



# Most Experts Agree ... But What About Other EEG Readers?

Epilepsy Currents

2020, Vol. 20(2) 78-79

© The Author(s) 2020

Article reuse guidelines:

[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)

DOI: 10.1177/1535759720901511

[journals.sagepub.com/home/epi](https://journals.sagepub.com/home/epi)

## Inter-Rater Reliability of Experts in Identifying Interictal Epileptiform Discharges in Electroencephalograms

Jing J, Herlopian A, Karakis I, et al. *JAMA Neurol*. Published online October 21, 2019. doi:<https://doi.org/10.1001/jamaneurol.2019.3531>

**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 physicians with at least 1 year of subspecialty training in clinical neurophysiology as participants. 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  $\kappa$ ) 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: 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% confidence interval [CI]: 67.0%-77.8%) and mean  $\kappa$  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  $\kappa$  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.

## Commentary

Electroencephalogram (EEG) interpretation is a complex skill. Although accepted definitions and terminologies are available, a gold standard is lacking and the final interpretation of EEG is subjective. Misinterpretation of EEG can have significant adverse consequences, particularly misdiagnosis of epilepsy predicated on benign variants, sharply contoured background activity, or artifacts mistaken for interictal epileptiform discharges (IEDs).<sup>1-3</sup> Common sense would indicate that most EEG reads must be reasonably accurate, given that EEG has proven clinical value. Nevertheless, studies dating back several

decades have reported interreader agreement on EEG findings ranging from poor to substantial but have been generally limited by factors including small EEG sample size, potentially biased selection of EEG samples (either too easy or too complex), and small numbers of readers from single centers.<sup>1,4,5</sup> Factors that have been correlated with improved interrater agreement for IEDs include fellowship EEG training, subspecialty board certification, academic versus private practice, and greater time devoted to EEG reading.<sup>1,5</sup> However, a recent study found that even expert readers with high confidence in their interpretation were often not in agreement with their



Creative Commons Non Commercial No Derivs CC BY-NC-ND: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 License (<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits non-commercial use, reproduction and distribution of the work as published without adaptation or alteration, without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

peers.<sup>6</sup> Therefore, despite the widespread use of EEG, one could still reasonably argue the reliability of an interpretation regardless of the qualifications of the reader.


In an effort to provide greater clarity on the issue of interrater reliability of identifying IEDs, Jing et al completed a study using a large sample size of routine scalp EEG recordings. Nine experts reviewed 1051 scalp EEG recordings of 30 to 60 minutes duration, rating them for the presence or absence of IEDs. A total of 991 consecutively obtained EEGs clinically interpreted as containing IEDs were used, along with 60 interpreted as IED-free. Both inpatient and outpatient studies were included. Experts had at least 1 year of fellowship training in clinical neurophysiology. Each study was first reviewed by one expert who marked potential IEDs. These potential IEDs were then sent to the remaining 8 experts for their review, with a total of 13 262 waveforms scored. Percentage agreement for interpretation of the presence or absence of IEDs in an entire EEG recording was 80.9% (mean  $\kappa$  of 69.4%), which is considered substantial. For individual IEDs, the percentage agreement was lower at 72.4% (mean  $\kappa$  of 48.7%), but this is still moderate agreement. Interestingly, individual readers were found to be consistent in under- or overcalling relative to the group. Compared to earlier studies looking at agreement in EEG interpretation, this study has many strengths. First, the EEG sample size was large, and full-length EEGs were included rather than single epochs. Because consecutive abnormal EEGs from the clinical practice at Massachusetts General Hospital were included, it is less likely that the samples were biased toward outliers that were atypically perfect or difficult. A potential limitation is that all but 2 of the expert readers trained at the same institution. Since EEG is taught from master to apprentice, readers from the same “guild” may be more likely to agree with each other than those who trained elsewhere.

While we can be reassured that experts generally agree on what represents a true IED on scalp EEG, the ultimate goal is to improve the skill and reliability of everyone who interprets EEG. Despite its importance as a core diagnostic test in neurology, there are no accepted standards of EEG training or proficiency. In a recent survey of graduating US neurology residents, only 37.3% of adult neurology trainees and 66.7% of child neurology trainees reported feeling confident about interpreting EEG independently.<sup>7</sup> The American Council of Graduate Medical Education competency target for EEG skills at the time of graduation includes interpreting common EEG abnormalities, recognizing normal EEG variants, and creating an EEG report. However, only two-thirds of graduating neurology residents are confident they had met these milestones, and 15% of program directors rated these milestones as unreasonable to achieve by the end of residency training.<sup>8</sup> We lack

standardized EEG training curricula, and some neurology residency programs have no requirement for an EEG rotation.<sup>8</sup> Improving EEG education is an achievable goal. Even a short web-based training session has been shown to improve interreader agreement on EEG pattern recognition.<sup>9</sup> Furthermore, if a validated sample of IEDs now exists, every EEG reader could potentially benefit from assessing the reliability of their interpretation, regardless of their level of confidence, training, or experience, just as we would hope to validate a computer spike detection algorithm. Until we have reliable ways to measure our individual skills as EEG readers, it may be wise to recall the adage that more damage is likely done by overinterpretation than underinterpretation. Electroencephalogram results should be considered in the context of everything that is known about a given patient. If the results don't make sense, a request for an additional review of the tracing or a repeat study may be reasonable.

By Katherine Noe 

## ORCID iD

Katherine Noe  <https://orcid.org/0000-0002-9328-8546>

## References

1. Williams GW, Luders HO, Brickner A, Goormastic M, Klass DW. Interobserver variability in EEG interpretation. *Neurology*. 1985; 35(12):1714-1719.
2. Benbadis SR, Tatum WO. Overinterpretation of EEGs and misdiagnosis of epilepsy. *J Clin Neurophysiol*. 2003;20(1):42-44.
3. Benbadis SR, Lin K. Errors in EEG interpretation and misdiagnosis of epilepsy. Which EEG patterns are overread? *Eur Neurol*. 2008; 59(1):267-271.
4. 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.
5. 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.
6. Grant AC, Abdel-Baki S, Weedon J, et al. EEG interpretation reliability and interpreter confidence: a large single-center study. *Epilepsy Behav*. 2014;32(1):102-107.
7. Mahajan A, Cahill C, Scharf E, et al. Neurology residency training in 2017. A survey of preparation, perspectives, and plans. *Neurology*. 2019;92(2):76-83.
8. Daniello KM, Weber DJ. Education research: the current state of neurophysiology education in selected neurology residency program. *Neurology*. 2018;90(15):708-711.
9. Gaspard N, Hirsch LJ, LaRoche SM, et al. Interrater agreement for critical care EEG terminology. *Epilepsia*. 2014;55(9):1366-1373.