A Stochastic Framework for Evaluating Seizure Prediction Algorithms Using Hidden Markov Models

Stephen Wong, Andrew B. Gardner, Abba M. Krieger and Brian Litt

J Neurophysiol 97:2525-2532, 2007. First published 4 October 2006; doi:10.1152/jn.00190.2006

Journal of Neurophysiology publishes original articles on the function of the nervous system. It is published 12 times a year (monthly) by the American Physiological Society, 9650 Rockville Pike, Bethesda MD 20814-3991. ISSN: 0022-3077, ESSN: 1522-1598. Visit our website at http://www.the-aps.org/.

J Neurophysiol 97: 2525–2532, 2007.
First published October 4, 2006; doi:10.1152/jn.00190.2006.

A Stochastic Framework for Evaluating Seizure Prediction Algorithms Using Hidden Markov Models

Stephen Wong,[1,2] Andrew B. Gardner,[1,2] Abba M. Krieger,[3] and Brian Litt[1,2]

[1]*Department of Neurology, Hospital of the University of Pennsylvania;* [2]*Department of Bioengineering, School of Engineering and Applied Science; and* [3]*Department of Statistics, Wharton School of Business, University of Pennsylvania, Philadelphia, Pennsylvania*

Submitted 21 February 2006; accepted in final form 30 September 2006

Wong S, Gardner AB, Krieger AM, Litt B. A stochastic framework for evaluating seizure prediction algorithms using hidden Markov models. *J Neurophysiol* 97: 2525–2532, 2007. First published October 4, 2006; doi:10.1152/jn.00190.2006. Responsive, implantable stimulation devices to treat epilepsy are now in clinical trials. New evidence suggests that these devices may be more effective when they deliver therapy before seizure onset. Despite years of effort, prospective seizure prediction, which could improve device performance, remains elusive. In large part, this is explained by lack of agreement on a statistical framework for modeling seizure generation and a method for validating algorithm performance. We present a novel stochastic framework based on a three-state hidden Markov model (HMM) (representing interictal, preictal, and seizure states) with the feature that periods of increased seizure probability can transition back to the interictal state. This notion reflects clinical experience and may enhance interpretation of published seizure prediction studies. Our model accommodates clipped EEG segments and formalizes intuitive notions regarding statistical validation. We derive equations for type I and type II errors as a function of the number of seizures, duration of interictal data, and prediction horizon length and we demonstrate the model's utility with a novel seizure detection algorithm that appeared to predicted seizure onset. We propose this framework as a vital tool for designing and validating prediction algorithms and for facilitating collaborative research in this area.

INTRODUCTION

Several years ago, implantable, responsive brain-stimulation devices to treat epilepsy entered human clinical trials (Kossoff et al. 2004). Preliminary safety and efficacy results are encouraging, but given responder rates (defined as a >50% reduction in seizures) between 35 and 43% (Worrell et al. 2005), there is still room for improvement. There is evidence that focal neurostimulation to abort seizures may be more effective when delivered early in seizure generation (Murro et al. 2003). This potential to improve device efficacy has increased interest in seizure prediction, now defined as reliably identifying periods of time in which there is increased probability of seizure onset (Litt and Echauz 2002; Litt and Lehnertz 2002). Participants in the First International Collaborative Workshop on Seizure Prediction compared algorithm performance from different laboratories on a shared set of continuous, intracranial electroencephalogram (IEEG) recordings from five major international epilepsy centers (Lehnertz and Litt 2005). This collaborative experiment was stimulated by concern that analyzing incomplete, "clipped" data sets might bias experimental results. In the end, no method convincingly demonstrated prospective seizure prediction accuracy sufficient for clinical application. A shortcoming that was identified was a lack of specificity of preictal (preseizure) changes on the IEEG. This difficulty rendered many methods, which were previously reported to herald seizure onset, much less useful. Additional challenges identified in the published consensus statement were the need to design experiments with increased statistical rigor and the need for agreement on statistical methods to perform this validation. It is in response to these issues that we present the following study.

Many early publications on seizure prediction are controversial because of statistical bias inherent in their study designs. Sources of selection bias in these studies include the use of clipped EEG segments containing only seizures and the immediate preseizure period, the exclusion of sleep EEG epochs, and few (or none) randomly chosen "baseline" segments (Le Van Quyen et al. 2001; Navarro et al. 2002). Without representative interictal periods, a seizure prediction algorithm may be operating along the low-specificity, high-sensitivity portion of its receiver–operator characteristic (ROC) curve, reducing the significance of its sensitivity measure. In addition, post hoc bias resulted from mining these limited EEG segments without a prospective validation data set to confirm the findings (Le Van Quyen et al. 2001; Martinerie et al. 1998; Navarro et al. 2002). More recent publications analyzed larger portions of the interictal EEG and provided false-positive rates (FPRs). However, the FPR measure alone, without a specified average false-positive duration, is still inadequate for statistical validation. For example, when comparing two algorithms with the same FPR, the algorithm with a longer average false-positive duration will cover a larger proportion of the EEG with positive predictions, resulting in an artificial bias toward higher sensitivity.

Underemphasis on evaluation of interictal periods may also indicate an implied assumption on the part of the investigators: that the putative preictal state occurs only before seizures and reflects a deterministic, transitional state that inevitably leads to seizure (Andrzejak et al. 2001; Martinerie et al. 1998). However, clinical observation is more compatible with the notion of a putative preictal state as a "permissive" state for seizure generation (e.g., increased seizure probability with fever, toxic-metabolic derangements, or natural cycles such as

Address for reprint requests and other correspondence: S. Wong, Department of Neurology, 2 Ravdin Penn Epilepsy Center, Hospital of the University of Pennsylvania, 3400 Spruce St., Philadelphia, PA 19104 (E-mail: swong@swong.org).

