RESEARCH ARTICLE

# Interrater Reliability of Expert Electroencephalographers Identifying Seizures and Rhythmic and Periodic Patterns in EEGs

Jin Jing, PhD, Wendong Ge, PhD, Aaron F. Struck, MD, Marta Bento Fernandes, PhD, Shenda Hong, PhD, Sungtae An, Safoora Fatima, MD, Aline Herlopian, MD, Ioannis Karakis, MD, PhD, MSc, Jonathan J. Halford, MD, Marcus C. Ng, MD, Emily L. Johnson, MD, Brian L. Appavu, MD, Rani A. Sarkis, MD, MSc, Gamaleldin Osman, MD, MS, Peter W. Kaplan, MBBS, FRCP, Monica B. Dhakar, MD, MS, Lakshman Arcot Jayagopal, MD, Zubeda Sheikh, MD, MS, Olga Taraschenko, MD, PhD, Sarah Schmitt, MD, Hiba A. Haider, MD, Jennifer A. Kim, MD, PhD, Christa B. Swisher, MD, Nicolas Gaspard, MD, PhD, Mackenzie C. Cervenka, MD, Andres A. Rodriguez Ruiz, MD, Jong Woo Lee, MD, PhD, Mohammad Tabaeizadeh, MD, Emily J. Gilmore, MD, Kristy Nordstrom, AS, Ji Yeoun Yoo, MD, Manisha G. Holmes, MD, Susan T. Herman, MD, Jennifer A. Williams, MB, BAO, BchFRCPI, Jay Pathmanathan, MD, PhD, Fábio A. Nascimento, MS, Ziwei Fan, MS, Samaneh Nasiri, PhD, Mouhsin M. Shafi, MD, PhD, Sydney S. Cash, MD, PhD, Daniel B. Hoch, MD, PhD, Andrew J. Cole, MD, Eric S. Rosenthal, MD, Sahar F. Zafar, MD, Jimeng Sun, PhD, and M. Brandon Westover, MD, PhD

**Correspondence**
Dr. Westover
mwestover@mgh.harvard.edu

*Neurology*® 2023;100:e1737-e1749. doi:10.1212/WNL.0000000000201670

## Abstract

### Background and Objectives

The validity of brain monitoring using electroencephalography (EEG), particularly to guide care in patients with acute or critical illness, requires that experts can reliably identify seizures and other potentially harmful rhythmic and periodic brain activity, collectively referred to as "ictal-interictal-injury continuum" (IIIC). Previous interrater reliability (IRR) studies are limited by small samples and selection bias. This study was conducted to assess the reliability of experts in identifying IIIC.

### Methods

This prospective analysis included 30 experts with subspecialty clinical neurophysiology training from 18 institutions. Experts independently scored varying numbers of ten-second EEG segments as "seizure (SZ)," "lateralized periodic discharges (LPDs)," "generalized periodic discharges (GPDs)," "lateralized rhythmic delta activity (LRDA)," "generalized rhythmic delta activity (GRDA)," or "other." EEGs were performed for clinical indications at Massachusetts General Hospital between 2006 and 2020. Primary outcome measures were pairwise IRR (average percent agreement [PA] between pairs of experts) and majority IRR (average PA with group consensus) for each class and beyond chance agreement (κ). Secondary outcomes were calibration of expert scoring to group consensus, and latent trait analysis to investigate contributions of bias and noise to scoring variability.

### Results

Among 2,711 EEGs, 49% were from women, and the median (IQR) age was 55 (41) years. In total, experts scored 50,697 EEG segments; the median [range] number scored by each expert was 6,287.5 [1,002, 45,267]. Overall pairwise IRR was moderate (PA 52%, κ 42%), and majority IRR was substantial (PA 65%, κ 61%). Noise-bias analysis demonstrated that a single

RELATED ARTICLES

**Editorial**
Putting the "Big" in Big Data: Learning to Be Just as (Un)certain as a Clinician at EEG
Page 799

**Research Article**
Development of Expert-Level Classification of Seizures and Rhythmic and Periodic Patterns During EEG Interpretation
Page 808

MORE ONLINE

**Class of Evidence**
Criteria for rating therapeutic and diagnostic studies
NPub.org/coe

**CME Course**
NPub.org/cmelist

From the Massachusetts General Hospital/Harvard Medical School Department of Neurology (J.J., W.G., M.B.F., S.S.C., A.J.C., D.B.H., E.S.R., S.F.Z., M.B.W.), MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC) (J.J., W.G., M.B.F., S.S.C., D.B.H., A.J.C., E.S.R., S.F.Z., M.B.W.), MA; University of Wisconsin-Madison Department of Neurology (A.F.S., S.F.); William S. Middleton Memorial Veterans Hospital Madison (A.F.S.), WI; National Institute of Health Data Science (S.H.), Peking University, Beijing, China; Georgia Institute of Technology (S.A.), College of Computing, Atlanta, GA; Yale University-Yale New Haven Hospital (A.H.), CT; Emory University School of Medicine (I.K.), GA; Medical University of South Carolina (J.J.H.), SC; University of Manitoba (M.C.N.), Canada; Johns Hopkins School of Medicine (E.L.J.), MD; University of Arizona College of Medicine (B.L.A.), AZ; Brigham and Women's Hospital (R.A.S.), MA; Mayo Clinic-Rochester (G.O.), MN; Warren Alpert School of Medicine of Brown University (M.B.D.), Providence, RI; University of Nebraska Medical Center (L.A.J.), NE; West Virginia University Hospitals (Z.S.), WV; University of Chicago (H.A.H.), Chicago, IL; Atrium Health (C.B.S.), NC; Université Libre de Bruxelles - Hôpital Erasme (N.G.), Belgium; Icahn School of Medicine, Mount Sinai (J.Y.Y.), NY; New York University (NYU) Grossman School of Medicine (M.G.H.), NY; Barrow Neurological Institute (S.T.H.), Phoenix, AZ; Mater Misericordiae University Hospital (J.A.W.), Dublin, Ireland; University of Pennsylvania (J.P.), PA; Beth Israel Deaconess Medical Center/Harvard Medical School (M.M.S.), MA; and University of Illinois at Urbana-Champaign (J.S.), College of Computing, Champaign, IL.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

Copyright © 2022 American Academy of Neurology



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

underlying receiver operating curve can account for most variation in experts' false-positive vs true-positive characteristics (median [range] of variance explained ($R^2$): 95 [93, 98]%) and for most variation in experts' precision vs sensitivity characteristics ($R^2$: 75 [59, 89]%). Thus, variation between experts is mostly attributable not to differences in expertise but rather to variation in decision thresholds.

### Discussion

Our results provide precise estimates of expert reliability from a large and diverse sample and a parsimonious theory to explain the origin of disagreements between experts. The results also establish a standard for how well an automated IIIC classifier must perform to match experts.

### Classification of Evidence

This study provides Class II evidence that an independent expert review reliably identifies ictal-interictal injury continuum patterns on EEG compared with expert consensus.

Identifying seizures and other types of seizure-like activities in electroencephalography (EEG) has long been an essential part of medical care for patients with epilepsy and has recently become integral to the care of patients with acute or critical illness.[1-3] Seizures occur in up to half of critically ill patients who undergo EEG[4-7] monitoring, are associated with worse outcomes,[5,7-14] and if recognized promptly can be managed with antiseizure medications and other interventions. Growing evidence suggests that other seizure-like events that exist along a continuum between clear-cut seizures and normal brain activity—the so-called ictal-interictal-injury continuum (IIIC)—can also damage the brain, particularly when prolonged. However, despite standardized definitions,[15-17] identification of seizures and other IIIC events can be challenging.[18-24] Errors can harm patients, with overcalling leading to overtreatment and undercalling leading to treatment delays and neuronal injury.[25-29]

How reliably specialists recognize IIIC events is unknown. Previous studies on expert interrater reliability (IRR) have been based on small samples and have largely focused on examples specially selected for certification examinations, leaving the reliability of EEG to guide neurologic care uncertain.[18-24]

Measurement of expert IRR is a prerequisite for developing automated IIIC event detection systems because expert identification is the accepted gold standard. Automated detection could extend the reach of brain monitoring beyond the small pool of experts with EEG subspecialty training and help expand brain monitoring to underserved areas. Although automated IIIC event detection software is commercially available,[30-32] without definitive studies of expert IRR and high-quality annotated benchmark data sets, how well these systems compare with experts remains uncertain.

