HHS Public Access

Author manuscript

Clin Neurophysiol. Author manuscript; available in PMC 2025 August 10.

Published in final edited form as:

Clin Neurophysiol. 2025 May ; 173: 138–146. doi:10.1016/j.clinph.2025.02.275.

Utility of the IFCN criteria for identifying interictal epileptiform discharges by experts: A decision hygiene approach to improve inter-rater reliability

Doyle Yuan[a,b,*], **Roohi Katyal**[c], **Irfan Sheikh**[a,b], **Ioannis Karakis**[e,f], **Selim Benbadis**[g], **Ushtar Amin**[g], **Kollencheri Puthenveettil Vinayan**[h], **Niravkumar Barot**[i], **Dan Weber**[j], **Adam Greenblatt**[k], **Sándor Beniczky**[l], **M. Brandon Westover**[d], **Fábio A. Nascimento**[k]

[a]Department of Neurology, University of Texas Southwestern Medical Center, Dallas, TX, USA

[b]Peter O'Donnell Jr. Brain Institute, University of Texas Southwestern Medical Center, Dallas, TX, USA

[c]Department of Neurology, Louisiana State University Health Sciences Center, Shreveport, LA, USA

[d]Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

[e]Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA

[f]University of Crete School of Medicine, Heraklion, Greece

[g]Department of Neurology, University of South Florida, Tampa, FL, USA

[h]Department of Pediatric Neurology and Amrita Advanced Center for Epilepsy, Amrita Institute of Medical Sciences, Cochin, Kerala, India

[i]Department of Neurology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

[j]Department of Neurology, Saint Louis University, Saint Louis, MO, USA

[k]Department of Neurology, Washington University in St. Louis, St. Louis, MO, USA

[l]Aarhus University, Danish Epilepsy Center and Aarhus University Hospital, Denmark

Abstract

Objective: To determine if implementing the IFCN criteria to define interictal epileptiform discharges (IEDs) improves expert inter-rater reliability (IRR) and diagnostic performance.

Methods: Nine EEG experts rated the same 200 candidate IEDs (100 expert-consensus, 100 epilepsy monitoring unit [EMU]-validated) as epileptiform or not, in random order, in two rounds separated by at least 30 days. During the second round, raters additionally selected the applicable IFCN criteria for each candidate IED.

[*]Corresponding author. doyle.yuan@utsouthwestern.edu (D. Yuan).

Declaration of competing interest

Dr. Westover is a co-founder, scientific advisor, consultant to, and has personal equity interest in Beacon Biosignals. All other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Results: Overall, there were no major differences in performance (AUC; 0.90 vs. 0.91) or IRR (AC1; 0.48 vs. 0.47) between both Parts; nor was there a major difference in calibration within the expert-consensus dataset (median absolute calibration index; 35.5 vs. 30.0). Similarly, there were no major differences in performance or IRR within either dataset. IRR was substantial within the EMU-validated dataset and only fair within the expert-consensus dataset. IRR was fair for criteria 2, 3, 5 and 6, and moderate for criteria 1 and 4.

Conclusions: Our findings suggest that the IFCN criteria to define IEDs may not significantly improve IRR, performance, or overall calibration among experts.

Significance: Increasing expert IRR for each criterion may enhance the utility of the IFCN criteria in clinical practice.

Keywords

EEG; Inter-rater reliability; IFCN criteria; Decision hygiene

1. Introduction

The value of electroencephalography (EEG) in the diagnosis and management of seizure disorders is well established (Amin & Benbadis, 2019; Benbadis et al., 2020; Fisher et al., 2014; Goodin & Aminoff, 1984; Pillai & Sperling, 2006; van Donselaar et al., 1992). In particular, the presence of interictal epileptiform discharges (IEDs) supports the diagnosis of epilepsy in patients for whom there is moderate clinical suspicion for epileptic seizures (Goodin & Aminoff, 1984) and establishes the diagnosis in patients presenting after a single unprovoked epileptic seizure (Fisher et al., 2014). Accurately and reliably identifying IEDs on EEG depends not only on the skill of the reader but also on their idiosyncratic preferences to over- and under-call candidate waveforms. While experts attain a high level of skill through specialized training in EEG, their over- and under-calling tendencies give rise to imperfect inter-rater reliability (IRR).

In their investigation of expert judgments across multiple disciplines, Kahneman and others proposed a framework for understanding so-called “noise”, or unwanted variability in judgments between experts, especially when the “correct answer” is undefined or not ascertainable. They divided noise into two components: (1) level noise, which arises from the relative overall judgment tendency of a particular expert (e.g., a lenient expert), and (2) pattern noise, which arises from the specific characteristics of a case or situation (e.g., a lenient expert who is unusually strict when judging particular types of cases). The authors coined the term “decision hygiene” to refer to strategies which systematically reduce noise (Kahneman et al., 2021).

This framework is relevant to the expert judgment of EEGs: each EEG is typically interpreted by only one reader (i.e., IRR is usually unknown), errors are additive (under-calling one EEG and over-calling another commits two diagnostic errors which do not “cancel out”), and a gold standard is not readily available. In the context of EEG interpretation, level noise corresponds to variation in individual preferences for classifying sharp transients as IEDs, with “under-callers” maintaining a high threshold and “over-

callers" maintaining a low one. Pattern noise, on the other hand, is the product of arbitrary reactions to features within and outside of the EEG and tends to be influenced by prior experience and training.

Comparable exposure to diverse EEG patterns during subspecialty training likely reduces pattern noise in expert ratings. In other words, pattern noise is inversely related to the rater's skill, which previous studies have confirmed to be predictably high among experts (Jing et al., 2020). Level noise has thus been hypothesized to be a larger contributing factor to suboptimal IRR in IED identification; in other words, the same EEG may be read differently depending on the reader's threshold (Nascimento et al., 2022).

We postulated that applying standardized operational criteria may reduce level noise among experts by ensuring each rater considers the same features of a candidate IED and by defining a uniform threshold in terms of the number of criteria satisfied. Notably, the use of analogous checklists has proven effective in reducing error across a variety of disciplines (Gawande, 2011), as they break down a complex decision into a series of more straightforward, independent determinations.

The International Federation of Clinical Neurophysiology (IFCN) has proposed six criteria in its operational definition of IEDs (Kane et al., 2017). In one study which enlisted experts to rate EEGs validated in epilepsy monitoring units, application of these criteria yielded excellent performance in accurately identifying IEDs (Kural et al., 2020a; McLaren et al., 2022). However, this study focused primarily on demonstrating the validity of the criteria against known EMU outcomes, rather than accessing their broader utility as a decision support tool. The present study aims to determine whether conscious application of the IFCN criteria would improve expert IRR and performance over a larger and more representative dataset.

2. Methods

2.1. Study design

We conducted a two-part prospective study to compare standard visual analysis with explicit consideration of the IFCN-proposed criteria in the identification of IEDs. We invited experts with subspecialty training in EEG interpretation from various institutions across primarily the United States to rate 200 EEG epochs in Part I and the same 200 epochs in Part II, in random order. Raters were not informed they would be rating the same EEGs twice. We additionally enforced a 30-day minimum waiting period between both parts to further reduce the effect of visual recall.

We deployed a novel, interactive, online platform ("EEGHub", supplemental Fig. 1) to display EEGs and collect responses from raters. Our EEG viewing application, delivered via web browser, offered several advantages over conventional static images, including the ability to switch montages (longitudinal bipolar, common average reference, transverse bipolar, physical reference [Cz], and ipsilateral ear reference), adjust sensitivities, and automatically optimize EEG rendering for a variety of screen resolutions. The platform

allowed raters to complete the task across multiple sessions, automatically saving their progress after each sitting.