Contact the authors for a copy of the deidentified data set for research purposes (~250 MB).The costs of publication of this article were defrayed in part by the payment of page charges. The article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

sleep or menstruation). A permissive state allows periods of increased probability of seizure onset to evolve back into the interictal state without producing a seizure. In this paradigm, false-positive detections can be a reflection of underlying stochastic physiological mechanisms and not simply a shortcoming of imperfect detection tools that label distinct events as similar.

A *hidden Markov model* is a mathematical tool that can be used to model a process that assumes that a series of observations are emitted from underlying "hidden" states that transition between each other in a stochastic fashion. The observations are also emitted probabilistically, conditional on the hidden state. Markovian dynamics have been successfully used to model the processes of seizure generation in the past, in experimental animals, and to detect antiepileptic drug compliance (Albert 1991; Hopkins et al. 1985; Le et al. 1992; Sunderam et al. 2001). The "hidden" feature of HMMs is attractive, in that we assume that observable EEG signals arise from underlying dynamical brain states. The analogy in the present case is that EEG signals processed with automated seizure prediction algorithms result in noisy observation sequences that reflect an underlying preictal state. Furthermore, because of the stochastic nature of the state transitions, HMMs have the flexibility to incorporate the notion of a "permissive" preictal period.

Using a discrete three-state HMM (baseline, detected, and seizure), we created a statistical framework to evaluate seizure prediction algorithms. We derived equations for type I (false-positive) and type II (false-negative) errors that defined validation thresholds, enabling us to test the power of predictive algorithms against a null hypothesis. Intuitive notions, such as the difficulty of statistical validation with small sample sizes or a large number of positive predictions, were reproduced. We also illustrate the method's utility in analyzing a seizure detection algorithm produced in our laboratory that appeared to have seizure predictive ability.

METHODS

All computation was performed with Microsoft Windows-based PCs equipped with dual-core Intel Xeon processors, running Matlab R14. Default Matlab programs, programs within the prepackaged statistical toolbox, and custom-designed programs were used for the analysis.

Data collection and IEEG classification

Data consisted of the binary output of a seizure detection algorithm that classified underlying IEEG into baseline or detected states (Gardner et al. 2006; see *An example validation* in RESULTS). The details of IEEG data collection are described in Gardner et al. (2006). All patient data were acquired with informed consent, deidentified, and processed under a protocol approved by the University of Pennsylvania Institutional Review Board (IRB), in accordance with University human research policy. The channel of IEEG data selected for use in each patient was the one that most often exhibited the earliest seizure onset, as determined by standard clinical methods. Two board-certified (American Board of Clinical Neurophysiology) electroencephalographers marked the period between seizure onset and termination as a third state. Unequivocal electrographic onsets (UEOs) described in Risinger et al. (1989) were used to mark the beginning of a seizure. The seizure termination point was defined as the point when clonic activity ceased or spread to >5 s between bursts.

Hidden Markov model creation and training

We assumed that the data to be validated come in the trinary form (binary detector outputs plus gold-standard human seizure markings) as outlined above. We trained a three-state HMM, with states 1, 2, and 3 denoting the baseline, detected, and seizure states, respectively. HMMs are described by their transition probability matrix (**A**) and symbol emission probability matrix (**B**). In this framework, both **A** and **B** are 3×3 for a three-state model with three corresponding observation symbols. Initial values for these probabilities were refined by training on the data with the Baum–Welch algorithm (Rabiner 1989), an implementation of the more general expectation-maximization (EM) algorithm, that converges on local probability peaks within HMMs (Dempster et al. 1977). The following estimations and constraints were imposed during training

$$a_{ii} = 1 - (1/D_i) \quad (1)$$

$$a_{13} < a_{23} \quad (2)$$

$$b_{11} > b_{12} \quad b_{21} < b_{22} \quad (3)$$

$$b_{33} = 1 \quad b_{13} = b_{23} = b_{31} = b_{32} = 0 \quad (4)$$

where D_i is the average duration of state i (in prediction interval time units), a_{ii} is the same-state transition probability, a_{ij} is the transition probability from the ith to the jth hidden state, b_{ij} is the emission probability of the ith observation variable while in the jth hidden state, and b_{ii} is the emission probability of the ith observation variable while in the ith hidden state.

Equation 1 is derived in Rabiner (1989); published average state durations (e.g., the length of a typical seizure) can be used to estimate these quantities. *Equation 2* indicates that the detected state has a tendency to precede seizures. *Equation 3* arises from the assumption that the prediction algorithm makes more correct than incorrect classifications of the underlying state. *Equation 4* arises from the fact that the "gold-standard" electroencephalographer's markings are noiseless. In practice, because of the Baum–Welch algorithm's convergence on local optima, uniform or random initial parameters are used and the models produced are tested with the forward algorithm to find the global optimum (Rabiner 1989). *Equations 2* and *3* reduce the number of mathematically symmetrical models produced by training.

Finally, let $\mathbf{S}^{\infty}$ denote the 3×1 vector of values s_1, s_2, and s_3 (where $s_1 + s_2 + s_3 = 1$), which represent the proportion of time spent in each state, at steady state. These values can be found by assuming steady state, which means that the probability of being in a particular state is invariant over time (e.g., by setting $s_1 = s_1a_{11} + s_2a_{21} + s_3a_{31}$), and solving the resultant system of linear equations.

Null hypothesis testing

Traditional statistical validation of an HMM involves calculating the ratio of the probabilities that the data originated from the trained HMM versus a null HMM. Null models can be created by mathematical manipulation of the trained HMM matrices; e.g., a null model in which the detector does not detect anything is created by setting $b_{12} = b_{11}$ and $b_{21} = b_{22}$. The null model of interest here is one in which detections bear no relationship to seizures. Unfortunately, this null HMM could not be created without introducing other unwanted differences in other parameters that would inevitably count as differences in log-odds scoring.