Therefore, we conducted a large-scale study of expert IRR for identifying IIIC events. To establish a diverse and representative set of well-annotated IIIC events, we recruited 30 experts to classify 50,697 events from 2,711 patients. We measured expert IRR by agreement between pairs of experts (pairwise IRR) and agreement with the consensus score (majority IRR). Finally, we assessed the relative contributions of noise and bias[33] to expert IRR.

Our results establish precise estimates of expert reliability based on a large and diverse sample and a parsimonious theory to explain the origin of disagreements between experts. The results also present a standard for how well an automated IIIC classification system must perform to match or exceed experts.

## Methods

### Standard Protocol Approvals, Registrations, and Patient Consents

We conducted the study following the "An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies" (STARD) guidelines.[33] The index test is independent EEG interpretation by neurologists with clinical neurophysiology fellowship training ("experts"). The reference standard is the consensus EEG interpretation among these same experts. The study was conducted prospectively, with data collection and analyses specified before the index test and reference standard were assessed. The study was approved by the Massachusetts General Hospital (MGH) IRB, which waived the requirement for informed consent.

### Participants

We selected 2,711 recordings from patients hospitalized between July 2006 and March 2020 who underwent EEG as part of clinical care at MGH in medical, neurologic, and surgical intensive care and general care units. EEG electrodes were placed according to the International 10–20 system (eTable 1, links.lww.com/WNL/C519). Patients were selected in 2 stages based on clinical notes mentioning IIIC events (eTable 2, links.lww.com/WNL/C519). We placed no restrictions on age (eTable 1, links.lww.com/WNL/C519). The large group was intended to ensure broad coverage of all variations of IIIC events encountered in practice.

### EEG Labeling

Labeling of 10-second EEG segments was performed using custom local-based and web-based interfaces in 2 stages


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

(eAppendix 2 in the Supplement, links.lww.com/WNL/C519). The first stage involved targeted annotations by small groups of independent experts. The second stage involved multiple labeling rounds by larger groups of independent experts guided by automated selection of new segments to be labeled. Experts could pan left and right 20 seconds before or after the target segment, change montages, and adjust the signal gain. A 10-minute spectrogram was provided for additional context. Raters were given a forced choice of 6 options: seizure (SZ), lateralized periodic discharges (LPDs), generalized periodic discharges (GPDs), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and "other" if none of those patterns was present. The final ground truth label assigned to each EEG segment was the category chosen most often for the segment by expert reviewers.

### EEG Raters

In total, 124 EEG raters from 18 centers labeled varying numbers of EEG segments, including 30 fellowship-trained physicians ("experts") and 94 technicians and trainees. Experts who participated in stage 2 of labeling and scored ≥1,000 segments were included in IRR analysis (eFigure 1, eTable 3 in the Supplement, links.lww.com/WNL/C519).

### Qualitative Analysis

We visually reviewed cases for each IIIC type that had high, intermediate, and low levels of agreement to identify factors that might explain expert disagreements.

### Statistical Calibration Analysis—Identification of Undercallers and Overcallers

We characterized scorers' statistical calibration for each IIIC event type[34,35] to identify overcalling and undercalling behavior (eAppendix 5 in the Supplement, links.lww.com/WNL/C519). For each IIIC, type segments were assigned to one of the 5 probability bins 0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100% based on the proportion of experts who classified them into that IIIC category. For each expert, we estimated their probability of classifying segments within each of these probability bins as that category (instead of the other 5 categories) by fitting a parametric model. For statistical stability, we limited calibration analysis to experts who scored ≥10 segments within each bin for each IIIC category. Based on model fit, we defined a calibration index ranging from −100% (maximal undercalling) to 100% (maximal overcalling), with 0% representing perfect calibration. We define significant overcalling and undercalling as having a calibration index above 20% or below −20%.

### Interrater Reliability (IRR)

We analyzed pairwise IRR in 2 ways: first, *percent agreement* (PA) is the average probability across pairs of raters that if one rater labeled an EEG segment pattern *C*, the other labeled the same segment *C*. Second, *chance-adjusted* agreement, that is, kappa, κ, is given by $\kappa = (PA - PC)/(1 - PC)$, where PC is *percent agreement achievable by chance*, defined as randomly guessing among the 6 IIIC categories with equal probability. We adopted standard naming conventions[36] for levels of reliability indicated by κ values: *slight* 0%–20%, *fair* 21%–40%, *moderate* 41%–60%, *substantial* 61%–80%, and *almost perfect* 81%–100%. We defined *majority IRR* as PA and κ with the IIIC category that received the highest number of votes. For statistical stability, we restricted majority IRR analysis to experts who scored ≥10 segments per IIIC category and segments that received scores from ≥10 experts, and restricted pairwise IRR analysis to experts who each scored ≥100 segments in common within each IIIC category. We report median, 25th, and 75th percentiles across experts (majority IRR) and expert pairs (pairwise IRR).

### Interpattern Conditional Probabilities (Confusion Matrices)

We also calculated pairwise conditional probability matrices, or "confusion matrices," $P(B|C)$, defined as the average probability that one expert labels a pattern *B*, given that another expert labels *C* (eAppendix 6, eFigure 2 in the Supplement, links.lww.com/WNL/C519). Confusion matrices go beyond the IRR metrics by showing not only how well raters agree but also the pattern of disagreement. The majority conditional confusion matrix is calculated similarly as the average probability that an expert labels a pattern *B*, given that the majority labeled it *C*.

### Analysis of Expert Noise and Bias

We developed a latent trait model to investigate contributions of noise and bias to interexpert disagreements[37-39] (eAppendix 8 in the Supplement, links.lww.com/WNL/C519). This model has 2 parameters ("latent traits") for each IIIC type: noise level σ, reflecting a rater's skill in classifying IIIC, and threshold θ, representing a rater's bias as an overcaller or undercaller. We hypothesized that most participating experts had similar skill (σ) and disagree mostly due to bias (different θ values)—we call this the "Similar Skill, Individualized Thresholds" (SSIT) model. We tested SSIT by how well it fits 3 performance statistics measured for all experts: false-positive rate (FPR), true-positive rate (TPR, aka sensitivity), and positive predictive value (PPV, aka precision), relative to the group consensus. We quantified goodness of fit as the percent variance explained by the model.

### Data Availability

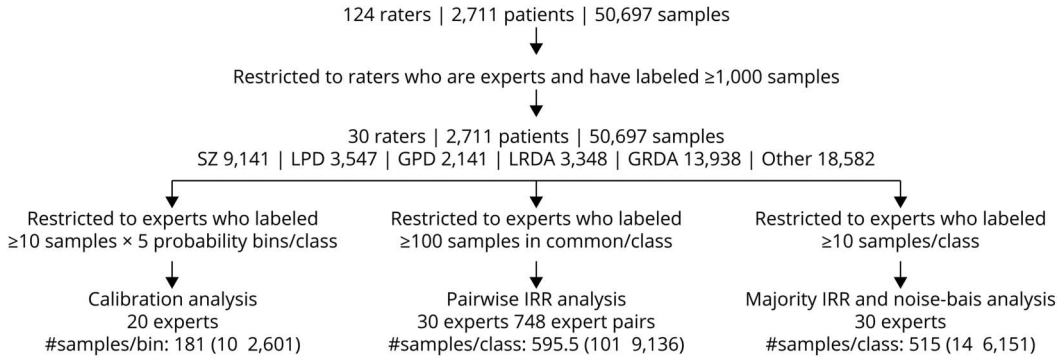
The data in this study will be available after approval of a data access agreement, pledging to not reidentify individuals or share the data with a third party. All data inquiries should be addressed to the corresponding author.

