Published in final edited form as:

Clin Neurophysiol. 2012 June ; 123(6): 1088–1095. doi:10.1016/j.clinph.2011.09.023.

High inter-reviewer variability of spike detection on intracranial EEG addressed by an automated multi-channel algorithm

Daniel T. Barkmeier¹, Aashit K. Shah², Danny Flanagan³, Marie D. Atkinson², Rajeev Agarwal⁴, Darren R. Fuerst², Kourosh Jafari-Khouzani⁵, and Jeffrey A. Loeb^{1,2}

¹The Center for Molecular Medicine and Genetics, Wayne State University School of Medicine, Detroit Michigan

²Department of Neurology, Wayne State University School of Medicine, Detroit Michigan

³Brain Research Institute, Florey Neuroscience Institutes (Austin), Melbourne, Victoria, Australia

⁴Center for Signal Processing and Communications, Department of Electrical and Computer Engineering, Concordia University, Montreal, QC

⁵Department of Diagnostic Radiology, Henry Ford Hospital, Detroit Michigan

Abstract

Objectives—The goal of this study was to determine the consistency of human reviewer spike detection and then develop a computer algorithm to make the intracranial spike detection process more objective and reliable.

Methods—Three human reviewers marked interictal spikes on samples of intracranial EEGs from 10 patients. The sensitivity, precision and agreement in channel ranking by activity were calculated between reviewers. A computer algorithm was developed to parallel the way human reviewers detect spikes by first identifying all potential spikes on each channel using frequency filtering and then block scaling all channels at the same time in order to exclude potential spikes that fall below an amplitude and slope threshold. Its performance was compared to the human reviewers on the same set of patients.

Results—Human reviewers showed surprisingly poor inter-reviewer agreement, but did broadly agree on the ranking of channels for spike activity. The computer algorithm performed as well as the human reviewers and did especially well at ranking channels from highest to lowest spike frequency.

Conclusions—Our algorithm showed good agreement with the different human reviewers, even though they demonstrated different criteria for what constitutes a ‘spike’ and performed especially well at the clinically important task of ranking channels by spike activity.

Significance—An automated, objective method to detect interictal spikes on intracranial recordings will improve both research and the surgical management of epilepsy patients.

© 2011 International Federation of Clinical Neurophysiology. Published by Elsevier Ireland Ltd. All rights reserved.

Corresponding Author: Jeffrey A. Loeb, M.D., Ph.D. Department of Neurology Wayne State University School of Medicine 3122 Elliman Building 421 East Canfield St. Detroit, MI 48201 jloeb@med.wayne.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

electrocorticography; interictal spike; epilepsy surgery; quantitative methods

INTRODUCTION

Interictal spikes are abnormal discharges that occur between seizures in patients with epilepsy. They are frequently generated in the same regions of the brain that initiate seizures (Asano et al. , 2003, Marsh et al. , 2010), but their exact role in epileptogenesis and the induction or protection from seizures remains controversial and requires further study. A significant barrier to such studies comes from the laborious nature of quantifying many thousands of interictal spikes that vary greatly over time (Marsh et al., 2010), together with the subjectivity and tremendous variability between EEG reviewers on exactly what constitutes a spike on subdural electrocorticography (ECoG) (DumpeImann and Elger, 1998, 1999).

To address these issues, we first conducted a study of inter-reviewer variability in marking interictal spikes in ECoG and, second, compared these results to an automated spike detection algorithm. While many algorithms have been developed for the detection of interictal spikes, most are designed for scalp recordings (Wilson and Emerson, 2002), and those few that were developed specifically for ECoG were either run on single channels or had humans validate spikes after they were detected by an algorithm (Brown et al. , 2007, DumpeImann and Elger, 1998, 1999, Valenti et al. , 2006). Here, we show surprisingly poor inter-reviewer agreement on what constitutes a spike in 10 patients and validated a novel automatic spike detection algorithm that generally agrees with human reviewers in that it emulates human reviewing methods by analyzing multi-channel ECoG sections. An automated spike detection method that produces good agreement with human reviewers will allow real-time measurements of interictal spiking and enable further quantitative studies of their clinical relevance.