Instead, we chose to use the Viterbi algorithm (Rabiner 1989; Viterbi 1967) to recover the single most likely state sequence from a sequence of observations associated with the trained HMM. By using the state sequence rather than the observation sequence, we can determine whether there is a larger-than-expected number of transitions from the detected state into the seizure state. Under the assumption of stationarity, the expected proportion of transitions from the

detected into the seizure state is: $s_2/(s_1 + s_2)$. If n states are used, and the mth state is thought to be the preictal state, this expected proportion would be $s_m/(s_1 + s_2 + \ldots + s_n)$, where $m < n$.

Type I and type II error derivations

Based on the above, we derived the following equations for significance testing (see APPENDIX for derivations):

Type I error

$$Error \geq \sum_{i=x}^{N} \binom{N}{i} \left(\frac{S_2}{S_1 + S_2}\right)^i \left(\frac{S_1}{S_1 + S_2}\right)^{N-1} \quad (5)$$

where N is the total number of 2→3 and 1→3 transitions and x is the minimum number of 2→3 transitions that satisfies the equation and causes us to reject H_0 in favor of H_1; the left-hand side of the equation is typically set to the $\alpha = 0.05$ significance level.

Type II error

$$Error = \sum_{i=0}^{x-1} \binom{N}{i} y^i (1 - y)^{N-i} \quad (6)$$

where x is the value found from *Eq. 5* for a significance level of $\alpha = 0.05$, N is the total number of seizures, and y is the minimum number of 2→3 transitions divided by N, which satisfies the equation and causes us to erroneously reject H_1 in favor of H_0; the left hand-side of the equation is typically set to the $\beta = 0.2$ type II level.

Figure 1 shows a schematic of the entire validation process.

RESULTS

Graphing parameter space: type I error

The surfaces corresponding to the limits of statistical validation under standard values of type I errors ($\alpha = 0.05$) are found by exploring parameter space. The results are shown in Fig. 2*A*. The region of a validateable parameter space is the volume above the surface.

Two qualitative findings that formalize intuitive notions regarding statistical validation can be inferred from the graphs. In particular, validation becomes difficult when the ratio of detected versus baseline state increases. This reinforces the intuitive notion that it is the false-positive proportion of EEG—not the false-positive rate—that is important for validation. The other regime where validation becomes difficult is when the number of seizures is very small. This is again intuitive: given a low sample size, statistical validation becomes difficult.

The sensitivity threshold for the $\alpha = 0.05$ significance level varies considerably with the proportion of detected state. This reflects the potential for selection bias to severely affect the sensitivity measure were one to use samples in which positive detections were more or less prevalent. As mentioned in the INTRODUCTION, early publications on seizure prediction claimed high sensitivity when comparing algorithm output between segments of data "distant" in time from seizures to a far lower proportion than data segments immediately before seizure onset. Later studies using longer segments of continuous IEEG data invalidated many of these early claims (Aschenbrenner-Scheibe et al. 2003; Mormann et al. 2005). This highlights the need to use representative proportions of interictal and preictal data, in addition to other benchmarks of performance, such as false-positive rates, false-positive durations, and P values calculated from quantitative comparisons with null models like the one provided here.

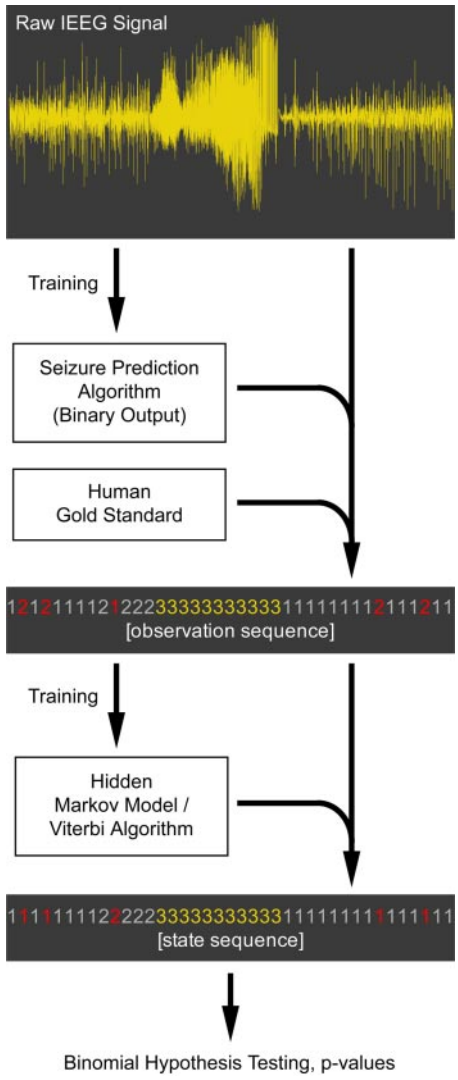

FIG. 1. Diagram of the statistical validation process. Electroencephalogram (EEG) is used to train a seizure prediction algorithm. This algorithm then converts the EEG to a binary sequence (baseline and detected). Human electroencephalographer markings of seizures are then further used to create a trinary observation sequence (baseline [1], detected [2], and seizure [3]). This sequence is used to train an hidden Markov model (HMM), which is in turn used to Viterbi-decode the original observation sequence into the hidden state sequence. Illustrative noise observations are indicated in the sequences in red. Transitions into the seizure state are then counted and used in hypothesis testing to determine whether a statistical association exists between the detected and seizure states.

Crafting a clinical study: type II error

Once the type I error and sensitivity are known, power calculations for clinical trial enrollment are done in the usual manner. Figure 2*B* shows the minimum type II error for the value of $\beta = 0.2$, based on the minimal sensitivities required to meet the $\alpha = 0.05$ significance level as seen in Fig. 2*A*.