## Results

Overall, 124 raters scored 50,697 EEG segments from 2,711 patients, including 49% women, with a median (IQR) age of 55 (41) years, and EEG duration of 18 (22) hours (eTable 1 in the Supplement, links.lww.com/WNL/C519). Limiting analysis to experts with training in clinical neurophysiology yielded 30 experts; all scored ≥1,000 EEG segments. Among these 30, there were 20 experts with enough data for calibration analysis and 30 for pairwise IRR, majority IRR, and noise-bias analyses. The


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

**Figure 1** Scoring Flowchart




In total, 124 raters (30 experts and 94 technicians or trainees) scored 50,697 segments from 2,711 patients' EEG recordings. The number of segments among these with consensus labels of seizure (SZ), lateralized or generalized periodic discharges (LPDs, GPDs), lateralized or generalized rhythmic delta activity (LRDA, GRDA), or none of those patterns ("other") are indicated. Constraints applied to ensure statistical stability for calibration analysis, pairwise interrater reliability (IRR) analysis, and majority IRR analysis are shown, together with the resulting number of experts' data, and the number of segments is shown. For calibration analysis, the number of segments available is expressed as the median [minimum, maximum] number of segments per probability bin. For pairwise and majority IRR, the number of segments is given as the median [minimum, maximum] number of segments per pattern class. For pairwise IRR analysis, the number of expert pairs among the 30 experts with sufficient jointly scored data for analysis is also shown.

median [range] of EEG segments scored per expert was 6,287.5 [1,002, 45,267], and the median number of labels per EEG segment was 15 [10, 29]. The consensus (majority) labels were Other for 37% (N = 18,582). Among the remaining 32,115 IIIC patterns, 28% (9,141) were SZ, 11% (3,547) were LPD, 7% (2,141) were GPD, 10% (3,348) were LRDA, and 43% (13,938) were GRDA. A flow diagram of the sample selection is shown in Figure 1.

## Qualitative Observations

We visually reviewed EEG segments for the 5 IIIC events with varying degrees of expert agreement (Figure 2, eFigure 3 in the Supplement, links.lww.com/WNL/C519). We identified 3 main types of events: (1) "Ideal" patterns: cases with high agreement tend to be clear examples that match standardized definitions.[15,17,40] (2) "Protopatterns": cases where votes split evenly/nearly evenly between "other" and one IIIC pattern tend to be partially/semiformed, having some but not all classic features. (3) "Edge" cases: cases where raters split between 2 IIIC patterns tend to have features of both classes. These observations suggest that classifying real-world IIIC patterns is challenging in part because they do not form distinct clusters. Instead, "ideal" IIIC patterns are connected by one or more continuous paths in "feature space" leading through a series of intermediate edge cases and protopatterns. This supports the concept that IIIC patterns lie along an "ictal-interictal continuum" as proposed by Chiappa et al.[41] and popularized by Hirsch et al.[42]

We further defined these categories quantitively using the following rules:

- Idealized: the top class received >80% of votes.
- Proto: the top 2 classes are an IIIC pattern and "other" (non-IIIC), and the difference in fractions of votes received by the top 2 classes is <10%.
- Edge: the top 2 classes are both IIIC patterns, and the difference in votes is <10%.
- In-between: any other samples beyond the 3 categories above that are ill defined.

Using these definitions on the samples with ≥10 votes (N = 11,474 samples), the percentages of patterns of each type were 35.2% idealized, 4.2% edge, 8.9% proto, and 51.7% in-between. This distribution pertains to our data but may not represent the natural frequencies in EEG in general.

## Calibration Curves

Calibration curves are shown in Figure 3A for the 20 experts included in the analysis. Overcalling and undercalling are common: the proportions of experts with calibration indices >20% for overcalling or undercalling (overcalling/undercalling, %), respectively, were SZ (11/11), LPD (37/21), GPD (32/16), LRDA (32/21), GRDA (32/26), and "other" (21/32). The IIIC class with the largest proportion of well-calibrated experts (undercalling and overcalling <20%) was SZ (79%), followed by GPD (53%), "other" (47%), LRDA (47%), LPD (42%), and GRDA (42%).

## Pairwise IRR

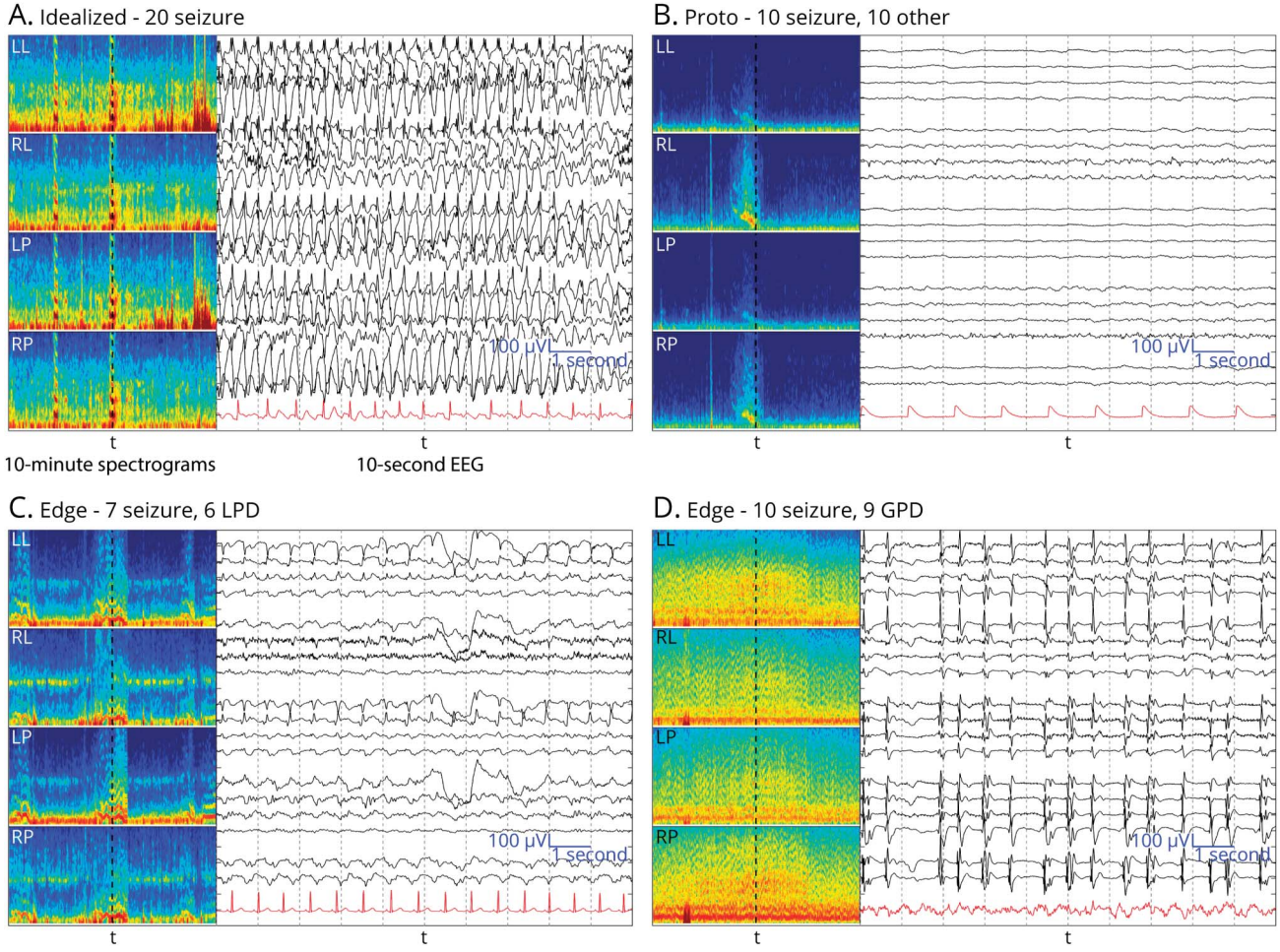
All 30 experts' data met inclusion criteria for pairwise IRR analysis, among whom 748 ordered pairs of experts scored ≥100 EEG segments in common (eTable 4 in the Supplement, links.lww.com/WNL/C519). Overall pairwise IRR (PA/κ, %) was *moderate* (52/42); pairwise IRR was *slight* for LRDA (34/20), *fair* for SZ (45/34) and GRDA (44/33), *moderate* for LPD (63/56) and GPD (55/45), and *substantial* for "other" (68/62) (eTable 4, links.lww.com/WNL/C519).

