JAMA Neurology | Original Investigation

Interrater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms

Jin Jing, PhD; Aline Herlopian, MD; Ioannis Karakis, MD; Marcus Ng, MD; Jonathan J. Halford, MD; Alice Lam, MD, PhD; Douglas Maus, MD, PhD; Fonda Chan, MD; Marjan Dolatshahi, MD, MSc; Carlos F. Muniz, MD, MSc; Catherine Chu, MD; Valeria Sacca, MS; Jay Pathmanathan, MD; WenDong Ge, PhD; Haoqi Sun, PhD; Justin Dauwels, PhD; Andrew J. Cole, MD; Daniel B. Hoch, MD; Sydney S. Cash, MD, PhD; M. Brandon Westover, MD, PhD

IMPORTANCE The validity of using electroencephalograms (EEGs) to diagnose epilepsy requires reliable detection of interictal epileptiform discharges (IEDs). Prior interrater reliability (IRR) studies are limited by small samples and selection bias.

OBJECTIVE To assess the reliability of experts in detecting IEDs in routine EEGs.

DESIGN, SETTING, AND PARTICIPANTS This prospective analysis conducted in 2 phases included as participants physicians with at least 1 year of subspecialty training in clinical neurophysiology. In phase 1, 9 experts independently identified candidate IEDs in 991 EEGs (1 expert per EEG) reported in the medical record to contain at least 1 IED, yielding 87 636 candidate IEDs. In phase 2, the candidate IEDs were clustered into groups with distinct morphological features, yielding 12 602 clusters, and a representative candidate IED was selected from each cluster. We added 660 waveforms (11 random samples each from 60 randomly selected EEGs reported as being free of IEDs) as negative controls. Eight experts independently scored all 13 262 candidates as IEDs or non-IEDs. The 1051 EEGs in the study were recorded at the Massachusetts General Hospital between 2012 and 2016.

MAIN OUTCOMES AND MEASURES Primary outcome measures were percentage of agreement (PA) and beyond-chance agreement (Gwet κ) for individual IEDs (IED-wise IRR) and for whether an EEG contained any IEDs (EEG-wise IRR). Secondary outcomes were the correlations between numbers of IEDs marked by experts across cases, calibration of expert scoring to group consensus, and receiver operating characteristic analysis of how well multivariate logistic regression models may account for differences in the IED scoring behavior between experts.

RESULTS Among the 1051 EEGs assessed in the study, 540 (51.4%) were those of females and 511 (48.6%) were those of males. In phase 1, 9 experts each marked potential IEDs in a median of 65 (interquartile range [IQR], 28-332) EEGs. The total number of IED candidates marked was 87 636. Expert IRR for the 13 262 individually annotated IED candidates was fair, with the mean PA being 72.4% (95% CI, 67.0%-77.8%) and mean κ being 48.7% (95% CI, 37.3%-60.1%). The EEG-wise IRR was substantial, with the mean PA being 80.9% (95% CI, 76.2%-85.7%) and mean κ being 69.4% (95% CI, 60.3%-78.5%). A statistical model based on waveform morphological features, when provided with individualized thresholds, explained the median binary scores of all experts with a high degree of accuracy of 80% (range, 73%-88%).

CONCLUSIONS AND RELEVANCE This study's findings suggest that experts can identify whether EEGs contain IEDs with substantial reliability. Lower reliability regarding individual IEDs may be largely explained by various experts applying different thresholds to a common underlying statistical model.

JAMA Neurol. 2020;77(1):49-57. doi:10.1001/jamaneurol.2019.3531
Published online October 21, 2019. Corrected on January 13, 2020.

Author Audio Interview
Related article page 103
Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: M. Brandon Westover, MD, PhD, Division of Clinical Neurophysiology, Department of Neurology, Massachusetts General Hospital, 55 Fruit St, Boston, MA 02114 (mwestover@mgh.harvard.edu).

Detecting interictal epileptiform discharges (IEDs) in electroencephalograms (EEGs) is a fundamental part of evaluating patients with suspected epilepsy.[1-3] Identifying IEDs helps explain recurrence following a first seizure,[1,2] classify epilepsy type,[4] localize ictal onset,[5,6] and manage anticonvulsants.[4,7-9] Identification is challenging because IEDs' morphologies vary and can resemble waves in normal background activity (eg, vertex waves in sleep) or artifacts (eg, extracerebral potentials from muscle, eyes, or heart).[10] Mistakes in recognizing IEDs are common and consequential among neurologists without subspecialty training.[11-15] False-negative findings delay treatment, whereas false-positive findings lead to inappropriate treatment and delay diagnosis of other disorders.[11]

How reliably specialists recognize IEDs is unknown. Several small studies report that interrater reliability (IRR) for IED detection may be poor among experts.[16-31] Those studies were based on small numbers of patients, IEDs, and experts and focused on specially selected sets of patients, leaving the reliability of EEG as a diagnostic test uncertain.

A definitive study of expert IRR regarding IEDs is a prerequisite for developing automated IED detection systems because expert identification represents the accepted criterion standard. Automated detection software holds promise for extending the reach of epilepsy diagnostic testing beyond the relatively small pool of experts with EEG subspecialty training and for expanding epilepsy care to underserved areas where epilepsy remains largely underdiagnosed and undertreated.[32-34] Although automated IED detection software is commercially available,[25] it is unclear how well these software systems compare with human experts.[35]

Therefore, we performed a large study to assess expert IRR for identifying IEDs. First, we measured the reliability of clinical neurophysiology experts in scoring IEDs using many EEGs, IEDs, and experts. Second, we studied factors underlying expert IRR. For this, we used logistic regression models to investigate how well morphological wave features account for expert IRR. Third, we hypothesized that, when experts disagree, they do not use different implicit models but rather apply different thresholds to the same underlying model. To test this hypothesis, we measured how well binary scores of individual experts are explained by a single model with individualized thresholds.

Key Points

Question What is the reliability of subspecialty-trained clinical neurophysiologists in detecting interictal epileptiform discharges in routine electroencephalograms?

Findings In this multicenter trial, 8 experts independently annotated 13 262 candidate interictal epileptiform discharges. Interrater reliability for individual interictal epileptiform discharges was fair (κ = 48.7), whereas that for whether a given electroencephalogram contained any interictal epileptiform discharges was excellent (κ = 69.4).

