Supplementary Material

I. Marking Study

Abstract

In the main body of this paper we presented an HFO detection and classification algorithm that requires no manual intervention. In this supplement, we compare the performance of the automated method with that of three expert human reviewers on an HFO verification task. Our main qualitative conclusion is that human reviewers are currently not in sufficient agreement about what constitutes an HFO to place high emphasis on ground truth data in detection benchmarking; and we find that the automated approach is statistically indistinguishable from humans in the classification task.

Introduction

Despite the proliferation of tools for automatically detecting seizures and other epileptiform activity, no algorithm yet exists for the fully automated extraction of 100-500 Hz transient high-frequency oscillations (HFOs) from intracranial EEG recordings. Several authors have reported on *semi-automated* approaches (Crépon et al. 2010; Csicsvari et al. 1999a; b; Staba et al. 2002) to HFO detection, which use intensive visual pre- and post-processing in conjunction with machine detection. As Gardner et al. discuss (Gardner et al. 2007), none of these groups presents formal validation data for their automated methods; acceptable detection performance is either implicit or simply asserted. Staba et al. (Staba et al. 2002), for example, state without demonstration that "during development of [their] technique, it was found that it was effective in detecting greater than 84% of putative oscillatory events observable with visual EEG analysis."

"Gold standards" for HFO detection

How concerned should one be, at this point in time, about the scarcity of formal validation data for published machine HFO detection algorithms? Below we argue that, given the current state of the field, it would be misguided to place too strong an emphasis on classical performance metrics for existing or proposed automated detectors. As justification for this opinion, we offer the following three points. First, even in the absence of rigorous direct validation, current methods are undoubtedly *useful*, as evidenced by their widespread acceptance: several research groups (Crépon et al. 2010; Schevon et al. 2009; Staba et al. 2007; Staba et al. 2002; Worrell et al. 2008) have adopted methods similar to those originally presented without formal validation by Csicsvari et al. (Csicsvari et al. 1999a), for instance, and have done so with success – where "success" means simply that detected events turned out to be related to outcome measures of scientific interest.

Second, as the results of Gardner et al. (Gardner et al. 2007) show, the task of having clinicians visually *identify* transient oscillations in iEEG – a requirement for generating ground truth data against which to evaluate automated methods – does not yield the complete set of events that is *verified* by the same reviewers when presented with a superset of their own markings containing those of a machine detector as well. Though the Gardner study involved oscillations in the gamma band, it seems likely that their conclusion that human reviewers tend to make many false negative errors would also apply to 100-500 Hz HFOs. In fact, one might expect the effect to be even larger, both because 100-500 Hz HFOs are less familiar to clinical reviewers and vigilance requirements for marking are even more demanding given the higher recording bandwidths. Findings like those of Gardner et al. call into question the sensibility of using a set of human markings as an absolutely rigid benchmark for automated detectors.

Lastly, insisting on perfect establishment of ground truth tends to raise the distracting existential question "what *is* an HFO?" HFOs are, after all, human constructs, employed because they presumably help us understand or communicate about the workings of the brain. Far more critical than pinning down specifically which waveforms *are* and *are not* HFOs, a priori, is using the concept of an HFO to probe the data for evidence of its validity and practical utility. Taking this empirical tact, we settle for a crude detector – one that has been vetted by clinical opinion for its ability to find at least *some* things resembling what would catch the eyes of human reviewers – and we analyze its imperfect outputs. If we find results of scientific importance, we can use them to refine our understanding of what the critical properties of HFOs are and subsequently to optimize our detector, in the hope that we will extract more, or different, information in the next iteration. This evolutionary view of detector design is a fundamental to the approach we have taken in this work.

What we should be asking of the earliest incarnations of fully automated methods, then, is not whether they meet premature and arbitrary performance specifications for detection, but whether they can approximate the successes of semi-automated methods *without data pre-selection and post-processing by humans*. The latter limit the scientific interpretability of conclusions about HFOs and seizure generation – including our ability to assess the generalizability and clinical utility of those findings – to a far greater degree than the odd percentage point of sensitivity or specificity.

The present study

At the same time we offer this lengthy caveat, we also appreciate that it is helpful for practitioners who wish to evaluate our algorithm to understand it within the context of human performance. In the work

we describe below, we asked three board-certified epileptologists to classify detected HFO candidates and compared their markings with the outputs of our automated classifier.