For each EEG, we highlighted a two-second interval containing the candidate IEDs to be considered. We provided the patient's age at the time of recording but otherwise did not provide any additional case-specific information. In Part I, we asked raters to make only a binary decision of "spike" (for any epileptiform waveform) or "non-spike" (non-epileptiform). Immediately before starting Part II, raters were presented with educational materials on the six IFCN criteria for defining IEDs and how to apply them to classify candidate IEDs, including a previously published infographic (Fig. 1), a video lecture by one of the senior authors (SB) endorsed by the International League Against Epilepsy (ILAE) (Kural et al., 2022), and links to a webcast hosted on YouTube created by the authors FN and MBW (EEGTalk.com). Then, in Part II, we asked raters to select and document for each candidate IED the applicable IFCN criteria, followed by the binary classification. While we explicitly stated the IFCN's recommendation that at least four of the six criteria be met to consider a pattern epileptiform, we permitted raters to make the final classification decision independent of their IFCN criteria choices.

To preserve anonymity of responses, we assigned each participant a unique identifier. During the study, only the platform administrator had access to the table linking each identifier to each invitee; this table was purged upon conclusion of the study.

Finally, at the end of the study, we surveyed all participants by email on their training background, years of EEG-reading experience, prior familiarity with the IFCN criteria, and viewership of the presented educational materials.

The study was conducted in the fall of 2023. Preparation of the data and sharing of deidentified candidate IEDs on EEGHub was conducted under IRB-approved protocols. The study data were deidentified and obtained from epileptologists who volunteered to participate in the study. Therefore, the study did not require IRB approval based on our review of local IRB policies.

2.2. EEG selection

We selected deidentified 11-to-15-second EEG epochs, each containing at least one candidate IED, from two previously published datasets: (1) an expert-rated collection of over 18,000 15-second epochs, each sampled at 128 Hz (Jing et al., 2020), and (2) an EMU-validated collection of 100 11- to 14- second epochs, each sampled at 200 Hz (Kural et al., 2020a). All EEGs were recorded using a standard 19-electrode array plus single-channel EKG; all but three EEGs in the second dataset additionally included inferior temporal electrodes (Seeck et al., 2017).

Within the first dataset, we considered only EEGs rated by at least 12 experts from diverse training and practice backgrounds, and we divided these EEGs into five "bins" based on the percentage of experts who rated the candidate IED as epileptiform (0–20 %, 20–40 %, 40–60 %, 60–80 %, and 80–100 %). We then randomly selected twenty epochs from each bin, yielding a total of 100 candidate IEDs with varying degrees of consensus. To establish

a gold standard, we defined epochs as epileptiform if more than 50 % of experts rated them as such and non-epileptiform if fewer than 50 % of experts rated them as epileptiform. For three epochs which exactly 50 % of experts rated as epileptiform, no gold standard was defined. The resulting subset contained a mix of inpatient and outpatient EEG, from patients aged 26–91 years (mean 73 years, median 74 years); rhythmic and periodic patterns were present in about half of the epochs and accounted for nearly all the gold-standard IEDs.

From the second dataset, we included all 100 epochs, of which 54 were recorded from patients aged 2.5–76 years (mean 32 years, median 29 years) with concordant epileptic seizures captured during an EMU admission who did not have concurrent non-epileptic events, and the remaining 46 were from patients aged 11–89 years (mean 38 years, median 34 years) with EEG sharp transients and EMU-confirmed non-epileptic events. Within this dataset, the EMU outcome defined the gold standard, with nearly all of gold-standard IEDs being sporadic sharp waves or spikes.

2.3. Data analysis

To measure performance, we computed the false positive rate (i.e., 1-specificity), true positive rate (i.e., sensitivity), and accuracy of consensus majority ratings assessed against the gold standard defined above, for the combined dataset of 200 EEGs and for each of the two datasets separately. We computed 95 % confidence intervals on these statistics using Wilson's method. Using the binormal model, we fit theoretical receiver operating characteristic (ROC) curves to these parameters and computed the area under the curve (AUC) as a measure of rater performance per part, per dataset. In addition, we fit an overall ROC curve to the observed true- and false- positive rates of each rater by using the method of least squares. We used the Wilcoxon signed rank test to compare the AUC values between Parts I and II.

For the expert-consensus dataset, we performed statistical calibration analysis for each rater to assess under-calling and over-calling tendencies. For each of the five bins described above, we plotted the percentage of the twenty candidate IEDs in that bin deemed epileptiform by each rater (the observed probability) against the midpoint of that bin (the expected probability: 10 %, 30 %, 50 %, 70 %, and 90 %). We fit a smooth parametric model to these points and computed respective calibration indices for each rater (Jing et al., 2023), which we compared between Parts I and II using the Wilcoxon signed rank test.

To quantify IRR, we computed Gwet's AC1 (Gwet, 2008) per part, per dataset. For Part II, we also computed AC1 for each of the six IFCN criteria separately. We interpreted these scores using the conventional Landis and Koch scale (Landis & Koch, 1977).

As raters were permitted to submit responses which deviated from the IFCN's operational definition of IEDs, we were interested in measuring the effect of this on performance and IRR. Thus, for each response in Part II, we additionally derived a hypothetical binary rating strictly from the number of selected IFCN criteria, using thresholds of four, five, and six criteria to define an epileptiform rating. We repeated the above analyses using this hypothetical rating and compared these results with those derived from the actual binary ratings in Part II.

To further estimate of the influence of considering the IFCN criteria on rater responses, we computed intra-rater agreement between both parts by Cohen's kappa, with lower agreement implying a greater effect of the IFCN criteria on decision-making.

Finally, to determine whether performance or intra-rater agreement was related to experience, we computed the Spearman correlation coefficient between each rater's self-reported number of years reading EEG with their AUC values from Part I and II as well as intra-rater agreement, for the combined dataset.

3. Results

3.1. Study participants

Of the thirteen experts we invited, twelve completed Part I, and nine completed Part II. All nine of these experts had completed post-residency fellowship training in neurophysiology and/or epilepsy before taking part in the study. In the follow-up survey, each reported between two and thirty years of experience reading EEG, with a mean of eleven years and median of seven years (Table 1). The experts trained at nine different institutions across the United States, except for one participant who trained in Japan.

Of all participants who finished Part I, most did so within a single day (range 13 min to 15 days, median 39 min, mean 44 h). Among those who completed Part II, the average time to completion was about two weeks (range 77 min to 43 days, median 14 days, mean 16 days). Data from the three experts who completed Part I but not Part II were excluded from the analysis.

3.2. Performance

Aggregate performance metrics are provided in Table 2. There was a small increase in specificity and decrease in sensitivity from Part I to Part II, although neither change was statistically significant. Accordingly, across the combined dataset, there was no statistically significant difference in the per-rater and overall AUC values between both parts: the numerical improvement seen within the EMU-validated subset was offset by the decline in AUC within the expert-consensus subset (Fig. 2). There was also no significant correlation between the per-rater AUC values and the number of years reading EEG in either Part I (rho −0.126, $p = 0.75$) or Part II (rho 0.176, $p = 0.65$).

We did not find hypothetical ratings derived from the number of selected IFCN criteria to perform better than the actual ratings submitted. Sensitivities and specificities of rater responses were largely in line with predicted ratings using a threshold of four satisfied IFCN criteria, with a median number of seven discrepant responses out of 200 (range 1–22, mean 9).

Overall performance as measured by the AUC was better within the EMU-validated dataset than within the expert-consensus subset (Fig. 2). This is similarly reflected in ROC curves constructed using varying thresholds of IFCN criteria, in aggregate (Fig. 3) and per rater (Fig. 4). Nonetheless, the average number of selected IFCN criteria within the

expert-consensus dataset did appropriately increase with increasing IED probability (Fig. 6), validating the scoring of the expert-consensus EEGs.

3.3. Calibration

Calibration values and fitted curves for each rater are illustrated in Fig. 5, and computed calibration indices are summarized in Table 3. While under-calling and over-calling were common in both parts, there was significantly more under-calling in Part II ($p = 0.0039$), with a median calibration index of −28.84, than in Part I, with a median calibration index of 0.0. However, this change did not translate into better overall calibration: comparing the absolute value of the calibration indices yielded no statistically significant difference between Part I and Part II (median 35.5 vs. 30.0; $p = 0.20$).

