EDITORIAL

Putting the "Big" in Big Data

Learning to Be Just as (Un)certain as a Clinician at EEG

Erik Kaestner, PhD, and William Stacey, MD, PhD

Neurology® 2023;100:799-800. doi:10.1212/WNL.0000000000207224

Correspondence
Dr. Kaestner
ekaestne@health.ucsd.edu

RELATED ARTICLES

Research Articles
Interrater Reliability of Expert Electroencephalographers Identifying Seizures and Rhythmic and Periodic Patterns in EEGs
Page 806

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

The deployment of machine learning to augment clinical judgment is an exciting avenue in the effort to improve patient care. In basic terms, machine learning can be thought of as an algorithm that learns how to distinguish patterns within data, and it can learn better with larger amounts of data. Distinguishing patterns is at the heart of many aspects of clinical care. EEG interpretation is a prime example: a trained, laborious task that might benefit from algorithms. The challenge, however, is to build an algorithm that is reliable enough to be useful and trustworthy in standard clinical care.

In this issue of *Neurology*®, 2 companion papers address multiple hurdles to deploying such automated algorithms with a new detector, Seizures, Periodic and Rhythmic pattern Continuum (SPaRCNet).[1,2] SPaRCNet is a deep convolutional neural network, a deep learning tool that was originally developed to characterize images. In this case, it learns how to discriminate between different EEG patterns by training on markings made by expert clinicians, that is, it learns to do with the neural network what the clinicians do by eye. However, there are 2 main challenges: (1) The experts do not always agree and (2) EEG algorithms to detect spikes or seizures have been around for decades, but epilepsy physicians are notoriously suspicious of them. What makes SPaRCNet unique is that it goes beyond standard seizure detection and seeks to identify 4 additional patterns within the "ictal-interictal-injury continuum" (IIIC).[3] It is difficult for clinicians to learn to distinguish these patterns, so how can an algorithm do it? The authors took a sequential approach. First, they collected a massive data set, recruited many epilepsy specialists, and learned how human reviewers score the records.[1] Next, they used those expert classifications as a training set to teach SPaRCNet how to distinguish the different patterns, using the majority vote of the experts as the benchmark.[2] The massive scale in these 2 works provides a level of rigor and interpretation that has never been possible before.

For the first paper, Jing et al.[1] reviewed many studies to collect a large data set of EEG epileptiform events for experts to classify. The study gathered more than 50,000 EEG segments from 2,711 patients, a much larger sample than previous efforts. Gathering this massive collection of EEGs required a combination of targeted and random sampling of events[4] that were marked by 30 experts in clinical neurophysiology from 18 institutions. This data set is perfectly designed to train a machine learning classifier. The authors first chose to examine human categorization patterns and found that agreement was often not strong among human reviewers (average pairwise agreement was 52%). To understand these disagreements, they used a latent trait model to show that the discrepancies were due to differences in where reviewers drew boundaries between categories, rather than lack of expertise. Given that these patterns are truly a continuum without clear distinctions, this insight into why reviewers disagree is a fascinating result on its own.

Armed with a labeled data set, the next step was to build and test a classifier. The comparison between human reviewers provided a metric for success—the algorithm simply had to be as good (or uncertain) as a typical expert. In the study described in the second paper, Jing et al.[2] trained and benchmarked SPaRCNet. The authors found that in an apples-to-apples comparison, SPaRCnet indeed performed as well as the human reviewers at differentiating IIIC patterns and had better prediction accuracy (known as calibration). Despite these results, SPaRCNet is a "black box"

From the Department of Psychiatry (E.K.), University of California San Diego; Department of Neurology (W.S.), Michigan Medicine, and Department of Biomedical Engineering (W.S.), University of Michigan, Ann Arbor.

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 editorial.

algorithm, with 2 potential problems to overcome. First, classification algorithms can be overtrained. In essence, if you give a sophisticated algorithm enough information, it can give you any answer you want, but only within the original data set. A key strength of their approach was that the algorithm was trained on one group of patients and then tested on different patients who were withheld from all training. This method doubles the number of patients needed, but also removes the concern of overtraining. A second concern is that black box algorithms provide no insight into their decisions, making it difficult to handle clinical "gray areas." To address this concern, Jing et al. provided a 2-dimensional projection of the IIIC classes (Figure 3).[2] The map looks like a starfish: a central hub ("other" patterns that include normal activity) that branches out to the 5 IIIC patterns. They show with examples how moving farther out on each arm makes the pattern more clearly identified, but the patterns meld as one moves in closer to the center; it is easy to see how reviewers could disagree about ambiguous patterns. This remarkably intuitive visualization is an excellent way to understand the "continuum" and how it leads to uncertainty.