Methods

Reviewer Labeling

Five thousand HFO candidates (~0.4%) were randomly selected among all those identified in stage 1, across all patients. As a conservative measure, we excluded all events from the two subjects (CT 01 and SZ 05) from whom a subset of data had been used to develop artifact-distinguishing features. The remaining 4,773 randomly selected events, across ten patients (nine epilepsy and one control)[1], were then used in a human labeling experiment.

Three board-certified neurologists independently marked all presented events as either valid (positive) or invalid (negative) HFOs, according to the following criteria for what constitutes a valid HFO: *"Any transient, quasi-periodic voltage variation with predominant frequency between 100 and 500 Hz, lasting on the order of tens of milliseconds, standing prominently apart from the background signal, and having apparently physiologic origin."* The criteria were intentionally somewhat vague to reflect the fact that there is currently no standard operational definition of an HFO.

Events were presented to reviewers via a custom Matlab graphical user interface (GUI), shown in Figure S1. The GUI was comprised of four complementary views of each HFO candidate. The bottom view displayed roughly 1 second (0.5 seconds on either side of the candidate) of 5 Hz–1 kHz[2] single-channel

[1] For CT 01, 1 of 1 data file was used in artifact training; for SZ 05, only 1 of 8 total files was used in training. Thus, the number of subjects from whom events were drawn for the labeling experiment is only one less than the total number of patients, not two.

[2] Display distortion was in practice negligible at this default timescale due to the relatively low signal power above the effective Nyquist frequency of the display. As reviewers were free to zoom in (but not out), this compromise allowed us to faithfully represent the full bandwidth across nearly all available time scales without the need for zoom-adaptive filtering.

iEEG, sampled at 2,713 Hz, with vertical scaling of 21 μV/mm. The event under consideration was delimited by red lines (solid, start; dotted, stop) and the view could be scrolled for 30 seconds on either side of the default display window. The top, left view was of the raw data (near DC—9 kHz, 32,556 Hz sampling rate) corresponding to the detection; the top, middle view was of the bandpassed data corresponding to the detection (100-500 Hz, 2713 Hz sampling rate); and the top, right view was a frequency-domain representation of the middle view. Unlike the bottom view whose vertical scaling was fixed, all top views were auto-scaled to fit their viewing windows. Reviewers were free to edit their markings until they had labeled every event and declared the task complete.

The human labeling task was binary, while the automated algorithm classified detections into one of five groups: four clusters, plus a fifth group ("Cluster 0") comprised of detections that were eliminated in stage 2. In order to compare human and machine performance directly, we took as machine-negative all events in cluster 0 and in cluster 2, whose centroid bore the closest qualitative resemblance to the artifacts we had designed features to identify. All other clusters were taken as machine-positive. This post-hoc labeling decision was made blinded to the human reviewers' markings; and while made manually for the present experiment, we note that it could readily be made automatically in the future if desired – for example by storing the coordinates of the cluster 2 centroid and assigning the negative HFO label to an automatic cluster whose centroid was sufficiently nearby.

Data Analysis

We use the chi-squared test of homogeneity to test whether HFO counts are distributed identically across populations (where "population" is analysis-dependent and clear from context below), and the chi-squared test of independence to test whether marker labels are independent. The chance model we use for markers assigns a positive label to each event with probability $p = N_p/N$, with N_p the total

number of events actually labeled positive by the marker and N the total number of marked events (4,773).

Results

In describing the results below, we use the term "reviewer" to refer specifically to humans and "marker," more generally, as a term that encompasses both humans and the machine algorithm. The terms "detection" and "event" are used synonymously.

Putative prevalence of valid HFOs by marker

Human reviewers were not in agreement about the overall prevalence of HFOs in the data set of candidates presented to them. The percentages of detections marked as positive HFOs by reviewers A, B, C, and the machine classifier (M) were 24.6%, 5.5%, 11.5%, and 13.0%, respectively. We rejected the null hypothesis that the proportion of detections marked as positive was independent of human reviewer ($\chi^2(2, N = 14{,}319) = 763.84$, $p << 0.0001$). Also apparent from these numbers is that the automated method's propensity to mark events as positive is not extreme relative to humans'.

