PHYSICAL REVIEW E **69**, 061915 (2004)

Measure profile surrogates: A method to validate the performance of epileptic seizure prediction algorithms

Thomas Kreuz,[1,2,*] Ralph G. Andrzejak,[1] Florian Mormann,[2] Alexander Kraskov,[1] Harald Stögbauer,[1] Christian E. Elger,[2] Klaus Lehnertz,[2] and Peter Grassberger[1]

[1]*John-von-Neumann Institute for Computing, Forschungszentrum Jülich, 52425 Jülich Germany*

[2]*Department of Epileptology, University of Bonn, Sigmund-Freud-Straße 25, 53105 Bonn, Germany*

(Received 21 August 2003; published 15 June 2004)

In a growing number of publications it is claimed that epileptic seizures can be predicted by analyzing the electroencephalogram (EEG) with different characterizing measures. However, many of these studies suffer from a severe lack of statistical validation. Only rarely are results passed to a statistical test and verified against some null hypothesis H_0 in order to quantify their significance. In this paper we propose a method to statistically validate the performance of measures used to predict epileptic seizures. From measure profiles rendered by applying a moving-window technique to the electroencephalogram we first generate an ensemble of surrogates by a constrained randomization using simulated annealing. Subsequently the seizure prediction algorithm is applied to the original measure profile and to the surrogates. If detectable changes before seizure onset exist, highest performance values should be obtained for the original measure profiles and the null hypothesis. "The measure is not suited for seizure prediction" can be rejected. We demonstrate our method by applying two measures of synchronization to a quasicontinuous EEG recording and by evaluating their predictive performance using a straightforward seizure prediction statistics. We would like to stress that the proposed method is rather universal and can be applied to many other prediction and detection problems.

DOI: 10.1103/PhysRevE.69.061915 PACS number(s): 87.19.La, 05.45.Tp, 05.45.Xt

I. INTRODUCTION

The hallmark of epilepsy is the occurrence of intermittent malfunctions of the brain known as seizures. In the electroencephalogram (EEG) most of these so-called *ictal* states are easily recognized by their rhythmic high-amplitude activity reflecting the abnormal synchronization of a large number of neurons [1]. With this in mind, the question arises as to whether it is also possible to discriminate the intervals preceding seizures (*pre-ictal* periods) from the intervals far away from any seizure activity (*inter-ictal* periods). Provided that the analysis of the EEG would allow one to reliably detect a pre-ictal state in a prospective setting, new therapeutic possibilities (e.g., seizure prevention strategies) can be envisaged [2].

Therefore, it is not surprising to find a very rich and diverse literature dealing with the prediction of epileptic seizures. Starting from earliest approaches based on pattern recognition [3] and spike detection [4,5], at first mostly univariate measures were employed, either linear [6,7] or nonlinear [8–11] in nature. Later efforts reporting the predictability of epileptic seizures by applying these two different kinds of univariate measures include Ref. [12] and Refs. [13,14], respectively. It is only recently that bivariate [15–17] or multivariate [18] measures have been added to the wide range of approaches reportedly being able to detect a pre-ictal state. The current impact of this topic is stressed by recent controversies about the relevance of nonlinear approaches for the prediction of epileptic seizures [19,20] and even more striking by studies raising doubts about the reproducibility of reported claims [21,22]. For an overview refer to Refs. [23–25].

Typically in a study on the predictability of epileptic seizures first a certain characterizing measure is calculated from multichannel EEG using a moving-window technique. The resulting measure profiles are then scanned for prominent features which can be related to the actual seizure times. These features might be drops or peaks (e.g., quantified as threshold crossings) or any other distinct pattern in the measure profile. In a second step the measures' capability to distinguish the pre-ictal from the inter-ictal interval is evaluated with a test statistics quantifying the occurrence of these features relative to the seizure times and resulting in some kind of performance value. If this performance is high, it might on the one hand reflect the existence of a pre-ictal state and the capability of the applied measure to detect it, but it might on the other hand also be due to statistical fluctuations or some (unknown) bias in the algorithm.