Meaning This study's findings suggest that experts can identify electroencephalograms containing interictal epileptiform discharges with substantial reliability and that disagreements about individual interictal epileptiform discharges can be largely explained by various experts applying different thresholds to a common underlying statistical model.

Methods

Study Design and Patients

We conducted the study following the Standards for Reporting of Diagnostic Accuracy (STARD) guidelines.[36] The index test was the independent EEG interpretation by neurologists with clinical electroencephalography fellowship training (hereinafter, "experts"). The reference standard is consensus EEG interpretation by 8 independent experts, all of whom have received at least 1 year of fellowship training in clinical neurophysiology. The study was conducted prospectively; the data collection and analytical methods were specified before the index test and the reference standards were assessed. The institutional review board at the Massachusetts General Hospital, Boston, approved the study and, because the study was considered to pose no risk to patients, waived the requirement for informed consent.

We selected 991 consecutive, noninvasive scalp EEG recordings, performed between 2012 and 2016 at the Massachusetts General Hospital, in which the medical record described one or more IEDs. Another 60 EEGs with no reported IEDs from the same period were randomly selected to serve as controls. Most EEGs were 30 to 60 minutes long, and those longer were clipped to 60 minutes. Electroencephalograms were performed in both inpatient and outpatient settings (**Table 1**). Electroencephalogram electrodes were placed according to the International 10-20 system.[37] After selecting EEGs, we performed our study in the following 2 phases (**Figure 1**).

Phase 1

Each of the 991 EEGs with potential IEDs was assigned on a first come, first served basis to 1 of 9 experts to identify candidate IEDs. When in doubt about a wave, the experts were instructed to include it for evaluation in phase 2 to encourage inclusion in the data of cerebral (eg, wicket spikes) and extracerebral (eg, lateral rectus spikes/movement and electrode pop artifacts, etc) transients that are occasionally confused with epileptiform discharges. Reviewers were instructed to annotate at least 100 IEDs or all present in cases containing fewer than 100 IEDs. Annotations were performed by using customized software, which allowed reviewers to adjust the gain and view the data using different montages.[38]

Phase 2

Interictal epileptiform discharge candidates from phase 1 were further independently annotated by 8 experts in phase 2. The number of experts was based on prior work[12,23,24] suggesting that 7 to 8 is the minimum number required. It was infeasible for experts to individually label all candidate IEDs from phase 1. However, many waves had nearly identical morphological features; therefore, we used a clustering method (eAppendix 1 in the Supplement) to reduce the collection into a smaller number of morphologically distinct candidate waveforms. We

Table 1. Patient and EEG Characteristics

	No. (%)				
Age, y	Total	Female	EEG	EMU	ICU
0-1	30 (2.85)	16 (53.33)	30 (100.00)	0	0
1-5	84 (7.99)	41 (48.81)	80 (95.24)	3 (3.57)	1 (1.19)
5-13	267 (25.40)	131 (49.06)	262 (98.13)	3 (1.12)	2 (0.75)
13-18	117 (11.13)	56 (47.86)	114 (97.44)	3 (2.56)	0
18-30	121 (11.51)	73 (60.33)	114 (94.21)	5 (4.13)	2 (1.65)
30-50	107 (10.18)	54 (50.47)	100 (93.46)	4 (3.74)	3 (2.80)
50-65	133 (12.65)	58 (43.61)	122 (91.73)	3 (2.26)	8 (6.02)
65-75	98 (9.32)	62 (63.27)	89 (90.82)	0	9 (9.18)
≥75	94 (8.94)	49 (52.13)	85 (90.43)	2 (2.13)	7 (7.45)
Total	1051 (100)	540 (51.38)	996 (94.77)	23 (2.19)	32 (3.04)

Abbreviations: EEG, electroencephalogram; EMU, epilepsy monitoring unit; ICU, intensive care unit.

also included 660 non-IED waves selected randomly from the 60 control EEGs as *catch trials* (11 from each EEG). Each expert independently reviewed all candidate IEDs, voting yes (IED) or no (non-IED) for each event. Review was performed using custom software manufactured in-house, NeuroBrowser, version 42 (eFigure 1 in the Supplement).[38]

Outcomes

The primary outcome measures were the mean percentage of agreement (PA) and beyond-chance agreement (Gwet κ) of pairs of experts in classifying EEG events as IEDs vs non-IEDs and classifying entire EEGs as containing IEDs or not. Secondary outcome measures were correlations between numbers of IEDs marked by experts across cases, calibration of individual expert scoring behavior to the group consensus, and receiver operating characteristic analysis of how well multivariate logistic regression models evaluate expert IED scoring behavior.

Statistical Analysis

Interrater Reliability

We measured beyond-chance agreement using the Gwet κ statistic,[39,40] which is calculated by estimating the percentage of agreement attributable to chance (PC), subtracting it from the observed percent agreement (PA), and dividing by the maximum possible beyond-chance agreement (1 − PC) as follows:

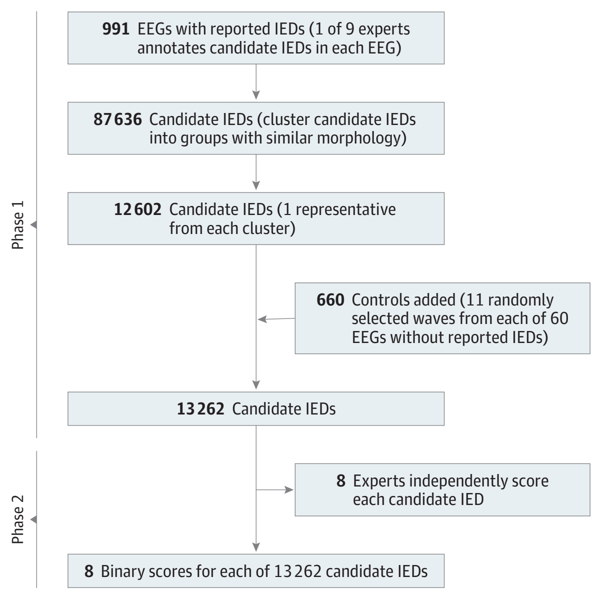
$$\kappa = (PA - PC)/(1 - PC).$$

We carried out IRR analysis on 2 levels: IED-wise and EEG-wise. For IED-wise analysis, we directly used the binary label for each spike provied by each expert. For EEG-wise analysis, we considered an EEG to be marked as containing IEDs if an expert marked at least 1 event in that EEG. We adopted standard conventions[41] for describing the strength of κ as follows: 0 to 0.20, slight; 0.21 to 0.40, fair; 0.41 to 0.60, moderate; 0.61 to 0.80, substantial; and 0.81 to 1.00, almost perfect.