An example validation

We tested this framework with a seizure detection algorithm produced in our laboratory based on a support vector machine

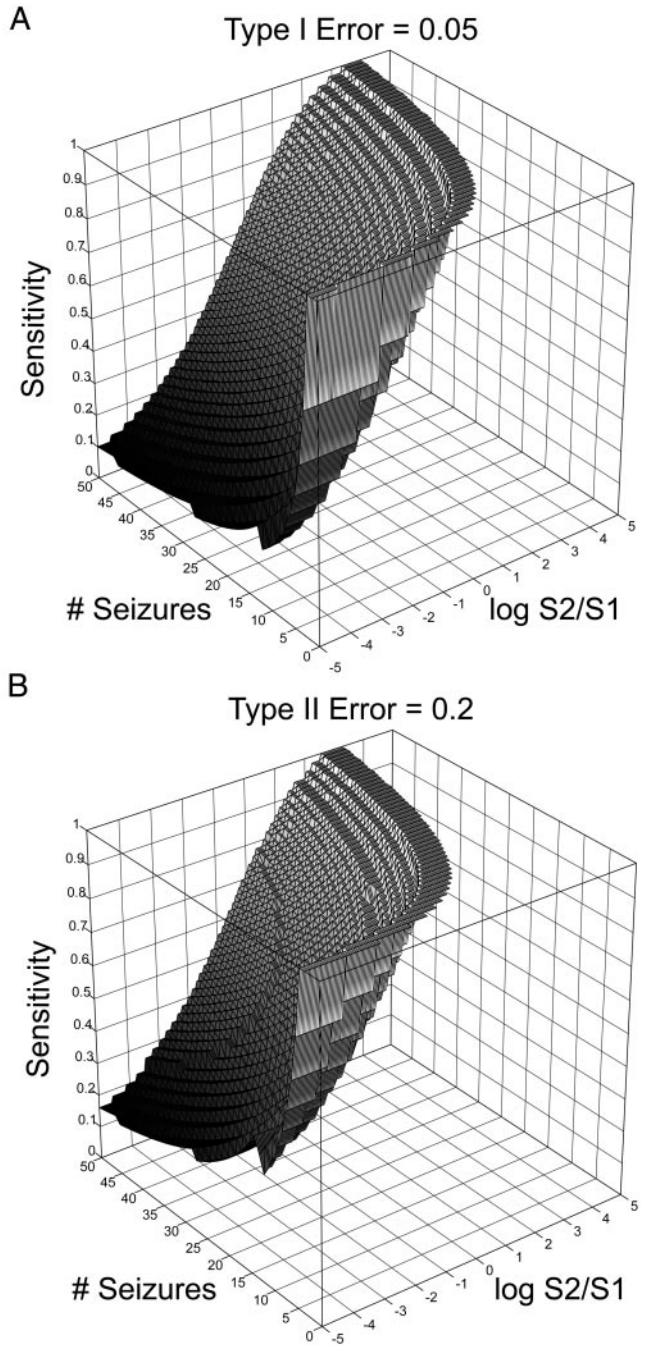

FIG. 2. *A*: minimum sensitivity required for statistical validation (z-axis) vs. the number of seizures in the data set (y-axis) vs. log base 2 of detected state/baseline state proportions (x-axis), for type I error = 0.05. For any given number of seizures and detected to baseline state proportion, a minimum sensitivity required for statistical validation at the 0.05 error level lies above the surface. Statistical validation is difficult in regions where the detected state constitutes a large proportion of the EEG, and in regions where the number of seizures is too low. (Note that the nonsmooth boundary results from the binomial equation, a discrete function.) *B*: minimum sensitivity required for statistical validation (z-axis) vs. the number of seizures in the data set (y-axis) vs. log base 2 of detected state/baseline state proportions (x-axis), for type II error = 0.2, assuming the minimal sensitivity calculated in *A* for a type I error of 0.5.

(SVM) (Gardner et al. 2006), which has promising potential for implementation in second-generation responsive stimulation devices for epilepsy. SVMs are statistical machine-learning algorithms commonly used as data classifiers. In this application, the SVM was crafted to detect local outliers (1-s IEEG frames with a 0.5-s advance) in an array of energy-based features derived from the IEEG. The system was retrained every 15 minutes. In addition to being an accurate and clinically useful seizure detector, the algorithm was able to detect seizures on average 7 s before human unequivocal electrographic seizure onset (UEO) markings. Because of this predictive performance, we chose to analyze it in this framework.

The IEEG data analyzed by the SVM-based detector consisted of nearly continuous recordings of IEEG activity from the electrographic seizure focus in five patients with mesial temporal sclerosis, with a combined total of 29 seizures in 515 h of recording. The raw detector output consisted of on-line, binary classifications as to whether each 0.5-s of IEEG was in the baseline or detected state. We used our HMM framework to analyze this raw detector output. A seizure was considered "predicted" if it arose during the detected state after HMM–Viterbi processing. The numbers of predicted versus unpredicted seizures were then substituted into our derived equations for significance testing. Note that in the original work by Gardner et al. (2006), the raw SVM outputs were further filtered by the following rule: if 16 or more of the prior 20 outputs were positive, then the detector would declare that frame, as well as the frames in the next 3 min, to be in the detected state; otherwise, that frame's output would be negative.

Using random starting points, the trained HMM with the highest probability of producing the raw SVM observation sequence (as calculated by the forward algorithm) is shown in Table 1. This HMM was found 78% of the time after training (39 of 50 trials), when random initial emission matrix probabilities were estimated according to the guidelines outlined in the METHODS section. This is consistent with prior observations that accurate estimation of the emission, rather than transition, probabilities plays a larger role in determining the final model (Rabiner 1989).