In the design of a seizure prediction algorithm there are many subtle points to be considered carefully. Typically the calculation of the measure as well as the later statistical evaluation involves the choice of certain parameters. In this context, much care needs to be taken to avoid in-sample optimization of these parameters. Certainly, what is true for a single measure holds also for a larger number of different measures. The application of a huge variety of measures to the EEG might yield a measure with seemingly good results just by chance (particularly on a limited database). Second, there are many degrees of freedom in the statistical evaluation. In the case of univariate measures often a best channel selection is performed, and for bivariate measures, which evaluate the dependences between two channels, there are even more channel combinations to choose from. Finally the

*Electronic address: t.kreuz@fz-juelich.de

same argument holds for different patients as well. Provocatively speaking, many (spurious) claims about the existence of a pre-ictal state might just be due to some "best parameter," "best measure," "best channel," and/or "best patient" selection.

Since usually these problems (which have, at least in part, also been addressed in Refs. [26,27]) cannot be solved during the design of a seizure prediction statistics, the question arises as to how to interpret a nonzero performance value. This value might correctly reflect the existence of a detectable pre-ictal state, but it might also be the spurious result of statistical fluctuations. Therefore, to assess the performance yielded by a seizure prediction algorithm, a method to judge its statistical validity is needed. The result should be verified against some null hypothesis and its level of significance should be estimated. This can be achieved using the concept of surrogates [28,29], in which the validity of a given test result is evaluated by applying the test not only to the original data but also to an ensemble of surrogate data generated by means of a Monte Carlo randomization. In our case the null hypothesis H_0 to test against can be stated as follows: The measure under investigation is not suited for seizure prediction. If this null hypothesis is fulfilled, it might be due to two different reasons. Either a pre-ictal state does not exist (and thus there is no measure suited for seizure prediction) or a pre-ictal state does exist, but the measure is not able to detect it. On the other hand, the null hypothesis can only be rejected if both inverse conditions are fulfilled: There are specific changes before a seizure and the measure is sensitive to these changes.

The performance of any seizure prediction algorithm crucially depends on whether the sequence of actual seizures is matched by some corresponding structure in the measure profiles. Therefore to test for statistical significance of a good performance by using the method of surrogates, any such structure should be destroyed by the randomization. Essentially, this can be done in two different ways. Andrzejak and colleagues [26] recently introduced the method of seizure time surrogates in which the seizure times are randomized, while the measure profiles are maintained. In this paper we propose the *method of measure profile surrogates*, a complementary approach, in which the seizure times are kept fixed and instead a constrained randomization of the measure profiles is performed using the method of simulated annealing.

The concept of surrogates as a means to test a null hypothesis is applied equivalently in both methods: The seizure prediction algorithm is run using the original measure profiles (seizure times) and its performance is compared to the results of the same algorithm using an ensemble of measure profile surrogates (seizure time surrogates). Provided that a pre-ictal state exists and the prediction algorithm is able to detect it, its performance should be highest for the original measure profiles (seizure times). In this case the null hypothesis could be rejected at the level of significance determined by the number of measure profile surrogates (seizure time surrogates).

Both methods are reasonable statistical approaches to address the correspondence between measure profiles and seizure times, but we argue that the method of measure profile surrogates is the more natural choice: Usually, within the method of surrogates the property to test for is destroyed in the surrogates. And in the present case the object under investigation is the measure rather than the sequence of seizures. More specifically, the aim is to test the measure for its capability to extract information from the EEG that enables the prediction of the original seizures and not to test the sequence of seizures whether they resemble the measure profiles.