METHODS

Patient selection and recordings

Intracranial EEG recordings for this study were obtained from 10 patients undergoing two-stage surgical resections at the Comprehensive Epilepsy Program at Wayne State University (#0860000MP2F, Wayne State University Human Investigation Committee). All patients were suffering from intractable epilepsy and had failed multiple antiepileptic drugs. Patients with tumors, malformations, or any brain pathology other than gliosis were excluded from this study. These patients represent an equal mix of children (8 months-14 years) and adults (20-54 years), and were selected from recent cases to include a wide range of activity levels and spike morphologies. Recordings were obtained with a 128-channel Stellate Harmonie Digital recorder (Stellate Inc., Montreal, PQ, Canada) sampled at 200Hz, from multiple arrays of subdural grid electrodes (PMC Cortec grid electrodes, PMC Corporation, MN, U.S.A., 4-mm diameter, 10 mm inter-electrode distance). Ten-minute samples were chosen by a fellowship-trained electroencephalographer not involved in the marking of spikes for this study from periods of quiet wakefulness at least six hours from a seizure. Periods of quiet wakefulness were chosen to provide a uniform state across patients and because this is the state used for interictal analysis for clinical decision making at our Comprehensive Epilepsy Program. Recordings were not screened for their technical quality. The average number of channels recorded per patient was 95.9 (SD $\pm$ 16.5; range 65-128). The initial development and testing of the algorithm was done on data from a separate set of 10 patients, whose sample recordings were chosen by the same criteria as the testing data. The

average age of the patients in the training set was 20.1 years ($SD \pm 15.7$; range 2-52) and the average number of channels per patient was 100.4 ($SD \pm 16.9$; range 78-128).

Human spike detection

Three trained electroencephalographers (two from the same institution and one from another) independently reviewed these sections in the same referential montage, with instructions to mark all spikes occurring on all channels (Table 1). Reviewer 1 is board-certified and fellowship-trained, Reviewer 2 has over 15 years of experience in reading EEGs and assisting in spike detection development, and Reviewer 3 is fellowship-trained. Stellate's Harmonie software was used to display the recordings and mark individual interictal spikes. Reviewers were instructed to mark all spikes occurring in all channels, including each spike occurring in a train of spikes and time-locked spikes occurring on nearby channels. Exceptions to marking included channels consisting of 60Hz electrical artifact and the sharp background activity encountered over primary motor/sensory/language cortex. Reviewers were provided with a line drawing of electrode grid placement to help identify these areas. Outside of these stipulations, reviewers were free to mark interictal spikes however they would for clinical purposes.

A Matlab script was used to tabulate the results of the interictal spike detections. Detections which occurred on the same channel within 115ms of another reviewer's mark were taken as detections of the same spike. This range was chosen because it allows for variation in the manual placement of marks without risking the inclusion of another nearby spike, and is similar to values used in previous studies (115ms (Dümpelmann and Elger, 1999); 100ms (Brown, 2007)). Varying the range by ± 10 ms did not affect results. As there is no true 'gold standard' for comparison, each reviewer was compared to the set of all marks made by the other two reviewers (Table 2). For example, a spike marked by Reviewer 1 would be considered a true positive if the same spike was also marked by either Reviewer 2 or Reviewer 3. We then used these results to calculate sensitivity (true positives/(true positives + false negatives)) and precision (true positives/(true positives + false positives)) for each human reviewer relative to the other reviewers. Finally, we used each reviewer's spike detections to generate a ranking of channels from highest spike frequency to lowest. This was done because of the importance of distinguishing brain regions based on activity levels that can be used as a means of determining the relationship between interictal spikes and surgical outcome through channel ranking (Asano, 2003, Hufnagel et al., 2000, Kim et al., 2010, Marsh et al., 2010). We compared rankings between reviewers in each patient by calculating Kendall's coefficient of concordance (W), which is a non-parametric test for assessing agreement among multiple raters that can be used on ranked data. The test statistic ranges from 0 (no agreement) to 1 (complete agreement). Spearman's rank correlation coefficient was also calculated for each reviewer pair in each patient, as this is a more common measure of non-parametric correlation, and so that human reviewer performance could be compared to algorithm performance.

Spike detection algorithm

The spike detection method described here fundamentally differs from previous methods by taking into account the background amplitudes of all channels in a recording, rather than examining single channels individually. By using a single scaling factor to adjust the amplitude of all channels equally, differences between high spike frequency and low spike frequency channels can be better preserved. The algorithm first uses a 20-50Hz filter on raw ECoG data to identify all potential spikes. It then uses the raw ECoG data, now filtered at 1-35 Hz, to determine a single scaling factor that is used to scale all channels in the block of ECoG. The scaled data is then used to develop amplitude and slope thresholds for the final detection of the spikes (a subset of the 'potential spikes'). The goal of this algorithm is to

mimic what an electroencephalographer ‘sees’ when looking at many channels simultaneously to find relative differences between nearby channels.

