Received: 20 June 2025 | Accepted: 8 January 2026

DOI: 10.1002/epd2.70181

Epileptic Disorders

ORIGINAL ARTICLE

Noise in the diagnosis of epilepsy by experts

Fábio A. Nascimento[1] | **John R. McLaren**[2] | **Wei Zhao**[1] | **Roohi Katyal**[3] | **Irfan S. Sheikh**[4] | **Wan Yee Kong**[5] | **Danah Aljaafari**[6] | **Nirav Barot**[5] | **Selim Benbadis**[7] | **Daniel Friedman**[8] | **Jay R. Gavvala**[9] | **Jonathan Halford**[10] | **R. Edward Hogan**[1] | **Peter W. Kaplan**[11] | **Ioannis Karakis**[12] | **Atul Maheshwari**[13] | **Rebecca Matthews**[12] | **Cormac O'Donovan**[14] | **Stefan Rampp**[15,16] | **Stephan Schuele**[17] | **Joseph Sirven**[18] | **William O. Tatum**[18] | **Jonathan Williams**[1] | **Elza Márcia Yacubian**[19] | **Doyle Yuan**[4] | **Sándor Beniczky**[20] | **Olivier Sibony**[21,*] | **M. Brandon Westover**[5]

Correspondence
Fábio A. Nascimento, Campus Box 8111, 660 South Euclid Avenue, St. Louis, MO 63110, USA.
Email: fabion@wustl.edu

Funding information
National Institutes of Health, Grant/Award Number: RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347 and R01NS130119; Veterans Affairs Office of Research and Development, Grant/Award Number: I01HX003107-01A2

Abstract

Objective: To measure the relative levels of signal and noise in expert diagnosis of epilepsy.

Methods: Twenty multinational epileptologists independently reviewed 50 vignettes of adult and pediatric patients presenting with suspected seizure(s) on two separate occasions with a ≥30-day washout period. Experts provided a diagnosis of epilepsy or non-epilepsy based on clinical information and, if requested, routine EEG and neuroimaging data. Cases had an established clinical diagnosis of epilepsy or non-epilepsy based on capture of habitual paroxysmal events on video-EEG or long-term clinical follow-up. Experts' judgments were analyzed to decompose variability into different sources: signal (objective differences between cases), level noise (experts' bias toward over/under-diagnosis), pattern noise (experts' idiosyncratic reactions to specific case features), and occasion noise (inconsistency across occasions).

Results: The probability of an expert making a different diagnosis for a given case on two different occasions was 16%. The probability of two different experts making a different diagnosis for the same case was 26%. Signal (case "difficulty") accounted for 66–69% of total variation, with 31–34% attributable to noise. Level noise was the largest contributor in the absence of EEG/neuroimaging results (23%), while pattern noise dominated when test results were available (24%). Occasion noise contributed relatively little (1%) but was still sufficient to cause diagnostic reversals in 16–22% between occasions.

Significance: The degree of noise in expert diagnosis of epilepsy is substantial, stemming primarily from physicians' idiosyncratic interpretations of case features and variable dispositions toward over- or under-diagnosis. Strategies to improve reliability are needed, including standardized data collection protocols and

Olivier Sibony and M. Brandon Westover are co-senior authors.

For affiliations refer to page 466.

structured decision algorithms. For "difficult cases," where expert reliability and accuracy are lowest, our findings support current clinical practice which favors early referral for video-EEG monitoring over reliance on diagnostic anchoring. This diagnostic pathway may become more accessible with advances in EEG technology (e.g., wearable devices) and artificial intelligence.

KEYWORDS

decision making, diagnostic error, epilepsy, epilepsy diagnosis, epilepsy misdiagnosis, noise

1 | INTRODUCTION

Patients expect consistency in the diagnoses given by doctors. However, humans are inconsistent decision-makers—and physicians are no exception. Judgments about diagnoses can vary from one physician to the next, even among specialists with similar training working from the same data.[1,2] This unwanted variability in judgments is referred to as noise.[3] Limiting noise is a prerequisite to improving diagnostic accuracy.

We performed a "noise audit" to identify and quantify systematic sources of diagnostic noise in epilepsy diagnosis by epileptologists beyond the expected variability due to case difficulty. This research leveraged the framework proposed by Kahneman et al.[3] to decompose variability into its component sources for the first time in epileptology, particularly regarding epilepsy diagnosis.

We recruited 20 epileptologists from different academic centers across the world, hereafter referred to as experts, to participate in this experiment. The primary aim was to measure the amounts of different types of noise in the diagnosis of epilepsy by experts. Each expert independently reviewed 50 case vignettes from real patients presenting with suspected seizure(s) and gave their diagnostic opinion as to whether the patient has epilepsy. The same physicians repeated the same exercise after a washout period of at least 30 days.

We found that roughly one-third of diagnostic variability is attributable to noise, with the remaining two-thirds stemming from expected differences in case difficulty. These findings provide the first quantitative framework for understanding diagnostic variability in epilepsy diagnosis by experts and identify specific targets for quality improvement interventions.

Key points

- There is a high level of noise in the diagnosis of epilepsy by experts.
- Roughly one-third of the total variation in experts' final diagnoses of epilepsy was due to noise.
- Noise stems from experts' idiosyncratic case interpretations (pattern noise) and tendencies to be over- or under-callers (level noise).
- Strategies to improve reliability include standardized collection source data and structured decision protocols or diagnostic algorithms.
- Measures to limit diagnostic noise are a requirement to improve diagnostic accuracy and, as a result, patient care.

2 | METHODS

Collection and preparation of cases, along with sharing of their deidentified data, were conducted under IRB-approved protocols from Massachusetts General Brigham (2013P001024), Beth Israel Deaconess Medical Center (2022P000481), Washington University in St. Louis (201301128), University of Texas Southwestern (27031), and Louisiana State University Health Shreveport (00002494). Case review was conducted by experts who volunteered to participate in the study; therefore, this portion of the study did not require IRB approval based on our review of local IRB policies.

2.1 | Cases

We selected 50 cases—adult and pediatric—seen for suspected seizure(s) (Table 1 and File S1) from case archives of four academic tertiary epilepsy centers in the US. Patients had to have an established diagnosis of epilepsy or non-epilepsy—hereafter referred to as "clinical diagnosis"—available at the time of the data collection confirmed either by (i) video-EEG monitoring data with suspected seizure(s) recorded or (ii) clinical follow-up of at least 6 months after initial evaluation.

Two authors (FN, MBW) categorized consecutive incoming cases into four levels of difficulty bins (easy,

TABLE 1 Characteristics across all 50 cases including clinical diagnosis, difficulty level, and metrics following review by 20 experts.