The trained HMM had mathematical properties consistent with an algorithm that was able to predict seizures. Specifically, there was more than a 10-fold greater probability for the detected state over the baseline state to give rise to seizures (10^{-4} vs. 8.08×10^{-6}). The a_{32} transition probability indicated that seizures were three times as likely to return to the

TABLE 1. *Most common trained HMM from raw detector output training sequences*

$$\mathbf{A} = \begin{bmatrix} 0.9594 & 0.0406 & 8.08 \times 10^{-6} \\ 0.2489 & 0.7511 & 0.0001 \\ 0.0010 & 0.0035 & 0.9955 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 0.9540 & 0.0460 & 0.0 \\ 0.0627 & 0.9373 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix}$$

$$\mathbf{S}^{\infty} = \begin{bmatrix} 0.8567 \\ 0.1397 \\ 0.0036 \end{bmatrix}$$

This HMM was found 78% of the time (39 of 50 random starts) with the constraints outlined in METHODS. This HMM was also the one associated with the highest probability of producing the training sequences. **A,** transition probability matrix; **B,** symbol emission probability matrix; $\mathbf{S}^{\infty}$, 3×1 vector of s_1, s_2, and s_3.

TABLE 2. *Results of significance testing after determining the state sequence with Viterbi analysis*

Output	Total Seizures, (*N*)	x*	y**	Predicted Seizures	S1/S2 Ratio	*P*-Value
HMM–Viterbi	29	8	0.3329	17	6.1344	2.9525×10^{-8}
Algorithm implementation	29	7	0.2963	5	8.4019	0.1898

Values represent the seizure prediction performance of the raw SVM-based seizure detector outputs *1*) filtered by HMM–Viterbi decoding vs. *2*) the particular implementation of the detector in Gardner et al. (2006). The expected S1/S2 ratio of baseline to detected state for HMM–Viterbi output was calculated by the trained HMM's transition matrix. The same ratio in the author-implemented case was determined empirically from the data by summation of baseline vs. detected outputs. *Minimal number of detected seizures for a type I error of 0.05; ** minimum proportion of detected seizures for a type II error of 0.2.

detected state rather than the baseline state (3.5×10^{-3} vs. 10^{-3}), reflecting possible seizure clustering, a common clinical phenomenon (Haut 2006), by remaining in a high-probability state. Interestingly, in accord with the hypothesis of a permissive preictal state, the a_{21} transition probability indicated that a detected state was far more likely to transition back into the baseline state than into the seizure state (0.2489 vs. 10^{-3}). Last, given the 1-s frame duration with 0.5-s advances, the value of the same-state transition probability for the detected state (0.7511) indicated that the SVM-detected state was of brief duration, on the order of seconds. This implied that the detected state was not the same preictal state of >10-min duration found by other techniques (Federico et al. 2005; Litt et al. 2001; Weinand et al. 1997). Instead, it suggested that it was either a distinct, brief preictal state seen by this particular technique or, more likely, that the SVM was able to detect quantitative changes in the IEEG associated with seizure onset in advance of human UEO markings.

Table 2 shows the results of significance testing after determining the state sequence with Viterbi analysis. The SVM–HMM–Viterbi output was able to detect seizures within seconds before their onset in a statistically significant manner, with 17 of the 29 seizures arising from the detected state, corresponding to a *P* value of 3×10^{-8}. Eight early detections constituted the threshold required for statistical significance ($\alpha = 0.041579$). For comparison, the performance of the SVM detector using the filtered output rule as originally implemented by the author is also shown.

The SVM detector using the filtered output rule was not able to forecast seizures with statistical significance for two reasons: *1*) the filtering rules substantially reduced the false-positive rate at the expense of inducing a delay in response time, causing most detection points to occur after the UEO, and *2*) the rules also used a refractory period of 3 min after detector firing, which increased the proportion of detected state. Although the refractory period increased the sensitivity for early detections (Gardner et al. 2006), the larger increase in false positives resulted in overall poorer statistical performance compared with the HMM–Viterbi-decoded output. Figure 3 shows both the raw and the filtered SVM outputs versus the HMM–Viterbi output for two typical seizure onsets. Although statistical significance was obtained, the brevity of the detected state highlights the fact that the algorithm was most likely detecting seizure onsets (as designed) earlier than the human-marked UEOs, rather than a preictal state.

It is important to note that retrospective bias exists in both the human gold standard as well as the HMM–Viterbi output itself. This is because expert readers have access to the entire EEG record when marking where the UEO occurs and the Baum–Welch training algorithm entails recurrent sweeps across the entire data set for parameter optimization. With this in mind, the above findings might represent the differences in change point detection between on-line/prospective and off-line/retrospective methods.