The spike detection algorithm was developed as a Matlab program, which interfaces directly with the Stellate ECoG files (Algorithm overview in Figure 1A). Initial development was done on a set of training patients entirely independent of the testing set on which statistical analysis was done, with input from a fourth fellowship-trained epileptologist not involved in marking the testing set. The following describes the algorithm in more detail:

1. The spike detection algorithm first extracts successive one-minute blocks of multi-channel data and removes artifact channels. Individual channels are considered artifactual for a given segment if the average slope is greater than 10 standard deviations outside the mean slopes of all channels. This method was developed on the training set of patients and has thus far worked without error. In the testing dataset analyzed here, only 3 out of 10 patients had individual channels that satisfied this criterion and were removed from further consideration. In all cases the algorithm was verified with visual inspection by each of the three reviewers. Briefer and more difficult to detect artifacts are not addressed by this method.
2. In the second step, the ECoG data are bandpass filtered between 20 and 50Hz to identify potential interictal spikes for further examination (similar to Brown, et al. 2007). This technique utilizes a classical definition of an interictal spike proposed by Niedermeyer (Niedermeyer and Lopes da Silva, 2005), where a sharp discharge must last between 20 and 70 milliseconds. In processing the training dataset it was our observation that the bandpass filter (20-50Hz) allows the peaks with durations of 20-50ms to stand out sharply from the background, while simultaneously preventing interference from 60Hz electrical artifact (Figure 1B). The absolute amplitudes of peaks greater than four standard deviations of the channel mean amplitude are noted as potential spike locations for further consideration.
3. Next, the raw ECoG data are bandpass filtered at 1-35Hz (second order digital Butterworth) and all channels in the one-minute block are scaled by a single scaling factor. The block scaling factor is determined by finding the value that will bring the median of all the channel amplitudes to an arbitrary, static value (70 μ V). We use the median of the average amplitudes rather than the mean because the mean of the extremely active channels tends to skew the scaling factor. Moreover, we have found that using the median for each patient, rather than the mean, provides more reliable results across multiple patients. To do this, the average value of the rectified amplitude for each channel is first calculated. Once the median value of the average channel amplitudes is calculated, all channels in this one-minute block are multiplied by the value that would bring this median amplitude up to 70 μ V. This method of scaling all channels together as one block was chosen because it most closely approximates how a human reviewer adjusts the display amplitude to review a recording and because scaling individual channels tends to cause over-marking channels with fewer spikes and under-marking of channels with many spikes (Figure 1C).
4. Once the data have been scaled, the amplitude and slope of each half-wave of the potential spikes identified previously in step 2 are calculated (Rakhade et al. , 2007) and the values are compared to static thresholds (Total amplitude of both half-waves > 600 μ V, slope of each half-wave > 7 μ V/ms, duration of each half-wave > 10ms). Potential spikes with half-waves that exceed these thresholds are marked as interictal spikes. This step incorporates another classic, albeit subjective, definition of an interictal spike, where a sharp wave must clearly stand out from the background activity (Cooper et al. , 1969). With this technique, all spikes on all

channels are marked individually. It does not attempt to spatially localize time-locked spikes occurring on adjacent channels.

Comparison to human reviewers and a commercial algorithm

Two measures were used to evaluate algorithm performance: the ability to correctly identify individual spikes marked by human reviewers and the ability to correctly rank channels by spike frequency as compared to human reviewers. We did the same comparison using the commercial spike detector included with Stellate's Harmonie software (version 6.1c) as a comparison to a widely-used algorithm. While originally developed for scalp EEGs, this FDA-approved software has subsequently been adapted for intracranial recordings and is used in many epilepsy centers. The Stellate detector compares the amplitude and sharpness of waveforms to the preceding five seconds of data on the channel being analyzed to see whether it stands out from background. We used the standard detector with the default amplitude threshold of '3', as we found this gave the best results. Sensitivity and precision of the algorithm were calculated in relation to each individual human reviewer, and then these values were averaged to give the final value. To evaluate the performance of the two automatic algorithms (ours and Stellate's) in ranking channels by spike frequency, Spearman's rank correlation coefficient was calculated between the channel ranking of each method and the average channel rank from the three human reviewers.

RESULTS

Highly trained human reviewers were asked to manually mark individual spikes on each of an average of 96 channels from 10 different patients. A total of 78,743 spike detections were made from highly variable ECoG patterns. Surprisingly, only 23.7% of these were marked by at least two reviewers and just 3.1% by all three reviewers (Table 1). Overall, human reviewers performed poorly in regard to identifying each other's spikes (Table 2). While reviewers 1 and 3 (from the same institution) had similar sensitivity (32% and 37%, respectively) and precision (43% and 37%, respectively), reviewer 2 marked many fewer spikes and as a result had markedly reduced sensitivity (14%), but increased precision (72%).