## Majority IRR

All 30 experts who scored met inclusion for majority IRR analysis (eTable 4 in the Supplement, links.lww.com/WNL/C519). Majority IRR was generally higher than pairwise IRR. Overall


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

**Figure 2** Selected EEG Examples for Class Seizure




(A) Example of idealized form of seizure (SZ) with uniform expert agreement. (B) Protopattern or partially formed pattern. About half of raters labeled these SZ and the other half labeled "other." (C, D) are edge cases (about half of raters labeled these SZ and half labeled them another IIIC pattern). For (B), there is rhythmic delta activity with some admixed sharp discharges within the 10-second raw EEG, and the spectrogram shows that this segment may belong to the tail end of a SZ; thus, disagreement between SZ and "other" makes sense. (C) 2 Hz lateralized periodic discharges (LPDs) showing an evolution with increasing amplitude evolving underlying rhythmic activity, a pattern between LPDs and the beginning of a SZ, an edge case. Panel D shows abundant generalized periodic discharges (GPDs) on top of a suppressed background with a frequency of 1–2 Hz. The average over the 10 seconds is close to 1.5 Hz, suggesting a SZ, another edge case.

majority IRR (PA/κ) was *substantial* (65/61); majority IRR was *moderate* for SZ (64/60), LRDA (55/50), and GRDA (60/56), and *substantial* for LPD (71/68), GPD (65/61), and "other" (73/70) (eTable 4, links.lww.com/WNL/C519).

## Confusion Matrices

We investigated patterns of disagreements between raters (pairwise IRR) using confusion matrices (Figure 3B). Most disagreements were over whether segments were IIIC vs "other." Beyond these, for EEG segments scored as SZ, diverging opinions were from most to least common: LPD (14%), then GPD (13%), and less likely to be LRDA (6%) or GRDA (4%). Diverging opinions for LPD were, from most to least common: GPD, SZ, LRDA > GRDA (5%, 5%, 5%, and 2%); for GPD: LPD > SZ > GRDA >> LRDA (9%, 8%, 5%, 1%); for LRDA: LPD > GRDA > SZ >> GPD (15%, 9%, 6%, 2%); and for GRDA: GPD > LRDA > LPD, SZ (7%, 6%, 3%, 3%). In summary, patterns that shared the property of being rhythmic, periodic, or had similar distribution (lateralized vs generalized) were more likely to be confused. Similar results are seen in the majority IRR confusion matrix (Figure 3C).

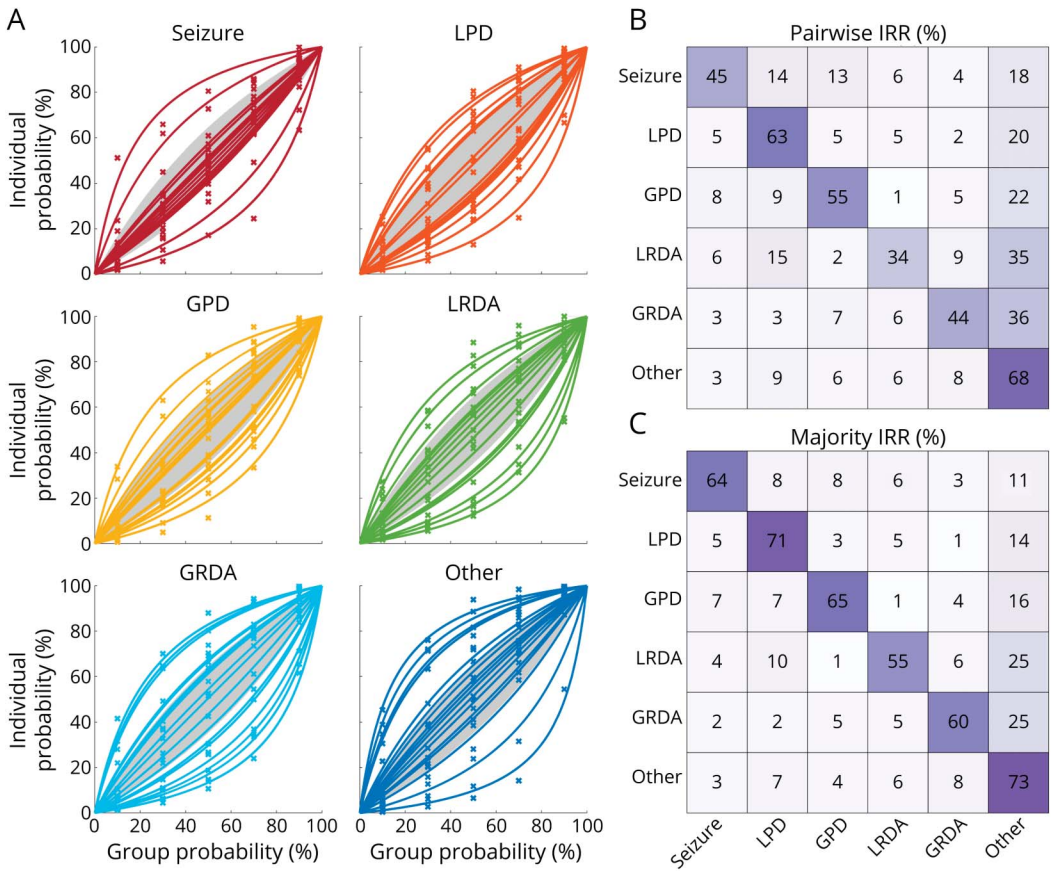
## Noise and Bias Underlying Expert IRR

Results of fitting the "Similar Skill, Individualized Thresholds" (SSIT) model are shown in Figure 4. Consistent with the SSIT model, experts' operating points lie near common receiver operating characteristic (ROC) curves for each IIIC pattern. The mean [range] percent of variance explained by the SSIT across IIIC types is 95 [93, 98] %. The area under these ROC curves ranges from 97% for SZ and GPD to 90% for "other." Across all IIIC patterns, experts tend to operate in a low false-positive rate (median 3.0 [0.8, 14.0] %) and a high-sensitivity (62 [28, 89] %) regime.

