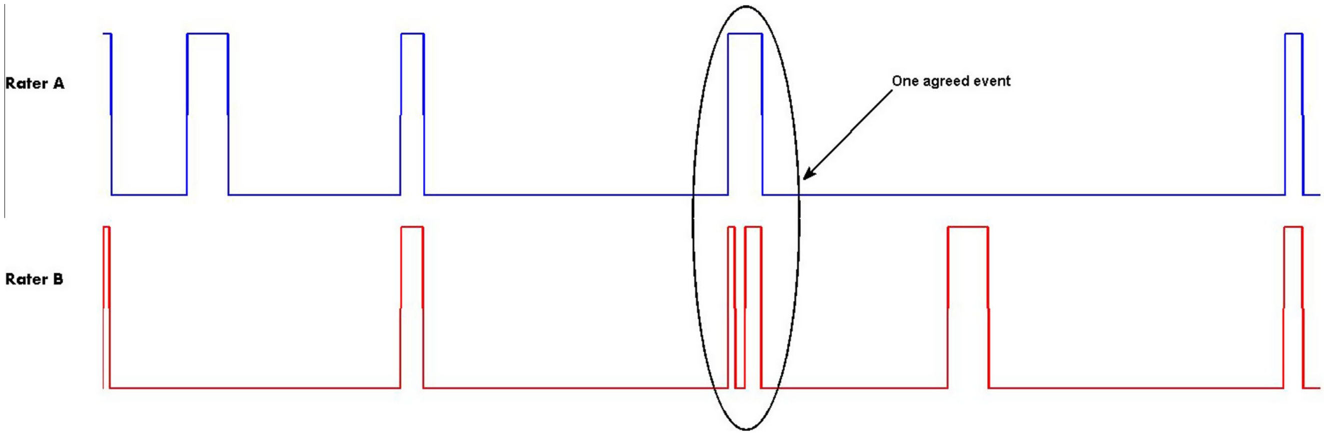

Fig. 2. A schematic example of FEA calculation.

5 events, whereas Rater B marked 6 events, among which two overlapped the fourth event marked by Rater A. Hence, the total number of the agreed events by Raters A and B is 4 and the number of disagreed events is 2, and therefore, the FEA between Raters A and B is 0.67.

In order to take into account the agreement on event durations, FEDA is calculated between two raters as the mean ratio of overlapped duration of an agreed event to the total duration of that event. Mathematically speaking, suppose that an event is marked by Rater A with a duration of X seconds and a (similar) event is marked by Rater B with duration of Y seconds, and the two marked events have an overlapped interval of Z seconds, then the FEA was calculated as $Z/(X + Y - Z)$. For cases that one marked event is overlapped with multiple events marked by the other rater, the calculation is extended to $(Z_1 + Z_2 + \cdots)/(X + Y_1 - Z_1 + Y_2 - Z_2 + \cdots)$. The FEDA gives a quantitation of how specific the agreement between two raters is with respect to timing and duration of the marked events.

Since both FEA and FEDA have a range between 0 and 1 and a perfect event agreement between the two raters will have a value of 1 for both FEA and FEDA, an additional IRA statistic can be calculated as (FEA + FEDA)/2, which also has a range between 0 and 1. This combined IRA statistic quantifies the agreement between two raters for both the occurrence and duration of each marked event.

2.5. Statistical inference

With the observed IRA statistics from rater pairs, we examined the distribution of IRA values and attempted to address the following questions: (1) What is the shape of the distributions of the IRAs? Are there any outliers? (2) Is there a statistically significant difference between the IRAs for marking seizures and PDs? (3) Is there a difference in the IRA of groups of pairs raters based on whether neither, only one, or both had passed the CCEMRC Certification Test? (4) Is the event-based IRA statistic we developed correlated statistically with the commonly used kappa statistic?

To answer the questions listed above: (1) The Kolmogorov–Smirnov test was utilized to test the normality of the distribution as well as for the comparisons between distributions and boxplots were used for checking outliers; (2) Landis and Koch's guidelines for kappa's magnitude and the Wilcoxon sign-rank test (nonparametric pair-*t*) were applied to measure the IRAs for marking seizures and PDs; (3) the Kruskal–Wallis rank sum test (nonparametric one-way ANOVA) was utilized to test if there was a difference between raters based on CCEMRS certification status; and (4) Pearson's correlation coefficients were calculated and compared with our event-based IRA statistic. The evidence from a test was considered significant only if the resulting *p*-value was less than 0.05.

3. Results

There was an average of 153.1 seizures (range 81–283) and 139.4 epochs of PDs (range 34–268) annotated by the 8 reviewers in the 30 EEG sample recordings, which is an average of 5.1 seizures and 4.6 PDs epochs per hour of recording. The total average percentage of the EEG epochs marked as including seizure activity was 81% (range 57–100%), and the average percentage of the recording marked as including PDs was 65% (range 27–90%).

