Seizure Detection: Interreader Agreement and Detection Algorithm Assessments Using a Large Dataset

Mark L. Scheuer,* Scott B. Wilson,* Arun Antony,† Gena Ghearing,‡ Alexandra Urban,† and Anto I. Bagić†

*Persyst Development Corporation, Solana Beach, California, U.S.A.; †University of Pittsburgh Comprehensive Epilepsy Center (UPCEC), University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, U.S.A.; and ‡Department of Neurology, University of Iowa, Iowa City, Iowa, U.S.A.

Purpose: To compare the seizure detection performance of three expert humans and two computer algorithms in a large set of epilepsy monitoring unit EEG recordings.

Methods: One hundred twenty prolonged EEGs, 100 containing clinically reported EEG-evident seizures, were evaluated. Seizures were marked by the experts and algorithms. Pairwise sensitivity and false-positive rates were calculated for each human–human and algorithm–human pair. Differences in human pairwise performance were calculated and compared with the range of algorithm versus human performance differences as a type of statistical modified Turing test.

Results: A total of 411 individual seizure events were marked by the experts in 2,805 hours of EEG. Mean, pairwise human sensitivities and false-positive rates were 84.9%, 73.7%, and 72.5%, and 1.0, 0.4, and 1.0/day, respectively. Only the Persyst 14 algorithm was comparable with humans—78.2% and 1.0/day. Evaluation of pairwise differences in sensitivity and false-positive rate demonstrated that Persyst 14 met statistical noninferiority criteria compared with the expert humans.

Conclusions: Evaluating typical prolonged EEG recordings, human experts had a modest level of agreement in seizure marking and low false-positive rates. The Persyst 14 algorithm was statistically noninferior to the humans. For the first time, a seizure detection algorithm and human experts performed similarly.

Key Words: EEG, Automated seizure detection, Artificial neural network, Deep learning, Noninferiority, Interreader agreement.

(J Clin Neurophysiol 2021;38: 439–447)

Achieving expert-level performance is now a primary goal in the development of EEG-based automated seizure detection algorithms. Human expert readings remain the gold standard for recognition of seizures in the EEG; no viable ground truth exists. But how well do experts agree in practice, and what metrics can be used to meaningfully compare algorithm performance to expert humans? This report details new work in this area, first reviewing existing literature concerning expert interreader agreement in seizure marking, then reporting new data from a large study of interreader agreement in epilepsy monitoring unit (EMU) recordings, and finally, quantitatively comparing expert performance to both a marketed and a new generation automated seizure detector.

The identification of seizures in an EEG is diagnostic of a seizure disorder, and their recognition is crucial when attempting to differentiate epileptic seizures from similar appearing nonepileptic clinical events (e.g., psychogenic nonepileptic episodes, convulsive syncope). An EEG-evident seizure's spatial distribution and morphologic characteristics help localize seizure foci and assist in determining a syndromic diagnosis.[1,2] Despite their importance, electrographic seizures are imprecisely defined. Confident seizure identification in an EEG usually requires the skills of a trained electroencephalographer. A seizure is defined as: "Phenomenon consisting of repetitive EEG discharges with relatively abrupt onset and termination and characteristic pattern of evolution lasting at least several seconds. These EEG patterns are seen during epileptic seizures. …The component waves or complexes vary in form, frequency, and topography. They are generally rhythmic and frequently display increasing amplitude and decreasing frequency during the same episode. When focal in onset, they tend to spread subsequently to other areas."[3] Another definition used in critical care research specifies that unequivocal seizures on EEG include the following: generalized spike–wave discharges at 3/s or faster; and clearly evolving discharges of any type that reach a frequency >4/s, whether focal or generalized.[4] Highly variable seizure characteristics, differences in background EEG, copious noncerebral artifact, reader fatigue, and variations in electroencephalographer training and experience can sometimes result in substantial differences between seizure events marked by different readers assessing the same EEG recording.[5–7] In clinical practice, the identification of electrographic seizure patterns boils down to, as Justice Potter remarked concerning obscenity, "I know it when I see it."[8] These factors lead to imperfect agreement between readers interpreting the same EEG segments. One recent study assessed agreement between multiple trained EEG readers evaluating 300 thirty-minute recordings and found that interreader agreement was only moderate even though a simple three-class rating (normal, ictal, and non-ictal abnormal) was used; despite this middling agreement, readers were highly confident in their interpretations, implying overconfidence.[9]

M. L. Scheuer and S. B. Wilson are full-time employees of Persyst Development Corporation. The other authors have no conflicts of interest to disclose.

The sponsor was not involved in the primary data collection or data marking.
This study was partially funded by Persyst Development Corporation.

Address correspondence and reprint requests to Mark L. Scheuer, MD, Persyst Development Corporation, 420 Stevens Ave, Suite 210, Solana Beach, CA 92075, U.S.A.; e-mail: scheuerml@persyst.com.