Within either of these methods there are certain properties of the original which should be preserved for the surrogates. In the case of seizure time surrogates it has been proposed to preserve the total number of seizures, the distribution of time intervals between consecutive seizures, and as the case may be, any clustering of the seizures [26]. This has been achieved by a random permutation of the original seizure intervals. As indicated already in Ref. [26], this approach is applicable only if the number of seizures and hence the number of possible permutations are large enough to allow the generation of the number of surrogates needed to obtain the desired significance. The number of possible permutations is even further diminished in the presence of recording gaps, since then permutations have to be discarded whenever one of the surrogate seizures falls into such a gap. To prevent a bias between the original and surrogates, also ictal and post-ictal intervals as well as all other events known to possibly cause changes in the EEG have to be avoided (for the sake of brevity, in the following these intervals will also be referred to as recording gaps). But even when a sufficient number of permutations remain, much care has to be taken to ensure that the inter-ictal interval as well as any possible pre-ictal interval are equally well represented in the original and in all of the seizure time surrogates.

In the method of measure profile surrogates these issues are easily addressed, since the original seizure times are not changed at all. Rather they are correctly considered as given conditions based upon which the measure profiles are probed for their predictive performance. But also in this method there exist some constraints—i.e., properties which should be extracted from the original measure profile and imposed on the surrogate measure profiles. First of all a suitable randomization should maintain all existing recording gaps. Furthermore, it is advisable not only to preserve the amplitude distribution but also to maintain essential parts of the autocorrelation function. The preservation of these features guarantees that, when regarded independently from the seizure times, the original as well the surrogate measure profiles can be considered as a possible original measure profile. The most important property that might remain different is the correspondence to the seizure times and this is exactly the property under investigation.

To illustrate our method, we use two different evaluation schemes to investigate the predictive performance of two bivariate measures of synchronization, the *mean phase coherence* as a measure for phase synchronization [15] and the recently proposed *event synchronization* [30]. These measures are calculated from the same quasicontinuous EEG recording of an epilepsy patient already analyzed in Ref. [26]. The seizure prediction statistics applied to the resulting measure profiles and their surrogates is straightforward, simply comparing amplitude distributions of pre-ictal and inter-ictal

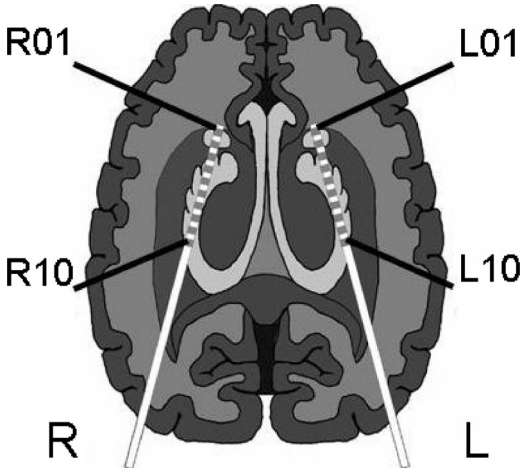

FIG. 1. Schematic view of implanted depth electrodes.

intervals [27]. The remainder of the text is organized as follows: First we describe the data (Sec. II A), the measures (Sec. II B), and the seizure prediction statistics (Sec. II C) used to demonstrate our method of measure profile surrogates. This method is introduced in Sec. II D. In Sec. III we show the results of our application, before we draw our conclusions in Sec. IV.

II. METHODS

A. Data

We analyzed quasicontinuous multichannel EEG recorded from an epilepsy patient over 5 days during which the patient had ten epileptic seizures. The EEG was recorded prior to and independently from the design of this study during the presurgical work-up [31]. Furthermore, the patient was not selected for this study according to any *a priori* knowledge of predictability or nonpredictability in the recordings. Using two implanted depth electrodes each equipped with ten separate contacts (denoted as L01,…,L10 and R01,…,R10), the EEG was measured directly within the brain (cf. Fig. 1) at a high signal-to-noise ratio. EEG data were sampled at 200 Hz using a 16-bit analog-to-digital converter and filtered within a frequency band of 0.5–85 Hz. The EEG contains one major and two minor recording gaps. In addition to the ten ictal and post-ictal intervals (defined from seizure onset until 30 min after seizure termination), four other events known to be associated with changes in the EEG (three subclinical seizures and one period of hyperventilation) took place during the acquisition.

