Supplemental Material

This document provides additional details about the following:

Number of Supplemental Tables: 4
Number of Supplemental Figures: 3

S1. Patient Characteristics

Characteristics of patients whose EEGs were included in the study are shown in **eTable 1**.

eTable 1. Patient and data characteristics.

Number of patients:	2,711
Number of experts (fellowship trained):	30
Length of training; (years), mean (range):	1.6 (1, 2)
Length of EEG practice after training:	8.9 (2, 35)
Patient age; (years), median (IQR):	55 (41)
0- ; # n (%):	124 (4.6%)
1-5:	84 (3.1%)
5-10:	103 (3.8%)
10-15:	82 (3.0%)
15-20:	92 (3.4%)
20-30:	244 (9.0%)
30-40:	200 (7.4%)
40-50:	234 (8.6%)
50-60:	387 (14.3%)
60-70:	523 (19.3%)
70-80:	396 (14.6%)
80-90:	201 (7.4%)
90 and above:	39 (1.4%)
Female; # n (%):	1,330 (49%)
IIIC on EEG; # n (%):	2,399 (88%)
Seizure:	1,373
LPD:	475
GPD:	429
LRDA:	480
GRDA:	1,035
EEG duration; (hours), median (IQR):	18.1 (21.5)
LTM; # n (%):	5,178 (85%)
EMU:	582 (10%)
routine EEG:	335 (5%)

The total number of EEGs with consensus labels from experts of one or more IIIC patterns (SZ, LPD, GPD, LRDA, GRDA) is 2,399. Note that the numbers for EEGs containing each type of IIIC pattern add to >2,399, because for some EEGs expert consensus labels include multiple IIIC patterns.

S2. Gathering labels from experts

eTable 2. EEG labeling rounds.

Stages of labeling	Sample selection	Annotation approach	Annotation platform	Round sub-index	# raters	# segments	Subtotal
Stage 1A	Clinical EEG reports	case-by-case	Local server	1	3	1,527	
				2	3	2,174	
				3	3	4,417	
				4	4	7,134	
				5	3	1,225	16,477
Stage 1B	Clinical EEG annotations	case-by-case	Local server		92+2*	17,081	17,081
Stage 2	Active learning	combined cases	AWS	1	27	956	
				2	29	1,512	
				3	28	1,384	
				4	24	1,425	
				5	32	1,902	
				6	27	1,808	
				7	12	1,734	
				8	8	1,756	
				9	9	1,677	
				10	32	1,891	
				11	25	1,094	17,139

*92: clinical raters provided clear annotations as part of the EEG file. 2: two raters (M.B. Westover, J. Jing) independently scored each segment.

We gathered annotations for 50,697 EEG segments, each 10-second long, from 124 independent raters. Among these, 30 subspecialty experts independently scored a minimum of 1,000 segments each (see **Figure 1**, and **eTable 3**, for details). In all rating tasks, raters were given a forced choice of six options: seizure, lateralized periodic discharges (LPD), generalized periodic discharges (GPD), lateralized rhythmic delta activity (LRDA), generalized rhythmic delta activity (GRDA), and "Other" if none of those patterns was present. Here we describe the approach that we adopted to labeling the data. Our approach consisted of two stages:

Stage 1: ***Hand-targeted sample selection****:* In the first stage we collected a large number of labels from relatively small groups of 3-4 raters. Not all raters involved in Stage 1 were fellowship-trained physicians, as raters also included EEG technicians, researchers, and fellows in training. The labels gathered in this phase were used to "seed" or initialize the active learning approach in Stage 2.

- **Stage 1A**: In this stage, cases were selected for labeling by small (3-4) groups of experts. Labeling was done on local servers via a graphical user interface (GUI) developed in house using the MATLAB (Natick) programming language[1]. The goal of this stage was to gather a large number of potential examples from each of the IIIC classes to be labeled by a larger set of independent experts in Stage 2. In all, 16,477 segments were labeled by 3-4 raters each in Stage 1A.