In order to provide uniformity as well as dramatically improve speed for quantitative analyses, we developed an automatic spike detection algorithm that uses frequency filtering followed by spike detection using the entire block of channels for each patient (Figure 1). The algorithm 'walks' through the ECoG in one minute intervals using a 4-step process described in detail in the methods section and the flow chart in Figure 1A. The first step removes artifactual channels that have significantly higher slopes compared to other channels. The second step uses a narrow bandpass filter (20-50Hz) to identify all potential spikes shown in Figure 1B. The third step creates a comparison of the entire block of channels (block scaling) using a 1-35Hz filter, and the fourth uses this block scaling and excludes spikes identified from the bandpass filtered data in step 2, whose amplitudes and slopes fall below an arbitrary threshold. These final spikes are then marked on the ECoG tracing to allow for easily visualization by human reviewers. An example of how spikes are marked when taking into account the entire block of channels as compared to scaling each channel individually is shown in Figure 1C, where the block scaling excludes low amplitude waveforms in channels 4 and 5 that would otherwise have been marked as spikes if individual channels were considered in isolation. The goal was to replicate the process used by a human reviewer who sees all channels simultaneously and identifies spikes based on how they stand out from the background as a whole.

We next compared this automatic algorithm to each of the individual reviewers (Table 3). While the algorithm detected more spikes than any single reviewer (50,315 vs. 39,802 for

highest reviewer), it performed with similar sensitivity and precision for each of the three human reviewers, and yielded more consistent results with the 3 human reviewers than a commercial detector that uses the sharpness of a waveform compared to the preceding 5 seconds (Harmonie V6.1c adapted from scalp EEG to ECoG) (Table 3). Sample channels selected from 4 patients showing widely different ECoG patterns, as well as sensitivities and specificities, are shown in Figure 2 to illustrate how our spike detection algorithm compares to each of the 3 human reviewers. It can be seen from this sampling how different ECoG patterns result in widely different spike detections by human reviewers. While the algorithm may not be perfect, by identifying more spikes in a defined fashion, it almost always agrees better with individual reviewers that show considerable variation.

Since the algorithm consistently detected more spikes than the human reviewers, we asked whether spikes picked up only by the algorithm were truly false positives or were simply missed by the reviewers. Ten of these spikes were randomly selected from each of the 10 patients in the testing set and then re-presented to each of the reviewers. They were then asked whether these detected spikes were real or not (Table 4). While not all were considered to be spikes by all reviewers, the large number of spikes detected only by the algorithm, but then later validated by human reviewers suggests that many of these are not 'false positives,' but 'real' spikes that were missed (Table 4. Mean: $31.3 \pm 30.4\%$; Reviewer 1: 65%; Reviewer 2: 6%; Reviewer 3: 23%). While the range here is consistent for each reviewer's pattern of spike identification, it shows that, unlike a computer algorithm, it is difficult for even an experienced human reviewer to consistently identify the same spikes from one reading to the next. On average, the algorithm detects about 40% more spikes that are not detected by any of the reviewers. Visual examination suggests that many of these detected spike events could indeed be considered spikes that would otherwise have been missed. It is impossible to know if spikes never identified by human reviewers are false positives or not.

In clinical practice, however, one could argue that identification of individual spikes is not important. What is critical for surgical decision-making is the ranking of channels from the highest to the lowest spiking. Based on the marked spikes from each human reviewer and both our new and the existing commercial algorithm, we compared channel ranking for each of the 10 patients. Unlike individual spike detection, channel ranking generally showed good agreement for most patients when simply ranking channel activity from highest to lowest spiking (Table 5). We used two different statistical measures of rank correlation. First, we found an overall Kendall's W correlation of 0.752 among the three human reviewers. Second we found that Spearman's rank correlation coefficient, a similar measure of ranking agreement for pairs of data, varied widely among pairs of reviewers in the 10 study patients. However, each reviewer had a similar average across the 10 patients that was comparable to our custom algorithm's value of 0.750 (Table 6). The commercial algorithm did less well with an average Spearman rank coefficient of 0.574.

As a means to illustrate these comparisons visually, we compared graphical summaries of the channel rankings from our custom algorithm to the three human reviewers (Figure 3). Heatmaps of the spike frequency are shown from highest in red to lowest in green and are superimposed on a patient's 3-dimensional brain rendering. These illustrations for patient 3 show that despite vastly different markings of individual spikes, the human reviewers in fact generally agree on the regions of highest and lowest spiking and that the custom algorithm does a nice job of predicting the averaged results from all 3 reviewers.