3.4. IRR of binary classification

Inter-rater reliability of IED classification, summarized in Table 2 and illustrated in Fig. 7, did not change significantly between Parts I and II. IRR was substantial within the EMU-validated dataset and only fair within the expert-consensus dataset, resulting in an overall moderate IRR across the combined dataset. Substituting hypothetical ratings as described above, we observed increasing IRR as the IFCN criteria threshold increased, reflecting more stringent requirements for classifying a candidate IED as epileptiform. This effect was most pronounced among the expert-consensus EEGs, where using a threshold of five criteria resulted in significantly better IRR compared to raters' own binary classification in either Part I or Part II. In fact, at a cutoff of five criteria (Kural et al., 2020a), we found IRR to be comparable between the expert-consensus EEGs (AC1 0.599; 95 % CI 0.512–0.686) and EMU-validated EEGs (AC1 0.589; 95 % CI 0.520–0.659); whereas using the cutoff of four criteria as recommended by the IFCN, IRR was lower within the expert-consensus subset (AC1 0.279; 95 % CI 0.190–0.367) than the EMU-validated subset (AC1 0.596; 95 % CI 0.516–0.675). However, imposing either cutoff yielded no statistically significant improvement in IRR across the combined dataset or within the EMU-validated subset.

3.5. IRR of IFCN criteria

The inter-rater reliability of each individual IFCN criterion is summarized in Table 4 and illustrated in Fig. 8. Criterion #4 (after-going slow wave) showed the highest IRR, followed by criterion #1 (sharp morphology), particularly within the EMU-validated dataset. IRR among the remaining criteria, including #2 (duration distinct from background activity), was only fair.

3.6. Intra-rater agreement

Intra-rater agreement between both parts is provided in Table 5. All but one rater demonstrated at least substantial agreement between their Part I and Part II responses. There was no significant correlation between intra-rater agreement and the number of years reading EEG (rho −0.126, $p = 0.75$). Interestingly, the rater with the lowest level of agreement was among the most experienced.

3.7. Survey results

On the follow-up survey, all raters affirmed familiarity with the IFCN criteria before entering the study. While most of them (88 %) indicated that they inspected the provided infographic, only a small minority (22–44 %) indicated they reviewed any of the additional video materials (Table 6). None of the experts reviewed all of the recommended educational materials.

4. Discussion

Contrary to our expectations, we did not find a significant improvement in overall expert performance or calibration from Part I, in which raters were expected to judge candidate IEDs using standard visual analysis, to Part II, in which raters were required to explicitly consider the IFCN criteria when making their judgments. The substantial intra-rater agreement between the two parts further suggests that systematically applying the IFCN criteria did not affect our raters' decision-making.

One possible explanation is that raters implicitly and perhaps subconsciously invoked the IFCN criteria for both parts. On our final survey, all experts did indicate they were already familiar with the IFCN criteria before beginning Part I. Even if they were unfamiliar with the exact criteria set forth by the IFCN, experts will have likely adopted an overlapping set of working criteria for what constitutes an IED, having undergone similar subspecialty training in EEG interpretation (Maulsby, 1971). Arguably, the IFCN criteria themselves, being grounded in expert opinion, represent an operational translation of this expertise.

An alternative and potentially concurrent possibility is that experts used their own judgment to classify candidate IEDs in both parts and then consciously or subconsciously "fit" the IFCN criteria in Part II to their predetermined judgments. Such decisions are difficult to quantify and may not even be obvious to raters themselves. One could also point to the low self-reported viewership of the supplementary videos as evidence that raters were not specifically trained on the IFCN criteria prior to beginning Part II and thus had a similar grasp of the IFCN criteria during both parts.

Previous studies which enlisted trainees rather than experts, and which involved completion of a hands-on educational program, have shown improvement in performance and IRR among trainees (Kural et al., 2022; Nascimento et al., 2024). However, within these studies, it is difficult to separate the effect of increased proficiency in standard visual analysis from the effect of learning the six criteria. Rather than to evaluate the effect of an educational intervention, our study instead aimed to measure the effectiveness of the criteria as a decision aid, while controlling for proficiency and experience by employing the same expert raters for both parts.

Our finding of comparable IRR between both parts suggests that using a checklist of standardized criteria was not effective in reducing noise, at least not in our cohort of nine experts. While introducing the IFCN criteria may have reduced a single complex decision into several smaller decisions, those smaller decisions proved to be noisy themselves: the IRR for each individual criterion was generally only fair (criteria 2, 3, 5, and 6) to moderate

(criteria 1 and 4). We speculate that the sum of this noise per criterion countervailed any level noise reduction from setting a common threshold to call a candidate IED epileptiform —leaving open the possibility that efforts to reduce inconsistencies in how each criterion is applied, such as focused training, may enhance overall decision reliability even among experts.

Traditionally, training in EEG interpretation relies on an empirical approach where trainees review numerous EEGs and gradually align their decision-making process with that of more experienced EEG readers. One study involving seven neurology trainees lends support to a complementary strategy: after receiving didactic and hands-on training on the IFCN criteria, participants demonstrated better performance and higher inter-rater agreement in classifying IEDs, reaching a level comparable to that of experts (Kural et al., 2022). These findings suggest that providing focused teaching of the IFCN criteria could be a viable approach to improve accuracy and reduce unwanted variability in how EEGs are interpreted.

One potential criticism of our study design is that only short EEG epochs emphasizing single discharges were provided for rating, which does not reflect clinical practice. In fact, Kural and others have proposed a framework in which the presence of multiple discharges in a single EEG could reduce the minimum number of operational criteria required to classify them as epileptiform (Kural et al., 2021). In contrast, our study aimed to evaluate the application of the IFCN criteria to single candidate IEDs, and the finding of less-than-perfect IRR even for each criterion separately reveals noise which may not “cancel out” when evaluating multiple candidate IEDs. Further studies are needed to characterize this noise in the setting of interpreting full-length EEGs.

Additionally, the lack of voltage (scalp potential) maps might have limited the evaluation of the sixth IFCN criterion, the presence of a physiologic field. A previous study did find the combination of criteria #1, #4, and #6, where a voltage map was provided to aid the evaluation of criterion #6, to have the highest specificity and sensitivity for IEDs (Kural et al., 2020b). In our study, while raters could toggle between multiple bipolar and referential montages, voltage maps were unavailable.

The difficulty of establishing a gold standard is a well-known limitation of studies of EEG interpretation. For candidate IEDs on a scalp recording, one might propose the gold standard of a corresponding cortical discharge; however, intracranial EEG data is unavailable for most EEGs encountered in clinical practice. Alternatively, since the value of identifying an IED lies in its prediction of heightened seizure risk (Fisher et al., 2014), a clinical gold standard could be defined based on the recording of an electrographic seizure in the same distribution as an IED. While this is what the EMU-validated dataset aims to achieve, EMU data is similarly unavailable in most routine cases, especially outside of academic referral centers and in low-resource settings.

In contrast, a more feasible gold standard may be one that reflects the majority opinion of a sufficient number of experts. This is the basis for consensus guidelines, the increased certainty of a medical diagnosis when it is reached by multiple independent diagnosticians, and other similar situations (Kahneman et al., 2021). For establishing a consensus gold

standard for IEDs, Halford and others determined an optimal number of raters of about seven, as a tradeoff between accuracy and obtainability (Halford et al., 2017). A related study corroborated the “wisdom of the crowd” effect, demonstrating that a computational model for IEDs reached peak performance when trained on the average ratings of at least ten experts (Bagheri et al., 2017). All candidate IEDs within our expert-consensus dataset were rated by at least twelve experts each (mean 15, maximum 18).

One caveat specific to a consensus-based gold standard, rather than an externally validated one, is the requirement to choose a consensus cutoff for performance analyses. For simplicity, we have chosen an unbiased cutoff of 50 % (simple majority), such that a candidate IED labeled by eight of fifteen experts as epileptiform would be defined as such, whereas a waveform labeled by only five of twelve experts as epileptiform would be considered non-epileptiform. Higher cutoffs exclude more equivocal waveforms from the analysis, resulting in better computed performance (supplemental Fig. 2) at the cost of potentially reduced generalizability, as both equivocal and unequivocal waveforms are routinely evaluated in clinical practice. However, it should be emphasized that since neither calibration or IRR depends on a gold standard, the above caveat does not apply to these analyses.