- **Stage 1B:** The goal of Stage 1B was to identify additional examples of less common IIIC patterns, to further diversify the sample set from Stage 1A, particularly SZ, LRDA, and GRDA. Cases were selected for labeling in Stage 1B following two steps. First, we identified time-stamped annotations in the clinical EEG recordings that indicated the presence of IIIC events. These clinical labels constituted the first label for these segments. In all, such "first" annotations came from 92 different individuals, including fellowship-trained experts, EEG technologists, and physicians in fellowship training. After this step, two of the authors (M.B. Westover, J. Jing) independently reviewed and labeled the candidate segments using the custom GUI (see above). In all, 17,081 segments received labels from at least 3 raters each in Stage 1B.

Stage 2: ***Automated sample selection using active learning:*** In this Stage, we aimed to increase the sample from Stage 1 to include a large number of segments, spanning a wide variety of each IIIC pattern type, from a large number of different patients, and from a large number of subspecialty trained experts. Selecting segments randomly from all 2,711 patients' EEGs assembled for this study would be a poor strategy, however, as the large majority of the EEGs are non-IIIC patterns. Thus naïve / random sampling would have led to a highly imbalanced selection of segments for experts to label. To avoid this, we used an active learning procedure that balances random sampling with targeted sampling to ensure diversity of EEG patterns across types of IIICs and patients, as described in our prior work[2]. In this approach, annotations from Stage 1 were used to initialize an iterative "active learning" (AL) process, adapted from our prior work[2]. Briefly, our approach consisted of these steps:

1) *Changepoint segmentation*: We augmented the set of 33,558 *labeled* segments from Stage 1 with a large set of *unlabeled* segments, taken from the entire set of 2,711 EEGs assembled for this project. Within these EEGs, the unlabeled segments that became candidates for labeling by experts were identified using an automated changepoint detection algorithm[46], which identified times when the total EEG power changed. We designed this process to yield segments of relatively homogenous EEG patterns. In all, 17,157,199 segments were identified across all EEGs. To reduce redundancy, we kept only the central 10 seconds of each segment ("changepoint centers"), which became candidates for labeling by expert.
2) *Training of "query" models:* Recruiting many experts to label all 17M changepoint centers would have been infeasible, and many EEG segments are highly similar. Therefore, to choose which segments to present to experts for scoring, we developed a series of "query" models, as described in prior work[3]. Each query model was trained on all labeled data collected thus far. The trained model was used to create a 2D "embedding map" based on the features learned by the model. This map was then used to select a diverse set of segments to present to experts for labeling, referred to as using the model to "query" the experts. We explored a range of approaches to using the model to query the experts. For further technical details, please see our prior publication[2].
3) *Further Sample Diversification Strategies:* To further ensure diversity within the sample we constrained labeling assignments to be as uniform as feasible across IIIC types (to avoid over-representation of the most common IIIC types), and across patients (to avoid most segments coming from a small minority of patients).
4) *Iteratively querying the experts*. EEG segments were presented to experts for labeling via a web-based graphical user interface (GUI) developed in house. Each EEG segment to be

scored was 10-second long. Experts could pan 20 seconds before and after the segment as context, could change the viewing EEG montages, and could adjust the signal gain. In addition, a 10-minute spectrogram was provided to give additional context. We conducted 11 rounds of expert labeling in total, with variable numbers of experts participating in and completing each round; numbers are given in **eTable 2**. Between each round of multi-expert labeling, we repeated steps (2) and (3) to select the next set of segments for labeling.

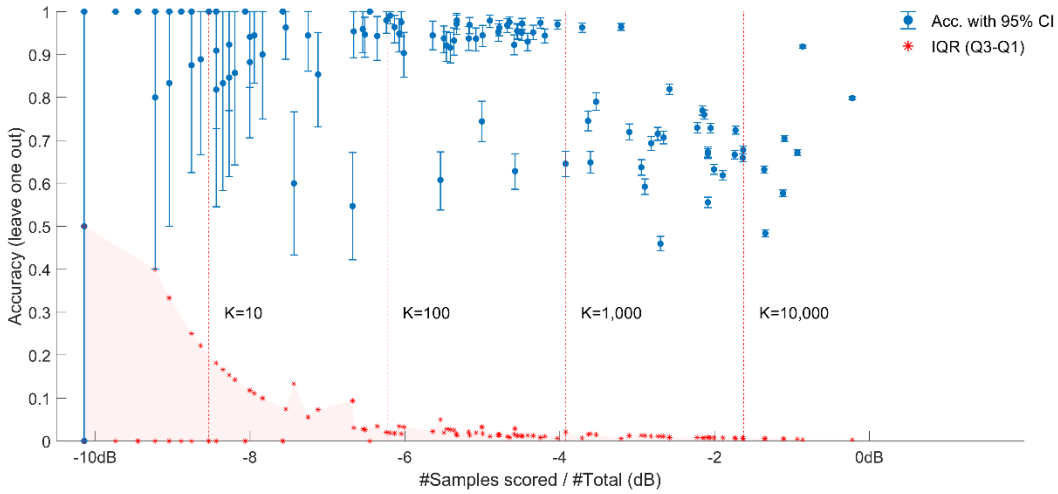
S3. Numbers of EEG segments scored by experts