DISCUSSION

This is perhaps the largest study of inter-reviewer variability in the identification of interictal spikes in intracranial EEG. There was surprisingly poor agreement (Table 1. $3.1 \pm 2.7\%$; range 0-7.4% for different patients) between three trained human reviewers in the identification of what constitutes an interictal spike. This poor agreement may be due to the lack of a rigid definition of an interictal spike, fatigue from such an extensive marking task, or the study design, which asked the reviewers to mark all spikes on all channels just as they would in clinical practice. This method is likely to produce more variability than studies using pre-selected spikes or evaluating only one or a few channels at a time. Perhaps less surprising is that the two reviewers from the same institution had better concordance (Reviewer 1 and Reviewer 3), suggesting an institutional bias in ECoG interpretation that could significantly affect surgical decisions at different programs. The sensitivity and precision results are somewhat lower than previous studies which compared reader variability using shorter files or individual channels (Black et al. , 2000, Brown, 2007, Dumpelmann and Elger, 1999, Webber et al. , 1993, Wilson et al. , 1996). Much of this is clearly due to the institutional bias shared by Reviewers 1 and 3, but absent in Reviewer 2, who reported significantly less spikes than the others. While there was significant disagreement on the identification of individual spikes for a given patient, the reviewers generally agreed on the ranking of channels from highest to lowest spike frequency. This is important because determination of the highest-spiking channels is often used clinically for surgical resection margins.

Towards this end, our detection algorithm was able to identify spikes with similar sensitivity and precision as the human reviewers, and does an excellent job of ranking channels based on interictal spiking activity. A common problem with automatically detecting interictal spikes is deciding which waveforms stand out from the 'background'. Most methods use measurements from a single channel at a time to determine the background and then ask whether spikes on that channel stand out from that measurement (Brown, 2007, Dumpelmann and Elger, 1998, 1999, Gotman and Gloor, 1976, Valenti, 2006). We have found that this approach leads to under-marking of channels containing numerous spikes, due to a higher measure of 'background' activity and over-marking of relatively inactive channels. Here we instead uniformly increase or decrease the amplitude of all channels in a recording as a block in order to preserve differences between high spike frequency and low spike frequency channels. Previous detection algorithms have been tested only against a small set of pre-extracted spikes or single channels from each patient, while here we show that our algorithm performs well when reviewing entire segments in a way that mimics how a human reviewer would see the ECoG.

The algorithm's code is fairly simple and fast to execute, meaning that it could easily be integrated with EEG recording software so that spikes could be marked each minute while a recording is being made (see Supplementary Material S1 for the code as a Matlab function). Other key advantages to this algorithm are that it requires no human intervention or tweaking of parameters and its performance was validated in a completely 'real world' environment. Since completing this study, we have now used this algorithm on longer stretches of ECoG in a variety of different states from a large number of patients. We have found that it generally performs as well, or better, than human reviewers and now gives a quantitative *and consistent* measure of spike frequency, which would not have been obtained otherwise because of the laborious nature of manual spike marking.

While poor agreement among reviewers may cast doubt on the importance of quantifying interictal spikes, numerous studies have shown the quantitation of interictal spikes to be clinically useful (Alarcon et al. , 1997, Asano, 2003, Bautista et al. , 1999, Holmes et al. ,

2000, Hufnagel, 2000, Hufnagel et al. , 1994, Kanazawa et al. , 1996, Marsh et al., 2010, McBride et al. , 1991, Pressler et al. , 2005). Our findings therefore underscore the need for automated methods of spike detection which function reliably across a wide range of ECoG types (Dümpelmann and Elger, 1998, 1999, Gotman, 2001, Wilson, 1996). A uniform method for spike identification will be critical not only for research into the importance of interictal spiking in epilepsy and behavior, but for any multicenter study that uses different electroencephalographers to read intracranial EEG, as well as for new electrical stimulation methods for epilepsy treatment that rely on rapid, unbiased detections.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Financial disclosure: This work was supported by grants from NIH/NINDS R01NS045207 and R01NS058802 (JAL) and a predoctoral fellowship from the Epilepsy Foundation of America (DTB).