Human reviewer preference by cluster

Reviewers had clear and differing cluster preferences. Figure S2A shows, for each human reviewer, all events falling into clusters 1-4 that were classified as positive HFOs. For reviewer A, the majority of such detections (57.1%) fell into cluster 4. The largest clusters for Reviewer B were 3 and 4, with the former (44.3%) favored over the latter (27.1%). Reviewer C displayed yet a third pattern, splitting a majority fairly evenly between clusters 1 (42.0%) and 4 (41.4%). For all three human reviewers, the smallest percentage was in cluster 2 (6.1%, 13.6%, and 1.9%, for A, B, and C, respectively), the putative artifact class. We reject the null hypothesis that the proportion of detections in each of the four clusters was the same across human reviewers ($\chi^2(6, N = 563) = 97.40$, $p << 0.0001$).

Figure S2B shows, for each human reviewer, all events falling into clusters 1-4 that were classified as negative HFOs. As expected, the putative artifact cluster dominates for all three reviewers: 43.3%, 36.1%, and 39.3% for reviewers A, B, and C, respectively. And as is the case for positive labels, we again reject the null hypothesis that the proportion of detections in each of the four clusters was the same across human reviewers ($\chi^2(6, N = 2197) = 41.47$, $p << 0.0001$). Not shown in figure S2 are events that were marked as "0" by the automated detector – detections that were never classified into clusters 1-4 due to elimination in stage 2. These events are accounted for below, where we give standard performance metrics for the automated classifier against a ground truth set derived from the human reviewers' markings.

Inter-rater agreement

A question of fundamental importance in defining ground truth data is to what degree independent human reviewers agree amongst themselves regarding what constitutes an HFO and what does not. Table S1 gives contingency tables, including the kappa score (Cohen 1960) and percentage agreement, for each of the three human-human marker pairs (top) and each of the three machine-human marker pairs. For all tables, we reject at the 5% significance level the null hypothesis that marker labels were independent, and the kappa values greater than one indicate that these difference were in the direction of agreement in all cases (AB: ($\chi^2(1, N = 4773) = 260.94$, $p << 0.0001$); AC: ($\chi^2(1, N = 4773) = 25.09$, $p << 0.0001$); BC: ($\chi^2(1, N = 4773) = 298.50$, $p << 0.0001$); MA: ($\chi^2(1, N = 4773) = 85.80$, $p << 0.0001$); MB: ($\chi^2(1, N = 4773) = 270.96$, $p << 0.0001$); MC: ($\chi^2(1, N = 4773) = 139.94$, $p << 0.0001$);). The average pairwise percentage agreement among human reviewers was 79%, while that for machine-human pairs was 80%. The average pairwise kappa score among human reviewers was 0.15, while that for machine-human pairs was 0.17. The latter average, however, and the individual kappa scores that comprise it are not straightforward to interpret given the different biases of the reviewers.

We also note that we have aggregated across subjects in computing these inter-rater agreement measures. Given the rarity of positive HFOs, sample sizes were too small to compute reliable statistics on an individual subject basis. But inspecting kappa scores leads us to hypothesize that the degree of inter-marker agreement and the differences between human-human and machine-human pairs may vary with patient. For example, average human-human kappa for SZ 05 (1627 events) was 0.21 while the machine-human value was 0.27; for SZ 07 (1448 events) average performance was near chance for both human-human (-0.07) and machine-human (0.01) pairs; and for SZ 03 (295 events) average human-human kappa was 0.37, while average machine-human kappa was 0.19. It would be instructive to investigate these differences more systematically by conducting another marking experiment in which larger random samples of equal sizes were drawn from each subject.

The main conclusion we reach is that a given human is no more consistent with another human in his markings than he is with the machine.

HFO ambiguity

Ground truth looks very different depending on which of several plausible defining rules is adopted. 33.7%[3] of all detections were marked by at least one human reviewer as positive HFOs, while 39.6%[4] of all detections were marked by at least one marker as positive. 6.0%[5] of events were marked by at least two human reviewers (i.e. majority consensus) as positive HFOs, while 10.3%[6] of all events were labeled positively by at least two markers. Only 2.0%[7] of events were marked by all three viewers (i.e.

[3] Chance, which should be higher = 36.9%.

[4] Chance, which should be higher = 45.1%.

[5] Chance, which should be lower = 4.5%.

[6] Chance, which should be lower = 8.7%.

[7] Chance, which should be lower = 0.16%.

unanimous consensus) as valid HFOs, while 2.5%[8] of all events were marked by at least three markers as positive[9]. The range of these values, which is affected by both the marginal probabilities displayed by each marker and the degree to which they tend to actually agree, gives one view of the general uncertainty among reviewers about what counts as an HFO.