The SSIT model can also account for variance in experts' precision and recall characteristics (PRC), with mean [range]


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

**Figure 3** Interrater Reliability Analysis




(A) Calibration curves: segments were binned for each of the 6 classes according to the percentage of experts who classified them as that class. Bins were chosen to be 0%–20%, 20%–40%, 40%–60%, 60%–80%, and 80%–100%. Calibration curves were calculated for each expert, and each pattern class based on the percentage of segments within each bin that the expert classified as belonging to that class, producing a set of 5 percentages (one for each bin). A single parameter curve (see eAppendix 5 in the Supplement, links.lww.com/WNL/C519) was fit to these percentages to characterize the experts' tendency to overcall and undercall. Experts with calibration curves >20% above the diagonal (above the shaded region) are considered overcallers. Experts with calibration curves >20% below the diagonal (below the shaded region) are considered undercallers. (B) and (C) Confusion matrices: these heatmaps show a pattern of disagreement between experts for IIIC (and "other") classes. These are presented as conditional probabilities (between 0% and 100%). For the pairwise IRR confusion matrix (panel B), the number in each square is the average (across pairs of experts) probability that a rater labels a pattern $A_1$ (the x-axis) if another rater had labeled it pattern $A_2$ (the y-axis). The sum of values within each row is 100%. The matrices are not symmetric, because $P(A_1 \mid A_2)$ does not equal the $P(A_2 \mid A_1)$, because there are differences in the underlying prevalence of the patterns. The diagonal is the "pattern" pairwise agreement shown in eTable 4 in the Supplement, links.lww.com/WNL/C519. For the majority IRR confusion matrix (panel C), the numbers are the average (across experts) probability that a rater labels a segment pattern $A_1$ (x-axis) if the majority label for that segment is $A_2$. GPD = generalized periodic discharges; GRDA = generalized rhythmic delta activity; IRR = interrater reliability; LPD = lateralized periodic discharges; LRDA = lateralized rhythmic delta activity.

variance explained for each IIIC pattern of 75 [59, 89] %. Areas under the PR curves ranged from 90% for "other" to 48% for LRDA. Experts tend to operate near the upper right elbow of these curves, in the high sensitivity, and in a high precision range, although sensitivity (62 [28, 89] %) and precision (median 65 [28, 87] %) vary widely.

### Classification of Evidence

This study provides Class II evidence that an independent expert review reliably identifies ictal-interictal injury continuum patterns on EEG compared with expert consensus.

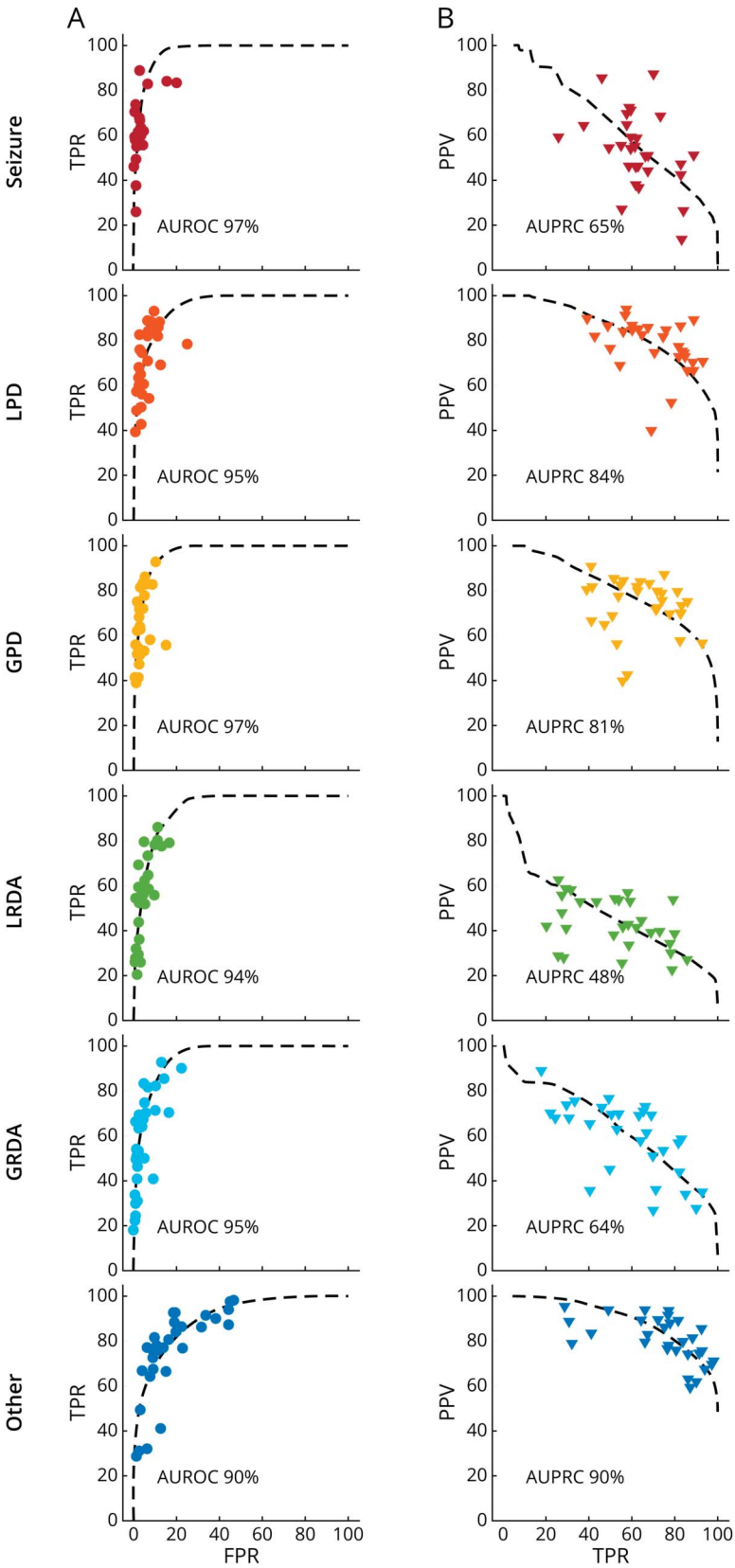
## Discussion

Our findings suggest that reliability for classifying seizures and rhythmic and periodic EEG patterns between experts is only *moderate* (overall pairwise IRR: PA 52%; κ 42%). Nevertheless, disagreement between experts can be largely explained without invoking ambiguities in the definitions of IIIC patterns or errors in judgement. Rather, disagreement can be largely explained by experts applying different decision thresholds, effectively drawing different classification boundaries on what is inherently an underlying continuum.

Previous studies of experts' IIIC interrater reliability have reached variable conclusions (Table 1, eAppendix 9 in the Supplement, links.lww.com/WNL/C519). We found 8 relevant previous studies. Gerber et al.[18] studied IRR among 5 experts who scored 58 ten-second EEG segments from 11 patients with subarachnoid hemorrhage and found moderate agreement for rhythmic or periodic patterns (κ = 68%). Ronner et al.[19] studied IRR among 5 experts who scored 90 ten-second segments from 30 ICU patients and concluded that expert identification of seizures is "not very reliable,"


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

**Figure 4** Bias vs Noise Analysis




We calculated 3 performance metrics for each expert based on the agreement of their scores with the consensus score for each EEG segment: The false-positive rate (FPR): the percentage of segments that do not belong to a given class that an expert incorrectly scores as belonging to the class; true-positive rate (TPR; aka sensitivity), the percentage of segments within a class that the expert correctly scores as belonging to the class; and the positive predictive value (PPV; aka precision), the percentage of segments scored by an expert as belonging to a given class that do in fact belong to that class. In (A), we plot TPR vs FPR. A receiver operating characteristic (ROC) curve from the SSIT (similar expertise, individualized thresholds) model is fit to experts' data for each IIIC category, shown as a dashed black line. The area under the ROC is shown in each plot. In (B), we plot the PPV vs TPR. A precision recall curve (PRC) is fit to experts' data for each IIIC category. The area under the PRC is shown in each plot. The goodness of fit for ROC and PRC curves is calculated using $R^2$ values (see text).