Case	Clinical diagnosis	Difficulty level[a]	Difficulty level per experts[b]	Experts who requested EEG and neuroimaging[c]	Accuracy[d] (relative to clinical diagnosis)	Accuracy[d] (relative to group consensus)	Experts who changed the diagnosis between occasions 1 and 2	Percent agreement[e]
1	Epilepsy	3	3	73%	63%	63%	35%	52%
2	Non-epilepsy	4	3	78%	20%	80%	10%	67%
3	Non-epilepsy	4	3	93%	25%	75%	30%	61%
4	Non-epilepsy	2	3	88%	48%	53%	15%	48%
5	Epilepsy	1	2	70%	100%	100%	0%	100%
6	Epilepsy	2	3	58%	65%	65%	40%	53%
7	Epilepsy	1	2	65%	98%	98%	5%	95%
8	Epilepsy	2	3	100%	100%	100%	0%	100%
9	Non-epilepsy	1	3	70%	33%	68%	5%	54%
10	Non-epilepsy	3	3	78%	78%	78%	35%	63%
11	Non-epilepsy	1	3	78%	75%	75%	20%	61%
12	Non-epilepsy	3	2	28%	88%	88%	5%	77%
13	Non-epilepsy	3	3	73%	98%	98%	5%	95%
14	Epilepsy	3	3	83%	55%	55%	30%	48%
15	Epilepsy	1	2	25%	98%	98%	5%	95%
16	Epilepsy	1	2	65%	95%	95%	10%	91%
17	Epilepsy	4	3	65%	23%	78%	25%	63%
18	Non-epilepsy	3	3	73%	100%	100%	0%	100%
19	Non-epilepsy	4	3	80%	68%	68%	15%	54%
20	Epilepsy	2	3	78%	90%	90%	20%	82%
21	Epilepsy	2	3	88%	98%	98%	5%	95%
22	Epilepsy	1	2	60%	90%	90%	10%	81%
23	Epilepsy	1	2	30%	95%	95%	10%	90%
24	Non-epilepsy	4	3	60%	18%	83%	25%	70%
25	Epilepsy	2	2	55%	98%	98%	5%	95%
26	Epilepsy	3	3	88%	65%	65%	30%	53%
27	Epilepsy	1	3	55%	93%	93%	5%	86%
28	Epilepsy	3	3	88%	15%	85%	20%	73%
29	Epilepsy	4	3	100%	55%	55%	30%	48%
30	Epilepsy	4	3	38%	10%	90%	20%	83%
31	Epilepsy	3	3	58%	70%	70%	20%	58%
32	Epilepsy	2	2	23%	95%	95%	10%	90%
33	Epilepsy	4	3	100%	40%	60%	20%	49%
34	Non-epilepsy	3	3	95%	90%	90%	20%	81%
35	Epilepsy	3	3	93%	15%	85%	20%	73%
36	Non-epilepsy	4	3	88%	85%	85%	0%	73%
37	Epilepsy	4	3	63%	25%	75%	30%	61%
38	Non-epilepsy	3	3	88%	85%	85%	20%	75%
39	Non-epilepsy	2	3	68%	85%	85%	20%	73%
40	Non-epilepsy	2	3	65%	90%	90%	20%	82%
41	Epilepsy	4	2	53%	50%	50%	30%	48%
42	Non-epilepsy	1	2	45%	100%	100%	0%	100%
43	Non-epilepsy	2	3	48%	85%	85%	20%	74%
44	Non-epilepsy	4	3	83%	90%	90%	10%	82%

(Continues)

TABLE 1 (Continued)

Case	Clinical diagnosis	Difficulty level[a]	Difficulty level per experts[b]	Experts who requested EEG and neuroimaging[c]	Accuracy[d] (relative to clinical diagnosis)	Accuracy[d] (relative to group consensus)	Experts who changed the diagnosis between occasions 1 and 2	Percent agreement[e]
45	Non-epilepsy	4	3	88%	73%	73%	25%	58%
46	Non-epilepsy	3	3	75%	60%	60%	10%	50%
47	Non-epilepsy	2	3	50%	95%	95%	0%	90%
48	Non-epilepsy	1	2	25%	98%	98%	5%	95%
49	Non-epilepsy	1	2	38%	88%	88%	25%	77%
50	Non-epilepsy	1	3	68%	68%	68%	35%	54%

[a]Difficulty level per pre-study evaluation by authors (scale of 1–4, where 1 is easy and 4 is hard).

[b]Mean difficulty level assigned to cases by participating experts across occasions 1 and 2 (scale 1–5, where 1 is easy and 5 is hard).

[c]Mean percentage across occasions 1 and 2.

[d]Mean accuracy taking the final diagnosis (question 1 and, if applicable, question 2) across occasions 1 and 2.

[e]Percent agreement (whether epilepsy or non-epilepsy) taking the final diagnosis (question 1 and, if applicable, question 2) across occasions 1 and 2.

moderate, difficult, very difficult) with the aim of creating a case mix typical for an epilepsy subspecialty practice. "Easy" and "very difficult" corresponded to cases they judged most experts would or would not diagnose correctly relative to the clinical diagnosis (epilepsy or non-epilepsy). Moderate and difficult cases fell in between. Case collection continued until reaching an equal number of epilepsy (25) and non-epilepsy (25) cases and comparable numbers of cases in each level of difficulty. Non-epilepsy cases were not further categorized into subtypes based on the specific etiology (e.g., psychogenic, physiologic). Among epilepsy cases, there were 13 confirmed by long-term follow-up and 12 by capture of seizure(s) on EEG. Among non-epilepsy cases, 19 were confirmed by long-term follow-up and 6 by capture of non-seizure events on EEG.

2.2 | Participants

We recruited experts via email. Authors who contributed cases were not eligible to participate in case review. For each case, participants read a vignette summarizing relevant history and physical examination data available at the initial patient evaluation and were asked the likelihood that the patient has epilepsy (*question 1*) according to ILAE diagnostic criteria.[4] Possible responses were "low," "high," and "intermediate."

If "intermediate" was selected, a representative sample of the patient's EEG was presented (10- to 20-second epoch on bipolar and common average referential montages) and, if available, a short summary of neuroimaging findings. The EEG epoch shown was selected by the authors to present any suspicious sharp transients, if present. Participants were then asked if the patient has epilepsy (*question 2*). Possible responses were "yes" or "no." Finally, participants were asked to rate the level of difficulty for each case (*question 3*) using a 5-point Likert scale ("very easy" to "very difficult"). A representative clinical vignette is shown in Table 2.

The first round of case review was *occasion 1*. After a washout period of at least 30 days, participants reviewed the same cases and answered the same questions again (*occasion 2*).