Our study is unique in its use of two datasets grounded on different gold standards, which likely accounts for their dissimilarities in performance and IRR. Namely, performance as measured by per-rater AUCs was lower within the expert-consensus dataset than the EMU-validated dataset, as was IRR on both the binary classification and the individual IFCN criteria. We speculate the greater amount of error within the expert-consensus dataset may be attributable to its composition: in curating a balanced number of EEGs from different levels of consensus, we sought to better capture the spectrum of equivocal to unequivocal IEDs encountered in clinical practice. In contrast, the substantially higher IRR in classifying the EMU-validated EEGs suggests that this dataset represents the less equivocal end of the spectrum of both epilepsy and non-epilepsy cases. While incorporating more equivocal cases may have provided a more challenging and representative benchmark, lower expert IRR may imply a less definitive gold standard against which to measure performance, though having a large enough number of raters is expected to compensate for this to some degree. Ultimately, we advocate the creation of an optimized gold standard merging expert consensus and an external, objective source (based on EMU data and/or long-term clinical follow-up), thus encompassing the breadth of patients with seizures and epilepsy.

5. Conclusion

In this two-part prospective study, we found no statistically significant improvement in diagnostic accuracy of experts when explicitly applying the six operational criteria for an IED proposed by the IFCN, compared to standard visual analysis alone. From the standpoint of decision hygiene, conscious application of the criteria also failed to reduce noise, or inter-rater variability, in expert judgments. Further studies are warranted on how to mitigate this noise, including at the percriterion level, to ensure reliable and accurate EEG interpretation and proper diagnosis and management of epilepsy.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

Amin U, Benbadis SR, 2019. The role of EEG in the erroneous diagnosis of epilepsy. J. Clinical Neurophys. : Official Publication of the American Electroencephalographic Soc 36 (4), 294–297. 10.1097/WNP.0000000000000572.

Bagheri E, Dauwels J, Dean BC, Waters CG, Westover MB, Halford JJ, 2017. Interictal epileptiform discharge characteristics underlying expert interrater agreement. Clinical Neurophys. : Official J. Int. Federation of Clinical Neurophysiology 128 (10), 1994–2005. 10.1016/j.clinph.2017.06.252.

Benbadis SR, Beniczky S, Bertram E, MacIver S, Moshé SL, 2020. The role of EEG in patients with suspected epilepsy. Epileptic Disorders : Int. Epilepsy J. with Videotape 22 (2), 143–155. 10.1684/epd.2020.1151.

Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, Engel J, Forsgren L, French JA, Glynn M, Hesdorffer DC, Lee BI, Mathern GW, Moshé SL, Perucca E, Scheffer IE, Tomson T, Watanabe M, Wiebe S, 2014. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 55 (4), 475–482. 10.1111/epi.12550. [PubMed: 24730690]

Gawande A, 2011. The checklist manifesto: How to get things right. St Martin's Press.

Goodin DS, Aminoff MJ, 1984. Does the interictal EEG have a role in the diagnosis of epilepsy? Lancet (London, England) 1 (8381), 837–839. 10.1016/s0140-6736(84)92281-5. [PubMed: 6143148]

Gwet KL, 2008. Computing inter-rater reliability and its variance in the presence of high agreement. Br. J. Math. Stat. Psychol 61 (Pt 1), 29–48. 10.1348/000711006X126600. [PubMed: 18482474]

Halford JJ, Arain A, Kalamangalam GP, LaRoche SM, Leonardo B, Basha M, Azar NJ, Kutluay E, Martz GU, Bethany WJ, Waters CG, Dean BC, 2017. Characteristics of EEG interpreters associated with higher interrater agreement. J. Clinical Neurophys. : Official Publi. American Electroencephalographic Soc 34 (2), 168–173. 10.1097/WNP.0000000000000344.

Jing J, Ge W, Struck AF, Fernandes MB, Hong S, An S, Fatima S, Herlopian A, Karakis I, Halford JJ, Ng MC, Johnson EL, Appavu BL, Sarkis RA, Osman G, Kaplan PW, Dhakar MB, Jayagopal LA, Sheikh Z, Westover MB, 2023. Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. Neurology 100 (17), e1737–e1749. 10.1212/WNL.0000000000201670. [PubMed: 36460472]

Jing J, Herlopian A, Karakis I, Ng M, Halford JJ, Lam A, Maus D, Chan F, Dolatshahi M, Muniz CF, Chu C, Sacca V, Pathmanathan J, Ge W, Sun H, Dauwels J, Cole AJ, Hoch DB, Cash SS, Westover MB, 2020. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. JAMA Neurol. 77 (1), 49–57. 10.1001/jamaneurol.2019.3531. [PubMed: 31633742]

Kahneman D, Sibony O, Sunstein CR, 2021. Noise: A flaw in human judgment. Little, Brown Spark.

Kane N, Acharya J, Benickzy S, Caboclo L, Finnigan S, Kaplan PW, Shibasaki H, Pressler R, van Putten MJAM, 2017. A revised glossary of terms most commonly used by clinical electroencephalographers and updated proposal for the report format of the EEG findings. Revision 2017. Clin. Neurophysiol. Pract 2, 170–185. 10.1016/j.cnp.2017.07.002. [PubMed: 30214992]

Kural MA, Aydemir ST, Levent HC, Ölmez B, Özer IS, Vlachou M, Witt AH, Yilmaz AY, Beniczky S, 2022. The operational definition of epileptiform discharges significantly improves diagnostic accuracy and inter-rater agreement of trainees in EEG reading. Epileptic Disorders : Int. Epilepsy J. with Videotape 24 (2), 353–358. 10.1684/epd.2021.1395.

Kural MA, Duez L, Sejer Hansen V, Larsson PG, Rampp S, Schulz R, Tankisi H, Wennberg R, Bibby BM, Scherg M, Beniczky S, 2020a. Criteria for defining interictal epileptiform discharges in EEG: A clinical validation study. Neurology 94 (20), e2139–e2147. 10.1212/WNL.0000000000009439. [PubMed: 32321764]

Kural MA, Qerama E, Johnsen B, Fuchs S, Beniczky S, 2021. The influence of the abundance and morphology of epileptiform discharges on diagnostic accuracy: How many spikes you need to spot in an EEG. Clinical Neurophysiology : Official J. of the Int. Federation of Clinical Neurophysiology 132 (7), 1543–1549. 10.1016/j.clinph.2021.03.045.

Kural MA, Tankisi H, Duez L, Sejer Hansen V, Udupi A, Wennberg R, Rampp S, Larsson PG, Schulz R, Beniczky S, 2020b. Optimized set of criteria for defining interictal epileptiform EEG discharges. Clinical Neurophysiology : Official J. the Int. Federation of Clinical Neurophysiology 131 (9), 2250–2254. 10.1016/j.clinph.2020.06.026.

Landis JR, Koch GG, 1977. The measurement of observer agreement for categorical data. Biometrics 33 (1), 159–174. [PubMed: 843571]

Maulsby RL, 1971. Some guidelines for assessment of spikes and sharp waves in EEG tracings. American J. EEG Technol 11 (1), 3–16. 10.1080/00029238.1971.11080808.

McLaren JR, Jing J, Westover MB, Nascimento FA, 2022. Journal club: Criteria for defining interictal epileptiform discharges in EEG. Neurology 99 (10), 430–432. 10.1212/WNL.0000000000200991. [PubMed: 35853743]

Nascimento FA, Beniczky S, 2024. Teaching the 6 criteria of the international federation of clinical neurophysiology for defining interictal epileptiform discharges on EEG using a visual graphic. Neurology Education 2 (2), e200073. 10.1212/NE9.0000000000200073.

Nascimento FA, Jing J, Beniczky S, Benbadis SR, Gavvala JR, Yacubian EMT, Wiebe S, Rampp S, van Putten MJAM, Tripathi M, Cook MJ, Kaplan PW, Tatum WO, Trinka E, Cole AJ, Westover MB, 2022. One EEG, one read - A manifesto towards reducing interrater variability among experts. Clinical Neurophysiology : Official J. the Int. Federation of Clinical Neurophysiology 133, 68–70. 10.1016/j.clinph.2021.10.007.