ISSN: 0736-0258/20/3805-0439
DOI 10.1097/WNP.0000000000000709

Little research has assessed the real-world level of expert agreement in identifying seizures on the EEG. Wilson et al.[5] assessed pairwise agreement between four readers for seizure identification in ten 8-hour EMU recordings selected to represent differing seizure types. They reported an average any-overlap sensitivity of 92% and a false-positive rate of 2.8 per day. Ronner et al.[10] evaluated 90 highly selected 10-second EEG clips from 23 comatose intensive care unit (ICU) patients reported to either have seizures or have no seizures. They found a moderate level of agreement (kappa of 0.5) for an assessment of seizure present or absent in these brief case examples. Benbadis et al.[11] evaluated event clips and selected interictal examples from 22 consecutive patients with recorded paroxysmal clinical events. Twenty-two neurologist/epileptologists reviewed these samples and returned a diagnosis of psychogenic, epileptic, or other nonepileptic but not psychogenic cause for each EEG. They reported substantial interreader agreement (kappa of 0.69) for the diagnosis of an epileptic cause of the EEG findings. Kelly et al.[12] reported a large study of 55 seizure-enriched (i.e., reported nonseizure background content was lessened by a factor of three to four) prolonged EMU records. Pairwise assessment of three readers evaluating 1,208 hours of EEG for seizures revealed substantial interrater agreement (kappa of 0.68). Their method of decreasing reader burden by enriching the dataset for seizures may have resulted in the loss of some less demonstrative ictal events, potentially enhancing agreement. Abend et al.[13] reported interrater agreement for the presence or absence of seizures in two 30-minute records from each of 37 pediatric patients undergoing EEG monitoring after cardiac arrest. They reported moderate interreader agreement (kappa of 0.4) for the binary classification of whether a record did or did not contain seizures. Gaspard et al.[14] reported on interrater agreement among 49 readers assessing brief (10–60 seconds) EEG clips. Only five seizure-containing clips were present in the dataset, but readers had nearly perfect agreement in identifying these. The criteria and methods used for selecting the seizure samples were not discussed. Any selection bias favoring inclusion of clearly demonstrable seizure patterns likely would have affected agreement. Halford et al.[6] reported moderate interreader agreement (kappa of 0.58) among eight EEG experts for seizure marking in 30 selected 1-hour seizure-dense ICU recording clips. Stevenson et al.,[15] in a large study of 4,066 hours of neonatal EEG recorded from 70 patients, approximately half of whom were originally reported to have electrographic seizures, assessed agreement between three experts who marked the entire dataset for seizures. The records were generally seizure-dense, with 2,555 marked seizures. They reported a mean interreader sensitivity (seizure agreement rate; any overlap method) of 77.9%, with a mean false-positive rate of 3.4 per day. In another large study assessing five expert readers' marking of 50 prolonged ICU EEG recordings (1,776 hours), Tu et al.[7] reported average pairwise sensitivity of 70.2% and a false-positive rate of about 2 per day. In subgroup analyses of "unequivocal seizures" (per the American Clinical Neurophysiology Society's research definitions), readers averaged 75.5% sensitivity and 0.8 false positives per day. For the subgroup of "equivocal" seizures (those which the reader believed were seizures but that did not meet the stricter definition's criteria), a much lower pairwise sensitivity of 34.6% and higher false-positive rate of 1.6 per day were documented.

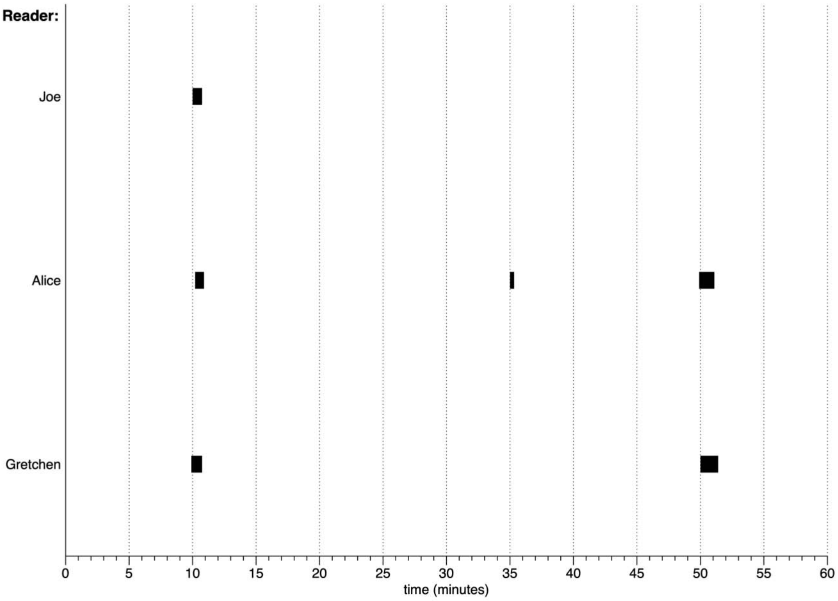
As is clear from the references cited above, expert readers clearly miss some seizures during standard waveform review. Many electroencephalographers consider marketed seizure detection algorithms to be useful adjunctive tools, but that the computerized detectors function at a level significantly below that of a capable reader.[16–18] An algorithm with expert-like seizure recognition capabilities would be useful in alerting care staff to the presence of seizures (which would assist in ictal testing paradigms), in more timely and complete recognition of seizures during inpatient and outpatient monitoring, and in helping to improve patient safety during and after seizures. It would allow better tasking and utilization of scarce and expensive expert human resources, and enable wider distribution of a higher level of care.

How do we know when an algorithm performs as well as humans? Previous studies of detection algorithms have been hampered by a variety of methodological shortcoming that prevent thorough comparison with human expert performance. Among these limitations are insufficiently populated datasets (both in number of patients and days of EEG assessed),[16] preselection of data using incompletely specified and potentially biased methods,[12] partial test dataset contamination by cases used in algorithm training,[17] assessments of records by single or multiple non-independent readers,[19] use of consensus readings (which favor easier to recognize, more demonstrative seizure patterns),[12] and circular use of algorithm seizure detection outputs to assist in initial identification of seizures so that algorithm performance can be assessed.[20] Ideally, algorithms should be evaluated on large datasets that have undergone little data selection (random or consecutive cases; no data reduction), represent the recording practices of multiple centers, and closely mirror cases encountered in clinical practice. Those data should be independently marked for seizures by at least several expert readers to adequately represent expected human performance variability.