Fig. 3 illustrates how the scoring data from one EEG epoch is used to calculate the IRA statistics. Event markings from 7 expert raters for EEG segment 21 (Rater #7 was involved in the collection of this EEG segment and therefore excluded) is shown. Note that, on the timing and duration of seizures, two groups of reviewers had a high level of agreement within the group (group #1 includes reviewers 1, 2, and 5 and group #2 includes reviewers 4 and 6); Reviewer 8 had moderate agreement with all other reviewers, and Reviewer 3 had no agreement with any other reviewers because this reviewer did not mark any seizures. There seemed to be little agreement on identification of PD epochs, except for a few regions marked by Reviewers 1, 4, 5, and 6.

Fig. 4 shows two IRA statistics (kappa and event-based) for each of the rater agreements for seizure and PD identification, for EEG segment 21 (as shown in Fig. 3). As observed from the visual analysis, for both statistics, the highest IRA values are in rater pairs composed of Reviewers 1, 2, 4, 5, and 6, followed by rater pairs including Reviewer 8, and no agreement between Reviewer 3 and the other reviewers (IRA = 0). For identification of PDs, IRA statistics suggest globally poor agreement for this EEG sample. Reviewer 5 and 6 had an agreed PD event with significant overlap in duration, and thus this rater pair had high event-based IRA statistic for PD marking as well as the highest kappa statistic.

It is worth noting that the kappa statistic suggested a higher "agreement on the location of seizures between Reviewers 1 and 6 than between Reviewers 4 and 6, whereas the event-based IRA statistic suggested the opposite, though both statistics showed negligible difference between the two rater pairs. One explanation is that, although Reviewers 1 and 6 had a different opinion in one event (event #3 in Reviewer 6's markings), it is relatively short, and the two Reviewers had better agreement (61.4%) on the durations for the events that they agreed on, compared to that between Reviewers 4 and 6 (54.2%). Therefore, they had a higher kappa

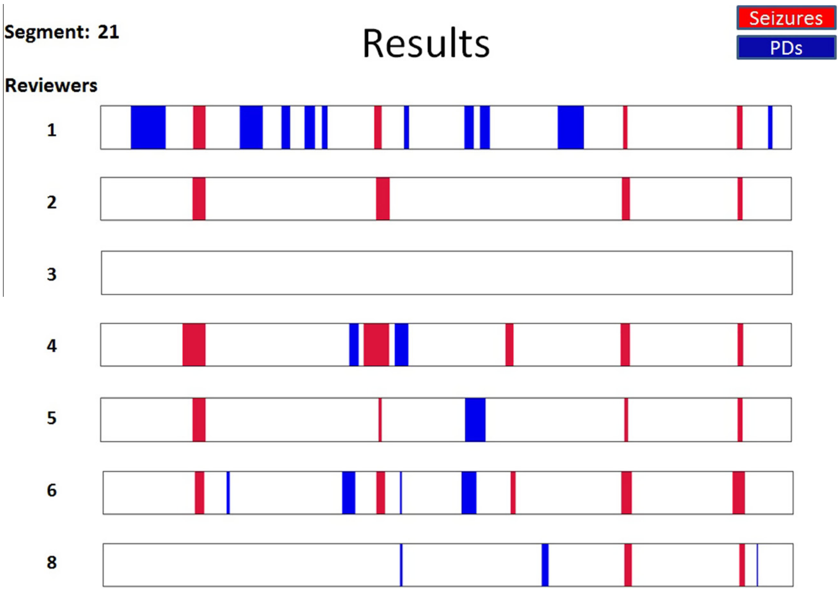

Fig. 3. Seizure and PD event markings by 7 expert raters in a 1-h EEG segment.