B. Measures

From these data two measures of synchronization were calculated using a moving-window technique with nonoverlapping segments of 20.48 s corresponding to N=4096 data points. In order to focus on local synchronization effects, in this study only the 18 neighboring channel combinations (L01-L02,…,L09-L10 and R01-R02,…,R09-R10) were analyzed.

1. Phase synchronization

The mean phase coherence R [15], a measure for phase synchronization, has already been applied in previous seizure prediction studies [17,20,27]. For its calculation first instantaneous phases $\phi_x(t)$ and $\phi_y(t)$ are extracted from two time series x and y of length N using the analytic signal approach [32,33]:

$$\phi_x(t) = \arctan\frac{\tilde{x}(t)}{x(t)}, \tag{1}$$

where

$$\tilde{x}(t) = \frac{1}{\pi}\ \mathrm{P.V.} \int_{-\infty}^{+\infty} \frac{x(t')}{t-t'}dt' \tag{2}$$

is the Hilbert transform ("P.V." denoting the Cauchy principal value). From this we obtain the mean phase coherence defined as

$$R = \left| \frac{1}{N}\sum_{j=1}^{N} e^{i[\phi_x(t_j)-\phi_y(t_j)]} \right| = 1 - V, \tag{3}$$

with V denoting the circular variance [34]. The mean phase coherence R is confined to the interval [0,1] with larger values indicating a higher degree of synchronization.

2. Event synchronization

The recently proposed event synchronization Q [30] quantifies the overall level of synchronicity from the number of quasisimultaneous appearances of certain events (here defined as local maxima and minima). In a first step the respective time series are scanned for these events, and the times of their occurrence are marked as t_i^x and t_j^y $(i=1,\ldots,m_x; j=1,\ldots,m_y)$ with m_x and m_y denoting the respective number of events. Allowing a maximum time interval τ_e in which two events are still regarded as simultaneous, Q is obtained by counting the number of times the same event (e.g., maximum or minimum) occurs simultaneously in both time series. To cover the interval [0,1] it is normalized by the number of events:

$$Q = \frac{\sum J_{ij}}{\sqrt{m_x m_y}} \tag{4}$$

with

$$J_{ij} = \begin{cases} 1 & \text{if } 0 \leqslant |t_i^x - t_j^y| \leqslant \tau_e, \\ 0 & \text{otherwise}. \end{cases} \tag{5}$$

C. Seizure prediction statistics

For the design of a seizure prediction statistics we follow our earlier work [27,35] by using a straightforward approach, simply comparing amplitude distributions of pre-ictal and inter-ictal intervals using receiver-operating characteristics (ROC's) (cf. [36]). This statistics will be applied to the original as well as to the surrogate measure profiles. Within this statistics, a threshold for amplitude values is continuously

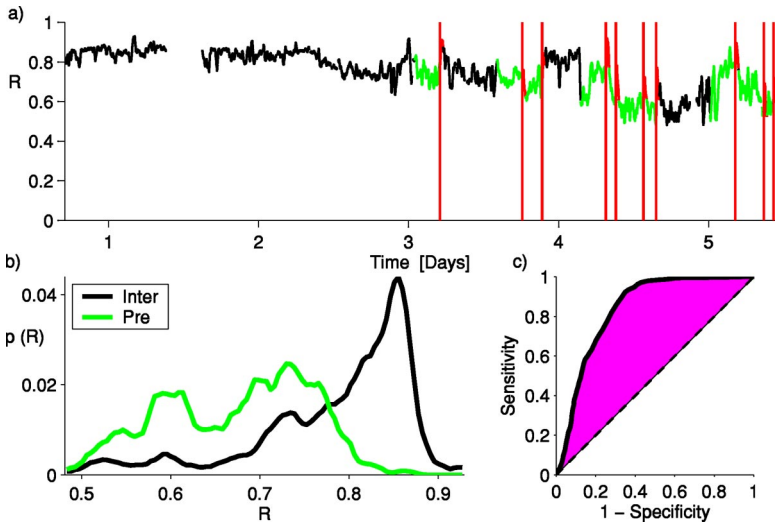

FIG. 2. Illustration of the seizure prediction statistics: (a) Original measure profile (smoothed using a 5-min moving average filter) of the best channel combination (R08-R09) for the mean phase coherence R with a pre-ictal interval of 240 min. Seizures are marked by vertical lines: day ticks denote midnight. Pre-ictal and inter-ictal intervals are depicted in bright and dark colors, respectively. (b) Distributions of values from these two intervals. (c) Corresponding ROC curve yielding the maximum performance value $A=0.68$.