REFERENCES

- Alarcon G, Garcia Seoane JJ, Binnie CD, Martin Miguel MC, Juler J, Polkey CE, et al. Origin and propagation of interictal discharges in the acute electrocorticogram. Implications for pathophysiology and surgical treatment of temporal lobe epilepsy. *Brain*. 1997; 120:2259–82. [PubMed: 9448581]
- Asano E, Muzik O, Shah A, Juhasz C, Chugani DC, Sood S, et al. Quantitative interictal subdural EEG analyses in children with neocortical epilepsy. *Epilepsia*. 2003; 44:425–34. [PubMed: 12614399]
- Bautista RE, Cobbs MA, Spencer DD, Spencer SS. Prediction of surgical outcome by interictal epileptiform abnormalities during intracranial EEG monitoring in patients with extrahippocampal seizures. *Epilepsia*. 1999; 40:880–90. [PubMed: 10403211]
- Black MA, Jones RD, Carroll GJ, Dingle AA, Donaldson IM, Parkin PJ. Real-time detection of epileptiform activity in the EEG: a blinded clinical trial. *Clin Electroencephalogr*. 2000; 31:122–30. [PubMed: 10923198]
- Brown MW, Porter BE, Dlugos DJ, Keating J, Gardner AB, Storm PB Jr. et al. Comparison of novel computer detectors and human performance for spike detection in intracranial EEG. *Clin Neurophysiol*. 2007; 118:1744–52. [PubMed: 17544322]
- Cooper, R.; Osselton, JW.; Shaw, JC. *EEG Technology*. 1st ed.. Butterworth; London: 1969.
- Dümpelmann M, Elger CE. Automatic detection of epileptiform spikes in the electrocorticogram: a comparison of two algorithms. *Seizure*. 1998; 7:145–52. [PubMed: 9627206]
- Dümpelmann M, Elger CE. Visual and automatic investigation of epileptiform spikes in intracranial EEG recordings. *Epilepsia*. 1999; 40:275–85. [PubMed: 10080505]
- Gotman, J. Computer-assisted data collection and analysis.. In: Wyllie, E., editor. *The treatment of epilepsy: principles and practice*. Lippincott Williams and Wilkins; Philadelphia: 2001. p. 209-24.
- Gotman J, Gloor P. Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. *Electroencephalogr Clin Neurophysiol*. 1976; 41:513–29. [PubMed: 61855]
- Holmes MD, Born DE, Kutsy RL, Wilensky AJ, Ojemann GA, Ojemann LM. Outcome after surgery in patients with refractory temporal lobe epilepsy and normal MRI. *Seizure*. 2000; 9:407–11. [PubMed: 10985997]
- Hufnagel A, Dümpelmann M, Zentner J, Schijns O, Elger CE. Clinical relevance of quantified intracranial interictal spike activity in presurgical evaluation of epilepsy. *Epilepsia*. 2000; 41:467–78. [PubMed: 10756415]
- Hufnagel A, Elger CE, Pels H, Zentner J, Wolf HK, Schramm J, et al. Prognostic significance of ictal and interictal epileptiform activity in temporal lobe epilepsy. *Epilepsia*. 1994; 35:1146–53. [PubMed: 7988503]

- Kanazawa O, Blume WT, Girvin JP. Significance of spikes at temporal lobe electrocorticography. *Epilepsia*. 1996; 37:50–5. [PubMed: 8603624]
- Kim DW, Kim HK, Lee SK, Chu K, Chung CK. Extent of neocortical resection and surgical outcome of epilepsy: Intracranial EEG analysis. *Epilepsia*. 2010; 51:1010–7. [PubMed: 20384767]
- Marsh ED, Peltzer B, Brown MW III, Wusthoff C, Storm PB Jr, Litt B, et al. Interictal EEG spikes identify the region of electrographic seizure onset in some, but not all, pediatric epilepsy patients. *Epilepsia*. 2010; 51:592–601. [PubMed: 19780794]
- McBride MC, Binnie CD, Janota I, Polkey CE. Predictive value of intraoperative electrocorticograms in resective epilepsy surgery. *Ann Neurol*. 1991; 30:526–32. [PubMed: 1789682]
- Niedermeyer, E.; Lopes da Silva, FH. *Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. 5 ed. Lippincott Williams and Wilkins; 2005.
- Pressler RM, Robinson RO, Wilson GA, Binnie CD. Treatment of interictal epileptiform discharges can improve behavior in children with behavioral problems and epilepsy. *J Pediatr*. 2005; 146:112–7. [PubMed: 15644834]
- Rakhade SN, Shah AK, Agarwal R, Yao B, Asano E, Loeb JA. Activity-Dependent Gene Expression Correlates with Interictal Spiking in Human Neocortical Epilepsy. *Epilepsia*. 2007; 48:86–95. [PubMed: 17910586]
- Valenti P, Cazamajou E, Scarpettini M, Aizemberg A, Silva W, Kochen S. Automatic detection of interictal spikes using data mining models. *J Neurosci Methods*. 2006; 150:105–10. [PubMed: 16122807]
- Webber WR, Litt B, Lesser RP, Fisher RS, Bankman I. Automatic EEG spike detection: what should the computer imitate? *Electroencephalogr Clin Neurophysiol*. 1993; 87:364–73. [PubMed: 7508368]
- Wilson SB, Emerson R. Spike detection: a review and comparison of algorithms. *Clin Neurophysiol*. 2002; 113:1873–81. [PubMed: 12464324]
- Wilson SB, Harner RN, Duffy FH, Tharp BR, Nuwer MR, Sperling MR. Spike detection. I. Correlation and reliability of human experts. *Electroencephalogr Clin Neurophysiol*. 1996; 98:186–98. [PubMed: 8631278]