Statistical Calibration

We compared scorers in terms of statistical calibration for IED-wise and EEG-wise labeling.[42,43] Calibration measures how well predicted probabilities agree with event frequencies. We defined the observed probability of each candidate IED as the proportion of experts who scored it as an IED; thus, each wave

Figure 1. Flow Diagram of the Study



EEG indicates electroencephalogram; IEDs, interictal epileptiform discharges.

was assigned to 1 of 9 probability bins (0, 1/8, ..., or 8/8) (eFigures 2-4 in the Supplement). We defined each expert's predicted probability for each bin as the proportion of candidate waves in that bin that the expert scored as IEDs. This method defines a calibration curve for each expert (y = predicted probability, x = observed probability), and the expert's calibration score is the mean absolute value of the difference between predicted and observed probabilities.

Analysis of Morphological Features Underlying Expert IRR

To investigate factors underlying expert IRR, we evaluated morphological features. These features were calculated by first identifying 5 fiducial points within the candidate wave, including a start point, peak, trough, slow-wave peak, and end point (eAppendix 2 in the Supplement). We then extracted 23 morphological features falling into 1 of the following 5 groups: (1) voltages, (2) durations, (3) slopes, (4) areas, and (5) across-channel correlation.

In univariate analysis, we investigated how well individual morphological features correlated with expert IRR by using the Spearman rank correlation coefficient.

In multivariate analysis, we first fit a single universal multivariate logistic regression (MLR) model to evaluate the binary IED scores of all 8 experts. Feature selection was accomplished using L1 regularization and 10-fold internal cross-validation. We reported the area under the receiver operating curve, estimated using 10-fold external cross-validation to provide an unbiased estimate of model performance.

To investigate which factors account for how experts score IEDs, we performed 2 sets of analyses. First, we identified individualized thresholds that, when applied to the universal model, best explain each expert's binary IED scores. We compared the accuracy of the universal model with individualized MLR models fit to each expert's scores. We then compared the accuracies of the 2 approaches by using statistical significance testing.

For all statistics, we estimated 95% CIs and 2-sided *P* values using 1000 rounds of bootstrapping. Statistical significance testing was performed using $\alpha = .05$. Analyses were performed using MATLAB, version R2018a (MathWorks). The statistical analysis was conducted from January 1, 2019, to January 31, 2019.

Results

Among the 1051 EEGs assessed in the study, 540 (51.4%) were those of females and 511 (48.6%) were those of males. In phase 1, 9 experts each marked potential IEDs in a median of 65 (interquartile range, 28-332) EEGs totaling 991 EEGs and 633 hours of data. The total number of IED candidates marked was 87 636. Clustering reduced the total to 12 602 morphologically distinct waves. To these we added 660 control waveforms (11 randomly selected waveforms from 60 cases reported to contain no IEDs) totaling 13 262 candidate IEDs. In phase 2, 8 experts independently assigned binary scores to all 13 262 candidates. Table 1 summarizes the demographics for the study cohort.

Figure 2A shows representative candidate IEDs arranged by number of expert votes received. Although the experts viewed the events in context (all channels, 10 seconds of EEG data), the examples include only a half-second from a single channel to allow many waves to be displayed together. Nevertheless, qualitative trends were evident. Events appeared more likely to be scored as IEDs when they were spikier, asymmetric, and/or included an after-going slow wave.

The mean PA for IEDs between pairs of experts was 72.4% (95% CI, 67.0%-77.8%), and the mean κ was 48.7% (95% CI, 37.3%-60.1%) (Figure 2B). For presence of IEDs in whole EEGs, the mean PA was 80.9% (95% CI, 76.2%-85.7%), and the mean κ was 69.4% (95% CI, 60.3%-78.5%) (Figure 2C).

Relative numbers of IEDs marked by experts were strongly correlated across cases (Figure 2D). The mean pairwise correlation between IED numbers was 0.96 (range, 0.86-0.99). Although rates for marking IEDs differ, experts tended to mark more or fewer IEDs in the same cases.

Calibration curves for expert IED detection are shown in Figure 2E and for whole EEGs in Figure 2F. Experts 1, 2, 3, and 5 showed good calibration, with curves close to the diagonal, whereas experts 6 and 7 tended to "overcall" and experts 4 and 8 tended to "undercall" relative to the group. The mean calibration score across all experts for IEDs was 0.18 (range, 0.08-0.36) and for whole EEGs, 0.20 (range, 0.10-0.37).

To investigate factors underlying expert IRR, we computed 23 morphological measures and correlated them with the proportion of experts who scored waves as IEDs (**Figure 3**A). The features that best correlated with the tendency of experts to score waves as IEDs were the slopes of the falling and rising phase of the half wave and 2 of the peak-to-peak voltage measurements (eTable in the Supplement and Figure 3B).

We hypothesized that disagreements about IEDs may be explained largely by experts applying individualized thresholds to the same underlying probabilistic model. To test this hypothesis, we fit a universal MLR model to evaluate scores of all experts combined by using the 23 morphological features, of which 10 were retained by the model-fitting procedure. The model assigned a probability for each waveform, converted to a binary score, by comparing it with a threshold. We identified individualized thresholds that maximized the accuracy with which the universal model explained each expert's binary scores. The universal model evaluated individual expert scores well, with a median accuracy of 80% (range, 73%-88%).

To assess whether individualized models may better account for the data than the universal model, we separated MLR models to each expert's labels and identified optimal thresholds for each. The individualized models evaluated experts' scores more accurately than did the universal model. Accuracy was identical to the universal model for all experts up to 2 decimal places. We concluded that IED scoring behavior of individual experts may be parsimoniously explained by a single model in which experts generally agree on the underlying probability but apply different thresholds that are more or less conservative to produce a binary yes or no score for each candidate IED.