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

**Table 1** Previous Studies of Expert Interrater Reliability for Detecting Seizures and IIIC Events

| | Types of IIIC | Patients[e] | Segments[e] | Centers[e] | Raters[e] | Segments/ rater[e] | Types of raters | Pairwise agreement | Types of patients |
|---|---|---|---|---|---|---|---|---|---|
| **Gerber et al. 2008[18]** | PD or RDA vs Other | 11 | 58 (1 sec)<br>30 (20–30 min) | 1 | 5 | 58<br>30 | CNP | κ 0.68 | ICU (SAH) |
| **Ronner et al. 2009[19]** | SZ vs Other | 30 | 90 (10 sec) | 1 | 9 | 90 | 5 Exp<br>4 in-Exp | κ 50 Exp<br>κ 29 in-Exp | ICU |
| **Mani et al. 2012[20]** | LPD<br>GPD<br>LRD<br>GRDA | <14 | 14[a] | 11 | 16 | 14 | CNP | κ 87 MT1<br>κ 92 MT2[b] | ICU |
| **Gaspard et al. 2014[21]** | SZ<br>LPD<br>GPD<br>LRDA<br>GRDA | 37 | 37 | N.A.[f] | 49 | 37 | CNP (25)<br>fellows (20) | κ 91.1 SZ<br>κ 89.3 MT1<br>κ 85.2 MT2 | ICU |
| **Westhall et al. 2015[22]** | SZ<br>LPD<br>GPD vs Other | 103 | 103<br>(~30 min[d]) | 4 | 4 | 103 | CNP | κ 72[c] | ICU (CA + coma) |
| **Halford et al., 2015[23]** | SZ vs PDs | 20 | 30 (1 hr) | 3 | 8 | 20–30 | CNP | κ 58 SZ<br>κ 38 PD | ICU |
| **Tu B et al., 2017[24]** | SZ vs Other | 50 | 50 EEGs<br>(avg 35.5 hr) | 1 | 5 | 2,093–2,085 | CNP | Sens: 70.2<br>Spec: 99.2 | ICU |
| **Present study** | SZ<br>LPD<br>GPD<br>LRDA<br>GRDA<br>Other | 2,711 | 6,095 EEGs<br>50,697 10sec | 18 | 30 | 1,002–45,267 | CNP (30)<br>others (94) | κ 34/60 SZ[#]<br>κ 56/68 LPD<br>κ 45/61 GPD<br>κ 20/50 LRDA<br>κ 33/56 GRDA<br>κ 62/70 Other | ICU<br>EMU<br>rEEG |

CA = cardiac arrest; CNP = clinical neurophysiologist; EMU = epilepsy monitoring unit; Exp = "experienced," neurologists with 5–10 years of experience in reading EEGs; GPD = generalized periodic discharges; GRDA = generalized rhythmic delta activity; ICU = intensive care unit; IIIC = ictal-interictal-injury continuum; in-Exp = "inexperienced" neurology residents; LPD = lateralized periodic discharges; LRDA = lateralized rhythmic delta activity; PD = periodic discharges (LPDs or GPDs); RDA = rhythmic delta activity (LRDA or GRDA); rEEG = routine EEG; Sens = sensitivity; Spec = specificity; SZ = seizure.
[a] Inferred; not stated explicitly in the article.
[b] MT1 = main term 1, that is, whether an IIIC pattern is generalized, lateralized, multifocal, bilateral independent or multifocal; MT2 = main term 2, that is, whether an IIIC pattern is RDA, PD, or sharp/spike-wave (SW).
[c] Result is for "malignant periodic or rhythmic patterns" (LPD, GPD, SZ).
[d] Samples were described as "full-length routine EEGs" (exact duration not reported).
[e] Result is reported in pair as "pairwise-κ/majority-κ."
[f] Multicenter study without mentioning the number of centers.

despite raters using the same strict criteria (κ = 50%). This study was partly responsible for efforts to further standardize nomenclature for rhythmic and periodic IIIC patterns.[15,17,40] Mani et al.[20] studied IRR among 16 experts who scored ~14 ten-second segments from ~14 ICU participants (exact number not specified) and concluded that agreement for rhythmic and periodic patterns was high (κ = 87–92%). Gaspard et al.[43] studied IRR among 25 experts (including 20 fellows) who scored 37 ten-second segments from 37 ICU patients and found excellent agreement for seizures (κ = 91%) and rhythmic and periodic IIIC (κ = 89–85%). Westhall et al.[22] studied IRR among 4 experts who scored 103 routine EEGs from 103 patients with comatose cardiac arrest, lumping periodic discharges (LPDs, GPDs) with seizures, and found κ = 72%. Halford et al.[23] studied IRR among 8 experts who scored 20–30 one-hour EEGs from 20 ICU patients and found κ = 58% for seizures and κ = 38% for periodic discharges. Bin Tu et al.[24] studied IRR among 5 experts who scored prolonged EEGs from 50 ICU patients and identified an average sensitivity for seizures of 70%. The relatively small numbers of patients and samples in these studies, and the differences in which patterns and how many pattern types raters evaluated, account for the variable IRR statistics across studies. Thus, previous studies leave the measurement of expert reliability for seizures and other epileptiform patterns open to questions of systematic and random error. Because of this gap, it is likewise unclear to what extent the existing commercial detectors are comparable with the performance of human experts. We hope that our results will serve as benchmarks for more rigorous testing of existing commercial software so that clinical users can make informed decisions about how to use available IIIC detectors and when to adopt new detectors because they become available for clinical use.

Our finding that disagreements among experts can be largely explained by differences in decision thresholds (the SSIT


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

model) has implications for clinical practice. First, the observed levels of disagreement are substantial and likely contribute to unwarranted variability in diagnosis and treatment. Second, it is nevertheless reassuring to find evidence of strong underlying levels of agreement about IIIC probabilities, as shown by the good fit of the underlying common receiver operating and precision recall curves (Figure 4). Third, efforts to improve IRR often focus on improving expertise (improving ROC and PRC curves) and on refining the definitions of "ideal" patterns. By contrast, our results suggest that educational efforts to improve IRR should also focus on harmonizing thresholds across experts, placing greater emphasis on the boundaries between patterns—"edge cases" and "protopatterns." Nevertheless, our findings do not prove that the SSIT model is correct or that it is the only possible explanation for our findings; thus, further studies are warranted to better define the relative contributions of variability in expertise and decision thresholds to IRR.

Our study has limitations. First, experts did not score all types of IIIC patterns; we omitted certain patterns such as bilateral independent periodic discharges (BiPDs) and spike-wave (SW) patterns (lateralized and generalized SW).[17,40] These patterns are similar to GPDs and LPDs but are encountered more rarely. Future efforts could pool these patterns from multiple centers for IRR analysis. Second, all EEGs were from a single institution. We believe this is unlikely to be a major limitation because our patient mix is similar to others in the literature. Third, although this is a large study with 2,711 patients, the number needed to truly represent all clinically important variations is not known. Fourth, experts did not review full EEG recordings. Nevertheless, experts were provided 20 seconds of EEG context before and after each segment, in addition to a 10-minute spectrogram, which we believe is sufficient for scoring. Although a full EEG review is preferable, it would have been infeasible for 20 experts to score 2,711 EEGs. We believe this likely does not substantially affect our results because, in practice, experts screen EEG background quickly and spend most review time deliberating about patterns suspicious for IIIC, similar to our study. Finally, we asked experts to assign IIICs to discrete categories. An ordinal scale (e.g., reporting confidence, as in the work of Wilson et al.[43]), or free text responses (e.g., about what experts believe when classifying patterns as "other"), would provide more information. However, this would have necessitated scoring a smaller number of IIIC and would depart from clinical practice. We felt it advantageous to score a larger set of candidate IIIC patterns.

Although scoring reliability for IIIC events is imperfect, scoring behavior is explained by a model that assumes experts assign very similar probabilities that a given EEG pattern belongs to a given IIIC category. Disagreements are largely over where to draw boundaries between patterns that exist along an underlying continuum. Our results establish precise estimates of expert reliability based on a large and diverse sample. Future efforts to increase expert reliability should focus on helping experts agree on the borders between patterns. The results also present a standard for how well an automated IIIC classification system must perform to match or exceed experts.