General classifier performance metrics

We formed a ground truth data set by labeling as positive all events marked positively by at least two human reviewers (i.e. majority human vote) and as negative all remaining events. The overall accuracy of the automated classifier against this benchmark was 86.7%. Sensitivity was a moderate 46.8%, reflecting the conservatism of Stage 2, which was designed to retain only events with large *spectral* dissimilarity from the background, a condition not explicitly enforced in the marking instructions for reviewers and to which we anticipated not all would adhere. Specificity was 89.2%, reflecting strong classification performance for negative events. Given the relatively high marginal probability of negative events, however, precision was 21.5%. The F_1-measure, the harmonic mean of precision and sensitivity, was 0.30.

The precision metric reported above for the automated procedure should be viewed in light of the sparseness of positive events and in terms of its improvement on Stage 1 alone. Moving from a data set that is 6% "pure"[10] to one that is 21.5% pure is an improvement of 258%. It is also important to remember that precision, as well as the other performance metrics we report, is highly dependent on our definition of ground truth. If we consider a ground truth data set whose positively labeled examples

[8] Chance, which should be lower = 0.74%.

[9] All reported values were significantly different at the 5% level (chi-squared test) from their chance values, which were computed using the marginal probabilities displayed by each marker. For brevity, we have omitted these results, as they are tangential to the point of the paragraph.

[10] Six percent is the probability that a given event emerging from stage 1 would be declared a positive HFO by at least two human reviewers.

191 are the union of all three human reviewers' positive markings, for example, the precision improves to
192 54.6% (with the F_1-measure improving slightly, indicating that this increase is not completely
193 counterbalanced by a decrease in sensitivity). Also, the precision metric reported above is an
194 aggregated measure with respect to the machine clustering. The precision for each of the four clusters
195 considered individually is different and in some cases higher than this aggregate measure, as we discuss
196 below.

197 Table S2 shows the performance results obtained when we modify our ground truth definition in a
198 manner consistent with the recommendations of Gardner et al. (Gardner et al. 2007). The modified
199 ground truth set considers any event marked by at least two markers, human or automated, to be a
200 positive HFO. The table compares the performance of each marker against this hybrid human-machine
201 ground truth, and also gives the difference between each metric and that expected under a chance
202 model. Chance values, which can be computed exactly, were for convenience generated by simulation
203 in the following way. For each rater, 100 random m x n marking matrices were generated, where m
204 was the total number of marked events (4,773) and n was the total number of markers (4). Random
205 marking matrices were drawn according to actual probability mass function displayed by each reviewer.
206 For each trial, performance metrics were computed using the modified ground truth rule described
207 above, and the 100 values in each performance metric category were averaged to yield a final expected
208 value for each. Values in parentheses in the table are the differences between the observed values and
209 these chance values.

210 *Machine cluster purity*

211 Given a machine-positive HFO cluster (i.e. 1, 3, or 4) the probability that one of its members was also
212 marked positive by human reviewers was dependent on cluster. Table S3 shows these results for two
213 cases, one in which ground truth positive is taken to be the union of all human reviewer positive

markings and one in which ground truth positive is taken to be a majority vote. For completeness, we also include the values computed for cluster 2. For both the majority ground truth ($\chi^2(2, N = 619) = 13.64, p = 0.0011$) and the union ground truth ($\chi^2(2, N = 619) = 33.88, p << 0.0001$), we reject the null hypothesis that the proportion of ground truth positive events occurring in each of the three machine positive clusters is the same.

Discussion

The results of this marking study strongly reinforce the idea that we are in the nascent stages of describing high frequency oscillations within the brain. Human reviewers do not agree on the prevalence of HFOs. Nor, relatedly, do they agree particularly well on what constitutes an HFO when they see one. Other results strongly suggest that, in addition to poor inter-rater agreement, intra-rater reliability is moderate at best (Gardner et al. 2007). Different reviewers demonstrate strong preferences for waveforms with differing characteristics. Nonetheless, the level of agreement does exceed chance – there is a core of commonality worth investigating more thoroughly. But the evidence makes it clear that, currently, "ground truth" HFO data are a false sense of security, and should be regarded as suggestive rather than authoritative.