The current study was designed to evaluate the performance of experts relative to each other and to assess performance of commercially marketed (Persyst 13, or P13) and new generation (Persyst 14 or P14) automated seizure detectors in comparison with multiple expert readers. Several goals directed development of the new P14 seizure detection algorithm. First, that it achieve near expert-level accuracy in seizure detection. Second, that it do so at much lower latency than the preceding P13 algorithm. Third, that it use current computer technology to accomplish seizure detection in real-time. To do this, many new features and methodologies were incorporated into the P14 detector. These include new artifact reduction technologies, by-channel processing to aid in the identification of low amplitude, focal seizure patterns, use of empirical null methodology to better capture the non-normal statistics of EEG patterns, improved and expanded training sets, incorporation of P14 spike detector data as an input, use of deep learning methods, and complete redevelopment of the algorithm architecture to minimize detection latency.

METHODS

This study used largely the same methodology as that reported by Scheuer et al.[21] concerning assessment of interreader



Pairwise analysis of Sensitivity and False Positive rate per page for the above markings:

Reference reader (temporary gold standard)	Test Reader	# Reference reader marks	# Test reader marks	# Reference reader seizure marks matched by Test reader	# Test reader marks not matching Reference reader marks (False Positives)	Sensitivity (Test Reader vs. Reference)	False Positive rate (Test Reader vs. Reference)
Joe	Gretchen	1	2	1	1	100% (1 of 1)	1 per hour
Joe	Alice	1	3	1	2	100% (1 of 1)	2 per hour
Alice	Gretchen	3	2	2	0	66% (2 of 3)	0 per hour
Alice	Joe	3	1	1	0	33% (1 of 3)	0 per hour
Gretchen	Alice	2	3	2	1	100% (2 of 2)	1 per hour
Gretchen	Joe	2	1	1	0	50% (1 of 2)	0 per hour

FIG. 1. Hypothetical seizure marking of a 1-hour segment of EEG by three expert readers with differing marking styles. The table shows the information leading to the six calculated pairwise sensitivity (using any overlap criterion) and false-positive rate outputs for this particular 1-hour segment. In practice, these numbers are generated for the entire EEG rather than a single hour.

agreement and algorithm performance in spike marking, but rather evaluated seizure marking. Pertinent differences or additions to that methodology are delineated here. In addition to the 100 consecutive seizure-containing records reported in the spike-marking study, 30 recordings reportedly without electrographic seizure activity (not necessarily normal) from 30 different individuals were also identified. All recordings had originally been visually analyzed for seizure activity by technologists, fellows, and attending clinical neurophysiologists; neither automated seizure detection algorithms nor quantitative EEG trending was used. Twenty-four hours of continuous EEG were retrieved for each recording, if available. Ten records from the seizure-free set were randomly withdrawn for other development purposes. Demographic data concerning subject age and gender were retained. Comments were removed from the recordings, and no video data were included. Thus, 120 de-identified recordings were presented to experts for seizure marking: 100 previously reported to contain seizures and 20 reported as seizure-free (to act as foils).

Three university faculty electroencephalographers (A.A., G.G., A.U.), each fellowship-trained (at different institutions) in clinical neurophysiology and board-certified in neurology with added qualifications in clinical neurophysiology by the American Board of Psychiatry and Neurology, each with three or more years of post-fellowship practice in epilepsy monitoring and EEG, and each proficient in the analysis of continuous EEG recordings for seizures, marked the entire test EEG dataset for seizures. Marking was performed independently and without consensus discussions. None of the marking readers participated in seizure detection algorithm development.

The experts carefully assessed the entirety of each recording using P13 EEG waveform review capabilities and without time constraint, and marked the earliest evident onset and stop point of every seizure they identified, per the standard definition and their best judgment. Readers were free to mark seizures of any duration, even brief events, if in their judgment the pattern represented an electrographic seizure. If a reader judged an EEG segment to contain excessive artifact that rendered a segment uninterpretable, then they could indicate such with a standard comment spanning the duration of the uninterpretable segment.

The P13 and P14 (new) seizure detection algorithms were compared with the human readers. None of the test data evaluated in this study were used during training of the algorithms.

The P14 algorithm was trained on records from 764 seizure-affected patients, collected from approximately a dozen institutions and marked by a half-dozen readers. In addition, hundreds of other records were used for training spike detection, artifact reduction, electrode artifact detection routines, etc. In total, the P14 algorithm was built on thousands of records. The P14 detector uses data from 10-20 system EEG recording electrodes in an 18-channel single distance anterior–posterior full bipolar montage plus channel F7–F8. Various software sensors and concepts are processed using advanced neural network technologies. Examples include assessments of power, frequency, bandwidth, and asymmetry by channel; segmentation and evolution of rhythmic activity in four frequency bands (1–4, 4–9, 9–16, and 16–25 Hz); probability that an electrode is generating signal of artifactual origin; vertical and lateral eye movement signals; chewing artifact probability; movement artifacts; myogenic artifacts; sleep stages; signal change points via use of empirical null statistics; and various seizure-related concepts including seizure, seizure onset, post-ictal changes, and identification of seizure candidates including seizure onset and cessation points. Thus, many features and concepts are evaluated for each channel. For example, the convolutional seizure neural network for channel C3-P3 has 118 inputs. A nested hierarchy of feed forward neural networks ultimately outputs seizure detections, each described by its onset, offset, and probability. The algorithm's analyses proceed in one-second increments, using information before the current one-second epoch to estimate the probability of seizure activity; seizure probability outputs can be updated for several minutes beyond a particular segment as more EEG data become available. (The algorithm does not "look ahead.")

