HHS Public Access

Author manuscript

Clin Neurophysiol. Author manuscript; available in PMC 2022 August 08.

Published in final edited form as:

Clin Neurophysiol. 2017 July ; 128(7): 1406–1407. doi:10.1016/j.clinph.2017.02.026.

What it should mean for an algorithm to pass a statistical Turing test for detection of epileptiform discharges

M. Brandon Westover[*],

Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

Jonathan J. Halford,

Department of Neurology, Medical University of South Carolina, Charleston, SC, USA

Matt T. Bianchi

Department of Neurology, Massachusetts General Hospital, Boston, MA, USA

Scheuer et al. from Persyst recently reported that their new proprietary automated spike detection algorithm, part of the Persyst 13 software package (P13) (Scheuer et al., 2017), detects epileptiform transients (ETs; spikes and sharp waves) as well as human experts. Authors of an accompanying commentary enthusiastically agreed (Webber and Lesser, 2017). We have informally tested P13 and are impressed. Nevertheless, we do not yet agree that P13 has passed a valid Turing test.

There is no explicit quantitative definition of ETs. Instead ETs may be defined qualitatively as "…distinctive waves or complexes…resembling those recorded in a proportion of human subjects suffering from epileptic disorders…" (Noachtar et al., 1999). Experts learn to detect ETs by practice, as apprentices to experienced clinical neurophysiologists. Consequently, the appropriate standard for judging automated ET detection is inter-rater agreement (IRA), aptly termed by Scheuer et al. a "statistical Turing test". An ET detection algorithm passes this test if expert-computer IRA is statistically indistinguishable from expert-expert IRA.

If P13 really does have sensitivity (i.e. will detect ETs when present) and specificity (i.e. will not report ETs when none are present) equivalent to well-trained human experts, then this represents a true "disruptive innovation", one that would warrant largely replacing expert review of diagnostic EEG studies by a computer algorithm (Table 1). However, in our opinion the statistical Turing test for ET detection remains an unmet challenge. The evidence provided thus far fails to satisfactorily address the following six key issues.

1. *Number of EEGs*. In Scheuer et al., EEGs from 40 patients were reviewed, containing 296–363 h of EEG and 2340–3229 ETs. (A range is given because the reviewers annotated different amounts of the same EEGs.) Although larger than many previous studies (Halford, 2009) we conservatively estimate the number of EEGs needed to train an expert EEGer as 500–1000, or 1–3 EEGs/day over

[*]Corresponding author at: MGH/Harvard Medical School, Department of Neurology, 290 Orchard Street, Massachusetts, Belmont, MA 02478, USA. mwestover@mgh.harvard.edu, mwestover@partners.org.

Conflict of interest

None.

1–2 years of fellowship training. It is unknown how many ET morphological "phenotypes" exist, but 40 cases probably does not provide nearly enough variety. We propose that a testing dataset should contain at least 1000 EEGs.

2. *Selection of EEGs*. The EEGs in Scheuer et al. were from patients of unreported age admitted to an epilepsy monitoring unit (EMU), selected to have ETs at a minimum of one per hour. This approach is guaranteed to produce an EEG dataset with a large number of ETs. But is this dataset representative of all ETs in diagnostic EEG recordings? This is far from clear. To include subjects of different ages, with different types of epilepsy, with differing antiepileptic medication responsiveness, and with different ET rates, the testing dataset would need to be much larger and drawn from inpatient and outpatient settings and pediatric and adult patient populations.

3. *Number of experts*. IRA varies widely between experts. Using a small number (n = 3 in the P13 study) risks making the statistical Turing test artificially "easy" if the raters selected happen to have low IRA. Using three experts is common since it meets minimal FDA requirements for product approval. However, larger numbers allow more precise determination of human expert IRA. Results from a recent study we performed (Halford et al., 2016), suggest the optimal number of expert ET raters is probably in the 5–10 range. Additionally, we have recently found (Halford et al., 2016) substantial differences in ET ratings between neurologists with and without EEG fellowship training.

4. *Qualifications of experts*. In the P13 study, the experts were three EEG technicians. A more convincing test would compare the algorithm to physicians who perform the final interpretation of clinical reports. Moreover, IRA for ET detection varies and is higher among experts who have undergone formal training in clinical neurophysiology (Halford et al., 2016). Ideally, experts should be pre-tested to determine their ET detection reliability, relative to peers, before participating in an ET validation study.