The automated algorithm we introduce performs similarly to humans at the task of culling positive exemplars from a large set of candidate HFOs. Humans agree no more with each other than they do with the machine. The second and third stages of the automated algorithm, taken together, offer at least a threefold improvement in positive predictive value over the stage 1 detector alone – more if we consider individual clusters, some of which seem to capture waveform features that are more saliently HFO-like to humans than others. The automated approach provides the further advantages of being perpetually consistent in its application of detection criteria and indefatigable in its marking effort.

237

238 The relative uncertainty among humans about what constitutes an HFO gives us confidence in framing
239 our work as exploratory, and in the value of studying the outputs of our algorithm on their own merits.
240 In future work, we will examine the relationship between the clusters our algorithm finds and putative
241 areas of seizure generation.

II. Other supporting material

In addition to the marking study detailed above, we provide several descriptions, figures, and tables that supplement other aspects of the main-body text.

Stage 2, additional detail

The principal components can be found by successively seeking out the spatial directions along which the lengths of the orthogonal projections of the data observations have maximal variance, subject to the constraint that each successive direction is orthogonal to its predecessors. These directions are exactly the eigenvectors of the covariance matrix, $\mathbf{C}$, of the data:

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{b}_n - \overline{\mathbf{b}}\right)\left(\mathbf{b}_n - \overline{\mathbf{b}}\right)^T \quad (1)$$

where $\mathbf{b}_n$ is the $P \times 1$ power spectral density representation of background segment n, described above, and $\overline{\mathbf{b}}$ is the mean of all N background segments associated with a given HFO candidate (for us, $\overline{\mathbf{b}} = 0$)[11]. The new coordinates for each background data segment are computed as:

$$\mathbf{X} = \mathbf{BU} \quad (2)$$

[11] Since the number of background segments is smaller than their dimensionality, P, the data lie in a linear subspace whose maximum dimension is $N - 1$. Therefore, at least $P - N + 1$ eigenvalues (projection variances) must be zero, and this fact is used to increase the efficiency with which the relevant eigendecomposition is performed (Bishop 2006).

where $\mathbf{X}$ is the new N x D data matrix of background-segment representations, $\mathbf{B}$ is the N x P matrix whose i^{th} row is $\mathbf{b}_i^T$, and $\mathbf{U}$ is the P x D matrix whose columns are the unit-normalized eigenvectors of $\mathbf{C}$ corresponding to the D largest eigenvalues (for us, $D = 2$, as mentioned above). The D-dimensional projection for the HFO candidate segment itself is then computed, after removing the mean of the background segments and dividing by their standard deviation, using the same matrix $\mathbf{U}$.

The BIC can be derived starting from the Laplace approximation to the "model evidence" (Bishop 2006) – the probability of the data given a particular model after marginalizing over all possible values of the parameters. Assuming a broad (nearly uniform) Gaussian prior distribution over the parameters and a Hessian matrix of the negative log-likelihood function (evaluated at the optimal parameter vector given the data) that is of full rank, the evidence for the i^{th} model ($\mathcal{M}^i$), denoted by $p\left(\mathbf{X}|\mathcal{M}^i\right)$, can be approximated by:

$$\ln p\left(\mathbf{X}|\mathcal{M}^i\right) \approx \ln p\left(\mathbf{X}|\mathbf{u}_{\mathrm{ML}}^i, \boldsymbol{\Sigma}_{\mathrm{ML}}^i, \pi_{\mathrm{ML}}^i\right) - \frac{1}{2} M^i \ln N \tag{3}$$

(Bishop 2006) where the subscript ML stands for the "maximum likelihood" estimates found via EM, and the constant M^i in the second term on the right is the number of free parameters in the i^{th} model; the latter term penalizes model complexity and hence guards against overfitting. The computation in (3) is the BIC, and we select the model for which it is largest.

The goal in stage 2 is to assign a given HFO candidate to one of two classes, $\mathcal{B}$ (background) or $\mathcal{A}$ (anomaly), while minimizing the misclassification rate. This is theoretically done by assigning $\mathbf{h}$ to $\mathcal{B}$ whenever

$$p(\mathcal{B}|\mathbf{h}) > p(\mathcal{A}|\mathbf{h}) \tag{4}$$

(Duda and Hart 1973), where $\mathbf{h}$ is the 2-D representation of the HFO candidate, discussed above. Applying Bayes's Theorem, this condition can be shown to be equivalent to

$$p(\mathbf{h}|\mathcal{B}) > \frac{p(\mathbf{h}|\mathcal{A})\, p(\mathcal{A})}{p(\mathcal{B})} \tag{5}$$

The GMM describing $\mathcal{B}$ allows estimation of the quantity on the left directly, but there is no such model describing $\mathcal{A}$, and one cannot reasonably be inferred given that there is at most a single observation from $\mathcal{A}$. The prior probabilities of $\mathcal{A}$and $\mathcal{B}$, respectively, are similarly unknown.

