Published in final edited form as:

Clin Neurophysiol. 2022 January ; 133: 68–70. doi:10.1016/j.clinph.2021.10.007.

One EEG, one read - A manifesto towards reducing interrater variability among experts

Fábio A. Nascimento[*],
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Jin Jing,
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Sándor Beniczky,
Department of Clinical Neurophysiology, Danish Epilepsy Center, Dianalund and Aarhus University Hospital, Aarhus, Denmark

Selim R. Benbadis,
Department of Neurology, University of South Florida and Tampa General Hospital, Tampa, FL, USA

Jay R. Gavvala,
Department of Neurology, Baylor College of Medicine, Houston, TX, USA

Elza M.T. Yacubian,
Department of Neurology and Neurosurgery, Universidade Federal de Sao Paulo, Sao Paulo, Sao Paulo, Brazil

Samuel Wiebe,
Department of Clinical Neurosciences, University of Calgary, Calgary, Alberta, Canada

Stefan Rampp,
Department of Neurosurgery, University Hospital Erlangen, Germany

Michel J.A.M. van Putten,
Department of Clinical Neurophysiology, University of Twente and Medisch Spectrum Twente, Enschede, the Netherlands

Manjari Tripathi,
Department of Neurology, AII India Institute of Medical Sciences, New Delhi, India

Mark J. Cook,

[*] Corresponding author at: 55 Fruit St., Wang 7th floor, Boston, MA, 02114, USA. fnascimento@mgh.harvard.edu, nascimento.fabio.a@gmail.com (F.A. Nascimento).

Declaration of Competing Interest
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Disclosures
Dr. Westover is co-founder of Beacon Biosignals, which played no role in this work. The other authors report no disclosures relevant to this manuscript. This work was not funded.

Department of Medicine, University of Melbourne, Melbourne, Victoria, Australia

Peter W. Kaplan,
Department of Neurology, Johns Hopkins University, Baltimore, MD, USA

William O. Tatum,
Department of Neurology, Mayo Clinic, Jacksonville, FL, USA

Eugen Trinka,
Department of Neurology and Neuroscience Institute, Christian Doppler University Hospital, Center for Cognitive Neuroscience, Paracelsus Medical University, Salzburg, Austria

Andrew J. Cole,
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

M. Brandon Westover
Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Electroencephalography (EEG) plays a major role in routine clinical care for patients with seizures and epilepsy. Accurate and reliable EEG interpretation are crucial as they guide clinical management in many circumstances (Fisher et al., 2014). Misinterpretation is a significant problem when EEGs are reviewed by neurologists without specialty training (Amin and Benbadis, 2019). Nonetheless, there is another important issue which has attracted less attention by our community: the same EEG may be interpreted differently by experts (i.e., neurologists with clinical neurophysiology and/or epilepsy fellowship training).

Expert interrater reliability (IRR) is imperfect in routine EEG interpretation (Jing et al., 2020). A recent multicenter study recruited eight experts to rate 13,262 candidate interictal epileptiform discharges (IEDs), extracted from routine 1,063 EEGs from patients of all ages, as IEDs or non-IEDs. Experts' IRR was fair (chance-corrected agreement, κ: 48.7%). Expert IRR appears to be better for determining whether an EEG contains any IEDs vs. none (κ: 69.4%). Nevertheless, overall reliability is limited by the quality of judgements regarding single IEDs, and the interpretation of an EEG in clinical practice often boils down to a single IED.

Reliability of expert interpretation of IEDs depends on two types of noise: *pattern noise* and *level noise* (Kahneman et al., 2021). Pattern noise is variability between experts in judgements about which probabilities to assign to candidate IEDs. When such variations are measured relative to a gold standard, pattern noise reflects experts' skill in discriminating IEDs from normal variations, benign variants, and artifacts[1]. Level noise, by contrast, is variability over where to set the threshold above which a candidate IED is considered epileptiform. We can understand the difference between pattern noise and level noise in terms of receiver operating characteristic (ROC) curves. Relative to a gold standard, each

[1] Kahneman et al. show how pattern noise can be measured even without a gold standard. In that case, pattern noise does not necessarily reflect skill, because the true values are unknown.

expert's performance can be quantified by two numbers: true positive rate (TPR, aka sensitivity) and false positive rates (FPR) (Fig. 1). The set of all possible TPR and FPR values an expert could in principle achieve by varying their choice of threshold is the expert's ROC curve; the area under this curve reflects the individual's pattern noise (lower pattern noise, higher area). By contrast, level noise arises from disagreements over where to place the threshold dividing positive and negative decisions (i.e., raters' different operating points on the ROC curve).