shifted across these distributions, and the fraction of amplitudes of the first distribution below this threshold is plotted against the respective fraction of the second distribution. With respect to the chosen hypothesis of separability (e.g., values from the pre-ictal distribution are lower than those from the inter-ictal distribution) this corresponds to plotting the sensitivity (ratio of true positives to total number of positives) against 1 minus the specificity (ratio of true negatives to total number of negatives). The capability of a measure to distinguish between the inter-ictal and the pre-ictal interval—i.e., its potential predictive performance—can then be quantified by the area between the resulting ROC curve and the diagonal. Identical distributions lead to a zero area, while for distributions that are completely nonoverlapping, ROC values of 0.5 or −0.5 are attained, depending on which hypothesis is used for the definition of sensitivity and specificity. To cover the range from $[-1,1]$ we here renormalize this area A by a factor of 2. Note that this definition differs from common practice in ROC statistics where values between 0 and 1 are used.

This is a rather simple statistics, but still its application involves, as usual, the choice of certain parameters. If computationally feasible, a common practice in such a case is to evaluate many different combinations of parameters and to choose the most successful one. Without statistical validation this is a typical example of "in-sample" optimization, but with a proper use of surrogates this is made legitimate. In this context, however, much care has to be taken to avoid any bias between the original and surrogates: i.e., exactly the same optimization should be applied to both.

In our case it is not known beforehand which are the prominent features to be extracted from our measure profiles (e.g., drops or peaks), at what times before a seizure, and in which channel combination they occur. The first point is addressed by testing for both a pre-ictal decrease as well as an increase of synchronization, thereby judging ROC values A by their absolute value. The length of the pre-ictal interval is set to 240 min, motivated by recently reported prediction times in the literature [12,17]. For every measure the evaluation of this statistics is then carried out twice, first regarding each channel combination separately and second after selecting the best channel combination. Thus in the first evaluation scheme we have two different values to choose from for each channel combination. Accordingly in the second scheme the final performance value for each measure is chosen as the maximum of $16\times 18=288$ different values. In Fig. 2 the seizure prediction statistics is illustrated using the measure profile, for which the maximum performance value is obtained when applying this optimization scheme.

D. Method of measure profile surrogates

To test against a certain null hypothesis via a constrained randomization of time series is a well-known concept within the framework of nonlinear time series analysis [29,37]. The original algorithm [28] and a number of expansions or refinements [38,39] are each designed to impose specific constraints on the surrogates and thus to address one particular null hypothesis. In contrast to these standard approaches the method of simulated annealing [40] provides a rather universal means for generating random time series with a wide variety of possible constraints and therefore allows testing of almost arbitrary null hypotheses. Furthermore, the standard algorithms act in the Fourier domain and therefore can produce artifacts because of their implicit assumption of periodic continuation. The resulting edge effect is due to the fact that when preserving the amplitude spectrum, according to the Wiener-Khinchin theorem only the periodic sample autocorrelation function is maintained. In contrast, the method of simulated annealing acts in the time domain and thus is able to preserve the original autocorrelation function. Simulated annealing is also clearly superior when it comes to the constrained randomization of data with recording gaps. Coping with these gaps is a nontrivial problem for Fourier-based randomization schemes. To treat each segment independently is not a good approach since it is desirable to preserve autocorrelations between different data sets as well. Interpolation schemes might offer a solution for quasicontinuous data sets, but become unfeasible when confronted with long recording gaps. Again, the method of simulated annealing offers a better approach since in the time domain the missing values due to the recording gaps can be set to zero and thus can be neglected in the autocorrelation function.