Highlights

- Human reviewers show poor agreement in identifying interictal spikes on intracranial EEG.
- An automated detection algorithm was developed which mirrors the way in which human reviewers detect interictal spikes.
- The automated algorithm performed more consistently than any of the human reviewers for each other's marks.

Figure 1. Spike detection using filtering and block scaling balances differences in human reviewers

(A) Flowchart overview of the detection algorithm. (B) The algorithm's main initial screening step involves bandpass filtering the data from 20-50Hz (bottom), which makes interictal spikes (blue arrows) stand out from background compared to more standard viewing filters (top). This method accentuates both large, obvious spikes (left) as well as those that may have otherwise been lost in larger slow wave background activity (right). (C) Scaling all channels together as a block preserves differences in spike detection between high spike frequency and low spike frequency channels. A set of 5 channels is shown from the original ECoG, after marking individual channels independently (channel scaling), and after block scaling. * shows the marked spikes.

Figure 2. Human reviewer and algorithm agreement across a range of patients

Patients chosen for this study displayed a wide range of interictal spike types as well as background activities on intracranial electrocorticography. Spike detections by the three human reviewers and the spike detection algorithm are shown for 4 of the 10 patients representing different patterns. The boxes above the spikes show where at least one reviewer or the automated spike detection algorithm marked a spike, denoted by the presence of a colored square within the box. Agreement was best when interictal spikes were very large and stood out strongly from background (Patients 7 and 10) and poorest when spikes are small and characterized primarily by the subsequent slow wave (Patient 3).

Figure 3. Heatmaps of interictal spike frequency

Heatmaps of interictal spike frequency superimposed on a patient's 3-dimensional brain rendering show that the three human reviewers rank channels similarly, but not identically (top). The spike detection algorithm, however, balances reviewer discrepancies and produces a similar pattern to the average of all three human reviewers (bottom).

Table 1

Total number of spikes detected by human reviewers

	Total number of detections			Percent of spikes marked by:	
	Reviewer 1	Reviewer 2	Reviewer 3	>=2 Reviewers	All 3 Reviewers
Patient 1	1159	7	7161	12.7%	0.0%
Patient 2	1425	371	757	30.9%	4.3%
Patient 3	1410	326	351	26.5%	4.9%
Patient 4	4500	750	6530	17.8%	0.5%
Patient 5	6534	9481	11935	39.2%	0.7%
Patient 6	1112	378	1985	18.1%	1.3%
Patient 7	5178	1508	4990	35.7%	5.9%
Patient 8	177	36	77	16.0%	2.9%
Patient 9	2485	1382	3576	32.3%	7.4%
Patient 10	722	0	2440	8.1%	0.0%
Mean	24702	14239	39802	23.7%	3.1%

Table 2
Sensitivity and precision of human reviewers for the union of the other human reviewers

	Reviewer 1		Reviewer 2		Reviewer 3	
	Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Patient 1	13.1%	81.0%	0.1%	100.0%	80.8%	13.1%
Patient 2	55.1%	39.9%	15.3%	73.6%	26.5%	53.9%
Patient 3	70.1%	29.6%	19.5%	93.9%	13.7%	56.1%
Patient 4	21.2%	32.5%	4.6%	58.7%	33.2%	26.0%
Patient 5	19.4%	49.5%	31.9%	51.8%	48.8%	64.2%
Patient 6	17.0%	33.1%	7.8%	56.6%	35.4%	25.5%
Patient 7	45.7%	49.8%	14.5%	75.8%	44.9%	53.2%
Patient 8	33.7%	19.2%	11.5%	75.0%	12.6%	31.2%
Patient 9	34.8%	61.1%	17.7%	61.5%	45.7%	41.2%
Patient 10	9.7%	32.7%	-	-	32.7%	9.7%
Mean	32.0%	42.8%	13.7%	71.9%	37.4%	37.4%

Reviewer 2 did not mark any spikes in Patient 10. A spike was considered a true positive if either or both of the other two reviewers also marked it.