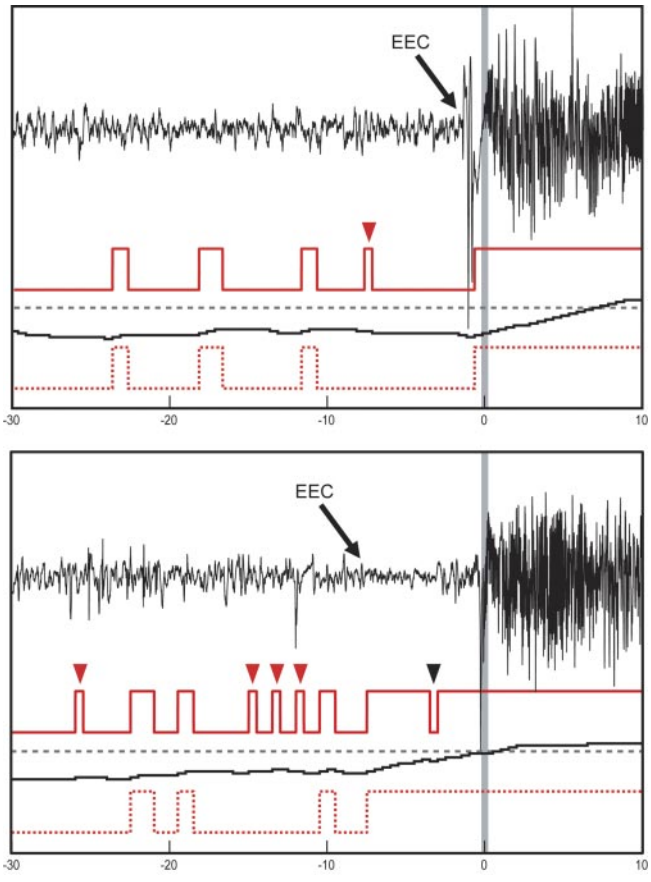

FIG. 3. Two typical seizure onsets, from different patients. *Topmost red traces*: raw binary support vector machine (SVM)–based seizure detection output. *Middle black traces*: running probability estimates determined by the particular implementation of the SVM algorithm in Gardner et al. (2006), with a dashed alarm threshold in gray. *Bottom red dotted traces*: HMM–Viterbi decoding of the raw binary SVM output into "state sequences." *x*-axis is time in seconds, with the unequivocal electrographic onset (UEO) indicated by a gray vertical bar set to *time = 0*. In both examples the onset of the detected state by the HMM–Viterbi decoded method occurs seconds before the UEO, apparently within the earliest electrographic change (EEC) to UEO period, consistent with the algorithm's original construction as an early seizure detector. The HMM–Viterbi decoded output transitioned to seizure before UEO in these 2 cases, but contained more false alarms, whereas the algorithm as implemented in Gardner et al. (2006) alarmed after UEO, but had many fewer false alarms. Raw detector's false-positive classifications, as determined by HMM–Viterbi analysis, are indicated by red arrowheads, whereas a false negative is indicated by a black arrowhead.

DISCUSSION

The HMM-based statistical framework outlined here is novel compared with existing statistical frameworks for seizure prediction (Andrzejak et al. 2003; Kreuz et al. 2004; Winterhalder et al. 2003) in that it is the first to incorporate an underlying stochastic model for seizure generation. The full connectedness of HMMs readily allows for modeling of a "permissive" preictal state that is able to return to the baseline state without giving rise to a seizure, in agreement with clinical observation, and suggests that events labeled as "false-positive" in EEG analyses may actually reflect physiologic dynamics intrinsic to the seizure generation process. Deterministic preictal states, if they exist, are also easily modeled because training may result in near-zero transition probabilities from the baseline to the seizure state.

Derivation of type I error within this framework allows performance benchmarking for comparison between algorithms validated on a standardized data set. Furthermore, derivations of type II error can also be used to calculate minimal enrollment and data collection requirements such that a prospective clinical trial, involving a specific seizure prediction algorithm, will be able to demonstrate efficacy.

The visualization of statistically valid parameter space derived from this framework highlights the difficulty of validation with small sample sizes or with disproportionate amounts of detected states versus interictal state data. Both of these intuitive notions are expected from any validation model and are reassuringly seen here. This statistical framework allows for evaluation even of clipped EEGs containing larger proportions of preictal EEG; the minimal required sensitivity simply rises accordingly, as seen on the graphs. However, accurate description of real-world performance requires the use of continuous EEG data because selection bias with clipped segments can alter sensitivity and specificity as shown in Fig. 2. This emphasizes the need to develop seizure prediction algorithms on long-term, continuous data, to ensure that published sensitivities and FPRs provide good estimates of real-world performance.

This HMM framework has the unique advantage that Viterbi-decoding of raw outputs provides a method for nonarbitrary autosegmentation of noisy observations into variable-length detection periods, which are consistent with a stochastic model of seizure generation. They are also compatible with the observation that the detection period length varies with the particular detector used, as well as other inconsistencies such as the electrode position relative to the signal generator. Nonarbitrary autosegmentation also provides the ability to aid in the interpretation of what might be occurring physiologically with a higher degree of confidence than with particular detector implementations, although in practice this is a complex procedure when one is ignorant of any meaning of the detected brain states. A straightforward approach would be to scrutinize the periods of time corresponding to the autosegmented states for electrophysiological or behavioral correlates. In addition, autosegmentation provides guidance for creation of rules for clustering of raw observations that can be tuned toward desired performance characteristics for prospective studies with the algorithm.

Caveats of this statistical framework

Validation by itself does not necessarily guarantee that a particular algorithm can detect and characterize a preictal state, or that a preictal state even exists. For example, this framework validates the SVM-based detection algorithm, which was not originally designed for seizure prediction. Although one interpretation is that a preictal brain state has been detected, on closer inspection, the detected state was brief in duration and at times began within the period between earliest electrographic change (EEC) and unequivocal electrographic onset (UEO) time periods (which is thought to be ictal). It is well known that a fair amount of subjectivity is involved in "gold-standard" human markings of seizure onsets, with a small degree of "jitter" surrounding the true event onset time. Because the EEC to UEO period is typically brief, the "predictive" findings seen here may be entirely explained by marking inaccuracies. In addition, the SVM algorithm functions by detecting sudden energetic departures from a previous 15 minute baseline. Given that the SVM-based detector treated any direction of change in the distribution its three energy-related features as a positive detection, it seemed unclear that it was identifying a consistent preictal state. These observations imply that the SVM was functioning as an early seizure detector (as it was originally designed to do), and not as a seizure predictor, despite its ability to identify seizure onsets before two expert human readers. This highlights the somewhat semantic distinction between seizure detection versus prediction and its fragile dependency on our knowledge (or ignorance) of the underlying neurophysiological process.

HMMs are able to flexibly accommodate additional states. Three is the minimum number required to model a permissive state that gives rise to seizures. This is the number we suggest using because we assume that human expert markings (seizures) and prediction algorithm outputs (baseline or detected) combine to produce a total of three states. The question is, how does one determine how many states are best? An increased number of states would be interesting if the additional states matched known physiological brain states, although this is not guaranteed. What we observed when fitting additional states (data not shown) was that additional states either split the baseline and detected states into similar states with modest difference or resulted in states that could be interpreted as detector noise with no known physiological consequence. Given our limited knowledge about the underlying brain states and the enormous data reduction imposed by limited sampling of the brain's activity through intracranial EEG, this is a precarious undertaking and is unnecessary for the general statistical validation of a binary detector.