Recent evidence suggests that pattern noise is similar among experts and is quite low, whereas level noise is comparatively high (Jing et al., 2020). Experts operate on (or near) a common high-area ROC curve, but with different operating points due to varying thresholds. Consequently, overall IRR is dominated by level noise. Experts with high thresholds (low FPR) have low TPR: "under-callers". Experts with low thresholds (high TPR) have high FPR: "overcallers". These differences lead to variability in EEG interpretation: one EEG, multiple reads. Such variability becomes problematic in scenarios where different interpretations of the same EEG determine whether a patient is diagnosed with epilepsy (Fisher et al., 2014).

EEG results should not vary depending on the interpreter's idiosyncratic over- or under-calling preferences. Reduction of level noise should be possible by implementing measures that help experts learn to use the same threshold. One approach consists of adopting standardized criteria to identify IEDs - such as the criteria proposed by the International Federation of Clinical Neurophysiology (IFCN) (Kural et al., 2020). An advantage of this approach is that it breaks the complex implicit judgment involved in IED identification into a series of simpler tasks, an approach that is often effective at reducing IRR in other fields (Kahneman et al., 2021). A disadvantage is that assessing the six IFCN features still requires subjective judgement and may not characterize all permutations of IEDs encountered in clinical practice. Another approach is to create a large, representative EEG database with robust external gold standard, and ask experts to annotate IEDs and classify EEGs, with instant feedback, until they achieve an acceptable error rate and converge to the same threshold. This approach has the advantage of not requiring an explicit (and possibly imperfect or incomplete) definition of IEDs. The possible disadvantage is that decisions based on implicit knowledge may be difficult to explain or defend. Therefore, a combination of these two approaches is likely best.

Defining a gold standard is also paramount. One avenue is to define it based on EEGs from patients with video-EEG data which captures their habitual spells. "True" IEDs can then be defined as sharp transients seen in patients with video-EEG-confirmed epilepsy, and non-IEDs as sharp transients seen in patients with nonepileptic events. A drawback of this approach is that most patients diagnosed with epilepsy do not undergo video-EEG monitoring. As a result, IEDs collected from video-EEG data may not encompass the full spectrum of patients with seizures nor the full range and variety of IEDs and non-IEDs in patients undergoing routine EEG as part of their evaluation for epilepsy. Thus, expert IRR measured in such a fashion may not reflect IRR among experts who interpret routine EEGs and may instead reflect interpretation of EEGs of patients with drug-resistant epilepsies.

An alternative is to define the gold standard based on expert consensus: obtain a collection of sharp transients from a large and representative group of patients who undergo routine EEGs, enlist a large and diverse group of experts to annotate them, and define IEDs as waveforms that the majority of experts considers epileptiform. This "wisdom of crowds" approach has the advantage of being unbiased by patient selection, but the disadvantage of being grounded in expert experience rather than an externally validated and objective source.

Reducing level noise is a critical part of reducing error in EEG interpretation. Noise reduction can be achieved by performing regular "noise audits" to assess the degree of variability within particular groups, and instituting interventions to increase IRR. On a larger scale, international groups responsible for accreditation should develop robust gold standards and valid pragmatic ways to measure and improve IRR. Such efforts will translate into less over- and under-diagnosis and ultimately better patient care.

Acknowledgements

We thank Dr. Daniel Kahneman and Dr. Olivier Sibony for invaluable discussions about level noise and pattern noise.

References

Amin U, Benbadis SR. The role of EEG in the erroneous diagnosis of epilepsy. J Clin Neurophysiol 2019;36(4):294–7. [PubMed: 31274692]

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. [PubMed: 24730690]

Jing J, Herlopian A, Karakis I, Ng M, Halford JJ, Lam A, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. JAMA Neurol 2020;77(1):49. 10.1001/jamaneurol.2019.3531. [PubMed: 31633742]

Kahneman D, Sibony O, Sunstein CR. Noise: a flaw in human judgement. New York: Little, Brown Spark; 2021.