eTable 3 provides the number of EEG segments scored by the 30 experts who contributed ≥1,000 annotations. The total number scored is shown in the second column ("Total # scored"). Segments received a variable number of scores. For some majority IRR noise-bias analyses, we limited analysis to segments labeled by ≥ 10 raters. Therefore, we also list the number of segments scored by each of the 30 experts that ultimately received additional labels from at least 9 other experts ("Segments with ≥ 10 labels"),

eTable 3. Numbers of EEG segments scored by experts.

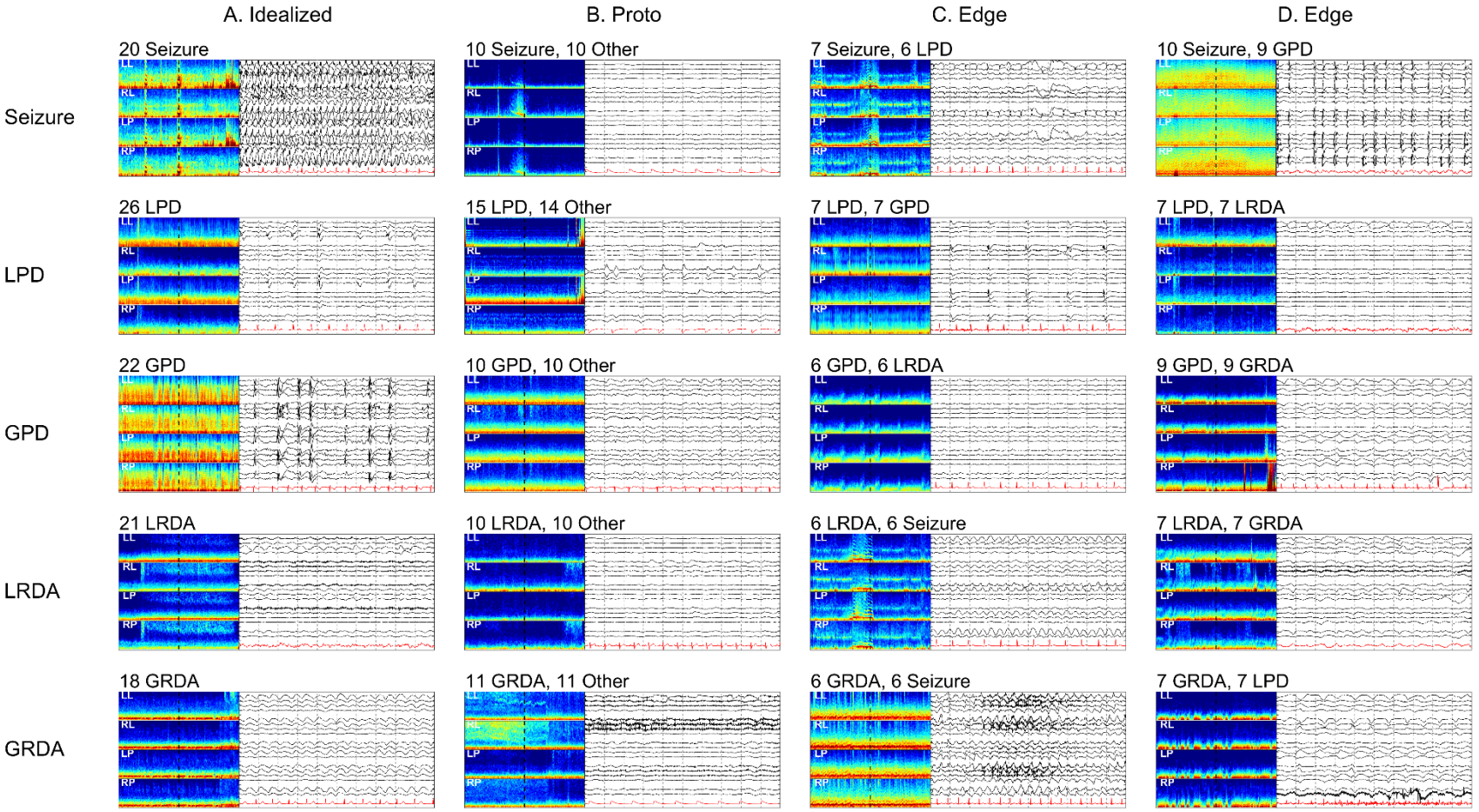
Expert #	Total # scored	Segments with ≥ 10 labels
1	45,267	11,470
2	19,987	10,301
3	16,902	11,474
4	16,601	11,473
5	13,210	11,474
6	13,029	10,479
7	10,768	4,327
8	9,923	9,923
9	9,864	9,864
10	8,992	8,566
11	8,892	8,892
12	6,810	6,810
13	6,527	6,527
14	6,289	6,289
15	6,289	6,289
16	6,286	6,286
17	6,001	6,001
18	5,859	5,859
19	3,838	3,838
20	3,554	3,554
21	3,410	3,410
22	3,293	3,293
23	3,026	3,026
24	2,852	909
25	2,663	487
26	2,281	2,256
27	1,486	1,486
28	1,381	1,341
29	1,339	1,339
30	1,002	996

