Journal Club: Criteria for Defining Interictal Epileptiform Discharges in EEG

Author(s):
John R. McLaren, MD1; Jin Jing, PhD1; M. Brandon Westover, MD, PhD1; Fábio A. Nascimento, MD1

Corresponding Author:
John R. McLaren, jmclaren2@mgh.harvard.edu

Affiliation Information for All Authors:
1. Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Neurology® Published Ahead of Print articles have been peer reviewed and accepted for publication. This manuscript will be published in its final form after copyediting, page composition, and review of proofs. Errors that could affect the content may be corrected during these processes.

Equal Author Contribution:

Copyright © 2022 American Academy of Neurology. Unauthorized reproduction of this article is prohibited
Introduction

Electroencephalogram (EEG) misinterpretation is a major contributor to epilepsy misdiagnosis.\(^1,2\) To help improve interictal epileptiform discharges (IED) identification, the International Federation of Clinical Neurophysiology (IFCN) proposed a set of 6 operational criteria\(^3\) (Table 1) that had not been clinically validated until Kural and collaborators provided Class III evidence that IFCN criteria in sensor space (visual analysis) and analysis in source space (spatial filtering method applied to transform scalp signals into regions of interest) have high specificity (>95%) and sensitivity (81-85%) for IED identification, similar to expert scorings.\(^4\)

Hypothesis and Design

The authors investigated the diagnostic accuracy and interrater agreement (IRA) for IED identification using the IFCN criteria in sensor space and source space against a benchmark specificity of 95% for clinical significance. The hypothesis was that these methods would provide high diagnostic accuracy and reduce EEG overinterpretation compared to non-
standardized visual analysis. The study was a large, blinded, retrospective review of EEGs from 100 patients with and without epilepsy.

**Methods**

The authors selected EEGs of consecutive patients undergoing video-EEG monitoring at two Danish centers between 2012 and 2017. Patients were included if they were age one year or older and had habitual paroxysmal events (epileptic or non-epileptic) with interictal sharp transients on EEG. For those with epileptic seizures, the sharp transients had to be concordant with the recorded ictal event. Patients were excluded if their recordings were diagnostically inconclusive or if they had both epileptic and non-epileptic events in the same recording. Interictal sharp transients, previously marked on initial diagnostic interpretation, were reviewed in 10-20 second epochs and selected for further evaluation by two authors who did not participate in the further in-study EEG analysis. These epochs were only included if both experts confirmed the sharp transients met selection criteria (transient, pointed peak, phase reversal in bipolar montage, and concordant with the recorded seizure, if epileptic). Seven separate experts then evaluated the samples in three different rounds more than one month apart, blinded to all other patient data.

In the **first round**, raters scored each marked transient as having the presence or absence of each IFCN criterion in sensor space. In the **second round**, raters analyzed EEG samples in source space in a different randomized order. IEDs were identified if the sharp transients met all three of the following: changes in time, distribution in space, and exclusion of artifacts and normal variants. In the **third round**, raters labeled the sharp transients as either epileptiform or non-epileptiform without any specific criteria, as if in their clinical practice. This was considered “expert scoring”.

Copyright © 2022 American Academy of Neurology. Unauthorized reproduction of this article is prohibited
Diagnostic accuracy of each extracted feature was achieved using the consensus majority scoring. For the IFCN sensor space criteria, a cutoff value of 2-to-6 fulfilled criteria for qualifying as an IED. Analysis in source space and expert scoring resulted in dichotomous scores, labeled as either IEDs or non-epileptiform sharp transients. The gold standard utilized to determine whether patients have epilepsy or non-epileptic events originated from video-EEG data recording these patients’ habitual spells.

**Statistical Analysis**

The receiver operating characteristic (ROC) curve was calculated based on cutoff values of 2 to 6 IFCN criteria in sensor space. ROC curves evaluate the performance of diagnostic tests and, more generally, the accuracy of a statistical model that classifies outcomes in a dichotomous fashion. The Wilson method was used to calculate 95% confidence intervals for sensitivity, specificity, and accuracy. These values were then compared between the different IED identification methods using the McNemar test. The interrater agreement (IRA) was calculated using Gwet agreement coefficient (AC1) and interpreted as poor (<0.02), fair (0.2-0.4), moderate (0.4-0.6), substantial (0.6-0.8), or almost perfect (>0.8). The IRAs of the IED identification methods were then compared using a bootstrap analysis.

**Results**

The AUC for IFCN criteria in sensor space was 0.977 for the consensus majority scoring and median 0.941 for individual raters. The ROC curve highlights that IFCN criteria in sensor space with cutoff values of 4 and 5 and analysis in source space performed similar to expert scoring with no statistically significant difference in accuracy ($p>0.157$). IFCN criteria with cutoff of 5
resulted in higher specificity ($p=0.025$) but lower sensitivity ($p=0.005$) compared to a cutoff of 4. The two methods that achieved specificity higher than expected 95% threshold were IFCN sensor space with cutoff of 5 and analysis in source space. Similarly, using the expert consensus as gold standard rather than patients’ video-EEG data, these two methods achieved the highest accuracies (93%). The three methods with the best diagnostic performance (IFCN sensor with cutoffs of 4 and 5 and source space) had an IRA that was moderate to substantial (AC1: 0.490–0.608) without a significant difference in IRA between them ($p>0.900$). IRA for individual criterion was found to be substantial for 1, 2, and 4, moderate for 6, and fair for 3 and 5 eTable 1 in the Supplement).