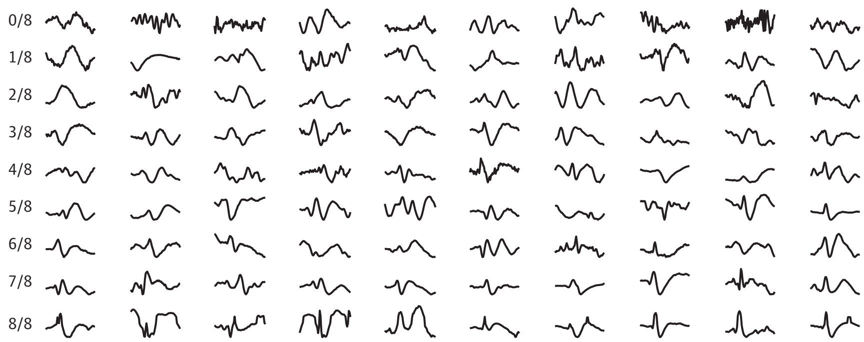
Discussion

Our findings suggest that expert reliability in scoring individual IEDs was fair (PA, 72.4%; κ, 48.7%). However, agreement regarding whether an EEG contained any IEDs was substantially higher (PA, 80.9%; κ, 69.4%). Moreover, differences between expert IED scoring may be largely explained by a single underlying probabilistic model based on a small number of morphological features but with individual experts applying more or less stringent thresholds to arrive at their binary decisions. Our results establish robust estimates for the reliability of experts for identifying IEDs in routine EEG recordings and provide a basis for evaluating automated IED detection systems.

Prior studies of expert IED detection reliability dating to the 1970s have generally concluded that expert IRR for IEDs is poor. These studies were limited by small samples and meth-

Figure 2. Interrater Reliability (IRR) for Interictal Epileptiform Discharges (IEDs) at the Level of Individual IEDs and Entire Electroencephalograms (EEGs)

A Candidate IEDs arranged by No. of expert votes



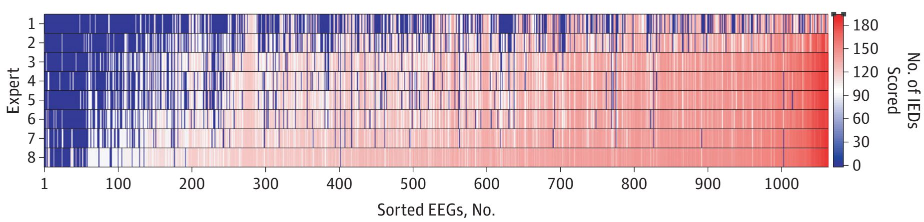
B IED-wise percentage of agreement

	E1	E2	E3	E4	E5	E6	E7	E8
E1	100	75.7	58.0	56.6	55.0	54.1	46.4	37.6
E2	75.7	100	72.3	70.6	69.3	69.3	62.8	55.5
E3	58.0	72.3	100	85.0	84.8	84.7	82.7	77.1
E4	56.6	70.6	85.0	100	86.2	83.8	82.8	78.3
E5	55.0	69.3	84.8	86.2	100	84.5	83.4	79.0
E6	54.1	69.3	84.7	83.8	84.5	100	84.2	80.5
E7	46.4	62.8	82.7	82.8	83.4	84.2	100	86.8
E8	37.6	55.5	77.1	78.3	79.0	80.5	86.8	100

C EEG-wise percentage of agreement

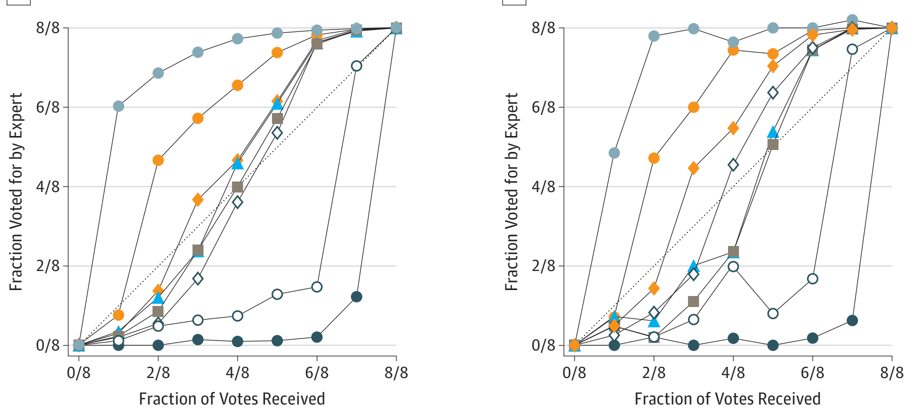
	E1	E2	E3	E4	E5	E6	E7	E8
E1	100	75.5	63.2	65.6	64.8	60.7	57.1	52.9
E2	75.5	100	80.5	82.3	81.4	78.9	76.7	73.2
E3	63.2	80.5	100	91.1	90.7	92.2	90.7	88.3
E4	65.6	82.3	91.1	100	91.9	89.7	88.7	86.2
E5	64.8	81.4	90.7	91.9	100	90.2	88.5	86.7
E6	60.7	78.9	92.2	89.7	90.2	100	92.3	90.9
E7	57.1	76.7	90.7	88.7	88.5	92.3	100	94.6
E8	52.9	73.2	88.3	86.2	86.7	90.9	94.6	100