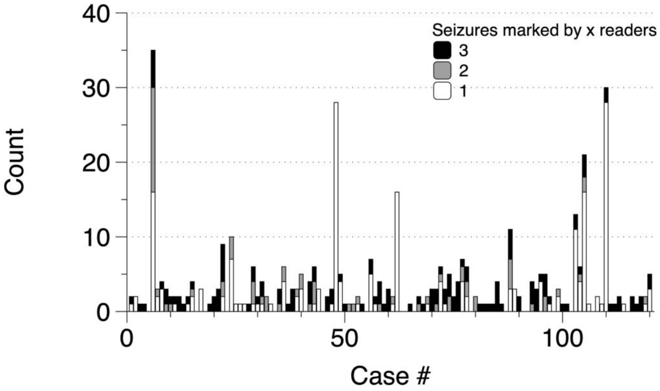


FIG. 2. Count of seizures by record, including number marked by one, two, or three readers.

Pairwise computations of sensitivity and false-positive rate were generated for each record, in a fashion analogous to that reported earlier for epileptic spike assessments by Scheuer et al.[21] Grand averages (across all EEGs) for these metrics were then calculated to avoid overweighting of records with many seizures or false positives. Seizure marks from a reader pair were designated a match (true positive) if they overlapped; non-overlapping marks constituted false positives for the test reader or algorithm (Fig. 1). The pairwise sensitivity and false-positive rates for the P13 and P14 algorithms were only determined with respect to the experts (i.e., P13 and P14 were compared with readers A, B, and C).

The P13 algorithm has no user-adjustable settings, whereas the P14 algorithm allows a user to adjust the threshold settings for seizure duration from 2 to 14 seconds and for seizure probability from 0.1 to 0.9. Receiver operating characteristic (ROC) plots for P14 were constructed by averaging its pairwise comparisons with the three experts, using duration thresholds of 2, 8, and 14 seconds and incrementing probability thresholds from 0.1 to 0.9 in increments of 0.1. For the final pairwise analyses, P14 duration and probability thresholds (8 seconds and 0.8, respectively) that most closely approximated the experts' pairwise marking styles were used for sensitivity and false-positive rate calculations.

Statistica (version 12; Dell Software), R v3.0.1 with "boot" library, and DataGraph (version 4.4; Visual Data Tools, Inc.) were used for analyses and graph production.

To evaluate the P13 and P14 seizure detection algorithms for possible noninferiority to expert reader performance (refer to Welleck[22] regarding noninferiority), the pairwise difference methodology specified by Scheuer et al.[21] was used. The P14 assessments were made using marking results generated at algorithm duration and probability threshold settings of 8 seconds and 0.8.

Persyst 14's detection performance for seizures marked by more than one expert was evaluated via creation of consensus markings of the experts. Three ROC plots for P14 reflecting seizures marked by at least one, at least two, or three experts were graphed, using a detection threshold of 8 seconds. Here, P14 detections not overlapping with expert consensus seizures were deemed false positives, even if a P14 event was also marked by a subconsensus number of experts.

To further explore the relative levels of performance of human readers and the P14 algorithm, their sensitivity and false-positive rates were assessed in comparison with consensus sets generated using the three combinations of two experts. This evaluation allowed further direct equivalent comparison of human and algorithm performance with respect to a subset of seizures for which a higher level of expert agreement was evident. Duration and probability thresholds of 8 seconds and 0.8, respectively, were used for the P14 algorithm assessments.

To assess the notification latency of the seizure detection algorithms, the time was measured between expert-marked seizure onset (determined during post hoc seizure marking) and the time at which enough seizure was identified by the algorithm to generate a detection notification. These measurements were

TABLE 1. Sensitivity and FPs for Readers and Two Seizure Detection Algorithms (Mean, SD, and Range)

Reader (Threshold)	Sensitivity % ± SD (Range)	FP/Day ± SD (Range)
A	84.9 ± 30.8 (0.0–100.0)	0.978 ± 2.914 (0.00–28.06)
B	73.7 ± 37.8 (0.0–100.0)	0.375 ± 1.800 (0.00–20.49)
C	72.5 ± 38.0 (0.0–100.0)	1.038 ± 3.417 (0.00–28.08)
Persyst 13	82.5 ± 33.2 (0.0–100.0)	11.32 ± 11.65 (0.00–61.10)
Persyst 14 (8 s, 0.8)	78.2 ± 33.9 (0.0–100.0)	0.974 ± 1.812 (0.00–11.02)

FP, false-positive rate.

conducted on the subset of seizures in which two or more experts marked a segment as a seizure and each seizure was marked by both the P14 and the P13 algorithms. The time of expert seizure onset was taken as the point at which two of three experts had marked the event. P14 duration and probability thresholds of 8 seconds and 0.8, respectively, were used. Median detection latency values were calculated for the P13 and P14 algorithms, and box plots of the results were graphed. Latency differences between P14 and P13 were statistically assessed using a two-tailed paired (by seizure) *t*-test assessment.

RESULTS

Three experts marked 120 long-term EEGs; 52% of the recordings were from women. Mean patient age was 39.8 years (range, 19–78 years; SD, 14.2 years). Mean EEG recording length was 23.4 hours (range, 6.4–24.2 hours). A total of 2,805 hours of EEG were marked by each reader.

Readers A, B, and C marked 275, 202, and 270 seizures, respectively. Combining overlapping seizure marks resulted in 411 individual seizures, of which 210 (51.1%) were marked by one reader, 68 (16.5%) by two readers, and 133 (32.4%) by all three readers (Fig. 2). Readers noted a variety of partial and generalized seizure types, but further characterization of the seizures was not performed.