2.3 | Signal strength and system noise

We performed a modified analysis of variance (ANOVA) to estimate the contributions of several underlying components to the total variation of expert responses. Technical details of the approach are provided in File S2. We refer to variation that reflects true differences in case difficulty as "signal," where some cases have clearer evidence for or against epilepsy than others. Signal strength represents the proportion of total diagnostic variation explained by these case differences, calculated from the variation across cases in the frequency of affirmative (epilepsy) diagnosis. Signal strength is 100% and there is zero noise when all diagnostic variation stems from case characteristics, with no contribution from noise. Conversely, signal strength is 0% when all variation is due to noise. We refer to the remaining variation after signal strength is accounted for as *system noise*. In other words, system noise is unwanted variance.

2.4 | Noise decomposition

We decomposed system noise into three subtypes: level noise (LN), pattern noise (PN), and occasion noise (ON). LN is the main effect of the expert. This reflects an expert's

TABLE 2 Representative clinical vignette used in the study.

Question 1

Clinical history

A 48-year-old right-handed female software engineer with a history of hypertension, anxiety, and depression developed 10-s spells beginning a year prior characterized by an aura of an indescribable feeling that a seizure was about to happen (and, at times, a warmth in her face). This was followed by staring and bilateral manual automatisms, with occasional hyperventilation. She had no postictal symptoms, and she was partly amnestic to the spells. She never had a convulsion.

She denied alcohol, tobacco, or recreational drug use. She had no history of febrile seizures, central nervous system infection, traumatic brain injury, or family history of seizures.

Her exam was unremarkable.

Based on the clinical data, what is the likelihood this patient has epilepsy?

- Low enough to conclude that this patient does not have epilepsy and an antiseizure drug should not be started
- Intermediate thus requiring further testing: neuroimaging and EEG
- High enough to conclude that this patient has epilepsy and an antiseizure drug should be started

If "intermediate" was selected, experts were taken to *Question 2*. If "low" or "high" were selected, experts were taken directly to *Question 3*.

Question 2

Neuroimaging and EEG

A brain MRI showed symmetrical hippocampi but slight increase in T2 signal within the left hippocampus.

A representative sample of this patient's EEG is shown below.

Based on the clinical data, neuroimaging, and EEG, do you think this patient has epilepsy?

- Yes: this patient has epilepsy and an antiseizure drug should be started
- No: this patient does not have epilepsy and an antiseizure drug should not be started

Question 3

Please rate the level of difficulty of this case (accounting for all the available data: clinical information, neuroimaging, EEG) in terms of determining whether or not this patient has epilepsy.

- Very easy
- Easy
- Neither easy nor difficult
- Difficult
- Very difficult

tendency to over- or under-diagnose epilepsy relative to other experts for a given case. PN measures the interaction between cases and experts (i.e., variability in the idiosyncratic but stable responses of a given expert to information about cases). For instance, a given expert might be more likely than others to give a diagnosis of epilepsy in cases with a history of peri-event bowel/bladder incontinence. ON is within-expert temporal variability, measurable because of repeated testing. This variability arises from transient factors that influence judgment, such as mood,[3,5,6] time since last meal,[3,7] and weather.[3,8,9] We assume experts would not remember cases between occasions because of the washout period of at least 30 days between *occasions 1* and *2*. These definitions are summarized in Table 3.

2.5 | Statistical analysis

Correlations are quantified using Spearman and biserial correlation coefficients. Statistical significance is defined based on an alpha level of .05. Data and code to reproduce all results and figures are available through a public data sharing repository at https://github.com/bdsp-core/Noise_in_Diagnosing_Epilepsy.

3 | RESULTS

Twenty experts completed all 50 cases on both *occasions*. They represented 18 academic centers across 5 countries. Experts reported regularly interpreting EEGs and seeing patients with seizures and epilepsy, both during clinical practice and fellowship training, for an average of 20 years (2.5–40; median 19; IQR 11.8–30.3). Results are summarized in Table 1.

3.1 | Variability between experts

Experts varied widely in their rates of making affirmative (epilepsy) diagnoses. For *question 1*, considering both *occasions*, the number of affirmative diagnoses varied from 0 to 17 of 50 (0–34% of cases). Among those who answered *question 2*, the range of affirmative diagnoses was 9 (of 27 cases answered) to 29 (of 45 cases answered) (Figure 1). Figure 2A,B shows how the percentage of affirmative diagnoses varies with signal strength. The probability of receiving the same final diagnosis from two different experts was 74%. Figure 2C,D shows how the percentage of experts agreeing varies with the signal strength, dipping to 50% for the most difficult cases.

3.2 | Variability across cases

Cases varied in the rate at which experts assigned a diagnosis of epilepsy. For *question 1*, this rate ranged from 0 to 16 of 20 experts (0–80%). For each case, the number who were undecided and asked for further neuroimaging/EEG information ranged from 4 to 20 (20–100%).

TABLE 3 Definitions of signal, system noise, and its decompositions.

Component	Definition
Signal (S)	• Variation of expert responses based on actual case difficulty differences • "Consensus variation"
Signal strength	• Proportion of variation of expert responses explained by case difficulty differences • Calculated from the variation across cases in the frequency of affirmative (epilepsy) diagnosis
System noise	• Proportion of variation of expert responses that cannot be explained by case difficulty differences • Decomposed into *level noise*, *pattern noise*, and *occasion noise*
Level noise (LN)	• Effect of the expert • Expert's tendency to over- or under-diagnose epilepsy relative to other experts
Pattern noise (PN)	• Effect of the interaction between the expert and cases • Expert's idiosyncratic but stable responses to information about cases
Occasion noise (ON)	• Effect of the interaction between the expert, cases, and time (within-expert temporal variability) • Expert's tendency to change the judgment when reviewing the same case on different occasions