D No. of IEDs marked by each of the 8 experts in each EEG



E Calibration: IED-wise

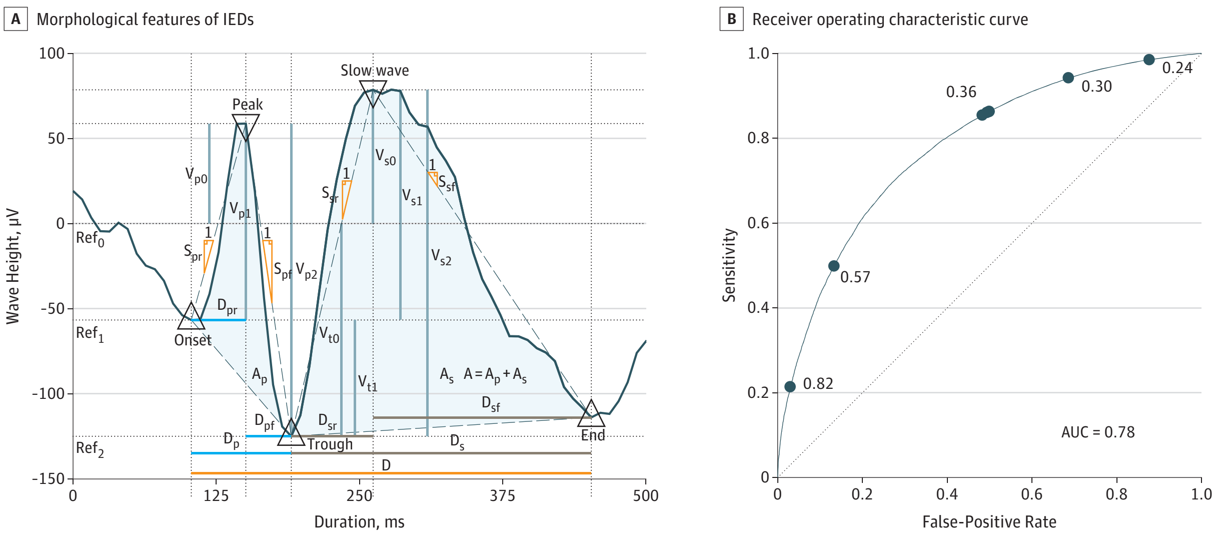
F Calibration: EEG-wise



A, The 8 rows each contain 10 randomly selected samples scored by 8 experts as being IEDs. Qualitatively, expert IRR increases in proportion to the degree that candidate waves exhibit the morphological features of IEDs as defined by IFSECN (International Federation of Societies for EEG and Clinical Neurophysiology) criteria. B, The IED-wise percentage of agreement between pairs of experts. C, The EEG-wise percentage of agreement among pairs of experts. D, The EEGs are arranged in order of the mean number of IEDs marked by all experts, and experts are arranged from top to bottom in order of the total number of IEDs they marked across all EEGs. E and F, The IEDs were binned according to the number of votes (0 through 8) that they received (termed the *reference standard*). The 8 calibration curves, 1 for each expert, indicate the probability of that expert marking events within a given bin as IEDs. These curves allow assessment of the variation among experts relative to the group consensus.

odological limitations (**Table 2**).[16-31] Hostetler et al[17] studied IRR among 5 experts who scored 20-minute EEGs from 5 patients and found unanimous agreement for only 18% of candidate IEDs. Webber et al[20] studied IRR among 8 experts on

Figure 3. Morphological Characteristics of the Interictal Epileptiform Discharges (IEDs)



A, For each candidate IED, 5 fiducial points are identified (triangles) corresponding to the IED peak, troughs preceding and following the peak, peak of the after-going slow wave, and trough following the slow-wave peak. These feature values are used to construct a single multivariate logistic regression (MLR) model to evaluate binary IED scores for all experts combined (combined model) and individualized MLR models to evaluate the IED scores of individual experts (individualized models). B, Receiver operating characteristic curve for the MLR model fit to the scores of all experts (universal model). The operating point (false-positive rate = 1 − specificity and sensitivity) of each expert, corresponding to the threshold in the combined model that best evaluates that expert's binary scores, is indicated by a solid circle. The number by each circle is the threshold that best evaluates that expert's binary scores. A indicates the area under the IED curve; A_p, area under the peak; A_s, area under the slow wave; AUC, area under the curve; D, duration of the IED candidate wave; D_p, duration of the peak; D_{pf}, duration of the falling half-wave of peak; D_{pr}, duration of the rising half-wave of peak; D_s, duration of the slow wave; D_{sf}, duration of the falling half-wave of the slow wave; D_{sr}, duration of the rising half-wave of the slow wave; Ref, reference; S_{pf}, slope of the falling half-wave of peak; S_{sf}, slope of the falling half-wave of the slow wave; S_{pr}, slope of the rising half-wave of peak; S_{sr}, slope of the rising half-wave of the slow wave; V_{p0}, peak voltage with respect to baseline 0 defined by V = 0; V_{p1}, peak voltage with respect to baseline 1 defined by V_{on}; V_{p2}, peak voltage with respect to baseline 2 defined by trough V_{tr}; V_{s0}, slow-wave peak voltage with respect to baseline 0 defined by $V = 0$; V_{s1}, slow-wave peak voltage with respect to baseline 1 defined by onset V_{on}; V_{s2}, slow-wave peak voltage with respect to baseline 2 defined by trough V_{tr}; V_{t0}, trough voltage with respect to baseline 0 defined by $V = 0$; and V_{t1}, trough voltage with respect to baseline 1 defined by onset V_{on}.

12 recordings of 3- to 5-minutes duration and found a mean of 52% pairwise agreement between readers. The probability that any single expert marked an event increased with the number of other readers marking the event. Wilson et al[16] studied IRR among 5 experts in 50 EEGs lasting 0.25 to 12 minutes and found 68% sensitivity in pairwise comparisons. Halford et al[24] studied 11 readers marking 30-second EEG samples from 100 patients and found a modest IRR (κ = 0.43) for IEDs. More recently, Bagheri et al[23] studied IED scoring among 18 experts on 30-second EEG samples from 200 patients and found an overall chance-corrected agreement of κ = 36%, although a subset of scorers had higher IRR.

Prior studies have involved carefully selected, mostly brief EEGs, biasing estimates of expert IRR. A partial exception is the study by Scheuer et al,[25] which involved a complete review of relatively long recordings, albeit by EEG technologists rather than by fellowship-trained clinical neurophysiologists. However, the study still involved a small number of patients,[41] most (88%) of whom were undergoing evaluation in an epilepsy monitoring unit. Thus, prior studies leave the measurement of expert reliability in the general clinical setting open to questions regarding systematic and random error.

Our results argue that much of the disagreement between experts over scoring of IEDs arises from the requirement to make binary decisions rather than assign probabilities. This is clear from our universal MLR model, which, using a small number of morphological features, was able to accurately evaluate binary scores of all 8 experts with high accuracy (80%). The same model was able, via thresholds calibrated to each expert, to accurately (73%-88%) evaluate each individual expert's binary IED scores. These findings are similar in spirit to earlier findings of Wilson et al,[16] which modeled expert IED perception (tendency to mark a given wave as an IED) by using an additive combination of simple morphological features.