Across both studies, category ambiguities are the biggest weakness—for both humans and the algorithm. Clinical reality has many gray areas, which the authors embraced. Rather than seeing this as a failing of the algorithm, they instead illustrated the continuum with its uncertain boundaries, with results that suggest that perhaps it is better to refer to probabilities and distances, rather than strict thresholds. This concept applies to other clinical decisions as well. For instance, for epilepsy surgery, there is the long-standing concept of the "seizure onset zone" being a gold standard, but it can vary drastically among different EEG experts.[5] Should surgical decisions depend on which clinician is reading on a given day? Perhaps our diagnoses and treatments should account for this uncertainty as well, rather than continuing in the dogmatic assignments of specific categories.

Overall, the main strength of the studies by Jing et al. lies in the power of "a large N." By including thousands of events and using data from hundreds of additional, withheld patients for testing, Jing et al. were able to provide robust and realistic results. Fields using algorithms to comb through large data sets, such as seizure detection algorithms (110 patients, 2,805 hours),[6] high-frequency oscillation detectors (121 patients, 3,000 hours),[7] as well as imaging to detect language impairment (82 diffusion-weighted images)[8] and disease (297 T1-weighted images)[9] must now look to this enviably raised standard. As seen in these 2 papers, having thousands of examples can lead to remarkable breakthroughs with modern Big Data tools.

A near-term application of this work could be deployment alongside clinicians using automated clustering and feature visualization. Jing et al.[2] explored this use-case finding that clinicians using their tools could review the model results in just a few minutes for a long EEG record. While full deployment will take more fine-tuning and has several challenges, the validation of this tool is already complete, far exceeding what is typically required for approval by governmental agencies. Although this tool does not replace full clinical review and more work is needed, in our opinion this tool would be a boon to have when monitoring critically ill patients with EEG.

Study Funding

K01NS12483 (E.K.) and R01NS094399 (W.S.).

Disclosure

E. Kaestner reports no disclosures. W. Stacey has a licensing agreement with Natus Medical, Inc. and consulting agreement with Neuronostics, Inc. Natus and Neuronostics had no involvement in the study. Go to Neurology.org/N for full disclosures.

Publication History

Received by *Neurology* January 24, 2023. Accepted in final form February 9, 2023.

References

1. Jing J, Ge W, Struck AF, et al. Interrater reliability of expert electroencephalographers identifying seizures and rhythmic and periodic patterns in EEGs. *Neurology*. 2023; 100(17):e1737-e1749.
2. Jing J, Ge W, Hong S, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*. 2023;100(17): e1750-e1762.
3. Hirsch LJ, Fong MWK, Leitinger M, et al. American Clinical Neurophysiology Society's standardized critical care EEG terminology: 2021 version. *J Clin Neurophysiol*. 2021;38(1):1-29. doi:10.1097/WNP.0000000000000806
4. Ge W, Jing J, An S, et al. Deep active learning for interictal ictal injury continuum EEG patterns. *J Neurosci Methods*. 2021;351:108966. doi:10.1016/j.jneumeth.2020.108966
5. Davis KA, Devries SP, Krieger A, et al. The effect of increased intracranial EEG sampling rates in clinical practice. *Clin Neurophysiol*. 2018;129(2):360-367. doi: 10.1016/j.clinph.2017.10.039
6. 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*. 2021;38(5):439. doi:10.1097/WNP.0000000000000709
7. Gliske SV, Irwin ZT, Chestek C, et al. Variability in the location of high frequency oscillations during prolonged intracranial EEG recordings. *Nat Commun*. 2018;9(1): 2155. doi:10.1038/s41467-018-04549-2
8. Kaestner E, Balachandra AR, Bahrami N, et al. The white matter connectome as an individualized biomarker of language impairment in temporal lobe epilepsy. *Neuroimage Clin*. 2020;25:102125. doi:10.1016/j.nicl.2019.102125
9. Gleichgerrcht E, Munsell B, Keller SS, et al. Radiological identification of temporal lobe epilepsy using artificial intelligence: a feasibility study. *Brain Commun*. 2022;4(2): fcab284. doi:10.1093/braincomms/fcab284

Neurology®

Putting the ''Big'' in Big Data: Learning to Be Just as (Un)certain as a Clinician at EEG

Erik Kaestner and William Stacey

Neurology 2023;100;799-800 Published Online before print March 6, 2023
DOI 10.1212/WNL.0000000000207224

This information is current as of March 6, 2023

Updated Information & Services	including high resolution figures, can be found at: http://n.neurology.org/content/100/17/799.full
References	This article cites 9 articles, 2 of which you can access for free at: http://n.neurology.org/content/100/17/799.full#ref-list-1
Subspecialty Collections	This article, along with others on similar topics, appears in the following collection(s): **EEG** http://n.neurology.org/cgi/collection/eeg_ **EEG; see Epilepsy/Seizures** http://n.neurology.org/cgi/collection/eeg_see_epilepsy-seizures **Epilepsy monitoring** http://n.neurology.org/cgi/collection/epilepsy_monitoring_
Permissions & Licensing	Information about reproducing this article in parts (figures,tables) or in its entirety can be found online at: http://www.neurology.org/about/about_the_journal#permissions
Reprints	Information about ordering reprints can be found online: http://n.neurology.org/subscribers/advertise

AMERICAN ACADEMY OF
NEUROLOGY®