Median seizure duration was 50 seconds (mean, 83.9 seconds; SD, 199 seconds; first quartile, 24.5 seconds; third quartile, 87 seconds; range, 3–3,000 seconds). There was a median of two seizures per record (mean, 3.4; range, 0–35).

Individual readers devoted a mean of about 68 hours evaluating the recordings. The mean sensitivity and false-positive rates for the human experts and algorithms, using pairwise comparisons, are shown in Table 1.

The ROC plots for the two algorithms (pairwise comparison with readers) are graphed in Fig. 3, along with the three average human expert comparisons (note that for the experts, their average result derives from two pairwise comparisons, whereas the algorithms' average results derive from three pairwise comparisons). Of the two algorithms, only P14 has both sensitivity and false-positive rate approximating the expert readers. In contrast, the earlier P13 algorithm has a much higher false-positive rate compared with the expert readers. The P14 algorithm can be compared with the P13 algorithm by choosing a P14 setting where their sensitivities are approximately equivalent. At that point, the P14 algorithm has a false-positive rate one-fifth that of the earlier algorithm.

Figure 4 shows the pairwise differences (by record) for the humans and algorithms. The sensitivity and false-positive rate were computed using accelerated bootstrap (BCa, $N = 3,000$) with results $\delta_{\text{sens}} = -22.0$ and $\delta_{\text{FP}} = 1.36$. For P14, using duration and probability thresholds of 8 seconds and 0.8, we found that the lower bound of the sensitivity differences confidence intervals are greater than -22.0 for all human comparisons. Also, the false-positive rate differences confidence intervals were less that 1.36. Meeting these two criteria resulted in a positive conclusion for the hypothesis that P14 is noninferior

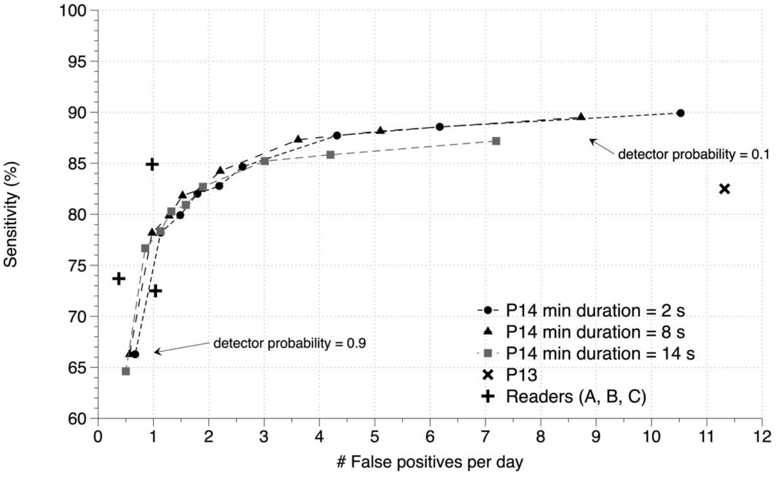


FIG. 3. For expert readers ($n = 3$), Persyst 13 algorithm, and Persyst 14 algorithm, plots of average false-positive rate versus average sensitivity with respect to reference expert markings ($n = 120$ records) using pairwise comparisons. The Persyst 13 algorithm has only one setting, whereas the Persyst 14 algorithm results depict three detector duration threshold settings (2, 8, and 14 seconds) and probability threshold settings varying from 0.1 to 0.9 (in 0.1 increments) at each duration setting.

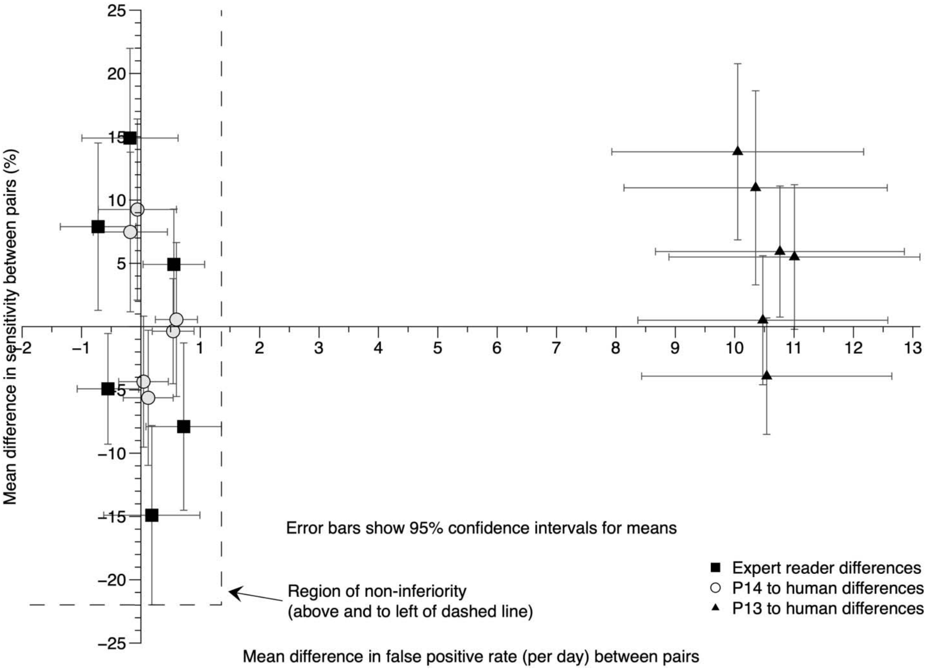


FIG. 4. Summary of sensitivity and false-positive rate pairwise *differences* per case, averaged over all cases ($n = 120$), between human expert pairs and algorithm–human pairs. See text for additional explanation.

to human readers. The P13 algorithm's false-positive rate did not meet the required criteria, resulting in a negative conclusion for the noninferior hypothesis.