Nascimento FA, Jing J, Traner C, Kong WY, Olandoski M, Kapur S, Duhaime E, Strowd R, Moeller J, Westover MB, (2024). A randomized controlled educational pilot trial of interictal epileptiform discharge identification for neurology residents. Epileptic Disorders : International Epilepsy Journal with Videotape. 10.1002/epd2.20229.

Pillai J, Sperling MR, 2006. Interictal EEG and the diagnosis of epilepsy. Epilepsia 47 (Suppl 1), 14–22. 10.1111/j.1528-1167.2006.00654.x.

Seeck M, Koessler L, Bast T, Leijten F, Michel C, Baumgartner C, He B, Beniczky S, 2017. The standardized EEG electrode array of the IFCN. Clinical Neurophysiology : Official J. the Int. Federation of Clinical Neurophysiology 128 (10), 2070–2077. 10.1016/j.clinph.2017.06.254.

van Donselaar CA, Schimsheimer RJ, Geerts AT, Declerck AC, 1992. Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. Arch. Neurol 49 (3), 231–237. 10.1001/archneur.1992.00530270045017. [PubMed: 1536624]

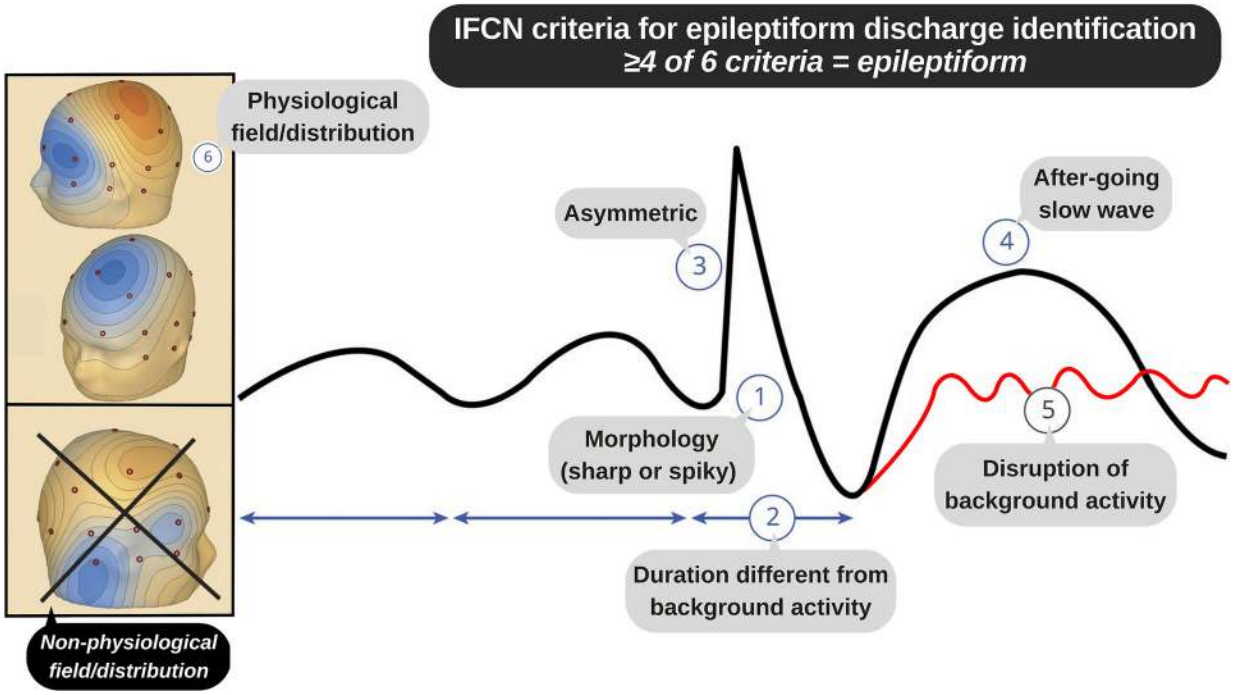

Fig. 1.
The six IFCN criteria for an IED (Nascimento & Beniczky, 2024).

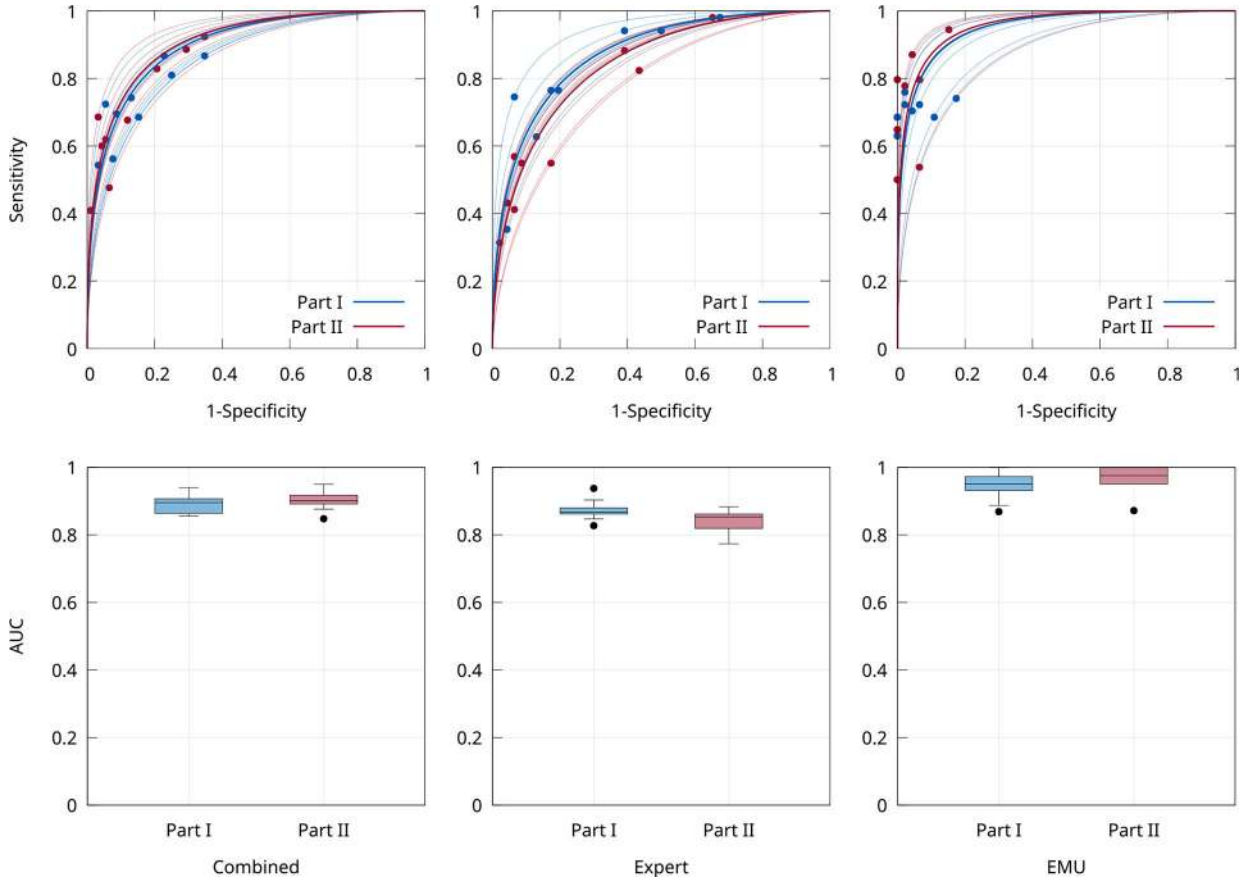

Fig. 2.
Per-rater and overall best-fit ROC curves and AUC values.