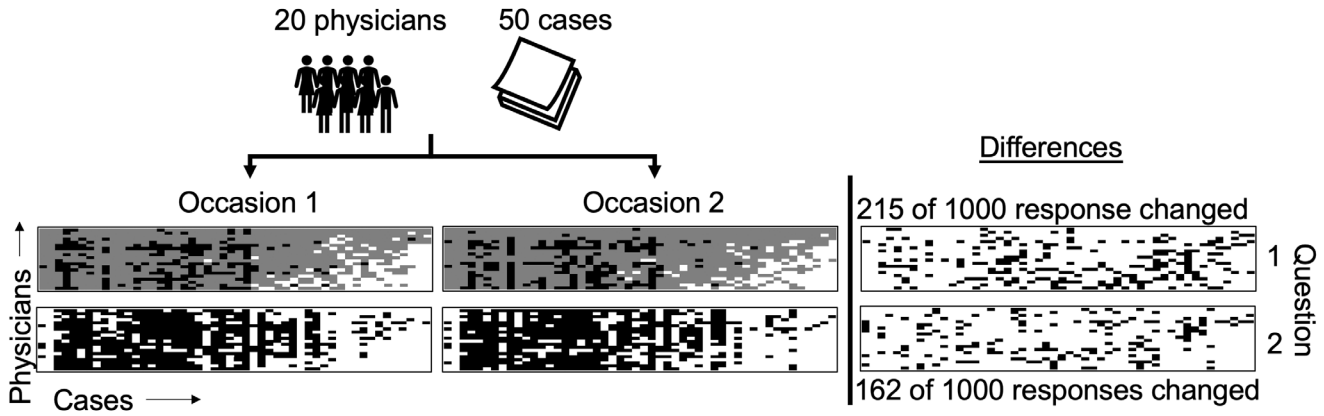

FIGURE 1 Expert response data. Twenty experts independently reviewed 50 case vignettes on two different occasions, at least 30 days apart (occasion 1, left; occasion 2, right). After reading the vignette, experts were asked Question 1: "Based on the clinical data, what is the likelihood this patient has epilepsy?" Answer options were: (1) No—"Low enough to conclude that this patient does not have epilepsy and an antiseizure drug should not be started"; (2) Yes—"High enough to conclude that this patient has epilepsy and an antiseizure drug should be started"; and (3) Unsure—"Intermediate, thus requiring further testing: Neuroimaging and EEG." Answers to Question 1 are indicated in the upper heatmap as shaded dots: Black ("no"), white ("yes"), and gray ("unsure"). The 20 rows of the heatmaps correspond to the 20 experts, and the 50 columns to the 50 cases. In cases where the expert answered "unsure," EEG and neuroimaging (if available) data were provided and the expert was asked Question 2: "Based on the clinical data, EEG, and neuroimaging (if available), do you think this patient has epilepsy?" (yes/no). The Question 2 panel displays the final diagnosis for each expert–case assessment. When Question 2 was not asked (because Question 1 was answered Low or High), the final diagnosis shown is the Question 1 decision; when Question 2 was asked, the final diagnosis shown is the Question 2 decision. Answer choices at this stage were limited to "yes" and "no." Differences between answers to the same case from the same expert on the two occasions are shown in the right-hand panel, with black dots indicating that the answer given on occasion 1 differed from that on occasion 2. Rows are sorted by experts' tendency to answer "yes," and columns by the tendency of cases to receive "yes" votes across the entire group of experts.

3.3 | Variability within experts (consistency)

Experts frequently gave different responses on *occasions 1* and *2*. Overall, the final diagnosis was reversed 162 times out of 1000. Thus, the probability of receiving a different final diagnosis when seeing the same expert on different occasions was 16%. On *question 1*, the average number of changed answers was 11 out of 50 (22%); two experts changed their minds in 16 cases (32%). Overall, for *question 1*, of the 1000 judgments on *occasion 1* (20 experts x 50 cases), 215 (21.5%) were different on *occasion 2*. For the final diagnosis (after *questions 1* and *2*), all experts changed their mind at least four times (8% of cases); the average number of changed answers was 8 (16% of cases), and 5 experts changed their mind 10 or

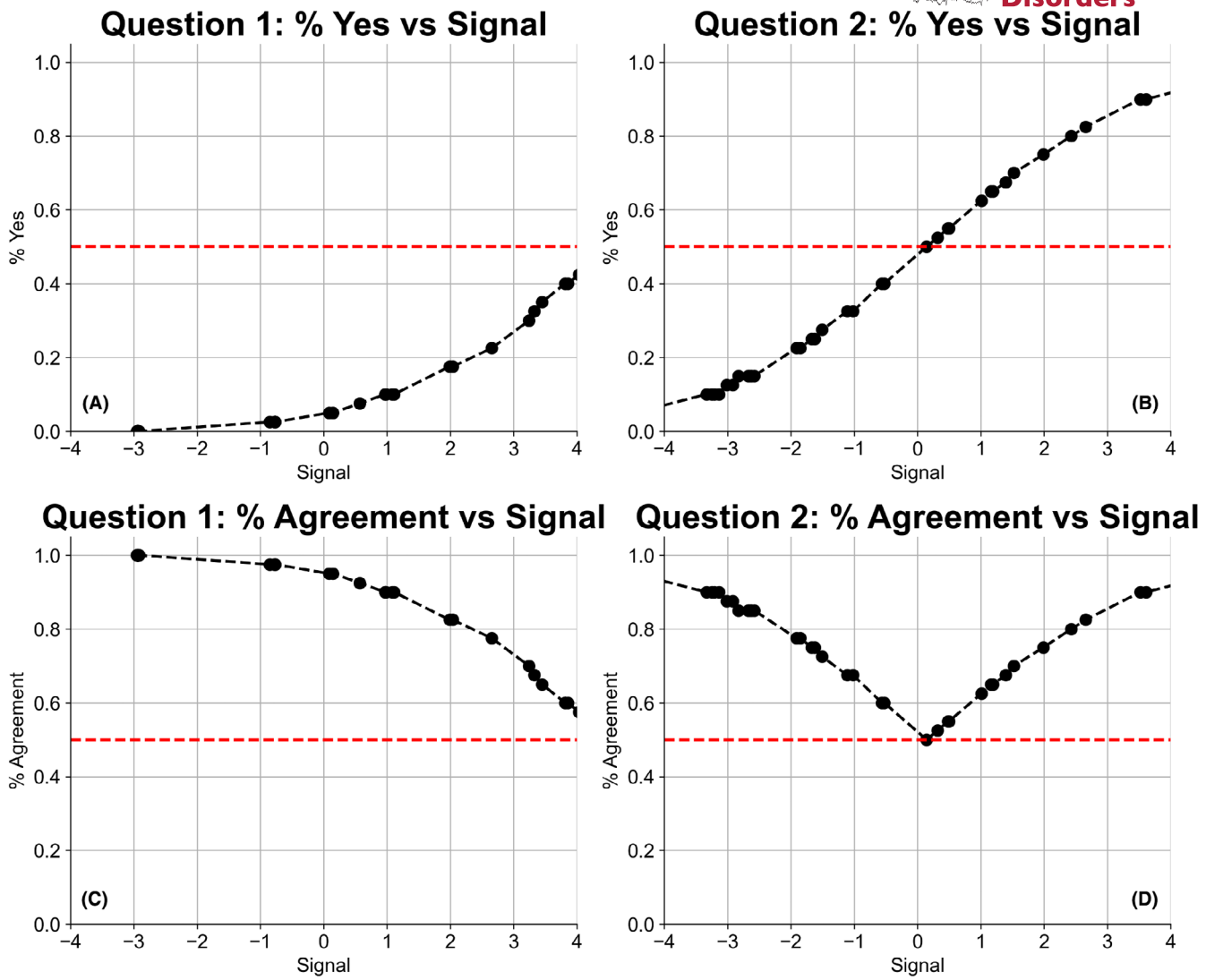

FIGURE 2 Expert agreement vs case difficulty. The panels show the percentage of experts answering "yes" to *question 1* (A) and to *question 2* (B), and the percent agreement (regardless of whether the answer is yes vs. not yes) for *question 1* (C) and *question 2* (D). Note that signal strength ranges from −4 to +4; magnitudes closest to zero represent difficult cases, whereas magnitudes closest to −4 or +4 represent easy non-epilepsy and easy epilepsy cases, respectively.

more times (≥20% of cases). Additional consistency data is summarized in File S3.