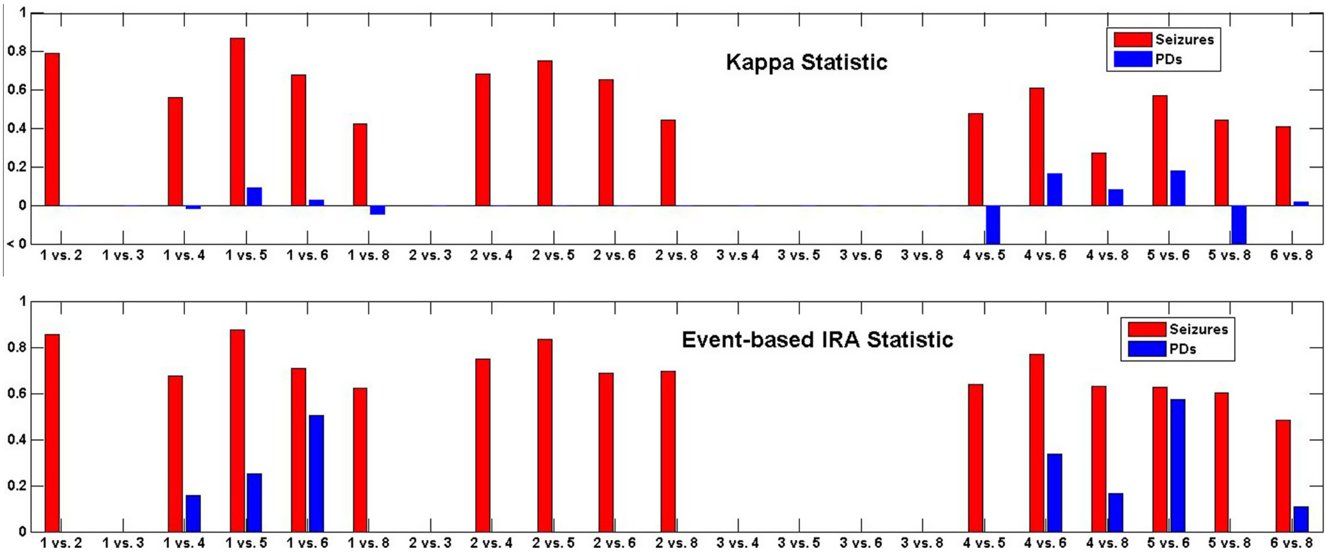

Fig. 4. Inter-rater agreement statistics derived from the event markings in Fig. 3.

value. Conversely, although Reviewers 4 and 6 had lower agreement on durations, their "inter-rater sensitivity" was 100% and thus had the higher value on the event-based IRA statistic.

For each of the rater pairs, the IRA statistics were averaged over all EEG files that were reviewed by both reviewers and in which at least one event was marked by at least one of the reviewers. If no events were marked by either reviewer, the IRA statistics on that EEG file could not be determined. Fig. 5 shows the distributions of IRA statistics (kappa and event-based) over all rater pairs.

According to one-sample Kolmogorov–Smirnov test (for normality), only the distribution of the pair-wise kappa statistic for PDs does not follow a normal distribution (rejected with $p = 0.03$). p-Values for kappa-for-seizures, event-based-IRA-for-seizures, and event-based-IRA-for-PDs are 0.453, 0.549, and 0.794, respectively, indicating that these measures are distributed normally. Based on the boxplot, there are outliers (horizontal lines outside the upper and lower whiskers) in IRAs for seizures (both kappa and event-based). For both seizure and PD identification, the standard deviations are smaller with event-based statistics. With a two-sample Kolmogorov–Smirnov test (for equality of two distributions), in both kappa and event-based statistics, the distribution of IRA statistics for identification of seizures is different from identification of PDs ($p < 0.001$). Wilcoxon signed-rank tests further confirm that the mean IRA for identification of seizures is significantly greater than that for PDs ($p < 0.001$ for both IRA statistics).

Based on Landis and Koch's guidelines, a mean kappa value of 0.58 for identification of seizures suggests an overall moderate agreement among raters. Among 28 rater pairs, 2 had fair agreement, 16 had moderate agreement, 8 had substantial agreement, and 2 had almost perfect agreement. For PD identification, a mean kappa value of 0.38 suggests an overall fair agreement among raters: 4 had no agreement, 4 had slight agreement, 7 had fair agreement, 12 had moderate agreement, and only 1 pair had substantial agreement. Fig. 6 shows the distributions in levels of agreement among rater pairs for recognition of seizures and PDs, respectively.