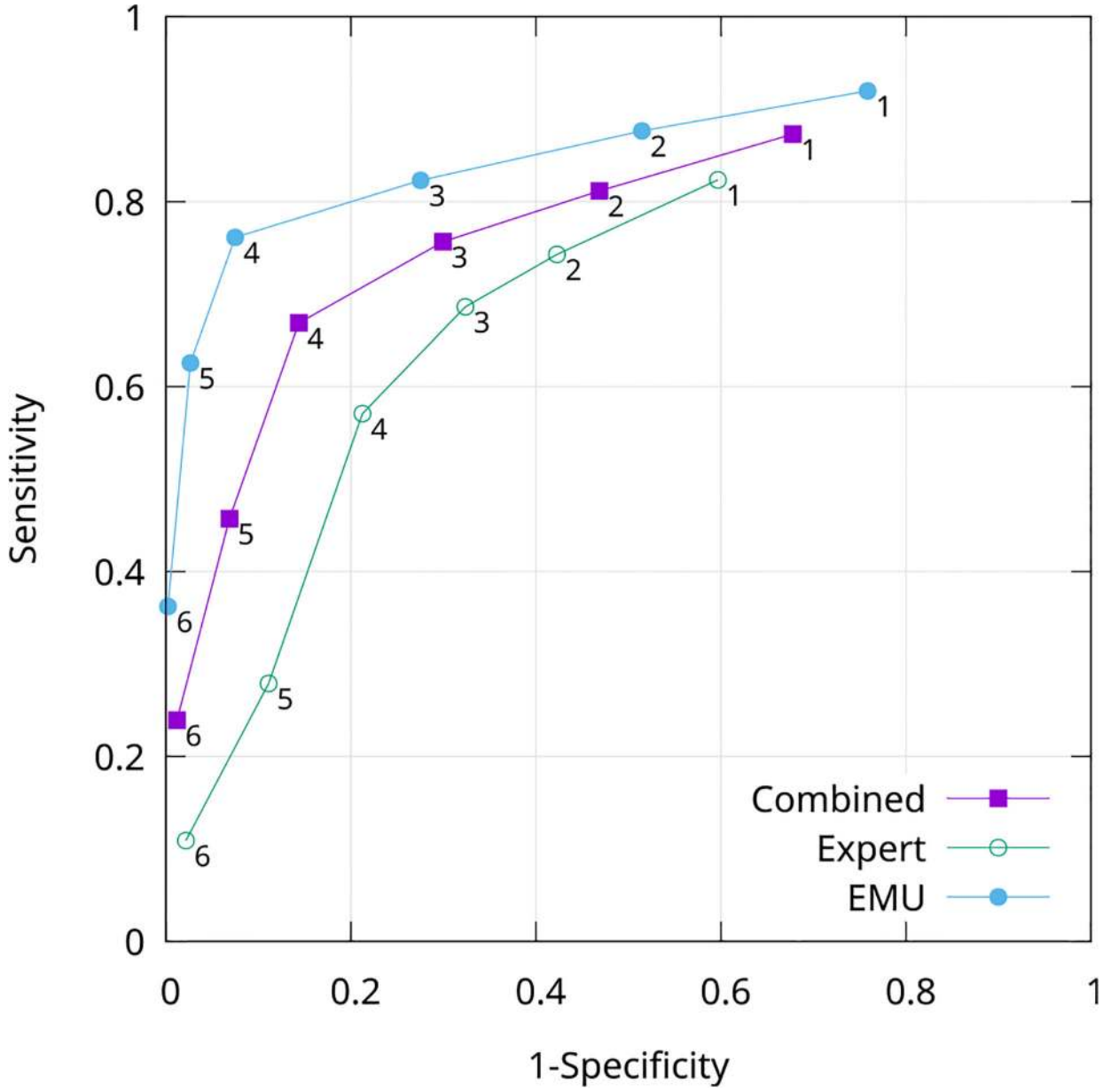

Fig. 3.
ROC curves constructed from number of selected IFCN criteria; comparison of datasets.

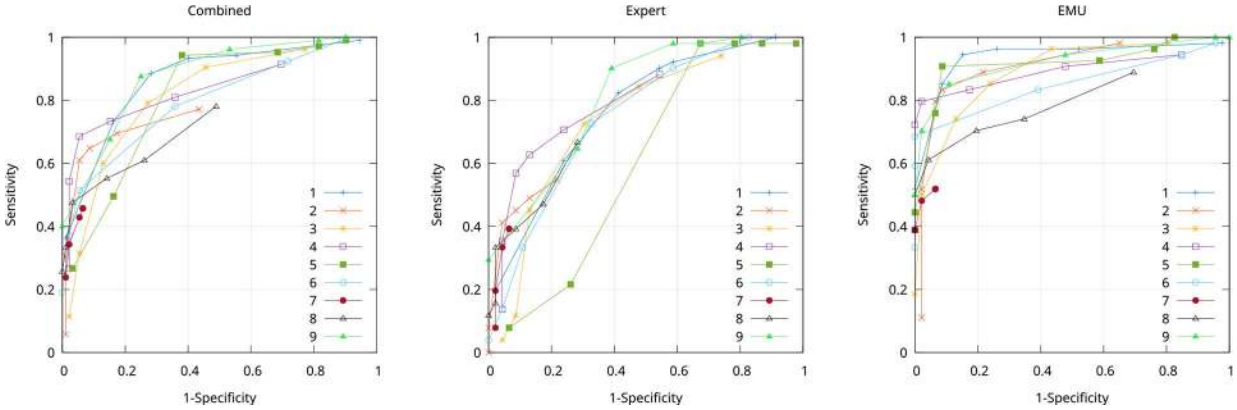

Fig. 4.
ROC curves constructed from number of selected IFCN criteria, per rater, per dataset.

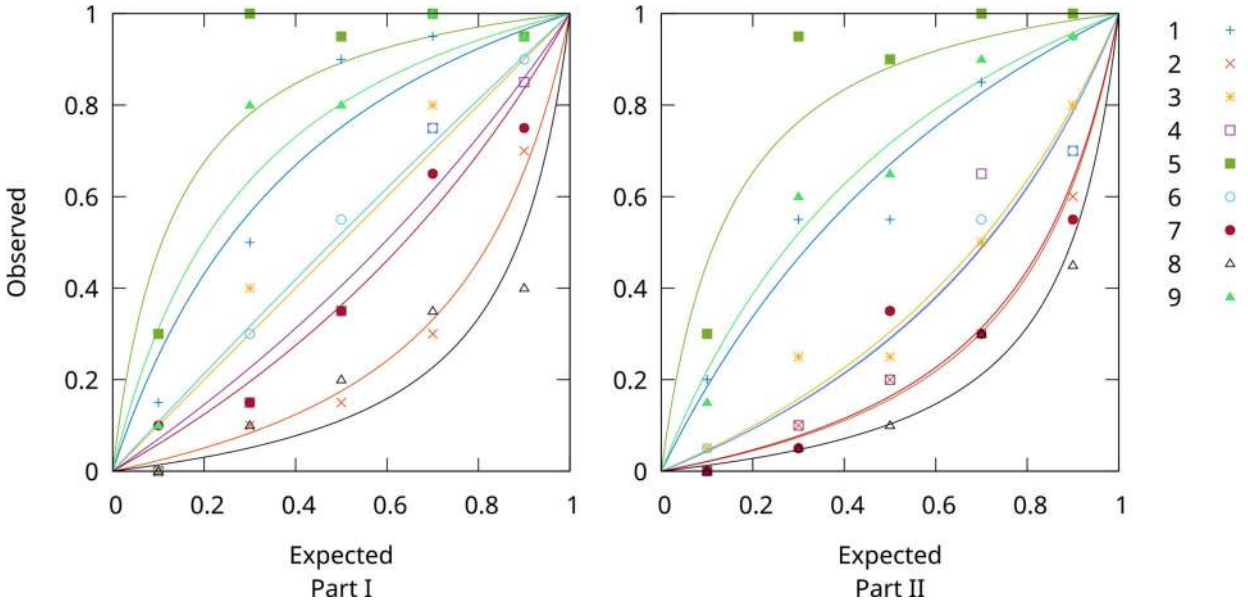

Fig. 5.
Per-rater calibration curves for the expert-consensus dataset.