3.4 | Diagnostic accuracy

After *question 1*, the mean accuracy of experts relative to the clinical diagnosis across both *occasions* was 60% (50–70%), whereas for the final diagnosis (after *question 2*), the mean accuracy was 71% (63–81%). There were no significant correlations between the accuracy of the final diagnosis and consistency ($r = -.09$; $p = .72$) or between accuracy and years in practice ($r = .18$; $p = .46$).

We also assessed diagnostic accuracy a second way, treating expert consensus with respect to the final diagnosis as truth. The two ways of defining the "true" diagnosis—clinical diagnosis and group consensus—were moderately correlated ($r = .55$, $p < .01$) and matched in 39 of 50 cases (78%). After *question 1*, the mean accuracy relative to group consensus was 64% (52–76%), whereas for the final diagnosis (considering *questions 1* and *2*), the mean accuracy was 81% (72–91%). Accuracy of the final diagnosis was correlated with consistency ($r = .46$; $p = .04$) but not with years in practice ($r = -.18$; $p = .45$).

Accuracy related to the final diagnosis varied with case difficulty (Figure S1). Considering the clinical diagnosis as truth, cases rated as easy and moderate per pre-study evaluation by the authors had a mean accuracy of 87%, whereas cases rated as difficult and very difficult had a mean accuracy of 56%. According to in-study rating by experts, averaged across both occasions, cases rated as

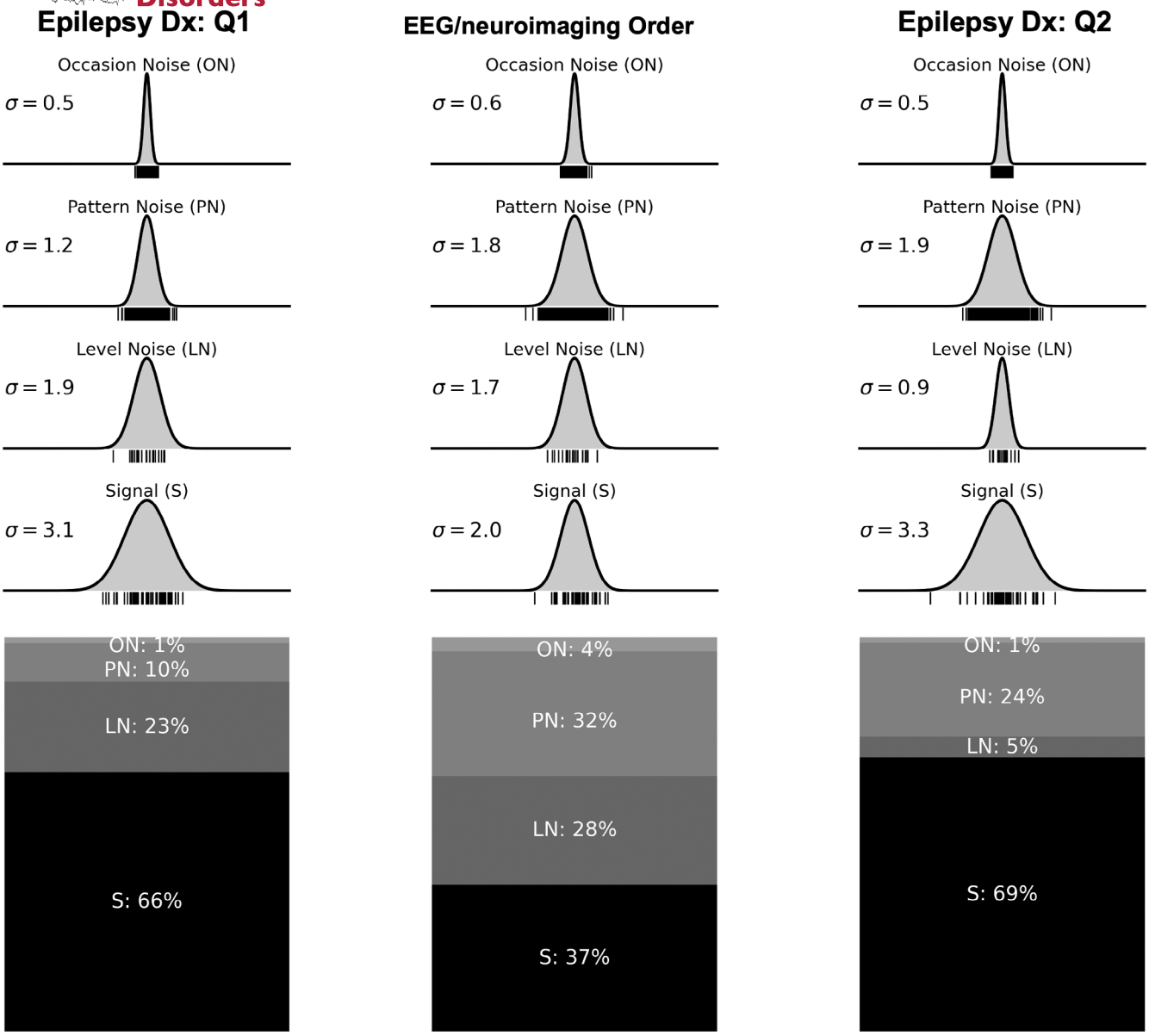

FIGURE 3 Noise decomposition: Quantitative summary. Noise decomposition results are shown for *question 1* (left), the decision of whether EEG and neuroimaging (if available) data was required (middle), and for *question 2* (right). Note that *question 2* refers to the final diagnosis. Noise is quantified by fitting normal distributions to the raw data from Figure S2 to estimate standard deviations, referred to as "signal" and "noise" levels. The stacked bar plots show the relative contributions of the signal and noise levels. S, signal; LN, level noise; PN, pattern noise; ON, occasion noise.

easy had a mean accuracy of 92%, whereas cases rated as neither easy nor difficult had a mean accuracy of 64%. Accuracy metrics considering the group consensus as truth, and percent agreement stratified by case level difficulty, are shown in Figure S1.

3.5 | Effect of EEG and neuroimaging

Given that experts had the option to request further testing with EEG and neuroimaging, we also analyzed judgments according to how many times extra testing was requested. These data are elaborated on in File S3.