To investigate whether there exist any systematic differences in rater performance based on number of samples scored, we display in **eFigure 1** the individual accuracy with 95% CI (bootstrapping 10K times, on all data relative to consensus with this rater's labels left out, i.e., "leave one rater out" consensus) vs. number of samples scored by each rater. As can be seen in **eFigure 1**, for raters scored too few samples (e.g., n<1,000 samples), the interquartile range (IQR) value is high indicating insufficient data to reliably estimate rater performance. When n increases beyond 1,000, IQR is stable, and accuracy does not correlate with # number of samples scored. We computed the correlation between # of samples scored and rater accuracy with n>1,000. The mean correlation coefficient is 0.036 [0.005, 0.068] indicating no significant correlation.



eFigure 1: Individual accuracy vs. number of samples scored.

We visually reviewed EEG segments for the five IIIC events with varying degrees of expert agreement (**Figure 2** and **eFigure 2**). We identified three main types of events: 1) "Ideal" patterns: Cases with high agreement tend to be clear examples that match standardized definitions. 2) "Proto-patterns": Cases where votes split evenly / nearly evenly between "Other" and one IIIC pattern tend to be partially / semi-formed, having some but not all classic features. 3) "Edge" cases: Cases where raters split between two 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 proto-patterns. This supports the concept that IIIC patterns lie along an "Ictal-Interictal Continuum".


A. Idealized
B. Proto
C. Edge
D. Edge
Seizure
LPD
GPD
LRDA
GRDA
20 Seizure
10 Seizure, 10 Other
7 Seizure, 6 LPD
10 Seizure, 9 GPD
26 LPD
15 LPD, 14 Other
7 LPD, 7 GPD
7 LPD, 7 LRDA
22 GPD
10 GPD, 10 Other
6 GPD, 6 LRDA
9 GPD, 9 GRDA
21 LRDA
10 LRDA, 10 Other
6 LRDA, 6 Seizure
7 LRDA, 7 GRDA
18 GRDA
11 GRDA, 11 Other
6 GRDA, 6 Seizure
7 GRDA, 7 LPD
LL
RL
LP
RP