Kural MA, Duez L, Sejer Hansen V, Larsson PG, Rampp S, Schulz R, et al. Criteria for defining interictal epileptiform discharges in EEG: a clinical validation study. Neurology 2020;94(20):e2139–47. [PubMed: 32321764]

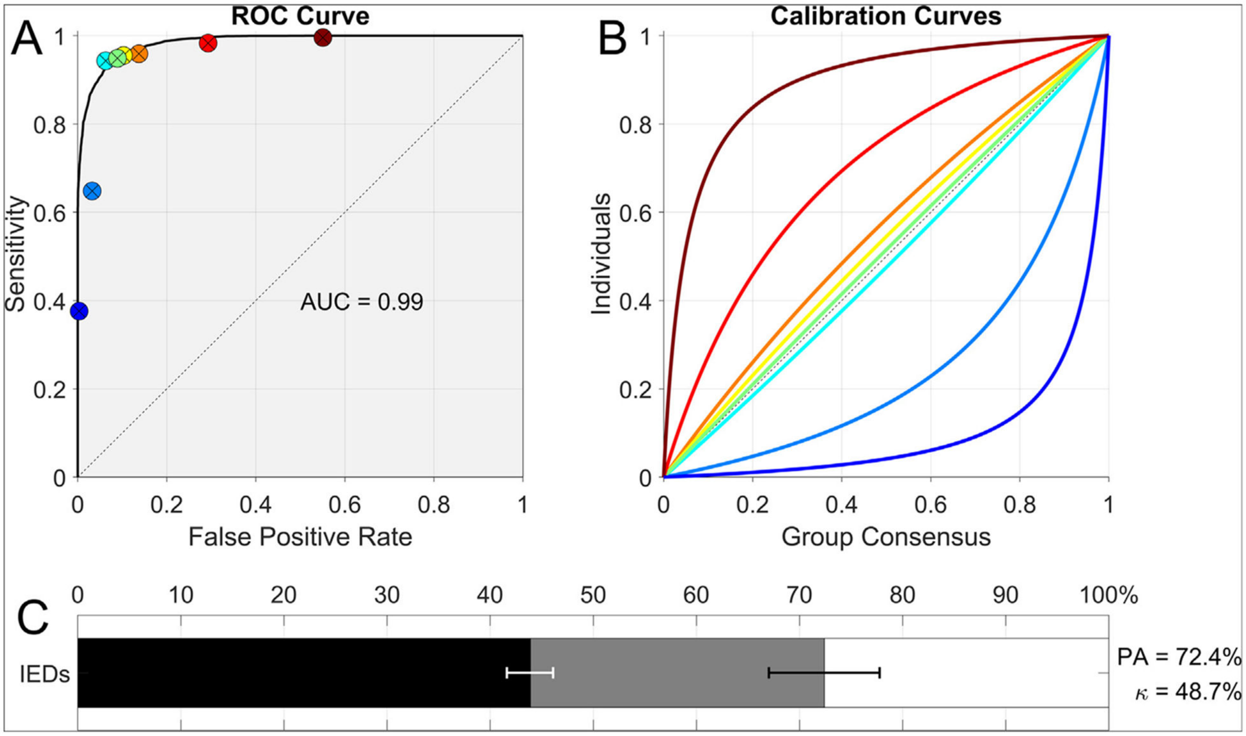

Fig. 1.

A: Receiver operating characteristic curve fit to all experts' scores, with the operating point (false-positive rate = 1 – specificity and sensitivity) of each expert indicated by a solid circle. B: Parametric calibration curve fit to the binary scores of each export, indicating the probability of that expert marking events within a given bin as IEDs. These curves allow assessment of the variation among experts relative to the group consensus. Colors are ordered from maximal under calling (blue) to maximal over calling (red). C: Inter-rater reliability (IRR): Kappa (ϰ) values in relation to percent agreement. Horizontal bars show the percent agreement (PA, black + gray bars, 95% CI in black error bar), relative to the maximal possible (100%, end of white bar). The length of the black bar shows the percent agreement by chance, PC (95% CI in white error bar). Mathematically, the chance-corrected IRR, ϰ, is the percentage of this possible beyond-chance agreement that is actually achieved, that is, ϰ = (PA − PC)/(100 − PC). Graphically, ϰ is represented as the fraction of the distance between 100% and the end of the black bar that is taken up by the gray bar.