An underlying agreement about the probabilities of IEDs is also evident from our finding that experts exhibited substantially higher reliability in investigating whether a given EEG contained IEDs than in assessing whether any single wave was an IED. This finding is important because, in clinical practice, experts typically provide an overall impression rather than count individual IEDs. Prior published studies of overall impressions of IEDs, although limited, have reported similar findings. Studies by Gotman et al[19] and Gotman and Wang[44] found 72% agreement between 2 raters on 100 EEGs. Agreement on overall impression was 85% in Houfek and Ellingson[45] and 92% in Struve et al.[46] Webber et al[20] also reported that EEG-wise agreement was higher than IED-wise agreement, although the results were not quantified. One case suggested by some authors[25] as an exception is the study of Black et al,[27]

Table 2. Prior Studies of Expert Interrater Reliability for Detecting Epileptiform Discharges

Source	No. Patients	No. Total Duration, h	No. IED Candidates	No. Experts	Types of Readers	Pairwise Agreement, %	Agreed by All to Be IED, %	κ Value, %	Type of Patients
Ehrenberg and Penry,[18] 1976	7	144	1447	3	CNP	NC	42.1	NC	Generalized epilepsy
Gotman et al,[19] 1978	110	4.3	2630	2	CNP	NC	72, 84[a]	NC	Normal, brain lesion, epilepsy
Whisler et al,[28] 1982	6	36	769	3	CNP	NC	48	NC	Generalized epilepsy
Guedes de Oliveira et al,[26] 1983	10	0.1	NC	8	CNP	NC	"Poor"	NC	IEDs and normal variants
Hostetler et al,[17] 1992	5	1.7	1626	6	CNP	NC	18	NC	Focal and generalized IEDs
Webber et al,[20] 1993	10	~ 1	1739	8	CNP	52	18	NC	EMU patients
Wilson et al,[16] 1996	50	~ 4.1	1952	5	CNP	68	NC	NC	Mixed cohort
Black et al,[27] 2000	106[b]	173	NC	3	CNP	NC	39[c]	NC	Focal and generalized IEDs
Stroink et al,[21] 2006	93	~ 47[d]	NC	2	CNP	NC	82[c]	63	Children with epilepsy
Nonclercq et al,[22] 2009	3	1.4	2500	3	CNP	NC	NC	80[e]	Children with ESES
Halford et al,[30] 2011	40	0.2	828	7	CNP	NC	NC	58	Adults with temporal lobe epilepsy, controls (1:1)
Halford et al,[24] 2013	100	0.8	2571	11	CNP	NC	NC	43	Adults with epilepsy
Scheuer et al,[25] 2017	40	253	5474	3	EEG technologists	45[f]	13.2	NC	EMU patients with epilepsy
Halford et al,[12] 2017	200	1.7	235	18	CNP	NC	NC	36	Normal, "difficult," benign variants (1:1:1)
Halford et al,[31] 2018	200	1.7	573	35	CNP[g]	NC	NC	81	Normal, "difficult," benign variants (1:1:1)

Abbreviations: CNP, clinical neurophysiology; EEG, electroencephalogram; EMU, epilepsy monitoring unit; ESES, electrical status epilepticus in sleep; IEDs, interictal epileptiform discharges; NC, not calculated (in original publication).

[a] The EEG-wise agreement (presence vs absence of any spikes in EEG) rather than spikewise agreement. Seventy-two percent is for review of EEG recordings on paper; 84% is for computer displays.

[b] The study included 521 EEGs, but only 106 were scored by all 3 scorers.

[c] The EEG-wise agreement (presence vs absence of any spikes in EEG) rather than spikewise agreement. The EEG-wise agreement was 85% when normal EEGs were included.

[d] Estimated assuming EEGs were 30 minutes (exact numbers were not reported by the authors).

[e] This κ statistic is calculated for the spike-wave index in children with ESES. This study involved several groups of EEGs from patients with ESES that were analyzed for different purposes. Interrater reliability results for IEDs are shown for the largest group.

[f] Agreement expressed as average pairwise sensitivity, that is, proportion of IEDs marked by one scorer that were marked by the other.

[g] All participants were neurologists, although only 27 of the 35 had received any specialty training in clinical neurophysiology.

which found only 39% EEG-wise agreement on 106 EEGs reviewed by 3 experts and 55% on 415 recordings reviewed by 2 experts. However, that analysis excluded EEGs with unanimous agreement that IEDs were absent. When those cases were included, thereby more closely reflecting clinical practice, unanimous agreement rose to 85% for EEGs read by 3 readers and 89% for those read by 2 readers.

Limitations

Our study has limitations. First, although EEGs were scored independently by 11 fellowship-trained experts, all except 2 experts trained at the same institution. Second, the optimal number of experts is unknown. Halford et al[12] and Bagheri et al[23] analyzed results from 35 scorers and concluded that 7 to 8 scorers is optimal and that IRR is significantly higher among neurologists with formal clinical neurophysiology fellowship training, as in our study. Third, in phase 2, our 8 experts reviewed examples selected in phase 1 rather than entire EEGs. Without this simplification, it would not have been feasible for 8 experts to score 991 EEGs. Nevertheless, we believe this modification does not substantially alter our results since, in practice, experts screen most of the EEG background quickly and spend most review time deliberating about waves that are suspicious for being IEDs,[47] similar to our study. Fourth, experts did not have access to patients' age. It is possible that making this information available may alter interpretation of some IEDs, particularly in pediatric patients. Fifth, we asked experts to score IEDs in a binary manner. An ordinal scale (eg, reporting confidence, as in Wilson et al[16]) may have provided more information per IED. However, that method would have necessitated experts scoring a smaller number of IEDs and would have been a departure from clinical practice. We believed it advantageous to score a larger set of candidate IEDs. Because of the size of our data set, we were able to perform calibration analysis and to provide nonbinary information.

Conclusions

Although scoring reliability for individual IEDs is limited, experts behave as if applying a common model to estimate the probability that a given waveform is an IED. Differences between experts' binary scores are largely attributable to different thresholds when making probabilistic assessments. Moreover, overall impressions regarding the presence or absence of IEDs in an EEG show substantial reliability. Our results establish precise estimates based on a large and unbiased sample of experts' IED-wise and EEG-wise reliability for IED detection. These results support the practice of expert interpretation of IEDs in routine EEGs for diagnosing and treating patients with established or suspected epilepsy. The results also present a standard for how well an automated IED detection system must perform to be considered comparable in skill to a human expert.

ARTICLE INFORMATION

Accepted for Publication: August 11, 2019.

Published Online: October 21, 2019. doi:10.1001/jamaneurol.2019.3531

Correction: This article was corrected on January 13, 2020, to add the middle initial and second degree to the name of the 10th author in the byline.