3.6 | Noise analysis

For each of the two *occasions*, the analysis of variance on the 1000 judgments showed a high level of noise. A visualization of the decomposition into signal and noise components is provided in Figure S2, and the findings are summarized in Figure 3.

Figure 3 shows that the main effect of cases, the consensus variation or "signal," accounts for 66% of total variation for *question 1*, and 69% for the final diagnosis.* The

*In the *Noise anaylsis* section, "question 2" refers to the final diagnosis.

remainder—34% of total variation for *question 1*, 31% for *question 2*—is system noise.

The main effect of expert diagnostic threshold (LN) accounts for 23% and 5% of response variation for *questions 1* and *2*, respectively. Pattern noise accounts for 10% of the response variation in *question 1*, and 24% for *question 2*. Occasion noise accounts for 1% of response variation for both *questions 1* and *2*. These findings suggest that most of the system noise initially derives from experts' differing level of willingness to make a diagnosis on the basis of history alone (LN), whereas after receiving test results, the total noise remains nearly the same (signal increased by just 3%) but derives primarily from experts' variable weighting of the available information (PN). Interestingly, despite the significant observed rate of diagnostic flip-flopping between *occasions* (22% for question 1, 16% for question 2; see Figure 1), occasion noise (ON) makes only a small contribution (1%) to the total variance. This suggests that most diagnostic reversals occur in cases for which the level of signal is weak and/or near that expert's point of equipoise, so that even small amounts of noise suffice. Figure 3 presents the same noise analysis for the decision to request further testing with EEG and neuroimaging.

Further analyses are presented in File S3 and Table S1.

4 | DISCUSSION

We identified a remarkably high level of noise in expert diagnoses of epilepsy. The probability of two different experts making a different diagnosis for the same case was 26%, and the probability of an expert making a different diagnosis for the same case on two different occasions was 16%. Noise accounted for one-third (31–34%) of the total variation across cases in experts' final diagnosis of a patient as having or not having epilepsy. These findings, however, should be interpreted considering study limitations discussed below, most notably that the experimental design did not fully capture real-world clinical practice.

The finding that expert judgments are noisy is remarkable, especially given the significant clinical repercussions that diagnostic inconsistency can have on patient care. However, noisy expert judgments are not unique to medicine. Within-expert noise (ON) has been documented in various fields. Professionals often contradict their own previous judgments when presented with the same data on different occasions.[3,10] When 22 physicians evaluated coronary angiograms on two different occasions, they disagreed with themselves 63% to 92% of the time.[11] When software developers were asked on two separate days to estimate the completion time for a task, the hours they projected differed by 71% on average.[12] Between-expert noise is likely even higher. In many tasks, experts' decisions exhibit significant unwanted variability: appraising real estate,[13] valuing stocks,[14] sentencing criminals,[15,16] auditing financial statements,[17] interpreting EEGs,[18] and more.[3] Clearly, professionals often make decisions that deviate substantially from their peers, their own prior decisions, and from the rules that they themselves aim to follow.

Our study has limitations. First, all cases were first presentations, thus experts did not have access to longitudinal data, additional testing if needed (e.g., epilepsy monitoring unit evaluation or ambulatory EEG for capture of habitual suspected seizure(s)), or data from smartphone footage of patients' suspected seizure(s), which can be helpful.[19] Further, experts were asked to render a binary diagnosis of epilepsy vs. non-epilepsy, and they did not have the option to express uncertainty or, as noted above, to request additional tests such as video-EEG monitoring. However, acquiring additional information is often a lengthy process and can involve referrals to other specialists, ancillary tests, and potentially trial and error with interventions. Additionally, even when the available information is incomplete or otherwise unsatisfactory (as is typically the case on the first visit), clinicians still need to render a working diagnosis of epilepsy vs. non-epilepsy after every visit. Consequently, the process itself imposes diagnostic delay, and misdiagnoses in either direction impact patient safety, quality of life, and healthcare utilization.

Another possible limitation is that the case vignettes contained limited information. In face-to-face patient visits, clinicians are free to obtain additional history. That said, it is possible that more information would have amplified noise.[3] EEGs were presented as screenshots, which precluded experts from interacting with the EEG. Similarly, experts were not able to review neuroimages themselves as we only presented a text summary of imaging findings. This may be a limitation, although it is possible that reviewing imaging findings and "second guessing" the radiologist's findings may add noise to the diagnostic process.[20]

The specific noise decomposition percentages observed in the study likely reflect specific details of our study. The values, therefore, may not directly generalize to other clinical scenarios. Replication studies will need to validate the relative contributions of signal and noise types in other contexts. Nonetheless, the methodological framework for decomposing diagnostic variability can be applied across different settings to quantify and qualify reliability as well as to identify improvement targets.

A further potential limitation is the deliberate design of our cohort. We aimed for equal representation of epilepsy and non-epilepsy cases, and an even distribution across difficulty levels. This approach was chosen over random selection to encompass the full spectrum of case

complexities encountered in a tertiary epilepsy center. Expert ratings, averaged across both occasions, classified 74% (37/50) of cases as "neither easy nor difficult" (difficulty level 3/5) and 26% (13/50) as "easy" (difficulty level 2/5). However, this distribution may or may not represent the typical case mix any given epilepsy expert encounters. Informal post-study discussions revealed mixed opinions among experts: some felt the case mix was representative, while others perceived a higher proportion of difficult cases compared to their usual practice. Average case difficulty likely varies across clinics due to factors such as patient population, referral patterns, and other variables. Our findings (Figure 2) indicate that most noise arises from "difficult cases," as expected. Consequently, the overall level of diagnostic noise may differ between centers and clinics based on their typical case complexity.

A final relative limitation is that our findings are primarily based on mid-career to senior epileptologists. Examining decision-making patterns among trainees and early-career epileptologists may reveal different noise profiles. Such research would be valuable for both clinical purposes and quality improvement initiatives—for instance, these data could inform targeted educational interventions for those early in their epilepsy training or clinical practice.

What prevents physicians from recognizing that their judgments are noisy? The answer lies in two familiar phenomena: professionals tend to have high confidence in their own judgments and they hold their colleagues' capabilities in high regard. This combination leads to an overestimation of agreement.

Why are the judgments of epilepsy experts so noisy? The ability to make highly reliable judgments can develop when professionals practice in a predictable environment with clear, immediate feedback (e.g., driving, professional sports, surgery). However, in many professions, including neurology, professionals learn to make judgments by listening to mentors and colleagues explain and critique (apprenticeship)—a less reliable source of knowledge than learning from one's own mistakes. Prolonged experience on a job increases people's confidence in their judgments, but in the absence of rapid feedback, confidence is no guarantee of either accuracy or consensus. Moreover, epilepsy specialists typically diagnose epilepsy based on indirect information about paroxysmal events which they have not observed, and this information is often uncertain and thus inherently challenging to interpret.