eFigure 2. Selected EEG examples. Rows are structured with the 1st row seizure, 2nd row LPDs, 3rd row GPDs, 4th row LRDA, and 5th row GRDA. Column-wise, examples of idealized forms of patterns are in the 1st column (**A**). These are patterns with uniform expert agreement. The 2nd column (**B**) are proto or partially formed patterns. About half of raters labeled these as one IIIC pattern and the other half labeled "Other". The 3rd and 4th columns (**C, D**) are edge cases (about half of raters labeled these one IIIC pattern and half labeled them as another IIIC pattern). For **B** row 1 (**B-1**) 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 seizure, thus disagreement between SZ and "Other" makes sense. **B-2** shows frontal lateralized sharp transients at ~1Hz, but they have a reversed polarity, suggesting they may be coming from a non-cerebral source, thus the split between LPD and "Other" (artifact) makes sense. **B-3** has diffused semi-rhythmic delta background with poorly formed low amplitude generalized periodic discharges with s shifting morphology making it a proto-GPD type pattern. **B-4** shows semi-rhythmic delta activity with unstable morphology over the right hemisphere, a proto-LRDA pattern. **B-5** shows a few waves of rhythmic delta activity with an unstable morphology and is poorly sustained, a proto-GRDA. **C-1** shows 2Hz LPDs showing an evolution with increasing amplitude evolving underlying rhythmic activity, a pattern between LPDs and the beginning of a seizure, an edge-case. **D-1** shows abundant GPDs on top of a suppressed background with frequency of 1-2Hz. The average over the 10-seconds is close to 1.5Hz, suggesting a seizure, another edge case. **C-2** is split between LPDs and GPDs. The amplitude of the periodic discharges is higher over the right, but a reflection is also seen on the left. **D-2** is tied between LPDs and LRDA. It shares some features of both; in the temporal derivations it looks more rhythmic whereas in the parasagittal derivations it looks periodic. **C-3** is split between GPDs and LRDA. The ascending limb of the delta waves have a sharp morphology, and these periodic discharges are seen on both sides. The rhythmic delta appears to be of higher amplitude over the left, but there is some reflection of the activity on the left. **D-3** is split between GPDs and GRDA. The ascending limb of the delta wave has a sharp morphology and there is asymmetry in slope between ascending and descending limbs making it an edge case. **C-4** is split between LRDA and seizure. It shows 2Hz LRDA on the left, and the spectrogram shows that this segment may belong to the tail end of a seizure, an edge-case. **D-4** is split between LRDA and GRDA. The rhythmic delta appears to be of higher amplitude over the left, but there is some reflection of the activity on the right. **C-5** is split between GRDA and seizure. It shows potentially evolving rhythmic delta activity with poorly formed embedded epileptiform discharges, a pattern between GRDA and seizure, an edge-case. **D-5** is split between GRDA and LPDs. There is generalized rhythmic delta activity, while the activity on the right is somewhat higher amplitude and contains poorly formed epileptiform discharges suggestive of LPDs, an edge-case. Note: Recording regions of the EEG electrodes are abbreviated as **LL** = left lateral; **RL** = right lateral; **LP** = left parasagittal; **RP** = right parasagittal.

S4. Inter-rater reliability statistics

As listed in **eTable 4**, we adopted standard conventions for describing IRR[37] based on κ values: *Slight* 0 to 20%, *Fair* 21 to 40% (shaded yellow), *Moderate* 41 to 60% (blue), *Substantial* 61 to 80% (green), *Almost-Perfect* 81 to 100%. For pairwise IRR, median and 25th and 75th percentile values are calculated across 802 pairs of raters among 30 experts who each mutually scored at least ≥1,000 of the same segments. For majority IRR, median and 25th and 75th percentile values are calculated across the 30 experts who each scored at least 1,000 segments.

eTable 4. Expert Inter-rater reliability for IIIC events.

	Pairwise IRR		Majority IRR	
	PA	κ	PA	κ
SZ	45 [19,77]	34 [21,47]	64 [40,84]	60 [52,68]
LPD	63 [35,95]	56 [42,71]	71 [44,86]	68 [54,82]
GPD	55 [25,78]	45 [31,62]	65 [42,97]	61 [46,74]
LRDA	34 [8,70]	20 [5,36]	55 [27,85]	50 [27,71]
GRDA	44 [11,90]	33 [14,51]	60 [25,75]	56 [41,71]
Other	68 [32,89]	62 [48,79]	73 [29,91]	70 [57,86]

Values shown are median [25th percentile, 75th percentile] across pairs of raters.