5. *Potential for bias*. The P13 study authors selected the EEGs, hired the technicians used to define expert-expert IRA, and performed the analysis. The EEGs and the ETs data are not publically available; thus replication is not possible. A more convincing approach would utilize or create a publically available benchmark database of EEGs annotated by a panel of independent clinical neurophysiologists.

6. *Algorithm transparency*. The ET detection algorithm in P13 is proprietary. When a diagnostic algorithm remains a trade secret, it is all the more critical that it be tested rigorously before users accept claims of equivalence with an existing gold standard.

In conclusion, we are not ready to let P13 take over the role of human experts in detecting ETs. The bar needs to be set higher. To pass a statistical Turing test for ET detection, we recommend that an algorithm meet the following criteria: (1) It must be evaluated on a publically available database consisting of EEG recordings from at least 1000 different

patients. These EEGs should be randomly selected and should come from more than one EEG lab, to reduce selection bias. These cases may not be used in developing or tuning the parameters of the algorithm being tested, to avoid overfitting. (2) The testing EEGs must be scored by at least seven physician experts each, all of whom have undergone formal training in clinical neurophysiology and who have been pre-tested to estimate their IRA for ET detection versus peers, to allow valid determination of expert-expert IRA; (3) the algorithm must achieve algorithm-expert IRA at least as high as expert-expert IRA. If an algorithm passes this test, then involving human experts in detecting ETs will no longer be necessary.

References

Halford JJ. Computerized epileptiform transient detection in the scalp electroencephalogram: obstacles to progress and the example of computerized ECG interpretation. Clin Neurophysiol 2009;120:1909–15. 10.1016/j.clinph.2009.08.007. [PubMed: 19836303]

Halford JJ, Arain A, Kalamangalam GP, LaRoche SM, Bonilha L, Basha M, et al. Characteristics of EEG interpreters associated with higher inter-rater agreement. J Clin Neurophysiol 2016. 10.1097/WNP.0000000000000344.

Noachtar S, Binnie C, Ebersole J, Mauguière F, Sakamoto A, Westmoreland B. A glossary of terms most commonly used by clinical electroencephalographers and proposal for the report form for the EEG findings. The International Federation of Clinical Neurophysiology. Electroencephalogr Clin Neurophysiol Suppl 1999;52:21–41. [PubMed: 10590974]

Scheuer ML, Bagic A, Wilson SB. Spike detection: inter-reader agreement and a statistical Turing test on a large data set. Clin Neurophysiol 2017;128:243–50. 10.1016/j.clinph.2016.11.005. [PubMed: 27913148]

Webber WRS, Lesser RP. Automated spike detection in EEG. Clin Neurophysiol 2017;128:241–2. 10.1016/j.clinph.2016.11.018. [PubMed: 27940048]

Table 1

Possible responses to a claim that an automated ET detection algorithm performs as well as human experts, based on adequacy of sensitivity (Se) and false positive rate (Fp).

1	*Unconvinced of either Se or Fp:* Continue to review the entire EEG, possibly using the algorithm's recommendations to enhance sensitivity. This approach would be warranted if the EEGer was not convinced that Se or Fp have been demonstrated to be sufficiently good to justify skipping any part of full EEG review, but was nevertheless convinced that the algorithm added some value.
2	*Convinced of Se but not of Fp:* Visually confirm all ETs detected by the algorithm, but if no ETs are detected then assume that none are present and forego manual review. This practice change would be rational if the EEGer was convinced that if ETs are present then they almost certainly will be detected (high Se), but that some detections might be false (unacceptably high Fp).
3	*Convinced of Fp but not of Se*: Do not review the EEG if the algorithm detected any ETs, but if no ETs are detected then go on to review the entire EEG to confirm that none have been missed. This practice change would be rational if the EEGer was convinced that any ETs detected are almost certainly correct (low Fp), but that the algorithm might miss some ETs (unacceptably low Se).
4	*Convinced of both Se and Fp:* Do not review the EEG. Instead, treat the algorithm's detections as equivalent to those of a human expert.