Table 3

Sensitivity and precision of automated detection algorithms for human reviewers

	Custom algorithm		Commercial algorithm	
	Sensitivity	Precision	Sensitivity	Precision
Patient 1	75.8%	24.0%	49.4%	23.3%
Patient 2	49.8%	41.8%	43.4%	21.4%
Patient 3	17.8%	33.5%	0.0%	0.0%
Patient 4	66.7%	26.6%	35.0%	41.9%
Patient 5	40.7%	52.7%	38.4%	48.0%
Patient 6	29.2%	36.3%	34.6%	38.7%
Patient 7	56.7%	49.4%	40.4%	48.8%
Patient 8	42.7%	22.4%	20.3%	5.8%
Patient 9	71.9%	15.3%	26.2%	11.7%
Patient 10	-	7.3%	-	50.0%
Mean	50.2%	30.9%	32.0%	29.0%

Reviewer 2 did not mark any spikes in Patient 10

Table 4

Spikes marked only by algorithm, but later verified by reviewers (out of 10 per patient)

	Reviewer 1	Reviewer 2	Reviewer 3
Patient 1	7	0	1
Patient 2	9	1	4
Patient 3	10	4	7
Patient 4	6	1	0
Patient 5	8	0	5
Patient 6	6	0	0
Patient 7	6	0	2
Patient 8	4	0	0
Patient 9	0	0	0
Patient 10	9	0	4
Mean	6.5	0.6	2.3

Table 5

Ranking of channels of individual reviewers compared to algorithm

Algorithm Rank	Patient 1			Patient 2			Patient 3			Patient 4			Patient 5			Patient 6			Patient 7			Patient 8			Patient 9			Patient 10		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
1	1	5*	2	2	1	1	1	1	4	23	2	1	34	1	1	1	4	2	1	1	1	1	11**	30	64	54	1	53	***	45
2	2	1	1	1	2	3	5	8	9	58	9	2	4	2	26	10	7	1	2	2	3	2	1	1	40	54	6	32	1	22
3	3	5	3	4	4	2	3	2	8	26	24	3	22	4	6	7	1	7	3	4	5	10	3	30	59	54	9	6	1	1
4	35	5	33	3	3	11	4	3	3	45	17	3	15	10	8	7	2	8	5	3	6	10	5	18	64	52	27	53	1	45
5	18	5	29	5	7	6	29	16	36	62	4	8	50	5	5	12	45	36	4	5	9	10	11	8	49	50	50	1	1	2
6	9	5	4	7	5	4	6	5	21	3	23	11	2	11	2	84	10	14	16	10	3	7	11	8	8	6	5	12	1	4
7	4	1	5	9	8	10	24	11	27	7	9	13	10	20	11	19	6	13	22	9	12	16	11	30	2	1	4	1	1	3
8	7	5	6	9	9	7	24	11	18	4	33	9	47	14	65	4	5	6	32	6	10	19	11	30	1	10	2	39	1	8
9	11	5	9	8	13	4	41	36	36	62	47	6	14	3	4	6	3	17	20	8	6	5	3	3	32	27	35	14	1	5
10	8	5	7	6	6	23	24	11	21	68	12	19	30	15	23	3	10	10	18	13	11	34	11	11	14	11	11	21	1	45

* R2 only marked spikes in 4 channels

** R2 only marked spikes in 10 channels

*** R2 marked no spikes in this recording

Table 6

Statistical comparison of ranked channels by interictal spiking activity

	Kendall's W	Mean Spearman Rank Coefficient			
		Reviewer 1	Reviewer 2	Reviewer 3	Commercial algorithm
Patient 1	0.653	0.560	0.284	0.529	0.850
Patient 2	0.795	0.705	0.708	0.758	0.493
Patient 3	0.811	0.747	0.747	0.680	0.551
Patient 4	0.802	0.652	0.687	0.755	0.949
Patient 5	0.882	0.796	0.822	0.849	0.901
Patient 6	0.767	0.561	0.667	0.706	0.834
Patient 7	0.838	0.712	0.744	0.823	0.915
Patient 8	0.622	0.416	0.448	0.460	0.583
Patient 9	0.858	0.821	0.806	0.737	0.697
Patient 10	0.490	0.478	-	0.478	0.726
Mean	0.752	0.645	0.657	0.678	0.750
					0.574

Reviewer 2 did not mark any spikes in Patient 10