S5. Calibration analysis

We characterized scorers' statistical calibration for each IIIC[4,5]. Calibration measures how well predicted probabilities agree with event frequencies. We defined the observed probability of each candidate EA type as the proportion of experts who scored it as such. To allow granularity in the analysis, we limited this analysis to segments scored by at least 10 individuals and assigned segments to one of 5 probability bins (0-20%, 20-40%, …, 80-100%). For each expert we estimated their tendency to score segments within each bin as the proportion of segments in that bin that the expert scored as that type of IIIC, using a parametric model. This defines a calibration curve for each expert (y = predicted probability, x = observed probability) for each IIIC type, and the expert's calibration score is the average absolute value of the difference between predicted and observed probabilities.

After assigning segments to bins based on votes received, for each expert and IIIC pattern we calculate the proportion of segments within each of the bins that the expert classified as that IIIC pattern. This provides up to 5 (one per bin) from which to estimate the expert's calibration curve y_i. Because of the finite sample size, these points will provide only a noisy indication of the expert's true calibration. Therefore, we fit a smooth parametric model to these data to estimate the expert's calibration curve. We do this as follows:

Let x_i be the center of probability bin i. Then we define the function $\hat{y}_i(x_i;\theta) = 1/(1+\exp(-z_i))$, where $z_i = \log(x_i/(1-x_i)) + \theta_i$. This calibration curve $\hat{y}_i(x_i;\theta)$ varies smoothly following a "football" shape (see **Figure 3A**) between the points (x,y) = (0,0), representing cases where all experts agree that a segment does not belong to a particular IIIC class, to (x,y) = (1,1), representing cases of unanimous agreement that a given segment is that type of IIIC. The one parameter θ determines whether the calibration curve $\hat{y}_i(x_i;\theta)$ follows the diagonal line $x = y$ ($\theta = 0$), goes above the diagonal line ($\theta > 0$, representing over-calling), or goes below the diagonal line ($\theta < 0$, under-calling). We fit the curve (determine the value of θ) via least squares, to minimize the mean squared error between the observations y_i and the parametric curve $\hat{y}_i(x_i;\theta)$.

In addition, we define a "calibration index" for each expert as the percentage of the maximal possible over- or under-calling that an expert's calibration curve exhibits. Specifically, for an extreme over-caller (a rater that scores all segments as belonging to a given IIIC class), the calibration curve will be maximally above the diagonal line on the calibration plot. The corresponding area of the region above the diagonal line will be 0.5. Similarly, a maximal under-caller will have an area below the diagonal line of -0.5. We thus define the calibration index as:

$c(\theta) = 100 \cdot (\hat{y}(x;\theta) - x)/0.5$, where we have introduced the scaling factor of 100 for convenience, such that the calibration index ranges between -100 (maximal under-calling) to 100 (maximal over-calling).

S6. Confusion matrices

Inter-Pattern Conditional Probabilities (Confusion Matrices): The pairwise conditional probability distribution was calculated, then averaged across pairs of experts, to obtain "confusion matrices", *P(B|C),* defined as the average probability that an expert labels a pattern as *B* given that another rater labels it *C*. The majority conditional confusion matrix is calculated similarly, as the average probability that an expert labels a pattern as *B*, given that the majority of experts labeled it as *C*.

To ensure statistical stability, we imposed the following inclusion requirements:

- For pairwise confusion matrices, we included scores from expert pair (i, j) if:
  - Expert i and expert j each scored at least $\geqslant$ 100 of the same segments
- For majority confusion matrices:
  - Each segment received scores from at least 10 experts
  - Each expert scored at least 10 segments within each IIIC category