Simulated annealing (for an overview see [41]) as a method for combinatorial minimization with false minima was introduced in Ref. [42] and was first applied to the generation of surrogates from time series by Schreiber and Schmitz [29]. In short, constraints are specified in terms of a cost function which is then minimized among all possible permutations of the original measure profile. This cost function can be interpreted as the energy E of a thermodynamic system which is annealed slowly towards the global minimum. In this process, starting from an initial random permutation of the original measure profile, randomly chosen pairs of values are exchanged repeatedly until a desired accuracy (i.e., a sufficiently low value of the cost function) is reached. In each iteration step the cost function is updated and depending on the present temperature T the exchange is accepted with probability

$$p(\Delta E,T)=\begin{cases} e^{-\Delta E/T}, & \Delta E>0, \\ 1, & \Delta E\leqslant 0. \end{cases} \tag{6}$$

Exchanges with increasing energy are also accepted with nonzero probability to allow escaping from local minima. Whenever a certain number of either tested or accepted exchanges has been performed, the temperature is slowly decreased according to some cooling scheme (e.g., $T_{new}=T_{old}\times\alpha$ with $1>\alpha\geqslant 0$).

In our application of this method the three different constraints mentioned in the Introduction can easily be imposed on the measure profile surrogates. First of all, recording gaps are preserved by excluding the missing values in the gaps from the permutation scheme. Since all surrogates are permutations of the original measure profile, the amplitude distribution is maintained by construction. The last constraint is the approximate preservation of the autocorrelation function

$$C(\tau)=\frac{1}{N-\tau}\sum_{n=0}^{N-\tau-1}x_{n+\tau}x_n,\ \tau\geqslant 0. \tag{7}$$

This constraint is formulated in the cost function

$$E=\sum_{\tau=1}^{N-1}\omega_\tau\left|C^{Surr}(\tau)-C^{Ori}(\tau)\right|, \tag{8}$$

with weights here defined as

$$\omega_\tau=\begin{cases} 1/\tau & \text{if } \tau\leqslant\tau_{\max}, \\ 0 & \text{otherwise}. \end{cases} \tag{9}$$

A proper choice of these weights is essential. First, they offer the possibility to define the part of the autocorrelation function that should be preserved. This crucially depends on the original autocorrelation function. Four typical examples for the measures and patient analyzed are depicted in Fig. 3. While the autocorrelation function of most channel combinations decays rather fast and does not show any long-range correlations, some channel combinations clearly seem to reflect the circadian rhythm resulting approximately in a 24-h periodicity. This different behavior can be judged as an essential property of the individual measure profiles worth preserving. To guarantee this, for each measure profile the maximum time lag $\tau_{\max}$ is set to 4600 windows, thereby ensuring that the first 26 h of the autocorrelation function (given a window length of 20.48 s) are maintained. Without such peculiarities present, a reasonable choice could have been the first zero crossing of the original autocorrelation function.

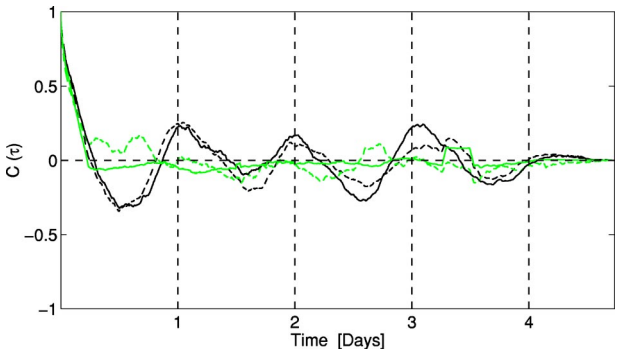

FIG. 3. Four exemplary autocorrelation functions of original measure profiles for the mean phase coherence R.