Figure 5 illustrates the ROC plots for P14 concerning seizures marked by at least one, at least two, or three experts, in turn. The algorithm is more sensitive to events deemed agreeable (consensus) by more readers. At an algorithm duration threshold of 8 seconds and probability threshold of 0.8 (the parameters yielding noninferior performance to experts in the earlier evaluations), 90% of seizures scored by all three experts were identified with 1.4 false positives per day. At those same thresholds, Fig. 6 shows P14 and expert performance using consensus sets consisting of seizures identified by the three combinations of two experts. Persyst 14's ability to detect these consensus-of-two seizures (91%, 88%, and 90%) was comparable with the experts (80%, 90%, and 97%).

Figure 7 shows box plots of measured detection latencies for P13 and P14 (161 paired consensus-marked cases). Median latency for P14 was 30 seconds and for P13 was 76 seconds ($P < 0.0001$).

DISCUSSION

This is the first study to thoroughly appraise seizure detection algorithms in comparison with expert humans using a large dataset of continuous EEGs independently and comprehensively assessed for seizures by several experts. Experts achieved an average pairwise sensitivity of 77% and false-positive rate of 0.8 per day. The P14 seizure detector's performance was statistically noninferior to the performance of this study's expert readers relative to one another. The performance of the older P13 seizure detection

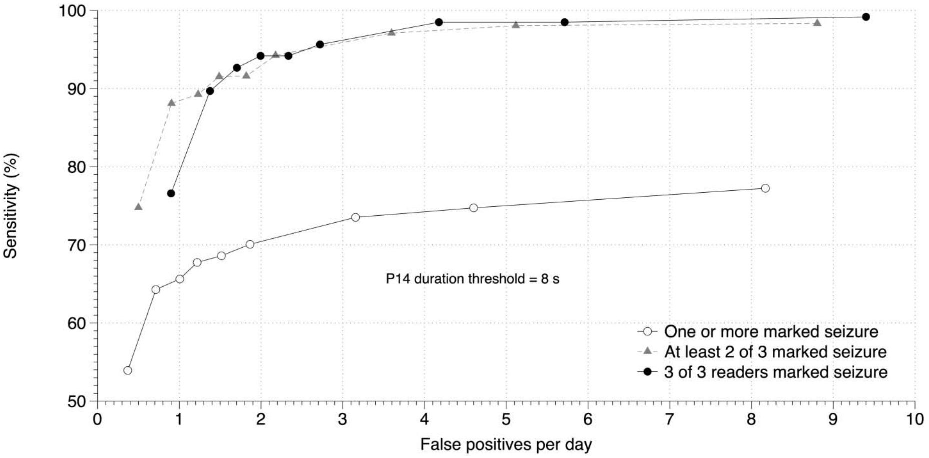


FIG. 5. False-positive rate versus sensitivity for the Persyst 14 seizure detection algorithm (duration threshold setting 8 seconds and probability threshold setting varying from 0.1 to 0.9), plotted for three levels of human reader consensus agreement: any seizure mark, at least two of three agreement, or three of three agreement.

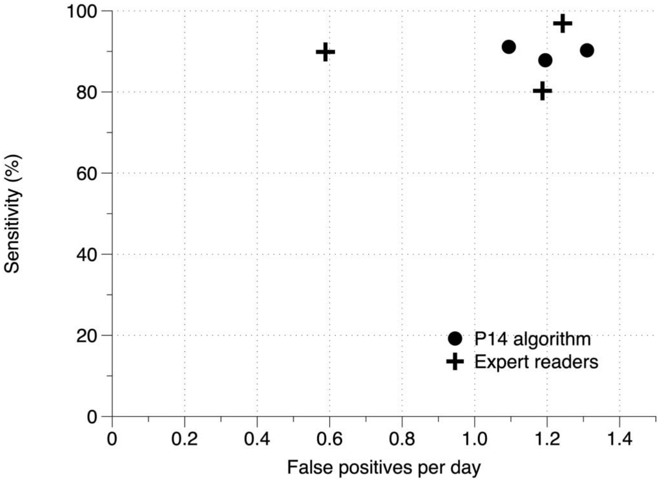


FIG. 6. False-positive rate versus sensitivity plotted for the Persyst 14 algorithm (duration = 8 seconds, $p = 0.8$) and human readers for sets of consensus seizures marked by the three combinations of two of three human readers.

algorithm did not approach that of the human experts because of a much higher false-positive rate.

The study design addressed some of the methodological issues evident in prior seizure detection studies. In comparison with previous seizure marking and detector algorithm studies, this study used a larger collection of multi-reader marked seizure-containing records,[6,12] a large number (100) of consecutive seizure-affected patients (the study by Furbass et al.[19] included 94 consecutive seizure-affected patients; Wilson et al.[17] assessed 426 seizure-affected patients but used briefer clips of EEG, most marked by single readers), and the second longest overall duration of carefully marked EEG (recent neonatal seizure marking study by Stevenson[15] evaluated more hours of EEG; a study by Gotman[20] did not include expert reader evaluation of the entire EEG). The consecutive nature of the seizure-affected records resulted in less preselection of overtly demonstrative electrographic seizure patterns, and the large number of records from different individuals favored a broader representation of seizure types typically encountered in the EMU. Also, the algorithm-to-human pairwise evaluation technique better accounted for interexpert variability in identifying seizures and, in comparison to consensus marking methods, does not require discarding less expert-agreeable events or the potentially biased assumption of improved accuracy of consensus-identified events.