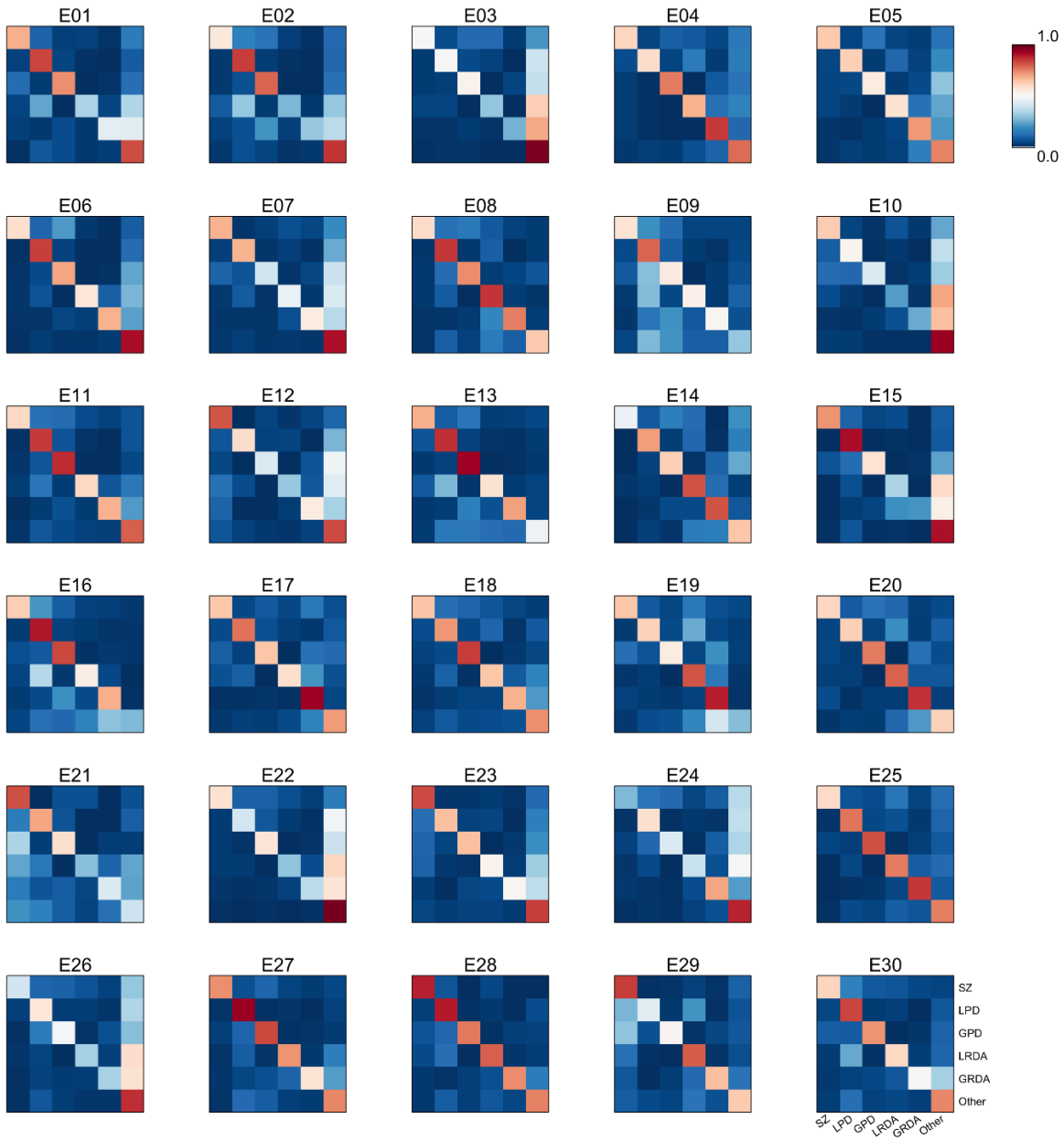
Let us index the ordered pairs of raters (i, j) who meet our inclusion criteria by $k = 1, 2, \ldots, N$, In our data, among the 30 raters (2×435 = 870 possible ordered pairs), there were 2×374 = 748 pairs that met the inclusion criteria. We denote the proportion of pairs jointly scored by the second rater in pair k scored as *B* given that the first rater scored *C* as $P_k(B|C)$. Note that this quantity forms a matrix as the values of B and C range over the different types of IIIC patterns. We take the average of these proportions across all pairs of experts as our estimate of the desired pairwise IRR confusion matrix:

$P(B|C) = \sum_{k=1}^{N} P_k(B|C)$.

Similarly, let i index the experts who met our inclusion criteria for inclusion in the analysis of majority IRR. Then we define the proportion of segments that expert i scores as *B* given that the majority score was *C* as $M_i(B|C)$. In our data, 30 experts met the inclusion criteria. We take the average of these matrices as our estimate of the majority IRR confusion matrix:

$M(B|C) = \sum_{i=1}^{N} M_i(B|C)$.