## Study Funding

M.B. Westover received funding from the Glenn Foundation for Medical Research and American Federation for Aging Research (Breakthroughs in Gerontology Grant); American Academy of Sleep Medicine (AASM Foundation Strategic Research Award); Football Players Health Study (FPHS) at Harvard University; Department of Defense through a subcontract from Moberg ICU Solutions, Inc; NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, and R01AG062989); and NSF (Award No. SCH-2014431). Dr. A.F. Struck received funding from the NIH (R01NS111022). S.S. Cash was funded by NIH NINDS R01 NS062092 and NIH NINDS K24 NS088568. M.B. Dhakar received funding from NIH NINDS NS11672601 and an American Epilepsy Society Infrastructure Research Award and got clinical trial support from Marinus Pharmaceuticals and Parexel Inc. J.A. Kim received support from NIH-NINDS (R25NS06574), AHA, and Bee Foundation. Dr. Olha Taraschenko was supported by research grants from the NIH (P20GM130447) and the American Epilepsy Society-NORSE Institute Seed grant. M.C. Cervenka receives or has received research grants from Nutricia, Vitaflo, Glut1 Deficiency Foundation, and BrightFocus Foundation; honoraria from Nutricia and Vitaflo/Nestle Health Sciences; royalties from Demos/Springer Publishing Company; and consulting fees from Nutricia and Glut1 Deficiency Foundation. The funding sources had no role in study design, data collection, analysis, interpretation, or writing of the report. All authors had full access to all data, and the corresponding author had final responsibility for the decision to submit for publication.


## Disclosure

The authors report no disclosures relevant to the manuscript. Go to Neurology.org/N for full disclosures.


## Publication History

Previously published with The Lancet on SSRN at http://ssrn.com/abstract=4063817. Received by *Neurology* April 13, 2022. Accepted in final form October 25, 2022. Submitted and externally peer reviewed. The handling editor was Associate Editor Barbara Jobst, MD, PhD, FAAN.


## Appendix Authors

| Name | Location | Contribution |
|---|---|---|
| **Jin Jing, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data |

Continued



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

## Appendix *(continued)*

| Name | Location | Contribution |
|---|---|---|
| **Wendong Ge, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data |
| **Aaron F. Struck, MD** | University of Wisconsin-Madison Department of Neurology; William S Middleton Memorial Veterans Hospital Madison, WI | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data |
| **Marta Bento Fernandes, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data |
| **Shenda Hong, PhD** | National Institute of Health Data Science, Peking University, Beijing, China | Drafting/revision of the manuscript for content, including medical writing for content; study concept or design |
| **Sungtae An** | Georgia Institute of Technology, College of Computing, Atlanta, GA | Drafting/revision of the manuscript for content, including medical writing for content; study concept or design |
| **Safoora Fatima, MD** | University of Wisconsin-Madison Department of Neurology | Drafting/revision of the manuscript for content, including medical writing for content; study concept or design |
| **Aline Herlopian, MD** | Yale University-Yale New Haven Hospital, CT | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Ioannis Karakis, MD, PhD, MSc** | Emory University School of Medicine, GA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Jonathan J. Halford, MD** | Medical University of South Carolina, SC | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Marcus C. Ng, MD** | University of Manitoba, Canada | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |

## Appendix *(continued)*

| Name | Location | Contribution |
|---|---|---|
| **Emily L. Johnson, MD** | Johns Hopkins School of Medicine, MD | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Brian L. Appavu, MD** | University of Arizona College of Medicine, AZ | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Rani A. Sarkis, MD, MSc** | Brigham and Women's Hospital, MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Gamaleldin Osman, MD, MS** | Mayo Clinic-Rochester, MN | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Peter W. Kaplan, MBBS, FRCP** | Johns Hopkins School of Medicine, MD | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Monica B. Dhakar, MD, MS** | Warren Alpert School of Medicine of Brown University, Providence, RI | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; Study concept or design |
| **Lakshman Arcot Jayagopal, MD** | University of Nebraska Medical Center, NE | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Zubeda Sheikh, MD, MS** | West Virginia University Hospitals, WV | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Olga Taraschenko, MD, PhD** | University of Nebraska Medical Center, NE | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Sarah Schmitt, MD** | Medical University of South Carolina, SC | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; Study concept or design |
| **Hiba A. Haider, MD** | University of Chicago, Chicago, IL | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |


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

## Appendix *(continued)*

| Name | Location | Contribution |
|---|---|---|
| **Jennifer A. Kim, MD, PhD** | Yale University-Yale New Haven Hospital, CT | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Christa B. Swisher, MD** | Atrium Health, NC | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Nicolas Gaspard, MD, PhD** | Université Libre de Bruxelles - Hôpital Erasme, Belgium | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Mackenzie C. Cervenka, MD** | Johns Hopkins School of Medicine, MD | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Andres A. Rodriguez Ruiz, MD** | Emory University School of Medicine, GA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Jong Woo Lee, MD, PhD** | Brigham and Women's Hospital, MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Mohammad Tabaeizadeh, MD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Emily J. Gilmore, MD** | | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Kristy Nordstrom, AS** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Ji Yeoun Yoo, MD** | Icahn School of Medicine, Mount Sinai, NY | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |

## Appendix *(continued)*

| Name | Location | Contribution |
|---|---|---|
| **Manisha G. Holmes, MD** | New York University (NYU) Grossman School of Medicine, NY | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Susan T. Herman, MD** | Barrow Neurological Institute, Phoenix, AZ | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Jennifer A. Williams, MB, BAO** | Mater Misericordiae University Hospital, Dublin, Ireland | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Jay Pathmanathan, MD, PhD** | University of Pennsylvania, PA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Fábio A. Nascimento, MD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; study concept or design |
| **Ziwei Fan, MS** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data |
| **Samaneh Nasiri, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; study concept or design |
| **Mouhsin M. Shafi, MD, PhD** | Beth Israel Deaconess Medical Center/Harvard Medical School, MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Sydney S. Cash, MD, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |

Continued



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

## Appendix *(continued)*

| Name | Location | Contribution |
|---|---|---|
| **Daniel B. Hoch, MD, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Andrew J. Cole, MD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Eric S. Rosenthal, MD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Sahar F. Zafar, MD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **Jimeng Sun, PhD** | University of Illinois at Urbana-Champaign, College of Computing, Champaign, IL | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design |
| **M. Brandon Westover, MD, PhD** | Massachusetts General Hospital/Harvard Medical School Department of Neurology, MA; Massachusetts General Hospital Clinical Data Animation Center (CDAC), MA | Drafting/revision of the manuscript for content, including medical writing for content; major role in the acquisition of data; study concept or design; analysis or interpretation of data |