How can we reduce noise in the diagnosis of epilepsy? One approach is to replace subjective human judgment with objective formal rules—algorithms—that use data about a case to produce a consistent prediction or a decision. There have been efforts to provide algorithms for the diagnosis of epilepsy,[21–25] but more work remains to validate the algorithms and to compare them to human judgment. Based on our results considering the clinical diagnosis as truth, an algorithm that correctly diagnoses patients with epilepsy with greater than ~75% accuracy may be useful in clinical practice for evaluating patients with possible seizures.

Another approach is to adopt procedures that promote consistency by ensuring personnel in the same role use similar methods to gather information, integrate it into a comprehensive view of the case, and translate that view into a decision. In this approach, judgments are structured into component parts. This takes the form of a scoring system, in which experts score each component separately, then mechanically combine the answers into an overall score.

Finally, in any approach adopted, judgment reliability can be improved by obtaining more information earlier in the diagnostic workup—particularly video-EEG monitoring, or even smartphone footage, capturing patients' habitual seizure(s). While this is not always possible or practical (e.g., in resource-limited settings, or when seizures occur infrequently), new EEG technology and artificial intelligence-assisted EEG analysis may make it possible to offer early video-EEG monitoring realistic to a wider range of patients. Early referral for video-EEG monitoring is distinctly important for "difficult cases," where expert reliability and accuracy are lowest.

Mitigating noise ultimately stands to help avoid epilepsy misdiagnosis. This is paramount because misdiagnoses have major negative consequences[26,27] for patients and healthcare systems.[26,28,29] The former include stigma, driving and employment restrictions, unnecessary investigations, inappropriate exposure to antiseizure medications (ASMs)[26] and even neurostimulation devices,[30] and failure to diagnose and address the actual etiology of patients' seizures.

We advocate development of validated diagnostic algorithms and structured judgment protocols to standardize the evaluation process. In addition, our work suggests the need for improved methods of ascertaining the true diagnosis ("gold standard") for future studies of epilepsy.

AFFILIATIONS

[1]Department of Neurology, Washington University School of Medicine, St. Louis, Missouri, USA
[2]Department of Neurology, Boston Children's Hospital, Boston, Massachusetts, USA
[3]Department of Neurology, Louisiana State University Health Sciences, Shreveport, Louisiana, USA
[4]Department of Neurology, University of Texas Southwestern Medical Center, Dallas, Texas, USA
[5]Department of Neurology, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[6]Department of Neurology, King Fahd Hospital of the University, College of Medicine, Imam Abdulrahman Bin Faisal University, Dammam, Saudi Arabia
[7]Department of Neurology, University of South Florida – Tampa

General Hospital, Tampa, Florida, USA
[8]Department of Neurology, New York University, New York, New York, USA
[9]Department of Neurology, McGovern Medical School, University of Texas Health Science at Houston, Houston, Texas, USA
[10]Department of Neurology, Medical University of South Carolina, Charleston, South Carolina, USA
[11]Department of Neurology, Johns Hopkins, Baltimore, Maryland, USA
[12]Department of Neurology, Emory University School of Medicine, Atlanta, Georgia, USA
[13]Department of Neurology, Baylor College of Medicine, Houston, Texas, USA
[14]Department of Neurology, Wake Forest University School of Medicine, Winston-Salem, North Carolina, USA
[15]Department of Neurosurgery and Neuroradiology, University Hospital Erlangen, Erlangen, Germany
[16]Department of Neurosurgery, University Hospital Halle (Saale), Halle (Saale), Germany
[17]Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA
[18]Department of Neurology, Mayo Clinic, Jacksonville, Florida, USA
[19]Department of Neurology and Neurosurgery, Universidade Federal de São Paulo, São Paulo, Brazil
[20]Department of Clinical Neurophysiology, Danish Epilepsy Center, Dianalund and Aarhus University Hospital, Aarhus, Denmark
[21]HEC Paris, Jouy-en-Josas, France

ACKNOWLEDGMENTS

We gratefully acknowledge intellectual contributions from Dr. Daniel Kahneman, who passed away on March 27, 2024. This work was supported by grants from the NIH (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119), the Veterans Affairs Office of Research and Development (I01HX003107-01A2), and NSF (2014431).

CONFLICT OF INTEREST STATEMENT

F. Nascimento serves as Associate Editor of Epileptic Disorders. S. Beniczky serves as Editor-in-Chief of Epileptic Disorders. M. Brandon Westover is a co-founder, scientific advisor, and consultant to Beacon Biosignals and has a personal equity interest in the company; the company was not involved in this work. The remaining authors have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from https://github.com/bdsp-core/Noise_in_Diagnosing_Epilepsy. Restrictions apply to the availability of these data, which were used under license for this study. Data are available from https://github.com/bdsp-core/Noise_in_Diagnosing_Epilepsy with the permission of https://github.com/bdsp-core/Noise_in_Diagnosing_Epilepsy.

ORCID

Fábio A. Nascimento https://orcid.org/0000-0002-7161-6385
John R. McLaren https://orcid.org/0000-0002-4463-9610
Roohi Katyal https://orcid.org/0000-0002-9147-3035
Selim Benbadis https://orcid.org/0000-0003-0214-5492
Jay R. Gavvala https://orcid.org/0000-0002-9392-6608
Ioannis Karakis https://orcid.org/0000-0001-5122-7211
Stephan Schuele https://orcid.org/0000-0003-3261-7120
William O. Tatum https://orcid.org/0000-0002-4536-3791
Sándor Beniczky https://orcid.org/0000-0002-6035-6581
M. Brandon Westover https://orcid.org/0000-0003-4803-312X