Another limitation of this work is that the relationship between the detections and the period when seizures are more likely to occur must be known for evaluation of a seizure prediction algorithm in this framework. For example, if an algorithm detects the "early preictal" state, and the period from which seizures immediately arise is indistinguishable from baseline, validation may fail despite possible substantial predictive power. This limitation is not unique to this framework, but is present in all published validation schemes; some schemes do not directly address this issue (Kreuz et al. 2004) and others use predetermined preictal horizon durations after detection (Winterhalder et al. 2003). It is possible to modify

our framework to accommodate these detection offsets, but the burden of the modifications falls on the algorithm designer. A screen for detection offsets can be performed by examining the temporal distribution of detections with respect to aligned seizure onset. A detection peak that coincides with seizure onset can be evaluated in the manner outlined herein. However, if the detection peak consistently occurs before seizure onset, an offset exists. In this case, one suggestion is to count the period from each detection onward as detections (regardless of whether they return to baseline), until an optimal number of seizures are "captured" within this period. Validation will then be determined by the amount of seizures captured in this manner versus the increase in detected state proportion.

The simplifying assumption of stationarity was used to derive the type I and type II error equations. IEEG records are unlikely to be stationary because all human data are acquired from patients in inpatient epilepsy monitoring units, most of whom undergo active titration of antiepileptic drugs. If the detection algorithm is sensitive to such changes (e.g., a detector that detects spikes can be rendered useless if an antiepileptic drug suppresses spikes), statistical validation cannot be guaranteed under all conditions. New data that could be made available from nonhospitalized patients with chronically implanted responsive antiepileptic devices may make this less of an issue. In addition, data can be selectively used from patients who are admitted to inpatient epilepsy units but did not require medication adjustment.

Finally, the Markov assumption, implicit in the framework above, states that the probability of a certain state at a time t depends only on what the state was at a time $t - 1$. Although this is an open question, the probability of entry into a distinct brain state likely exhibits longer-term dependencies, with probabilities of entering particular states dependent on the particular sequences of a number of past states. For example, if a postictal antiseizure effect exists, a long-duration seizure state may lower the probability for subsequent entry into the preseizure state from the baseline state.

In summary, the presented framework is useful in several concrete ways: *1*) it allows investigators to quantify the performance of a seizure prediction algorithm against a null hypothesis, in such a way that outputs of a seizure prediction algorithm can be analyzed in a transparent fashion, and *2*) it yields measures that readily facilitate the process of calculating data requirements in clinical trials involving seizure prediction algorithms, an eminently useful and practical application in the clinical setting. Finally, this model redefines the preictal period as a stochastic, probabilistic state out of which seizures might arise. It is this last contribution that distinguishes the above work from other recent papers on statistical validation of seizure prediction. The notion of a permissive preictal state in seizure generation modeling has the potential to improve implanted clinical antiepileptic devices. For example, a measure of the magnitude of the probability of seizure onset can be used to control dosages of therapeutic interventions such as focal brain stimulation or focal drug delivery.

As the field of seizure prediction has begun incorporating clearer, statistically driven methods into its studies, it appears to have been retreating from initial claims of success in what is turning out to be a very difficult but tantalizing problem. Far from engendering pessimism, we interpret this call for rigor as validation of the importance of this research, with the even greater potential to result in new and effective therapies for our patients.

APPENDIX

Derivation of type I and type II errors

Let p be the probability of making a transition from the detected state to the seizure state. In N seizures the probability $P(t \mid N, p)$ of obtaining t transitions from the detected state to the seizure state as a function of the parameters N and p is governed by the binomial distribution

$$p(t|N, p) = \binom{N}{t} p^t (1 - p)^{N-t} \tag{A1}$$

We want to test the null hypothesis H_0: $P \leq p_0$ versus the alternative H_1: $P > p_0$. We should therefore reject the null hypothesis if the number of transitions from detected state to seizure is at least x. Because we want to control the probability of a type I error (rejecting the null hypothesis when it is true) at level α this amounts to finding x that satisfies

$$\sum_{t=x}^{N} p(t|N, p) = \sum_{t=x}^{N} \binom{N}{t} p^t (1 - p)^{N-t} \geq \alpha \tag{A2}$$

The general approach is to find the smallest x, denoted by x_0, that satisfies *Eq.* A*2*.

The probability of a type II error (retaining the null hypothesis when it is false) can be calculated as a function of p as

$$\sum_{t=0}^{x-1} p(t|N, p) = \sum_{t=0}^{x-1} \binom{N}{t} p^t (1 - p)^{N-t} \tag{A3}$$

One way to proceed is to fix a value of p, say $p_1 > p_0$, and evaluate *Eq.* A*3*, thereby producing the probability of a type II error. Another approach is to fix the probability of a type II error that we are willing to tolerate at β and find the value of p that achieves this value. This is tantamount to solving the equation

$$\sum_{t=0}^{x-1} p(t|N, p) = \sum_{t=0}^{x-1} \binom{N}{t} p^t (1 - p)^{N-t} = \beta \tag{A4}$$

The value of p that satisfies *Eq.* A*4* is denoted by y in the paper.

ACKNOWLEDGMENTS

The authors thank Dr. John Pollard, Dr. Douglas Maus, and K. Essien, who contributed critiques, encouragement, and ideas that served as inspiration for the material in this paper. We also thank Dr. Leif Finkel for unwavering support.