Author Affiliations: Division of Clinical Neurophysiology, Department of Neurology, Massachusetts General Hospital, Boston (Jing, Herlopian, Lam, Maus, Chan, Dolatshahi, Muniz, Chu, Pathmanathan, Ge, Sun, Cole, Hoch, Cash, Westover); School of Electrical and Electronics Engineering, Nanyang Technological University, Singapore (Jing, Dauwels); Department of Neurology, Yale School of Medicine, New Haven, Connecticut (Herlopian); Department of Neurology, Emory University School of Medicine, Atlanta, Georgia (Karakis); Department of Neurology, University of Manitoba, Winnipeg, Manitoba, Canada (Ng); Department of Neurology, Medical University of South Carolina, Charleston (Halford); Department of Neurology, Department of Medical and Surgical Sciences, University "Magna Graecia" of Catanzaro, Italy (Sacca); Department of Neurology, Hospital of the University of Pennsylvania, Philadelphia (Pathmanathan).

Author Contributions: Drs Jing and Herlopian are considered co–first authors. Drs Jing and Westover had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Jing, Herlopian, Dauwels, Cash, Westover.
Acquisition, analysis, or interpretation of data: Jing, Herlopian, Karakis, Ng, Halford, Lam, Maus, Chan, Dolatshahi, Muniz, Chu, Sacca, Pathmanathan, Ge, Sun, Cole, Hoch, Cash, Westover.
Drafting of the manuscript: Jing, Herlopian, Cash, Westover.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Jing, Herlopian, Ge, Sun, Westover.
Obtained funding: Cole, Westover.
Administrative, technical, or material support: Jing, Karakis, Ng, Halford, Maus, Chu, Sacca, Ge, Sun, Cole, Hoch, Cash, Westover.
Supervision: Cole, Cash, Westover.

Conflict of Interest Disclosures: Dr Muniz reported being issued US patent 10,349,888. Dr Chu reported receiving grants from the National Institutes of Health (NIH) and being a paid consultant to Alliance Family of Companies, Biogen, and SleepMed. Dr Westover reported receiving grants from the NIH. No other disclosures were reported.

Funding/Support: Dr Lam was supported by grants K23 NS101037 and R25 NS065743 from the NIH's National Institute of Neurological Disorders and Stroke (NINDS) and the American Academy of Neurology. Dr Cash was supported by grants R01 NS062092 and K24 NS088568 from the NIH NINDS. Dr Westover was supported by research grants 1K23NS090900, 1R01NS102190, 1R01NS102574, and 1R01NS107291 from the NIH NINDS and the Glenn Foundation for Medical Research.

Role of the Funder/Sponsor: The funders had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Seidel S, Pablik E, Aull-Watschinger S, Seidl B, Pataraia E. Incidental epileptiform discharges in patients of a tertiary centre. *Clin Neurophysiol*. 2016;127(1):102-107. doi:10.1016/j.clinph.2015.02.056

2. van Donselaar CA, Schimsheimer RJ, Geerts AT, Declerck AC. Value of the electroencephalogram in adult patients with untreated idiopathic first seizures. *Arch Neurol*. 1992;49(3):231-237. doi:10.1001/archneur.1992.00530270045017

3. Fountain NB, Freeman JM. EEG is an essential clinical tool: pro and con. *Epilepsia*. 2006;47(suppl 1):23-25. doi:10.1111/j.1528-1167.2006.00655.x

4. Smith SJM. EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry*. 2005;76(suppl 2):ii2-ii7. doi:10.1136/jnnp.2005.069245

5. Pellegrino G, Hedrich T, Chowdhury R, et al. Source localization of the seizure onset zone from ictal EEG/MEG data. *Hum Brain Mapp*. 2016;37(7):2528-2546. doi:10.1002/hbm.23191

6. Barkley GL, Baumgartner C. MEG and EEG in epilepsy. *J Clin Neurophysiol*. 2003;20(3):163-178. doi:10.1097/00004691-200305000-00002

7. Panayiotopoulos CP. Optimal use of the EEG in the diagnosis and management of epilepsies. In: *The Epilepsies: Seizures, Syndromes, and Management*. Oxfordshire, UK: Bladon Medical Publishing; 2005:chap 2. https://www.ncbi.nlm.nih.gov/books/NBK2601/. Accessed September 17, 2019.

8. Tatum WO, Rubboli G, Kaplan PW, et al. Clinical utility of EEG in diagnosing and monitoring epilepsy in adults. *Clin Neurophysiol*. 2018;129(5):1056-1082. doi:10.1016/j.clinph.2018.01.019

9. Sánchez Fernández I, Chapman KE, Peters JM, Harini C, Rotenberg A, Loddenkemper T. Continuous spikes and waves during sleep: electroclinical presentation and suggestions for management. *Epilepsy Res Treat*. 2013;2013(1):583531. doi:10.1155/2013/583531

10. Kane N, Acharya J, Benickzy S, et al. 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*. 2017;2:170-185. doi:10.1016/j.cnp.2017.07.002

11. Binnie CD, Stefan H. Modern electroencephalography: its role in epilepsy management. *Clin Neurophysiol*. 1999;110(10):1671-1697. doi:10.1016/S1388-2457(99)00125-X

12. Halford JJ, Arain A, Kalamangalam GP, et al. Characteristics of EEG interpreters associated with higher interrater agreement. *J Clin Neurophysiol*. 2017;34(2):168-173. doi:10.1097/WNP.0000000000000344

13. Tatum WO. How not to read an EEG: introductory statements. *Neurology*. 2013;80(1)(suppl 1):S1-S3. doi:10.1212/WNL.0b013e318279730e

14. Miller JW, Henry JC. Solving the dilemma of EEG misinterpretation. *Neurology*. 2013;80(1):13-14. doi:10.1212/WNL.0b013e318279755f

15. Benbadis SR. "Just like EKGs!" should EEGs undergo a confirmatory interpretation by a clinical neurophysiologist? *Neurology*. 2013;80(1)(suppl 1):S47-S51. doi:10.1212/WNL.0b013e3182797539

16. Wilson SB, Harner RN, Duffy FH, Tharp BR, Nuwer MR, Sperling MR. Spike detection, I: correlation and reliability of human experts. *Electroencephalogr Clin Neurophysiol*. 1996;98(3):186-198. doi:10.1016/0013-4694(95)00221-9

17. Hostetler WE, Doller HJ, Homan RW. Assessment of a computer program to detect epileptiform spikes. *Electroencephalogr Clin Neurophysiol*. 1992;83(1):1-11. doi:10.1016/0013-4694(92)90126-3

18. Ehrenberg BL, Penry JK. Computer recognition of generalized spike-wave discharges. *Electroencephalogr Clin Neurophysiol*. 1976;41(1):25-36. doi:10.1016/0013-4694(76)90212-1

19. Gotman J, Gloor P, Schaul N. Comparison of traditional reading of the EEG and automatic recognition of interictal epileptic activity.