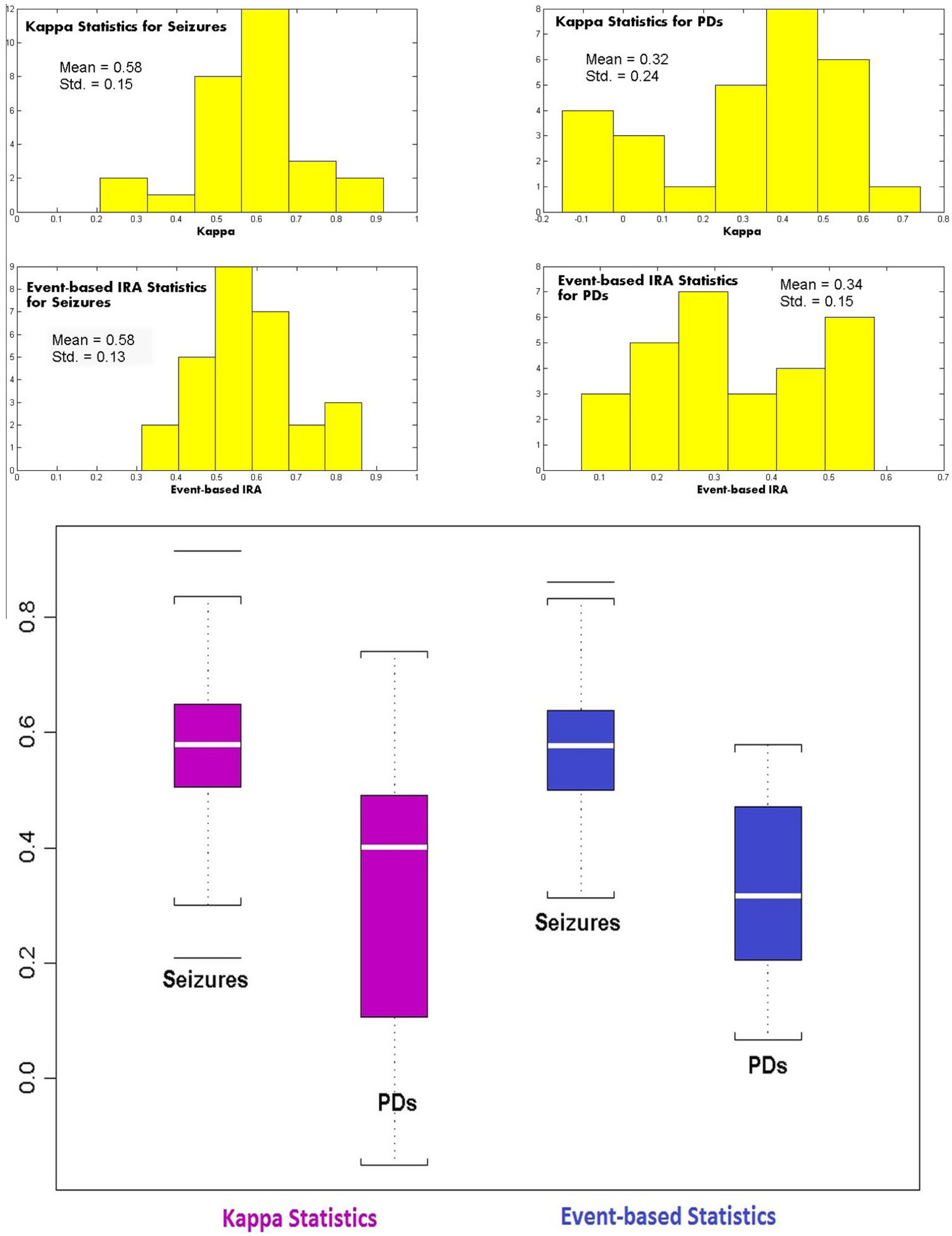

Fig. 5. Empirical distributions of IRA statistics over rater pairs. Top: Distributions of pair-wise kappa statistics and event-based IRA statistics for seizure markings and PD markings. Bottom: Boxplots (with outliers) generated from the four distributions shown in the top and middle panels. The white band inside the box is the median; top and bottom of the box are the third quartile (Q3) and first quartile (Q1), respectively; top and bottom whiskers are Q3 + 1.5(Q3 − Q1) and Q1 − 1.5(Q3 − Q1), respectively; the horizontal lines above or below the whiskers are outliers.

For both seizure and PD identification, mean kappa statistics are not significantly different from kappa with an event-based approach ($p = 0.35$ and 0.43 for seizures and PDs, respectively). Furthermore, Pearson's correlation tests suggest that the two IRA statistics are highly correlated: the correlation coefficient between the two IRA statistics in seizure markings is 0.98 ($p < 0.001$) and is 0.88 in PD markings ($p < 0.001$).

As described in Section 2, at the time of the study, four raters had taken and passed the ICU EEG Certification Test developed by CCEMRC, which mostly focused on the identification of PDs and rhythmic delta activities. To investigate if this test certification increased agreement in recognition of seizures and PDs, the rater pairs were categorized into three groups: both passed the test (G1), only one passed the test (G2), and neither passed the test (G3). The mean IRA statistics (kappa/event-based) for identifying seizures was: 0.66/0.66, 0.57/0.57, and 0.53/0.51 in G1, G2, and G3, respectively, and for identifying PDs was: 0.47/0.45, 0.32/0.34, and 0.15/0.23 in G1, G2, and G3, respectively. Using the bar plots, Fig. 7 shows the group comparisons in each of the four IRA analyses. It is clear that all four IRA statistics had the same trend: G1 > G2 > G3 and the differences among the groups identifying PDs were greater. A one-way ANOVA test suggests that at least one group is significantly different ($p = 0.04$) from the others in event-based IRA statistics for PD identification. Tukey's multiple comparison test further revealed that G3 rater pairs had significantly lower agreement than G1 rater pairs.

We also calculated the positive agreement percentage between raters as to whether either a seizure or PD is present in each