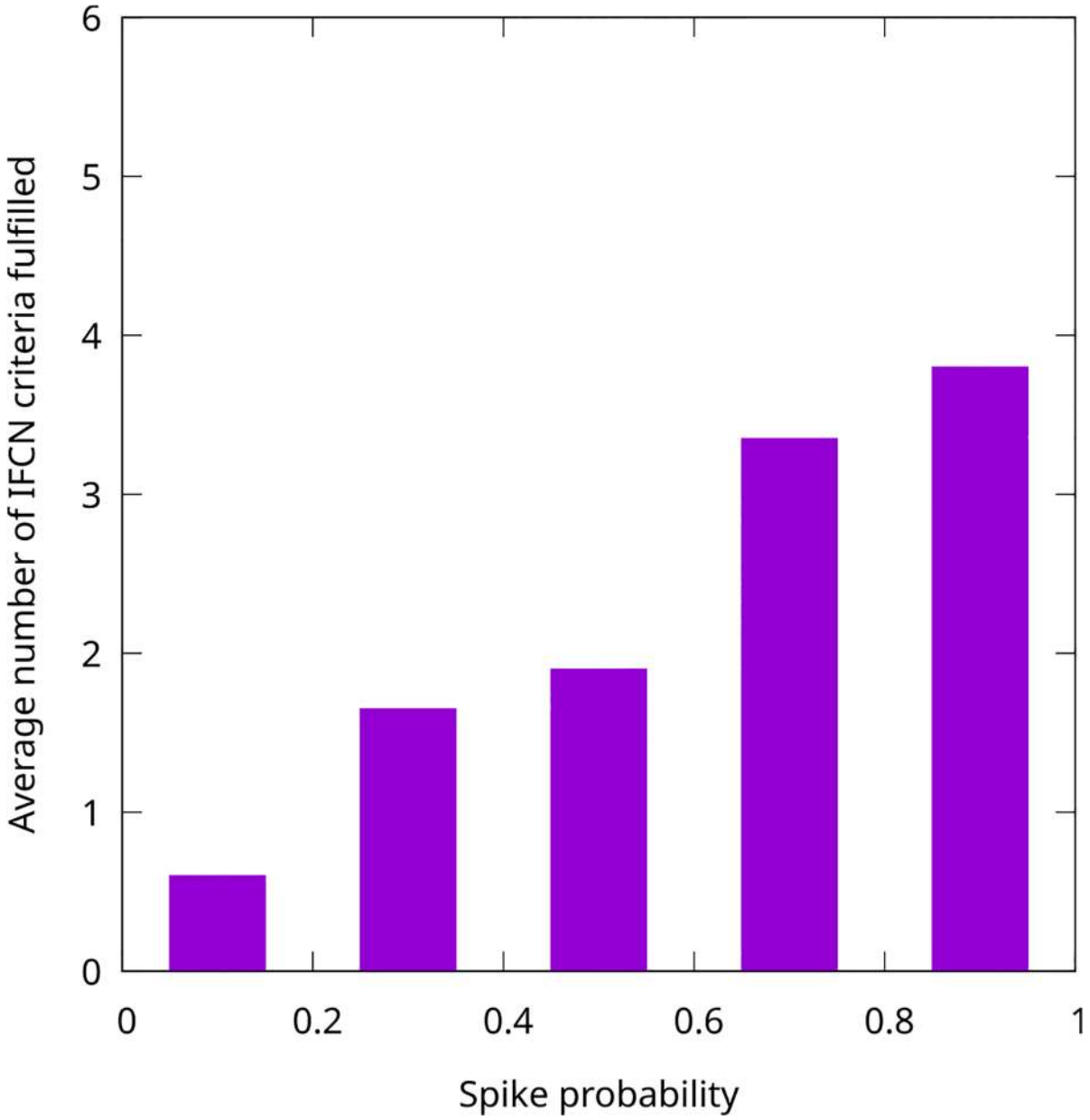

Fig. 6.
Average number of IFCN criteria selected in Part II for each bin of expert-consensus EEGs.

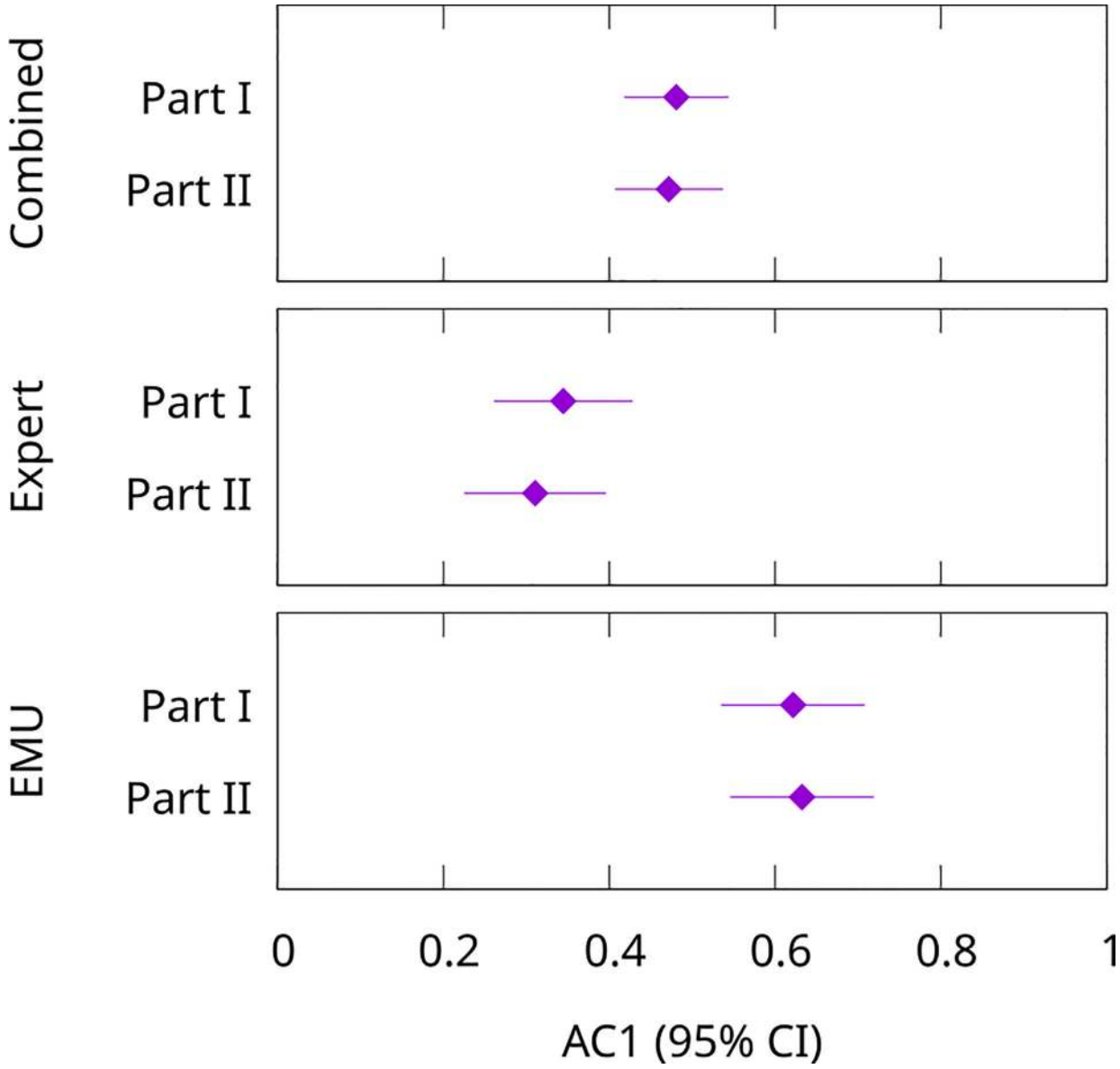

Fig. 7.
Inter-rater agreement on candidate IED classification.