Using pairwise comparisons, the three expert readers in this study only agreed with one another, on average, in 77% of events marked as seizures. This is slightly better than the 70% level of sensitivity among several readers reported by Tu et al.[7] for an ICU dataset, possibly reflecting somewhat more conventional ictal patterns in the EMU as compared with the neuro-ICU. The level of interreader sensitivity reported here is less than that reported by Wilson et al.[5] (92%), using a set of EMU seizure-containing records, likely because that dataset was subject to preselection favoring demonstrative ictal patterns and was much smaller in size than the current test dataset, thus easing the burden of review.

We believe, given the large size of this dataset and its careful marking by experts for research purposes, that the 77% level of average interreader sensitivity is likely a high-end estimate of the level of performance that could be expected during standard expert clinical waveform review of prolonged EEG recordings. The readers here were not facing significant time constraints, and they were aware that their markings would be compared with other experts. This would favor more thorough and considered marking. In the standard clinical review setting, where reading time is more constrained, distractions are more prevalent, and fatigue is common, it is likely that significantly more seizures are missed than was evident here. Multiple layers of independent human review, or the use of assistive technologies such as automated seizure detection and quantitative EEG trending, would be expected to improve on the baseline performance of a single expert reader.

The consensus seizure subset analyses conducted in this study demonstrated that, once two readers were in agreement regarding the presence of a seizure, a third reader was more likely to also identify such events as seizures. Readers averaged a sensitivity of 89% for these consensus-of-two events. The P14 seizure detector demonstrated a sensitivity of 89.7% for the

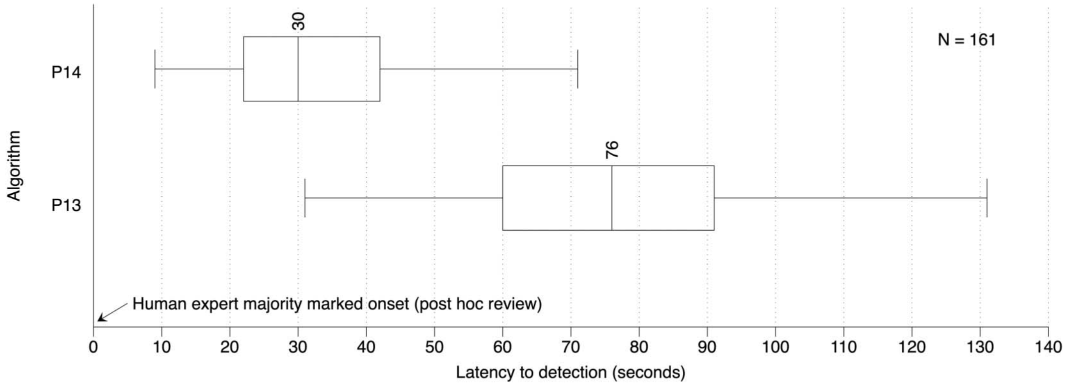


FIG. 7. Box plots of Persyst 13 and Persyst 14 seizure detection latencies with respect to seizure onset marked by expert majority rule in post hoc analysis. Box indicates 25th to 75th percentiles; whiskers show nonoutlier range.

same-consensus-of-two seizures, which was closely comparable with the human readers' average performance.

The P14 algorithm, ostensibly by a fourth reader, was also evaluated against the subset of seizures marked by all readers. There it performed at a slightly higher sensitivity of approximately 90% with a concomitant false-positive rate of about one per day (where any event not identified by all experts was designated a false positive). If, during clinical review, a P14 detection algorithm sensitivity was chosen that resulted in an average of four false positives per day, the sensitivity for three of three consensus seizure events would be 98.5% in this dataset. These data suggest that P14 performs similarly to expert readers in identifying "agreeable" seizures.

The median latency to seizure detection for the P14 algorithm (30 seconds) was much shorter than that of P13 (76 seconds). This indicates that automated online seizure alerts occur in a timelier fashion using P14. Of note, the latencies reported here are worst-case estimates relative to real-world observation by a trained electroencephalographer. When a trained person watches the EEG in real time, there is usually a lag between the time they recognize a seizure in progress and the onset time determined during careful post hoc review of the waveforms. Indeed, seizures are often missed during real-time observation, and only later identified during EEG review. We did not attempt to determine the typical lag time for skilled real-time human observation.

This study had several limitations. Its EEGs originated from a single center and were obtained from a group of definitively diagnosed patients with epilepsy. At least one clinician identified seizures in the majority (83%) of the EEGs assessed in this study. These selection biases possibly increased the average likelihood of seizure recognition by the experts. The quantity of individual records and volume of EEG evaluated were large but still do not fully represent the variability and edge cases found in the universe of EMU EEGs. A larger dataset sampled randomly or sequentially from many centers, incorporating more EEG recording settings, would improve data quality and strengthen the results of similar analyses. Although trained at different fellowship programs, the expert electroencephalographers all worked together for several years. The readings in this study were performed independently, but the readers' long-term shared work environment might have fostered reading style homogenization and marking agreement. We had no ground truth by which to validate the expertise of the readers. Their marking styles, though statistically recognizable (results not shown), seemed overall comparable. Increasing the number of readers would yield an improved statistical performance profile for expert humans. It should be noted, though, that the addition of readers would probably broaden the limits of interreader variability. In principle, were results from a sufficient number of experts available, new readers or algorithms could be evaluated with respect to the performance of the middle of an expert distribution.

It is possible that the use of by-case sensitivity and false-positive rate calculations for the various experts and algorithms, and accelerated bootstrap statistical methodology, could have resulted in an underestimate of the effects of extreme outlier low sensitivity or high false-positive rates on the results. However, an analysis of the effects of omitting the bootstrap estimate procedure showed that it had a very small absolute effect on the performance metrics and no effect on the ultimate result of noninferiority for P14 compared with the expert readers (results not shown). Assessment of individual extreme outlier results could potentially identify significant edge-case differences between human experts or human experts and the algorithms, and such information could prove useful in framing future algorithm training modifications that would further harmonize algorithm and expert results. We plan on conducting further analyses along these lines.