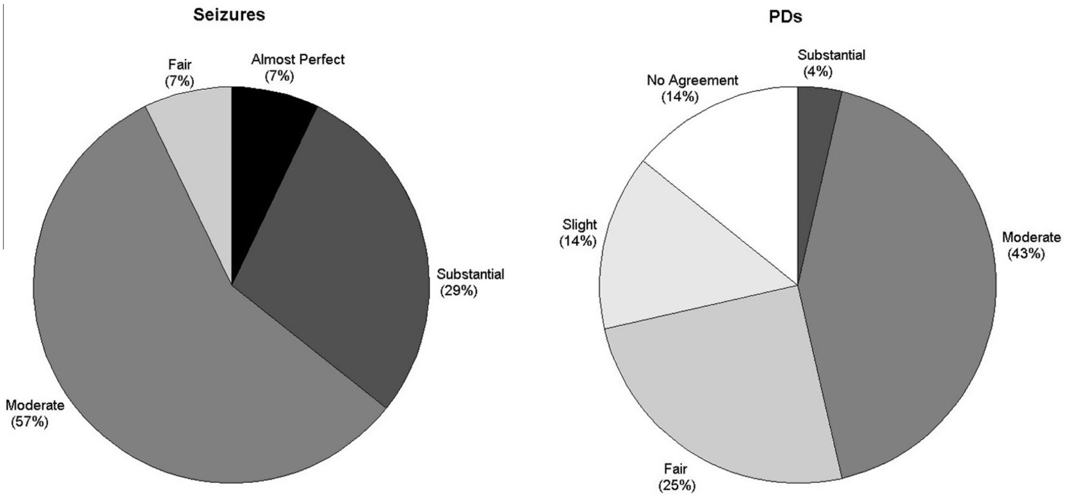

Fig. 6. Distribution of levels of agreement in marking seizures and PDs.

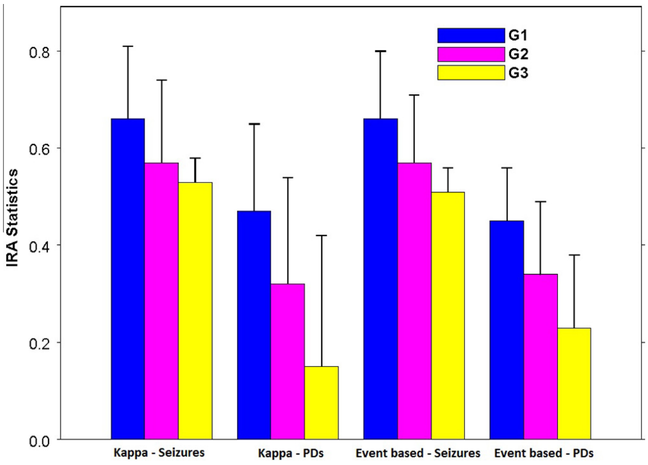

Fig. 7. Group comparisons in each of the four IRA analyses.

Table 1
Pairwise positive agreement percentage on the presence of **seizures** in each recording.

Rater	1	2	3	4	5	6	7	8
1		63%	50%	67%	73%	70%	60%	90%
2			50%	67%	73%	70%	55%	90%
3				53%	57%	45%	50%	67%
4					73%	60%	60%	100%
5						95%	67%	100%
6							50%	100%
7								100%
8								

Table 2
Pairwise positive agreement percentage on the presence of **PDs** in each recording.

Rater	1	2	3	4	5	6	7	8
1		43%	27%	70%	63%	70%	70%	60%
2			27%	47%	47%	35%	45%	60%
3				27%	27%	20%	25%	35%
4					70%	75%	80%	65%
5						65%	75%	65%
6							80%	60%
7								70%
8								

recording (Tables 1 and 2). This is important because clinical decisions may be made in many cases not based on the number of seizures/PDs present, but based on their presence or absence. In this study, the mean positive agreement percentage on the presence of seizures is 70% (std. = 18%), which is significantly greater ($p < 0.001$) than that for the presence of PDs (54%, std. = 19%). As expected, with both seizures and PDs, the recording-based mean positive agreement percentage is significantly larger than the individual event-based mean fraction of event agreement (FEA) (seizure: 0.70 vs. 0.57, $p = 0.004$; PD: 0.54 vs. 0.31, $p < 0.001$). In other words, raters had much more agreement in determining whether seizures/PDs are present in a recording than in determining the presence of a single event.

4. Discussion

Clinical decisions are often made based on the detection of seizures and PDs in continuous EEG monitoring studies of critically-ill patients. This study shows that there exists significant variability in the IRA for the detection of seizures and PDs in these types of recordings. IRA for the detection of seizures was 0.58, indicating moderate agreement. This was considerably higher than the IRA for the detection of PDs (which was 0.38), which is reassuring since clinical decisions are generally driven more by the presence of seizures than PDs. It is unclear why the IRA for marking PDs was lower than for seizures. One possibility is that since PDs were much more prevalent in the recording than seizures and since a significant amount of the EEG recordings contained PDs, the reviewers' concentration wavered intermittently, which might have resulted in accidental failure to label all of the PD epochs. Another possibility is that some of the PDs were of lower amplitude, leading some reviewers to ignore them since they may not have thought that these PDs were as clinically-relevant. Another possibility is that the criteria for determining a PD, even with the ACNS terminology, are still more subjective and/or difficult to apply than those for determining a seizure. Finally, at least one reviewer stated they manually selected representative examples of the PDs in the files, or occasionally brief runs of PDs but did not label whole sections of PD activity. This would produce the multiple to one comparison as shown in Fig. 2 and cause a reduced agreement in the PDs.