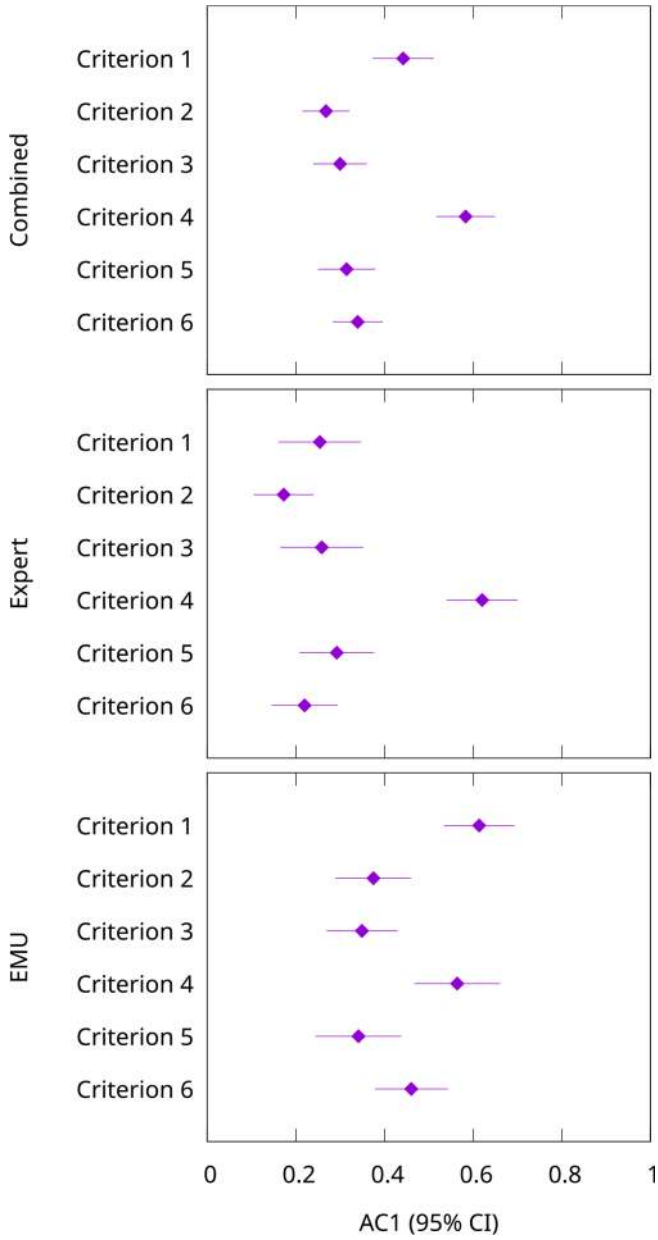

Fig. 8.
Inter-rater agreement on individual IFCN criteria.

Table 1

Demographic information of experts, **sorted by years of experience.**

Years of experience	CNP[a]	Epilepsy[a]
2	N	N
3	N	Y
3	N	Y
7	Y	Y
7	Y	N
10	Y	Y
14	Y	Y
24	N	N
30	Y	N

[a]Board certification status.

Table 2

Performance and inter-rater agreement.

Combined dataset				
	Sensitivity % (95 % CI)	**Specificity % (95 % CI)**	**Accuracy % (95 % CI)**	**AC1 (95 % CI)**
Part I	81.0 (72.4,87.3)	92.4 (85.1,96.3)	86.3 (80.8,90.4)	0.481 (0.418,0.544)
Part II	74.3 (65.2,81.7)	95.7 (89.3,98.3)	84.3 (78.5,88.7)	0.472 (0.407,0.537)
≥4 [a]	71.4 (62.2,79.2)	94.6 (87.9,97.7)	82.2 (76.3,86.9)	0.434 (0.372,0.497)
≥5 [a]	44.8 (35.6,54.3)	98.9 (94.1,99.8)	70.1 (63.3,76.0)	0.589 (0.520,0.659)
=6 [a]	15.2 (9.6,23.3)	100.0 (96.0,100.0)	54.8 (47.8,61.6)	0.784 (0.737,0.831)
Expert-consensus subset				
	Sensitivity % (95 % CI)	**Specificity % (95 % CI)**	**Accuracy % (95 % CI)**	**AC1 (95 % CI)**
Part I	82.4 (69.7,90.4)	84.8 (71.8,92.4)	83.5 (74.9,89.6)	0.344 (0.261,0.428)
Part II	66.7 (53.0,78.0)	93.5 (82.5,97.8)	79.4 (70.3,86.2)	0.311 (0.225,0.396)
≥4 [a]	60.8 (47.1,73.0)	91.3 (79.7,96.6)	75.3 (65.8,82.8)	0.279 (0.190,0.367)
≥5 [a]	15.7 (8.2,28.0)	100.0 (92.3,100.0)	55.7 (45.8,65.2)	0.599 (0.512,0.686)
=6 [a]	0.0 (0.0,7.0)	100.0 (92.3,100.0)	47.4 (37.8,57.3)	0.868 (0.827,0.909)
EMU-validated subset				
	Sensitivity % (95 % CI)	**Specificity % (95 % CI)**	**Accuracy % (95 % CI)**	**AC1 (95 % CI)**
Part I	79.6 (67.1,88.2)	100.0 (92.3,100.0)	89.0 (81.4,93.7)	0.622 (0.535,0.708)
Part II	81.5 (69.2,89.6)	97.8 (88.7,99.6)	89.0 (81.4,93.7)	0.632 (0.546,0.719)
≥4 [a]	81.5 (69.2,89.6)	97.8 (88.7,99.6)	89.0 (81.4,93.7)	0.596 (0.516,0.675)
≥5 [a]	72.2 (59.1,82.4)	97.8 (88.7,99.6)	84.0 (75.6,89.9)	0.595 (0.492,0.699)
=6 [a]	29.6 (19.1,42.8)	100.0 (92.3,100.0)	62.0 (52.2,70.9)	0.682 (0.591,0.773)

[a] Using hypothetical ratings derived from the specified threshold number of IFCN criteria satisfied.

Table 3

Per-rater calibration indices for the expert-consensus dataset.

Rater	Part I	Part II
1	35.5	23.33
2	−47.85	−51.17
3	0	−26.94
4	−12.75	−29.18
5	61.48	59.3
6	2.61	−28.84
7	−18.65	−49.99
8	−60.55	−62.52
9	43.9	30.02

Table 4

Inter-rater agreement on individual IFCN criteria.

Criterion	Combined[a]	Expert-consensus[a]	EMU-validated[a]
1	0.442 (0.373,0.511)	0.254 (0.161,0.347)	0.613 (0.534,0.693)
2	0.268 (0.215,0.321)	0.173 (0.105,0.240)	0.375 (0.289,0.460)
3	0.300 (0.239,0.360)	0.258 (0.165,0.352)	0.349 (0.269,0.429)
4	0.583 (0.517,0.649)	0.620 (0.540,0.700)	0.564 (0.467,0.660)
5	0.314 (0.250,0.379)	0.292 (0.208,0.377)	0.341 (0.244,0.438)
6	0.339 (0.283,0.396)	0.220 (0.145,0.294)	0.460 (0.378,0.543)

[a]All reported values are AC1 with 95% confidence intervals.

Table 5

Intra-rater agreement between Parts I and II.

Rater	Cohen's κ (95 % CI)
1	0.680 (0.577,0.784)
2	0.634 (0.52,0.748)
3	0.605 (0.493,0.716)
4	0.686 (0.583,0.79)
5	0.847 (0.77,0.925)
6	0.674 (0.569,0.778)
7	0.483 (0.365,0.602)
8	0.669 (0.553,0.785)
9	0.718 (0.62,0.815)

Table 6

Summary of survey responses.

Question	Yes responses (%)
Q1. Before completing Part 1, were you familiar with the IFCN operational criteria published in the revised glossary in 2017?	9 (100 %)
Q2. Before starting Part 2, did you inspect the infographic we provided?	8 (88 %)
Q3. Before starting Part 2, did you watch the video, EEG Talk Curriculum - Spike Operational Criteria (Part 1), on YouTube?	4 (44 %)
Q4. Before starting Part 2, did you watch the video, EEG Talk Curriculum - Spike Operational Criteria (Part 2), on YouTube?	2 (22 %)
Q5. Before starting Part 2, did you watch the lecture recording?	2 (22 %)