In future studies using this dataset, we hope to delineate some of the causes of disagreement between experts. If such causes can be identified, then it may be possible to minimize some of them and so increase expert agreement through training or additional adjunctive methods of data analysis. The other option of trying to define classes of features that foster agreement, while ignoring the not infrequent gray area EEG patterns that lead to differences in interpretation, is of questionable utility in clinical practice.

Experts' mean pairwise seizure sensitivity for these data, approximately 77%, indicates that well-trained humans are imperfect in their assessment of prolonged EEGs for seizures. The P14 seizure detector was statistically noninferior to this study's experts. The P13 algorithm was inferior. The P14 algorithm also detects seizures much sooner than P13. The pairwise comparative data indicate that the P14 detector has passed a modified Turing test,[21,23] showing that the computerized detector's mean performance across many recordings is statistically noninferior to that of the expert readers who assessed the same data. This does not mean that the P14 algorithm will detect all seizures recognized by human experts, particularly for edge cases not well-represented in its training sets, but the algorithm should be quite useful as an adjunct to the existing visual observation methods of seizure identification, sometimes enabling earlier seizure recognition, improving review efficiency, and enhancing overall seizure recognition.

REFERENCES

1. Krishnan V, Chang BS, Schomer DL. The application of EEG to epilepsy in adults and the elderly. In: Schomer DL, Lopes Da Silva F, eds. Niedermeyer's electroencephalography: basic principles, clinical applications, and related fields. 7th ed. Philadelphia: LWW, 2018; 521–535.
2. Koubeissi MZ, So EL, Nordli DR. EEG in adult epilepsy. In: Current practice of clinical electroencephalography. 4th ed. Philadelphia: Lippincott Williams & Wilkins, 2014; 315–337.
3. 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. Electroencephalogr Clin Neurophysiol Suppl 1999;52:21–41.
4. Hirsch LJ, LaRoche SM, Gaspard N, et al. American Clinical Neurophysiology Society's standardized critical care EEG terminology: 2012 version. J Clin Neurophysiol 2013;30:1–27.
5. Wilson SB, Scheuer ML, Plummer C, Young B, Pacia S. Seizure detection: correlation of human experts. Clin Neurophysiol 2003;114:2156–2164.
6. Halford JJ, Shiau D, Desrochers JA, et al. Inter-rater agreement on identification of electrographic seizures and periodic discharges in ICU EEG recordings. Clin Neurophysiol 2015;126:1661–1669.
7. Tu B, Young GB, Kokoszka A, et al. Diagnostic accuracy between readers for identifying electrographic seizures in critically ill adults. Epilepsia Open 2017;2:67–75.
8. Jacobellis v. Ohio, 378 U.S. 184 (1964).

9. Grant AC, Abdel-Baki SG, Weedon J, et al. EEG interpretation reliability and interpreter confidence: a large single-center study. Epilepsy Behav 2014;32:102–107.
10. Ronner HE, Ponten SC, Stam CJ, Uitdehaag BMJ. Inter-observer variability of the EEG diagnosis of seizures in comatose patients. Seizure 2009;18:257–263.
11. Benbadis SR, LaFrance WC, Papandonatos GD, et al. Interrater reliability of EEG-video monitoring. Neurology 2009;73:843–846.
12. Kelly KM, Shiau DS, Kern RT, et al. Assessment of a scalp EEG-based automated seizure detection system. Clin Neurophysiol 2010;121:1832–1843.
13. Abend NS, Gutierrez-Colina A, Zhao H, et al. Interobserver reproducibility of electroencephalogram interpretation in critically ill children. J Clin Neurophysiol 2011;28:15–19.
14. Gaspard N, Hirsch LJ, LaRoche SM, Hahn CD, Westover MB; Critical Care EEG Monitoring Research Consortium. Interrater agreement for critical care EEG terminology. Epilepsia 2014;55:1366–1373.
15. Stevenson NJ, Clancy RR, Vanhatalo S, Rosén I, Rennie JM, Boylan GB. Interobserver agreement for neonatal seizure detection using multichannel EEG. Ann Clin Transl Neurol 2015;2:1002–1011.
16. Pauri F, Pierelli F, Chatrian GE, Erdly WW. Long-term EEG-video-audio monitoring: computer detection of focal EEG seizure patterns. Electroencephalogr Clin Neurophysiol 1992;82:1–9.
17. Wilson SB, Scheuer ML, Emerson RG, Gabor AJ. Seizure detection: evaluation of the Reveal algorithm. Clin Neurophysiol 2004;115:2280–2291.
18. Kamitaki BK, Yum A, Lee J, et al. Yield of conventional and automated seizure detection methods in the epilepsy monitoring unit. Seizure 2019;69:290–295.
19. Fürbass F, Ossenblok P, Hartmann M, et al. Prospective multi-center study of an automatic online seizure detection system for epilepsy monitoring units. Clin Neurophysiol 2015;126:1124–1131.
20. Gotman J. Automatic seizure detection: improvements and evaluation. Electroencephalogr Clin Neurophysiol 1990;76:317–324.
21. Scheuer ML, Bagic A, Wilson SB. Spike detection: inter-reader agreement and a statistical Turing test on a large data set. Clin Neurophysiol 2017;128:243–250.
22. Wellek S. Testing statistical hypotheses of equivalence and noninferiority. 2nd ed. Boca Raton: CRC Press, 2010.
23. Turing AM. Computing machinery and intelligence. Mind 1950;59:433–460.