To address these issues, a heuristic criterion is employed, based on the squared Mahalanobis distances, Δ_k^2, from the HFO candidate to the center of each GMM component, given by:

$$\Delta_k^2 = (\mathbf{h} - \mathbf{u}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{h} - \mathbf{u}_k) \tag{6}$$

The squared Mahalanobis distances of a random sample drawn from a multivariate normal distribution (computed using the unbiased sample covariance matrix) will be distributed approximately as central chi-squared with D degrees of freedom, where D is the dimensionality of the data (McLachlan 1999). Using the assumption that $p(\mathbf{h}|\mathcal{A})$ is a monotonic decreasing function of $p(\mathbf{h}|\mathcal{B})$, so that the latter is high wherever the former is low, it is estimated that

$$p\left(\mathcal{B}|\Delta_1^2,\ldots,\Delta_K^2\right) \approx \sum_{k=1}^{K} \pi_k \int_{\Delta_k^2}^{\infty} q\left(t\right) \mathrm{d}t \tag{7}$$

where $q(x)$ is the central chi-squared density function with ν degrees of freedom:

$$q(x) = \begin{cases} \dfrac{1}{2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)} x^{\frac{\nu}{2}-1} \exp\left(-\frac{1}{2}x\right) & x > 0 \\ 0 & x < 0 \end{cases} \tag{8}$$

The function $\Gamma(z)$ is the Gamma function $\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t)\mathrm{d}t$.

A similar estimate was used by Roberts (Roberts 2000).

Procedurally, the percentage of the central chi-squared density lying to the right of the calculated Mahalanobis distance from each mixture component is found. A weighted average of these percentages, with weights equal to those of the corresponding mixture components is then computed. If the resultant estimated probability exceeds 5%, the HFO candidate is considered to have been generated by the local background process and it is removed from candidacy. All candidates for which the calculation in (7) – computed with respect to the candidate's unique local background model – falls below the 5% threshold are passed on to the final clustering stage.

330

331 **Figure Legends**

332 **Figure S1.** HFO marking tool. Screen shot of custom GUI used to present detections to clinical reviewers.

333 **Figure S2.** Human reviewer cluster preferences. Each whole pie represents the total number of
334 positively marked HFOs by a human reviewer (A: reviewer A; B: reviewer B; C: reviewer C) that were
335 classified by the machine as belonging to clusters 1, 2, 3, or 4. Pie wedges represent the proportion of
336 such marks falling into each cluster (blue: cluster 1; green: cluster 2; red: cluster 3; cyan: cluster 4).

337 **Figure S3.** Cluster 1 sample events in context. Raw (left) and 100-500 Hz bandpassed (right) voltage
338 traces of the five randomly selected events from cluster 1 that appear in figure 5B. Detected events are
339 demarcated by red lines (solid = start; dotted = stop) and shown within the context of 0.5 seconds of
340 data on either flank. Events are arranged on the vertical axis in the same order as in figure 5B. In the
341 abbreviations above each trace, the first letter indicates whether the recording comes from a macro-
342 (M) or microelectrode (m). RT = right temporal; LPO = left parietooccipital; MR = motor. Note that
343 identical labels do not imply identical electrodes, only that the electrodes have the same lobar location.

344 **Figure S4.** Cluster 2 sample events in context. Raw (left) and 100-500 Hz bandpassed (right) voltage
345 traces of the five randomly selected events from cluster 2 that appear in figure 5B. Detected events are
346 demarcated by red lines (solid = start; dotted = stop) and shown within the context of 0.5 seconds of
347 data on either flank. Events are arranged on the vertical axis in the same order as in figure 5B. In the
348 abbreviations above each trace, the first letter indicates whether the recording comes from a macro-
349 (M) or microelectrode (m). LIF = left inferior frontal; LT = left temporal; AT = anterior temporal. Note
350 that identical labels do not imply identical electrodes, only that the electrodes have the same lobar
351 location.