## References


1. Hill CE, Blank LJ, Thibault D, et al. Continuous EEG is associated with favorable hospitalization outcomes for critically ill patients. *Neurology* 2019;92(1):e9–e18. doi: 10.1212/WNL.0000000000006689
2. Westover MB, Gururangan K, Markert MS, et al. Diagnostic value of electroencephalography with ten electrodes in critically ill patients. *Neurocrit Care* 2020;33(2):479-490. doi: 10.1007/s12028-019-00911-4
3. Zafar SF, Subramaniam T, Osman G, Herlopian A, Struck AF. Electrographic seizures and ictal-interictal continuum (IIC) patterns in critically ill patients. *Epilepsy Behav EB* 2020;106:107037. doi: 10.1016/j.yebeh.2020.107037
4. Westover MB, Shafi MM, Bianchi MT, et al. The probability of seizures during EEG monitoring in critically ill adults. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol.* 2015;126(3):463-471. doi: 10.1016/j.clinph.2014.05.037
5. Claassen J, Jetté N, Chum F, et al. Electrographic seizures and periodic discharges after intracerebral hemorrhage. *Neurology* 2007;69(13):1356-1365. doi: 10.1212/01.wnl.0000281664.02615.6c
6. Oddo M, Carrera E, Claassen J, Mayer SA, Hirsch LJ. Continuous electroencephalography in the medical intensive care unit. *Crit Care Med.* 2009;37(6):2051-2056. doi: 10.1097/CCM.0b013e3181a00604
7. Kurtz P, Gaspard N, Wahl AS, et al. Continuous electroencephalography in a surgical intensive care unit. *Intensive Care Med.* 2014;40(2):228-234. doi: 10.1007/s00134-013-3149-8
8. Claassen J, Hirsch LJ, Frontera JA, et al. Prognostic significance of continuous EEG monitoring in patients with poor-grade subarachnoid hemorrhage. *Neurocrit Care* 2006;4(2):103-112. doi: 10.1385/NCC:4:2:103
9. Ribeiro A, Singh R, Brunnhuber F. Clinical outcome of generalized periodic epileptiform discharges on first EEG in patients with hypoxic encephalopathy postcardiac arrest. *Epilepsy Behav EB* 2015;49:268-272. doi: 10.1016/j.yebeh.2015.06.010
10. De Marchis GM, Pugin D, Meyers E, et al. Seizure burden in subarachnoid hemorrhage associated with functional and cognitive outcome. *Neurology* 2016;86(3):253-260. doi: 10.1212/WNL.0000000000002281
11. Zafar SF, Postma EN, Biswal S, et al. Effect of epileptiform abnormality burden on neurologic outcome and antiepileptic drug management after subarachnoid hemorrhage. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol.* 2018;129(11):2219-2227. doi: 10.1016/j.clinph.2018.08.015
12. Tabaeizadeh M, Aboul Nour H, Shoukat M, et al. Burden of epileptiform activity predicts discharge neurologic outcomes in severe acute ischemic stroke. *Neurocrit Care* 2020;32(3):697-706. doi: 10.1007/s12028-020-00944-0
13. Zafar SF, Rosenthal ES, Jing J, et al. Automated annotation of epileptiform burden and its association with outcomes. *Ann Neurol.* 2021. doi: 10.1002/ana.26161
14. Payne ET, Zhao XY, Frndova H, et al. Seizure burden is independently associated with short term outcome in critically ill children. *Brain* 2014;137(Pt 5):1429-1438. doi: 10.1093/brain/awu042
15. Hirsch LJ, Brenner RP, Drislane FW, et al. The ACNS subcommittee on research terminology for continuous EEG monitoring: proposed standardized terminology for rhythmic and periodic EEG patterns encountered in critically ill patients. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2005;22(2):128-135. doi: 10.1097/01.wnp.0000158701.89576.4c
16. Beniczky S, Hirsch LJ, Kaplan PW, et al. Unified EEG terminology and criteria for nonconvulsive status epilepticus. *Epilepsia* 2013;54(suppl 6):28-29. doi: 10.1111/epi.12270
17. Hirsch LJ, Fong MWK, Leitinger M, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2021 version. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2021;38(1):1-29. doi: 10.1097/WNP.0000000000000806
18. Gerber PA, Chapman KE, Chung SS, et al. Interobserver agreement in the interpretation of EEG patterns in critically ill adults. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2008;25(5):241-249. doi: 10.1097/WNP.0b013e318182ed67
19. Ronner HE, Ponten SC, Stam CJ, Uitdehaag BMJ. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. *Seizure* 2009;18(4):257-263. doi: 10.1016/j.seizure.2008.10.010
20. Mani R, Arif H, Hirsch LJ, Gerard EE, LaRoche SM. Interrater reliability of ICU EEG research terminology. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2012;29(3):203-212. doi: 10.1097/WNP.0b013e3182570f83
21. Gaspard N, Hirsch LJ, LaRoche SM, Hahn CD, Westover MB. Critical care EEGMRC. Interrater agreement for critical care EEG terminology. *Epilepsia* 2014;55(9):1366-1373. doi: 10.1111/epi.12653
22. Westhall E, Rosén I, Rossetti AO, et al. Interrater variability of EEG interpretation in comatose cardiac arrest patients. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol.* 2015;126(12):2397-2404. doi: 10.1016/j.clinph.2015.03.017
23. Halford JJ, Shiau D, Desrochers JA, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol.* 2015;126(9):1661-1669. doi: 10.1016/j.clinph.2014.11.008
24. Tu B, Young GB, Kokoszka A, et al. Diagnostic accuracy between readers for identifying electrographic seizures in critically ill adults. *Epilepsia Open* 2017;2(1):67-75. doi: 10.1002/epi4.12034
25. Kaplan PW. Assessing the outcomes in patients with nonconvulsive status epilepticus: nonconvulsive status epilepticus is underdiagnosed, potentially overtreated, and confounded by comorbidity. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 1999;16(4):341-352. discussion 353.
26. Hillman J, Lehtimäki K, Peltola J, Liimatainen S. Clinical significance of treatment delay in status epilepticus. *Int J Emerg Med.* 2013;6(1):6. doi: 10.1186/1865-1380-6-6
27. Williams RP, Banwell B, Berg RA, et al. Impact of an ICU EEG monitoring pathway on timeliness of therapeutic intervention and electrographic seizure termination. *Epilepsia* 2016;57(5):786-795. doi: 10.1111/epi.13354
28. Sutter R, Marsch S, Fuhr P, Kaplan PW, Rüegg S. Anesthetic drugs in status epilepticus: risk or rescue? A 6-year cohort study. *Neurology* 2014;82(8):656-664. doi: 10.1212/WNL.0000000000000009
29. Amorim E, McGraw CM, Westover MB. A theoretical paradigm for evaluating risk-benefit of status epilepticus treatment. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2020;37(5):385-392. doi: 10.1097/WNP.0000000000000753
30. Scheuer ML, Wilson SB, Antony A, Ghearing G, Urban A, Bagić AI. Seizure detection: interreader agreement and detection algorithm assessments using a large dataset. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc Published Online May* 2020;27. doi: 10.1097/WNP.0000000000000709
31. Kamousi B, Karunakaran S, Gururangan K, et al. Monitoring the burden of seizures and highly epileptiform patterns in critical care with a novel machine learning method. *Neurocrit Care* 2021;34(3):908-917. doi: 10.1007/s12028-020-01120-0
32. Koren JP, Herta J, Fürbass F, et al. Automated long-term EEG review: fast and precise analysis in critical care patients. *Front Neurol.* 2018;9:454. doi: 10.3389/fneur.2018.00454
33. Bossuyt PM, Reitsma JB, Bruns DE, et al. Stard 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ* 2015;351:h5527. doi: 10.1136/bmj.h5527
34. 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
35. 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




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

36. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;33(1):159-174. doi: 10.2307/2529310
37. Kahneman D, Rosenfield A, Gandhi L, Blaser T. *Noise Harv Bus Rev Published Online* 2016:38-46.
38. Kahneman D, Sibony O, Sunstein CR. *Noise: A Flaw in Human Judgment*: Little; 2021.
39. Embretson SE, Reise SP. *Item Response Theory*. Psychology Press; 2013.
40. Hirsch LJ, LaRoche SM, Gaspard N, et al. American clinical neurophysiology society's standardized critical care EEG terminology: 2012 version. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2013;30(1):1-27. doi: 10.1097/WNP.0b013e3182784729
41. Pohlmann-Eden B, Hoch DB, Cochius JI, Chiappa KH. Periodic lateralized epileptiform discharges–a critical review. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 1996;13(6):519-530.
42. Chong DJ, Hirsch LJ. Which EEG patterns warrant treatment in the critically ill? Reviewing the evidence for treatment of periodic epileptiform discharges and related patterns. *J Clin Neurophysiol Off Publ Am Electroencephalogr Soc.* 2005;22(2):79-91.
43. Wilson SB, Scheuer ML, Plummer C, Young B, Pacia S. Seizure detection: correlation of human experts. *Clin Neurophysiol Off J Int Fed Clin Neurophysiol.* 2003;114(11):2156-2164. doi: 10.1016/s1388-2457(03)00212-8


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