**Interpretation**

The design where EEGs were pre-selected before they were serially rated by experts in an independent fashion provides a solid foundation to address the authors’ research question. The use of ROC graphs is appropriate as they recognize the inherent tradeoff between sensitivity and false positive rate when making binary judgements in the face of uncertainty (as compared with studies that artificially impose a particular decision threshold as being optimal). For IRA calculations, the authors utilized the Gwet agreement coefficient (AC) to measure the amount of beyond-chance agreement between experts. The Gwet AC has advantages over the more commonly used Cohen’s kappa IRA statistic, which suffers from counterintuitive behavior (“Cohen’s Kappa Paradox”) in the presence of high levels of agreement. They then used a bootstrap method to compute p-values, testing whether differences in IRAs achieved using the various IED identification methods were statistically significant.
This study has additional strengths. It was well powered (>0.8) with 100 total patients, achieving specificity of 95% with a significance level of 5%. EEGs included in the study originated from consecutive patients, minimizing selection bias. Achieving a pre-test probability of approximately 50%, similar numbers of patients with epileptic and non-epileptic events were included. Regarding expert consensus, there were seven raters with substantial EEG experience (median 14 years). All experts were blinded to clinical data, reducing history bias. The authors who initially selected the EEG samples for in-study review did not participate in the rating process, removing potential confirmation bias. Lastly, EEG analytic methods were compared against classification of the patients’ habitual paroxysmal events based on video-EEG recordings. This gold standard has the significant advantages of being objective and externally validated. However, it also has disadvantages - namely concerning patient selection bias. This method is unable to account for EEG data from a significant number of patients evaluated for seizures—those who never require video-EEG monitoring. Additionally, patients without habitual episodes captured during video-EEG monitoring and those with inconclusive LTM were not included in this study. We believe that utilizing a combined gold standard including both patients’ video-EEG data and expert consensus would be advantageous. The latter has the advantage of being immune to patient selection; however, it carries a disadvantage of being grounded in expert experience rather than an externally validated and objective source. Ultimately, both methods complement each other and should be considered as a highly reliable gold standard for future studies.

Other limitations of the study include the lack of assessment of the pattern of repetition throughout the recordings since the IFCN criteria do not include such a criterion. Moreover, it was shown that the IRA associated with some IFCN criteria (namely 3, 5, and 6) was fair to
moderate, potentially compromising the utility of these criteria in identifying IEDs and their
global implementation. Another limitation was the lack of randomization between review
rounds, as all raters underwent the same sequential order of review (sensor space, then source
space, then expert review), exposing raters to potential confirmation bias. This effect was
minimized by presenting the samples in randomized order, spacing each round of analysis by one
month, and blinding the raters to the gold standard until all rounds were completed.
Several studies have since expanded our understanding of the use of the IFCN criteria. For
example, the combination of IFCN criteria 1, 4, and 6 generates high accuracy and IRA upon
identifying IEDs, and it is required that discharges meeting fewer IFCN criteria repeat in the
EEG study to maintain optimal diagnostic accuracy.
Lastly, teaching the IFCN criteria to trainees improves their diagnostic accuracy and IRA.
In summary the authors validated the use of the IFCN criteria (4 or 5/6) and source space
analysis to identify IEDs with an accuracy similar to expert scoring. The implementation of the
IFCN criteria and source analysis in clinical practice should contribute to decreased error and
misdiagnosis of epilepsy especially in challenging cases and/or in scenarios where IRA is
suboptimal. Given known educational gaps in neurology residency EEG training, we believe
residents should also receive dedicated training on how to use the IFCN criteria and source
analysis. This training would be based on objective criteria thus potentially bypassing the
occasional issue faced by some institutions where there is insufficient faculty time to teach.
Potential challenges to an implementation of these methods include the need to thoroughly train
readers on their use as well as the subjective judgement required to score each IFCN criterion.
References:


Table 1: IFNCN proposed operational definition of IEDs to help standardize EEG visual analysis

| 1. Di- or tri-phasic waves with sharp or spiky morphology (i.e. pointed peak) |
| 2. Different wave duration than the ongoing background activity either shorter or longer |
| 3. Asymmetry of the waveform: a sharply rising ascending phase and a more slowly decaying descending phase, or vice versa |
| 4. The transient is followed by an associated slow after-wave |
| 5. The background activity surrounding IEDs is disrupted by the presence of the IEDs |
| 6. The distribution of the negative and positive potentials on the scalp suggests a source of the signal in the brain, corresponding to a radial, oblique, or tangential orientation of the source. |