To illustrate these calculations, the component matrices $M_i(B|C)$ included in the calculation of the average $M(B|C)$ are illustrated graphically in **eFigure 3**. Axis labels are omitted to reduce clutter; however, the format is identical to **Figure 3C**, which shows the labels.



eFigure 3. Component majority IRR matrices for 30 experts.

S7. Noise and bias analysis

To investigate the reasons for differences in expert rating behavior, we created a "latent trait" model that quantifies the relative contributions of noise and bias[4–6]. Noise is assumed to account for imperfections in an expert's perception about the probability that a segment belongs to a given IIIC category, say GPD, while bias accounts for the threshold applied to the perceived probability above which the expert assigns the segment to an IIIC category. We express the perceived probability on the logit scale (to allow positive and negative values) as $z' = z + n$, where n is Gaussian noise with standard deviation σ, and $z = \log(p/(1-p))$ is the logit transformation of the true probability (proportion of experts scoring the given segment as GPD). The expert scores this

segment as GPD if $z' > \theta$, where θ is a threshold (bias) that quantifies that expert's bias for a particular IIIC category.

This model has two unknown parameters or "latent traits" for each IIIC category: the noise level σ, reflects a rater's skill in recognizing IIIC (for an ideal rater, $\sigma = 0$), and the threshold θ, represents the rater's bias as an over- or under-caller (for a rater who neither over- nor under-calls, $\theta = 0$). We hypothesized that most experts have similar levels of skill (similar σ) but disagree primarily because of bias (different θ values) – a "Similar Skill, Individualized Thresholds" (SSIT) model. To test this hypothesis, we fit the latent trait model for each expert by adjusting a single noise parameter (σ) value for all experts, and an individualized bias (θ) parameter for each expert to match three performance statistics that we measured for each expert: the expert's false positive rate (FPR), true positive rate (TPR, aka sensitivity), and positive predictive value (PPV, aka precision), relative to the group consensus, which we took as the correct answer. We quantified the degree to which the SSVT model accounted for the entire set of expert performance metrics (FPR, TPR, PPV) by calculating the percent variance explained by the model.

S8. Literature review methodology

We conducted a systematic search of the literature for publications that have studied inter-rater reliability between experts for identification of seizures and IIIC in EEGs from adult patients in the setting of critical or acute illness.

Our inclusion criteria were:

- EEGs scored by multiple experts for adult patients
- Quantifies expert inter-rater reliability for identifying seizures, or rhythmic and periodic epileptiform patterns.

We excluded studies exclusively about non-ICU or acutely ill patients; neonatal or intracranial EEG; IRR of personnel other than trained clinical neurophysiologists (e.g. medical residents).

Based on a list of key words and manuscripts that we knew covered this topic, we developed the following search in PubMed to identify articles of interest.

(interobserver OR interrater OR readers OR raters) AND
(correlation OR reliability OR agreement OR accuracy) AND
(EEG OR electroencephalography) AND
(expert OR seizure OR critical OR ICU)

Running this query on July 30, 2021 yielded 280 publications. We reviewed these manually to identify those that met our criteria. Our search finally identified 8 relevant prior studies. These are summarized in the main text in **Table 1**.

REFERENCES

1. Jing, J. *et al.* Rapid annotation of seizures and interictal-ictal-injury continuum EEG patterns. *J. Neurosci. Methods* **347**, 108956 (2021).
2. Ge, W. *et al.* Deep active learning for Interictal Ictal Injury Continuum EEG patterns. *J. Neurosci. Methods* **351**, 108966 (2021).
3. Jing, J. *et al.* Rapid Annotation of Seizures and Interictal-ictal Continuum EEG Patterns. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.* **2018**, 3394–3397 (2018).
4. Kahneman, D., Rosenfield, A., Gandhi, L. & Blaser, T. Noise. *Harv. Bus Rev* 38–46 (2016).
5. Kahneman, D., Sibony, O. & Sunstein, C. R. *Noise: a flaw in human judgment*. (Little, Brown, 2021).
6. Embretson, S. E. & Reise, S. P. *Item response theory*. (Psychology Press, 2013).