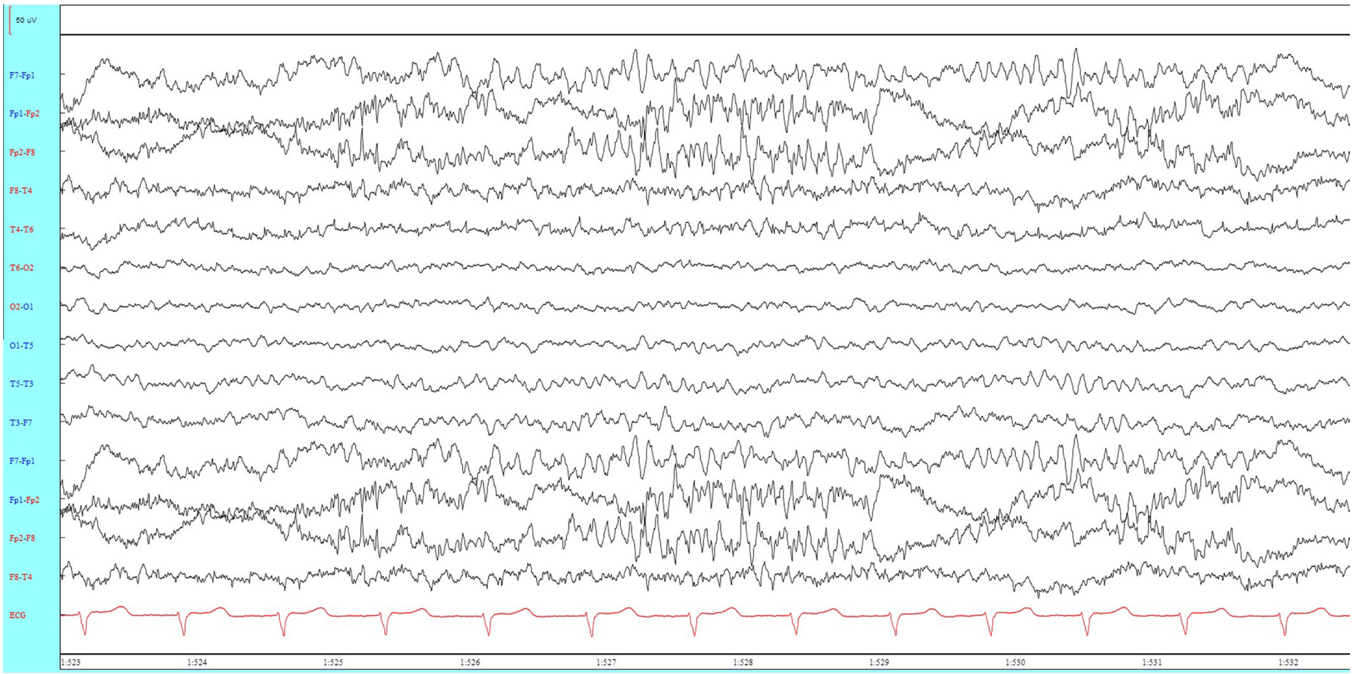

Fig. 8. An example of brief rhythmic discharges (BRDs).

This study also demonstrated the effectiveness of using a web-based platform (EEGnet) for EEG review and annotation, which enables the collection of a sizable amount of expert opinion without having to mail out EEG recordings, install software on clinical workstations, and collect annotation results on paper or in electronic documents from reviewers. All annotation actions of the reviewers were collected on a single server in real-time, which made data collection and tabulation easy. Hopefully EEGnet and similar web-based research platforms like it will enable many future studies to collect data on how experts label patterns in EEG recordings.

This study showed that the IRA for the identification of seizures, and even more so for PDs, was improved if reviewers had passed the ICU EEG Certification Test developed by CCEMRC. This may be because the reviewers who passed this test had studied the educational materials on the new proposed diagnostic criteria for labeling seizures and PDs (Hirsch et al., 2013), which led to a more uniform view of the morphology and application of the terminology used to describe these patterns. This result suggests that education programs that promote standardized criteria for interpretation can improve the agreement between readers in consistently applying terminology for events reported in cEEG data, such as PDs and seizures. The EEGnet system, which was used in this study to collect expert opinion, could also be used as a convenient web-based platform for educating neurologists and EEG technologists on how to label these patterns using a standardized terminology, since it allows users to manipulate typical EEG review station controls such as montage, filter, and sensitivity settings providing the opportunity to practically apply the proposed standards to real world EEG data.

In this study we created new metrics for measuring IRA between experts that incorporated the amount of temporal overlap between detections. These metrics showed similar results to those calculated with traditional kappa measures. For cases such as the current study in assessing agreement on marking transient events in continuous recordings, there are a few advantages in this new event-based IRA statistic: (1) it can be calculated directly from the markings and durations of the detections, and therefore avoids using an arbitrarily chosen epoch length in calculating the kappa statistic, (2) it can be combined with different weights on the agreement of the event occurrence and of duration, depending on the clinical questions, and (3) the interpretation is more straightforward for most clinical researchers.