Electroencephalogr Clin Neurophysiol. 1978;44(1):48-60. doi:10.1016/0013-4694(78)90104-9

20. Webber WR, Litt B, Lesser RP, Fisher RS, Bankman I. Automatic EEG spike detection: what should the computer imitate? *Electroencephalogr Clin Neurophysiol*. 1993;87(6):364-373. doi:10.1016/0013-4694(93)90149-P

21. Stroink H, Schimsheimer R-J, de Weerd AW, et al. Interobserver reliability of visual interpretation of electroencephalograms in children with newly diagnosed seizures. *Dev Med Child Neurol*. 2006;48(5):374-377. doi:10.1017/S0012162206000806

22. Nonclercq A, Foulon M, Verheulpen D, et al. Spike detection algorithm automatically adapted to individual patients applied to spike-and-wave percentage quantification. *Neurophysiol Clin*. 2009;39(2):123-131. doi:10.1016/j.neucli.2008.12.001

23. Bagheri E, Dauwels J, Dean BC, Waters CG, Westover MB, Halford JJ. Interictal epileptiform discharge characteristics underlying expert interrater agreement. *Clin Neurophysiol*. 2017;128(10):1994-2005. doi:10.1016/j.clinph.2017.06.252

24. Halford JJ, Schalkoff RJ, Zhou J, et al. Standardized database development for EEG epileptiform transient detection: EEGnet scoring system and machine learning analysis. *J Neurosci Methods*. 2013;212(2):308-316. doi:10.1016/j.jneumeth.2012.11.005

25. Scheuer ML, Bagic A, Wilson SB. Spike detection: inter-reader agreement and a statistical Turing test on a large data set. *Clin Neurophysiol*. 2017;128(1):243-250. doi:10.1016/j.clinph.2016.11.005

26. Guedes de Oliveira P, Queiroz C, Lopes da Silva F. Spike detection based on a pattern recognition approach using a microcomputer. *Electroencephalogr Clin Neurophysiol*. 1983;56(1):97-103. doi:10.1016/0013-4694(83)90011-1

27. Black MA, Jones RD, Carroll GJ, Dingle AA, Donaldson IM, Parkin PJ. Real-time detection of epileptiform activity in the EEG: a blinded clinical trial. *Clin Electroencephalogr*. 2000;31(3):122-130. doi:10.1177/155005940003100304

28. Whisler JW, ReMine WJ, Leppik IE, McLain LW Jr, Gumnit RJ. Machine detection of spike-wave activity in the EEG and its accuracy compared with visual interpretation. *Electroencephalogr Clin Neurophysiol*. 1982;54(5):541-551. doi:10.1016/0013-4694(82)90039-6

29. Wilson SB, Turner CA, Emerson RG, Scheuer ML. Spike detection II: automatic, perception-based detection and clustering. *Clin Neurophysiol*. 1999;110(3):404-411. doi:10.1016/S1388-2457(98)00023-6

30. Halford JJ, Pressly WB, Benbadis SR, et al. Web-based collection of expert opinion on routine scalp EEG: software development and interrater reliability. *J Clin Neurophysiol*. 2011;28(2):178-184. doi:10.1097/WNP.0b013e31821215e3

31. Halford JJ, Westover MB, LaRoche SM, et al. Interictal epileptiform discharge detection in EEG in different practice settings. *J Clin Neurophysiol*. 2018;35(5):375-380. doi:10.1097/WNP.0000000000000492

32. Carpio A, Hauser WA. Epilepsy in the developing world. *Curr Neurol Neurosci Rep*. 2009;9(4):319-326. doi:10.1007/s11910-009-0048-z

33. Newton CR, Garcia HH. Epilepsy in poor regions of the world. *Lancet*. 2012;380(9848):1193-1201. doi:10.1016/S0140-6736(12)61381-6

34. Levira F, Thurman DJ, Sander JW, et al; Epidemiology Commission of the International League Against Epilepsy. Premature mortality of epilepsy in low- and middle-income countries: a systematic review from the Mortality Task Force of the International League Against Epilepsy. *Epilepsia*. 2017;58(1):6-16. doi:10.1111/epi.13603

35. Westover MB, Halford JJ, Bianchi MT. What it should mean for an algorithm to pass a statistical Turing test for detection of epileptiform discharges. *Clin Neurophysiol*. 2017;128(7):1406-1407. doi:10.1016/j.clinph.2017.02.026

36. Bossuyt PM, Reitsma JB, Bruns DE, et al; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. *Ann Intern Med*. 2003;138(1):W1-12. doi:10.7326/0003-4819-138-1-200301070-00010

37. Homan RW, Herman J, Purdy P. Cerebral location of international 10-20 system electrode placement. *Electroencephalogr Clin Neurophysiol*. 1987;66(4):376-382. doi:10.1016/0013-4694(87)90206-9

38. Jing J, Dauwels J, Rakthanmanon T, Keogh E, Cash SS, Westover MB. Rapid annotation of interictal epileptiform discharges via template matching under Dynamic Time Warping. *J Neurosci Methods*. 2016;274:179-190. doi:10.1016/j.jneumeth.2016.02.025

39. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol*. 2008;61(pt 1):29-48. doi:10.1348/000711006X126600

40. Gwet KL. *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Gaithersburg, MD: Advanced Analytics LLC; 2010:ix, 197.

41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. doi:10.2307/2529310

42. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

43. Ferri C, Hernández-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. *Pattern Recognit Lett*. 2009;30(1):27-38. doi:10.1016/j.patrec.2008.08.010

44. Gotman J, Wang LY. State dependent spike detection: validation. *Electroencephalogr Clin Neurophysiol*. 1992;83(1):12-18. doi:10.1016/0013-4694(92)90127-4

45. Houfek EE, Ellingson RJ. On the reliability of clinical EEG interpretation. *J Nerv Ment Dis*. 1959;128(5):425-437. doi:10.1097/00005053-195905000-00006

46. Struve FA, Becka DR, Green MA, Howard A. Reliability of clinical interpretation of the electroencephalogram. *Clin Electroencephalogr*. 1975;6(2):54-60. doi:10.1177/155005947500600202

47. Mark HL. *Practical Approach to Electroencephalography*. Philadelphia, PA: Elsevier/Saunders; 2010.