REFERENCES

1. Miller M, Ackerman AB. How accurate are dermatologists in the diagnosis of melanoma? Degree of accuracy and implications. Arch Dermatol. 1992;128(4):559–60.
2. Robinson PJ, Wilson D, Coral A, Murphy A, Verow P. Variation between experienced observers in the interpretation of accident and emergency radiographs. Br J Radiol. 1999;72(856):323–30. https://doi.org/10.1259/bjr.72.856.10474490
3. Kahneman DSO, Sunstein CR. Noise: a flaw in human judgement. New York: Little Brown Spark; 2021.
4. Fisher RS, Acevedo C, Arzimanoglou A, Bogacz A, Cross JH, Elger CE, et al. ILAE official report: a practical clinical definition of epilepsy. Epilepsia. 2014;55(4):475–82. https://doi.org/10.1111/epi.12550
5. Forgas JP. On being happy and mistaken: mood effects on the fundamental attribution error. J Pers Soc Psychol. 1998;75(2):318–31. https://doi.org/10.1037//0022-3514.75.2.318
6. Forgas JP. Affect and cognition. Perspect Psychol Sci. 2008;3(2):94–101. https://doi.org/10.1111/j.1745-6916.2008.00067.x
7. Danziger S, Levav J, Avnaim-Pesso L. Extraneous factors in judicial decisions. Proc Natl Acad Sci U S A. 2011;108(17):6889–92. https://doi.org/10.1073/pnas.1018033108
8. Simonsohn U. Clouds make nerds look good: field evidence of the impact of incidental factors on decision making. J Behav Decis Mak. 2007;20(2):143–52. https://doi.org/10.1002/bdm.545
9. Heyes A, Saberian S. Temperature and decisions: evidence from 207,000 court cases. Am Econ J Appl Econ. 2019;11(2):238–65.
10. Ashton RH. A review and analysis of research on the test–retest reliability of professional judgment. J Behav Decis Mak. 2000;13(3):277–94.
11. Detre KM, Wright E, Murphy ML, Takaro T. Observer agreement in evaluating coronary angiograms. Circulation. 1975;52(6):979–86. https://doi.org/10.1161/01.cir.52.6.979
12. Grimstad S, Jørgensen M. Inconsistency of expert judgment-based estimates of software development effort. J Syst Softw. 2007;80(11):1770–7.
13. Adair A, Hutchison N, MacGregor B, McGreal S, Nanthakumaran N. An analysis of valuation variation in the UK commercial property market: Hager and Lord revisited. J Prop Valuat Invest. 1996;14(5):34–47.

14. Slovic P. Analyzing the expert judge: a descriptive study of a stockbroker's decision process. J Appl Psychol. 1969;53:255–63.
15. Clancy K, Bartolomeo J, Richardson D, Wellford C. Sentence Decisionmaking: the logic of sentence decisions and the extent and sources of sentence disparity. J Crim Law Criminol. 1981;72(2):524–54.
16. Anderson JM, Kling JR, Stith K. Measuring Interjudge sentencing disparity: before and after the Federal Sentencing Guidelines. J Law Econ. 1999;42(1):271–308.
17. Colbert JL. Inherent risk: an investigation of auditors' judgments. Account Organ Soc. 1988;13(2):111–21.
18. Nascimento FA, Jing J, Beniczky S, Benbadis SR, Gavvala JR, Yacubian EMT, et al. One EEG, one read – a manifesto towards reducing interrater variability among experts. Clin Neurophysiol. 2022;133:68–70. https://doi.org/10.1016/j.clinph.2021.10.007
19. Tatum WO, Hirsch LJ, Gelfand MA, Acton EK, LaFrance WC Jr., Duckrow RB, et al. Assessment of the predictive value of outpatient smartphone videos for diagnosis of epileptic seizures. JAMA Neurol. 2020;77(5):593–600. https://doi.org/10.1001/jamaneurol.2019.4785
20. Struck AF, Westover MB. Variability in clinical assessment of neuroimaging in temporal lobe epilepsy. Seizure. 2015;30:132–5. https://doi.org/10.1016/j.seizure.2015.06.011
21. Holden EW, Grossman E, Nguyen HT, Gunter MJ, Grebosky B, von Worley A, et al. Developing a computer algorithm to identify epilepsy cases in managed care organizations. Dis Manag. 2005;8(1):1–14. https://doi.org/10.1089/dis.2005.8.1
22. McInnis RP, Ayub MA, Jing J, Halford JJ, Mateen FJ, Brandon Westover M. Epilepsy diagnosis using a clinical decision tool and artificially intelligent electroencephalography. Epilepsy Behav. 2023;141:109135. https://doi.org/10.1016/j.yebeh.2023.109135
23. Sheldon R, Rose S, Ritchie D, Connolly SJ, Koshman ML, Lee MA, et al. Historical criteria that distinguish syncope from seizures. J Am Coll Cardiol. 2002;40(1):142–8. https://doi.org/10.1016/s0735-1097(02)01940-x
24. Wardrope A, Jamnadas-Khoda J, Broadhurst M, Grünewald RA, Heaton TJ, Howell SJ, et al. Machine learning as a diagnostic decision aid for patients with transient loss of consciousness. Neurol Clin Pract. 2020;10(2):96–105. https://doi.org/10.1212/CPJ.0000000000000726
25. Kerr WT, Janio EA, Chau AM, Braesch CT, le JM, Hori JM, et al. Objective score from initial interview identifies patients with probable dissociative seizures. Epilepsy Behav. 2020;113:107525. https://doi.org/10.1016/j.yebeh.2020.107525
26. Oto MM. The misdiagnosis of epilepsy: appraising risks and managing uncertainty. Seizure. 2017;44:143–6. https://doi.org/10.1016/j.seizure.2016.11.029
27. Josephson CB, Rahey S, Sadler RM. Neurocardiogenic syncope: frequency and consequences of its misdiagnosis as epilepsy. Can J Neurol Sci. 2007;34(2):221–4. https://doi.org/10.1017/s0317167100006089
28. LaFrance WC, Benbadis SR. Avoiding the costs of unrecognized psychological nonepileptic seizures. Neurology. 2006;66(11):1620–1. https://doi.org/10.1212/01.wnl.0000224953.94807.be
29. Juarez-Garcia A, Stokes T, Shaw B, Camosso-Stefinovic J, Baker R. The costs of epilepsy misdiagnosis in England and Wales. Seizure. 2006;15(8):598–605. https://doi.org/10.1016/j.seizure.2006.08.005
30. Arain AM, Song Y, Bangalore-Vittal N, Ali S, Jabeen S, Azar NJ. Long term video/EEG prevents unnecessary vagus nerve stimulator implantation in patients with psychogenic nonepileptic seizures. Epilepsy Behav. 2011;21(4):364–6. https://doi.org/10.1016/j.yebeh.2011.06.003

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nascimento FA, McLaren JR, Zhao W, Katyal R, Sheikh IS, Kong WY, et al. Noise in the diagnosis of epilepsy by experts. Epileptic Disord. 2026;28:457–469. https://doi.org/10.1002/epd2.70181

Test Yourself

1. Which of the following subtypes of "noise" reflects an expert's tendency to over- or under-diagnose epilepsy relative to other experts?
   A. Pattern noise
   B. Level noise
   C. Occasion noise
   D. None of the above
2. Which of the following subtypes of "noise" reflects an expert's idiosyncratic but stable responses to specific case features?
   A. Pattern noise
   B. Level noise
   C. Occasion noise
   D. None of the above
3. Noise in the diagnosis of epilepsy by experts stems primarily from which subtype(s) of noise?
   A. Pattern noise
   B. Level noise
   C. Occasion noise
   D. None of the above

Answers may be found in the supporting information