Figure S5. Cluster 3 sample events in context. Raw (left) and 100-500 Hz bandpassed (right) voltage traces of the five randomly selected events from cluster 3 that appear in figure 5B. Detected events are demarcated by red lines (solid = start; dotted = stop) and shown within the context of 0.5 seconds of data on either flank. Events are arranged on the vertical axis in the same order as in figure 5B. In the abbreviations above each trace, the first letter indicates whether the recording comes from a macro- (M) or microelectrode (m). A = anterior; LAMT = left anterior mesial temporal. Note that identical labels do not imply identical electrodes, only that the electrodes have the same lobar location.

Figure S6. Cluster 4 sample events in context. Raw (left) and 100-500 Hz bandpassed (right) voltage traces of the five randomly selected events from cluster 4 that appear in figure 5B. Detected events are demarcated by red lines (solid = start; dotted = stop) and shown within the context of 0.5 seconds of data on either flank. Events are arranged on the vertical axis in the same order as in figure 5B. In the abbreviations above each trace, the first letter indicates whether the recording comes from a macro- (M) or microelectrode (m). RF = right frontal; LPO = left parietooccipital; RT = right temporal. Note that identical labels do not imply identical electrodes, only that the electrodes have the same lobar location.

References

Bishop CM. *Pattern Recognition and Machine Learning*. New York, NY: Springer Science+Business Media, LLC, 2006, p. 738.

Cohen J. A Coefficient of Agreement for Nominal Scales. *Education and Psychological Measurement* 20: 37-46, 1960.

Crépon B, Navarro V, Hasboun D, Clemenceau S, Martinerie J, Baulac M, Adam C, and Le Van Quyen M. Mapping Interictal Oscillations Greater Than 200 Hz Recorded With Intracranial Macroelectrodes In Human Epilepsy. *Brain* 133: 33-45, 2010.

Csicsvari J, Hirase H, Czurko A, Mamiya A, and Buzsaki G. Fast Network Oscillations in the Hippocampal CA1 Region of the Freely Behaving Rat. *The Journal of Neuroscience* 19: 1999a.

Csicsvari J, Hirase H, Czurko A, Mamiya A, and Buzsaki G. Oscillatory Coupling of Hippocampal Pyramidal Cells and Interneurons in the Behaving Rat. *The Journal of Neuroscience* 19: 274-287, 1999b.

Duda RO, and Hart PE. *Pattern Classification and Scene Analysis*. New York: John Wiley and Sons, 1973.

Gardner AB, Worrell GA, Marsh E, Dlugos D, and Litt B. Human and Automated Detection of High-Frequency Oscillations in Clinical Intracranial EEG Recordings. *Clinical Neurophysiology* 118: 1134-1143, 2007.

McLachlan GJ. Mahalanobis Distance. *Resonance* 4: 20-26, 1999.

Roberts SJ. Extreme Value Statistics for Novelty Detection in Biomedical Signal Processing. *IEEE Proceedings Science, Technology, & Measurement* 47: 363-367, 2000.

Schevon CA, Trevelyan AJ, Schroeder CE, Goodman RR, McKhann Jr. G, and Emerson RE. Spatial characterization of Interictal High Frequency Oscillations in Epileptic Neocortex. *Brain* 132: 3047-3059, 2009.

Staba RJ, Fighetto L, Behnke EJ, Mathern GW, Fields T, Bragin A, Ogren J, Fried I, Wilson CL, and Engel J, Jr. Increased Fast Ripple to Ripple Ratios Correlate with Reduced Hippocampal Volumes and Neuron Loss in Temporal Lobe Epilepsy Patients. *Epilepsia* 48: 2130-2138, 2007.

Staba RJ, Wilson CL, Bragin A, Fried I, and Engel J, Jr. Quantitative Analysis of High-Frequency Oscillations (80–500 Hz) Recorded in Human Epileptic Hippocampus and Entorhinal Cortex. *Journal of Neurophysiology* 88: 1743-1752, 2002.

Worrell GA, Gardner AB, Stead SM, Hu S, Goerss S, Cascino GJ, Meyer FB, Marsh R, and Litt B. High-Frequency Oscillations in Human Temporal Lobe: Simultaneous Microwire and Clinical Macroelectrode Recording. *Brain* 131: 928-937, 2008.