This study provides a more accurate assessment than previous studies of IRA for the identification of seizures and PDs for four reasons. First, EEGnet web-based EEG review system accurately simulated a typical clinical review system. Two of three previous studies had used only screen-capture images of EEG (Ronner et al., 2009; Mani et al., 2012). Secondly, the samples of EEG viewed by experts were longer than in previous studies (which had used 30 s to 20 min samples), and therefore more reflective of daily clinical practice. Third, the time of onset and offset of seizures and PDs were labeled in this study, unlike previous studies which assessed the presence or absence of events, which allowed more accurate measurement of event identification. Fourth, the location of both seizures and PDs were identified, unlike previous studies which had focused on either one or the other (Ronner et al., 2009; Gerber et al., 2008; Mani et al., 2012).

There are significant limitations to this study. Only portions of prolonged EEG recordings were reviewed, so it is not clear what the IRA would be for the multi-day continuous EEG studies, which are common in clinical practice. Perhaps the IRA would be lower in a study that asked experts to review a 12–24 h EEG for each subject (more consistent with daily clinical practice) since EEG review would have to be performed more quickly, leading experts to accidentally overlook additional seizures and PDs. It is possible that the IRA found in this study could have been higher had we expended more effort toward educating reviewers on how to mark seizure and PD epochs by providing examples and testing each reviewer on brief epochs before the study began. We will consider doing this in the future. Several reviewers also commented that there was no category for marking brief rhythmic discharges (BRDs, an example shown in Fig. 8), which are part of the ACNS nomenclature. The absence of this diagnostic category could have led to some confusion about certain patterns in one EEG sample that lay on the borderline between seizures and PDs. Another limitation in this study is that there was no measure of intra-rater agreement, so it is unclear how much of the difference of agreement between reviewers was a true difference in opinion about the morphology of seizures or PDs could represent inconsistency of one or more raters in performing the task. Finally, there may have been a bias in the selection of the experts in this study since experts who had taken the CCEMRC training and test may have had more common knowledge about identification of seizures and PDS in ICU EEG recordings due to their involvement in previous CCERMC research projects.

Because continuous EEG is likely to be an important diagnostic measure of outcome for future treatment trials for non-convulsive seizures and status epilepticus, it is important to understand the degree of IRA for seizure detection and develop methods for improving it. Future studies of IRA for the detection of seizures and PDs could include an educational module and test designed in EEGnet that a subset of reviewers could take before IRA is measured again. Reviewers could be asked to perform the same task twice, separated by an interval of a month or more, to measure intra-rater agreement. Due to technical limitations with the EEGnet system, review of long EEG recordings would have been difficult in this study because experts were only able to view up to 20–50 s of EEG per second. Our new version of EEGnet, recoded to use the WebGL application programming interface, allows much quicker review of EEG data, which would allow a more realistic evaluation of IRA since experts could view EEG at a typical review speed in much longer EEG recordings. Recent studies have shown that trending tools, such as digital spectral array, can aid in the visual identification of seizures (Pensirikul et al., 2013; Williamson et al., 2014). Since some periodic patterns can be missed with trending, incorporation of trending tools into EEGnet could allow measurement of IRA for experts using visualization of trends. We hope this advancement will lead to more precise measurement of IRA for seizures and PDs and better training tools in the future.

Acknowledgment

This work was supported by 1R43NS064647 and 1R43NS076045-01A1 to D.S. from NIH-NINDS.

Conflict of interest: The authors have no conflict of interest to disclose.

References

Claassen J, Baeumer T, Hansen HC. Continuous EEG for monitoring on the neurological intensive care unit. New applications and uses for therapeutic decision making. Nervenarzt 2000;71:813–21.

Claassen J, Mayer SA, Kowalski RG, Emerson RG, Hirsch LJ. Detection of electrographic seizures with continuous EEG monitoring in critically ill patients. Neurology 2004;62:1743–8.

Claassen J, Jette N, Chum F, Green R, Schmidt M, Choi H, et al. Electrographic seizures and periodic discharges after intracerebral hemorrhage. Neurology 2007;69:1356–65.

DeLorenzo RJ, Waterhouse EJ, Towne AR, Boggs JG, Ko D, DeLorenzo GA, et al. Persistent nonconvulsive status epilepticus after the control of convulsive status epilepticus. Epilepsia 1998;39:833–40.

Friedman D, Claassen J, Hirsch LJ. Continuous electroencephalogram monitoring in the intensive care unit. Anesth Analg 2009;109:506–23.

Gerber PA, Chapman KE, Chung SS, Drees C, Maganti RK, Ng Y-T, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. J Clin Neurophysiol 2008;25:241–9.

Halford JJ, Pressly WB, Benbadis SR, Tatum 4th WO, Turner RP, Arain A, Pritchard PB, Edwards JC, Dean BC. Web-based collection of expert opinion on routine scalp EEG: software development and interrater reliability. J Clin Neurophysiol 2011;28:178–84.

Halford JJ, Schalkoff RJ, Zhou J, Benbadis SR, Tatum WO, Turner RP, et al. Standardized database development for EEG epileptiform transient detection: EEGnet scoring system and machine learning analysis. J Neurosci Methods 2013;212:308–16.