GRANTS

This work was supported by National Institute of Neurological Disorders and Stroke Grants R01-NS-048598-01 and R01-NS-041811-01, Klingenstein Foundation, Whitaker Foundation, Dana Foundation, American Epilepsy Society, Citizens United for Research in Epilepsy, Partnership for Pediatric Epilepsy, and The Epilepsy Therapy Development Project.

REFERENCES

Albert PS. A two-state Markov mixture model for a time-series of epileptic seizure counts. *Biometrics* 47: 1371–1381, 1991.

Andrzejak RG, Mormann F, Kreuz T, Rieke C, Kraskov A, Elger C, Lehnertz K. Testing the null hypothesis of the nonexistence of a preseizure state. *Phys Rev E* 67: 010901, 2003.

Andrzejak RG, Widman G, Lehnertz K, Rieke C, David P, Elger CE. The epileptic process as nonlinear deterministic dynamics in a stochastic environment: an evaluation on mesial temporal lobe epilepsy. *Epilepsy Res* 44: 129–140, 2001.

Aschenbrenner-Scheibe R, Maiwald T, Winterhalder M, Voss HU, Timmer J, Schulze-Bonhage A. How well can epileptic seizures be predicted? An evaluation of a nonlinear method. *Brain* 126: 2616–2626, 2003.

Dempster AP, Laird N, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39: 1–38, 1977.

Federico P, Abbott DF, Briellmann RS, Harvey AS, Jackson GD. Functional MRI of the pre-ictal state. *Brain* 128: 1811–1817, 2005.

Gardner AB, Kreiger AM, Vachtsevanos G, Litt B. One-class novelty detection for seizure analysis from intracranial EEG. *J Machine Learn* 7: 1025–1044, 2006.

Haut SR. Seizure clustering. *Epilepsy Behav* 8: 50–55, 2006.

Hopkins A, Davies P, Dobson C. Mathematical models of patterns of seizures: their use in the evaluation of drugs. *Arch Neurol* 42: 463–467, 1985.

Kossoff EH, Ritzl EK, Politsky JM, Murro AM, Smith JR, Duckrow RB, Spencer DD, Bergey GK. Effect of an external responsive neurostimulator on seizures and electrographic discharges during subdural electrode monitoring. *Epilepsia* 45: 1560–1567, 2004.

Kreuz T, Andrzejak RG, Mormann F, Kraskov A, Stögbauer H, Elger CE, Lehnertz K, Grassberger P. Measure profile surrogates: a method to validate the performance of epileptic seizure prediction algorithms. *Phys Rev E* 69: 161915, 2004.

Le ND, Leroux BG, Puterman ML. Exact likelihood evaluation in a Markov mixture model for seizure counts. *Biometrics* 48: 317–323, 1992.

Lehnertz K, Litt B. The first international workshop on seizure prediction: summary and data description. *Clin Neurophysiol* 116: 493–505, 2005.

Le Van Quyen M, Martinerie J, Navarro V, Boon P, D'Have M, Adam C, Renault B, Varela F, Baulac M. Anticipation of epileptic seizures from standard EEG recordings. *Lancet* 357: 183–188, 2001.

Litt B, Echauz J. Prediction of epileptic seizures. *Lancet Neurol* 1: 22–30, 2002.

Litt B, Esteller R, Echauz J, D'Alessandro M, Shor R, Henry T, Pennell P, Epstein C, Bakay R, Dichter M, Vachtsevanos G. Epileptic seizures may begin hours in advance of clinical onset. *Neuron* 30: 51–64, 2001.

Litt B, Lehnertz K. Seizure prediction and the preseizure period. *Curr Opin Neurol* 15: 173–177, 2002.

Martinerie J, Adam C, Le Van Quyen M, Baulac M, Clemenceau S, Renault B, Varela FJ. Epileptic seizures can be anticipated by non-linear analysis. *Nature Med* 4: 1173–1176, 1998.

Mormann F, Kreuz T, Rieke C, Andrzejak R, Kraskov A, David P, Elger CE, Lehnertz K. On the predictability of epileptic seizures. *Clin Neurophysiol* 116: 569–587, 2005.

Murro A, Park Y, Bergey G, Kossoff E, Ritzl E, Karceski S, Flynn K, Choi H, Spencer D, Duckrow R, Seale C. Multicenter study of acute responsive stimulation in patients with intractable epilepsy (Abstract). *Epilepsia* 44, *Suppl*. 9: 326, 2003.

Navarro V, Martinerie J, Le Van Quyen M, Clemenceau S, Adam C, Baulac M, Varela F. Seizure anticipation in human neocortical epilepsy. *Brain* 125: 640–655, 2002.

Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc IEEE* 77: 257–286, 1989.

Risinger M, Engel JJ, VanNess P, Henry T, Crandall P. Ictal localization of temporal seizures with scalp-sphenoidal recordings. *Neurology* 39: 1288–1293, 1989.

Sunderam S, Osorio I, Frei M, Watkins J. Stochastic modeling and prediction of experimental seizures in Sprague–Dawley rats. *J Clin Neurophysiol* 18: 275–282, 2001.

Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inform Theory* 13: 260–267, 1967.

Weinand ME, Carter LP, el-Saadany WF, Sioutos PJ, Labiner DM, Oommen KJ. Cerebral blood flow and temporal lobe epileptogenicity. *J Neurosurg* 86: 226–232, 1997.

Winterhalder M, Maiwald T, Voss HU, Aschenbrenner-Scheibe R, Timmer J, Schulze-Bonhage A. The seizure prediction characteristic: a general framework to assess and compare seizure prediction methods. *Epilepsy Behav* 4: 318–325, 2003.

Worrell G, Wharen R, Goodman R, Bergey G, Murro A, Bergen D, Smith M, Vossler D, Morrell M. Safety and evidence for efficacy of an implantable responsive neurostimulator (RNS™) for the treatment of medically intractable partial onset epilepsy in adults (Abstract). *Epilepsia* 46, *Suppl*. 8: 226, 2005.