Hirsch LJ. Continuous EEG monitoring in the intensive care unit: an overview. J Clin Neurophysiol 2004a;21:332–40.

Hirsch LJ. Brain monitoring: the next frontier of ICU monitoring. J Clin Neurophysiol 2004b;21:305–6.

Hirsch LJ. Urgent continuous EEG (cEEG) monitoring leads to changes in treatment in half of cases. Epilepsy Curr 2010;10:82–5.

Hirsch LJ, Kull LL. Continuous EEG monitoring in intensive care unit. Am J Electroneurodiagnostic Technol 2004;44:137–58.

Hirsch LJ, LaRoche SM, Gaspard N, Gerard E, Svoronos A, Herman ST, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2012 version. J Clin Neurophysiol 2013;30:1–27.

Hyllienmark L, Amark P. Continuous EEG monitoring in a paediatric intensive care unit. Eur J Paediatr Neurol 2007;11:70–5.

Jette N, Claassen J, Emerson RG, Hirsch LJ. Frequency and predictors of nonconvulsive seizures during continuous electroencephalographic monitoring in critically ill children. Arch Neurol 2006;63:1750–5.

Jirsch J, Hirsch LJ. Nonconvulsive seizures: developing a rational approach to the diagnosis and management in the critically ill population. Clin Neurophysiol 2007;118:1660–70.

Jordan KG. Continuous EEG and evoked potential monitoring in the neuroscience intensive care unit. J Clin Neurophysiol 1993;10:445–75.

Jordan KG. Neurophysiologic monitoring in the neuroscience intensive care unit. Neurol Clin 1995;13:579–626.

Jordan KG. Nonconvulsive status epilepticus in acute brain injury. J Clin Neurophysiol 1999a;16:332–40.

Jordan KG. Continuous EEG monitoring in the neuroscience intensive care unit and emergency department. J Clin Neurophysiol 1999b;16:14–39.

Kaplan PW. The EEG of status epilepticus. J Clin Neurophysiol 2006;23:221–9.

Kilbride RD, Costello DJ, Chiappa KH. How seizure detection by continuous electroencephalographic monitoring affects the prescribing of antiepileptic medications. Arch Neurol 2009;66:723–8.

Kull LL, Emerson RG. Continuous EEG monitoring in the intensive care unit: technical and staffing considerations. J Clin Neurophysiol 2005;22:107–18.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.

Mani R, Arif H, Hirsch LJ, Gerard EE, LaRoche SM. Interrater reliability of ICU EEG research terminology. J Clin Neurophysiol 2012;29:203–12.

Norman RG, Scott MA. Measurement of inter-rater agreement for transient events using Monte Carlo sampled permutations. Stat Med 2007;26:931–42.

Oddo M, Carrera E, Claassen J, Mayer SA, Hirsch LJ. Continuous electroencephalography in the medical intensive care unit. Crit Care Med 2009;37:2051–6.

Pandian JD, Cascino GD, So EL, Manno E, Fulgham JR. Digital video-electroencephalographic monitoring in the neurological–neurosurgical intensive care unit: clinical features and outcome. Arch Neurol 2004;61:1090–4.

Pensirikul A, Beslow LA, Kessler SK, Sanchez SM, Topjian AA, Dlugos DJ, et al. Density spectral array for seizure identification in critically ill children. J Clin Neurophysiol 2013;30:371–5.

Privitera M, Hoffman M, Moore JL, Jester D. EEG detection of nontonic–clonic status epilepticus in patients with altered consciousness. Epilepsy Res 1994;18:155–66.

Ronner HE, Ponten SC, Stam CJ, Uitdehaaq BM. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. Seizure 2009;18:257–63.

Towne AR, Waterhouse EJ, Boggs JG, Garnett LK, Brown AJ, Smith Jr JR, et al. Prevalence of nonconvulsive status epilepticus in comatose patients. Neurology 2000;54:340–5.

Vespa PM, Nuwer MR, Nenov V, Ronne-Englstrom E, Hovda DA, Bergsneider M, et al. Increased incidence and impact of nonconvulsive and convulsive seizures after traumatic brain injury as detected by continuous electroencephalographic monitoring. J Neurosurg 1999;91:750–60.

Vespa PM, O'Phelan K, Shah M, Mirabelli J, Starkman S, Kidwell C, et al. Acute seizures after intracerebral hemorrhage: a factor in progressive midline shift and outcome. Neurology 2003;60:1441–6.

Waterhouse EJ, Vaughan JK, Barnes TY, Boggs JG, Towne AR, Kopec-Garnett L, et al. Synergistic effect of status epilepticus and ischemic brain injury on mortality. Epilepsy Res 1998;29:175–83.

Williamson CA, Wahlster S, Shafi MM, Westover MB. Sensitivity of compressed spectral arrays for detecting seizures in acutely ill adults. Neurocrit Care 2014;20:32–9.