The second issue to be considered when choosing appropriate weights is the computational cost. Typically the number of iterations is quite large and in each iteration step an update of the cost function has to be performed. Fortunately this only requires the recalculation of those terms of the autocorrelation function to which the two values of the exchanged pair actually contribute. These can be further reduced by setting every other weight to zero. Given the smoothness of the autocorrelation function, the omitted terms are then adjusted automatically. To avoid periodicity artifacts the very first weights are not set to zero. In order to give higher importance to small lags, the remaining terms are weighted by $1/\tau$. Many further tricks to reduce the high computational cost can be found in Refs. [29,40].

Using this method of simulated annealing for each measure profile from every channel combination, an ensemble of 19 different measure profile surrogates is generated. Subsequently the seizure prediction statistics is applied to the original as well as to the measure profile surrogates. As stated already in Sec. II C, the evaluation of this statistics is carried out twice. Since each measure profile surrogate is generated by a constrained randomization of a single measure profile from one channel combination, in the first evaluation scheme the performance of the two synchronization measures is compared for each channel combination separately. For the original measure profile as well as for each of the 19 surrogates exactly the same optimization is performed, thereby choosing the one out of two different parameters (pre-ictal increase and decrease) that yields the maximum performance. In the second evaluation scheme for each measure the best channel combination is selected additionally. Here each measures' final performance value is thus chosen as the maximum value out of a set of 36 different possibilities. In each of the two schemes the respective null hypothesis can be rejected with a significance level of $p=0.05$ if highest values are yielded for the original measure profiles.

Both evaluation schemes test the general null hypothesis: that the measure under investigation is not suited for seizure

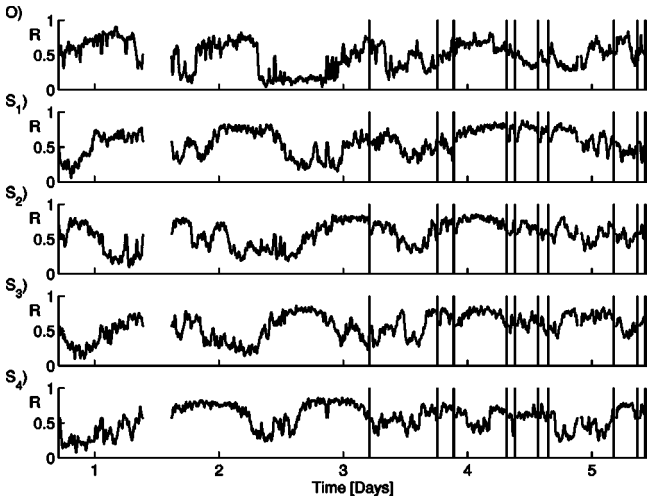

FIG. 4. Original measure profile (O) of the mean phase coherence R for the first channel combination L01-L02 and four exemplary surrogates (S_1–S_4), all of them again smoothed using a 5-min moving-average filter. Seizures are marked by solid vertical lines.

prediction. But actually they can be regarded as conceptually different tests with different extended null hypotheses, since they are not based on the same assumptions. Looking at the single channel combinations corresponds to testing for a possible predictive feature consisting of a significantly high number of local effects. Selecting the best channel combination, on the other hand, is aiming at prediction by a maximum local effect. Apart from these two many other evaluation schemes are conceivable [27]. Averaging over all channel combinations, to name one further example, would test for a global effect. In fact, the choice of an evaluation scheme for the surrogate test constitutes a new degree of freedom which has to be considered carefully. The respective scheme could, in principle, also be incorporated in the null hypothesis: e.g., the extended null hypothesis for the second scheme H_0^{II} could read as follows: The measure under investigation is not suitable to find maximum local effects predictive of epileptic seizures.

III. RESULTS

For an exemplary channel combination the original measure profile of the mean phase coherence R and four surrogates are depicted in Fig. 4. By construction all measure profiles are identical in certain characteristic properties (i.e., the recording gaps, the amplitude distribution, and the autocorrelation function up to the maximum time lag) and in this respect each of them can be regarded as a possible original measure profile. However, the surrogates can clearly be distinguished from the original measure profile as well as from one another by the temporal distribution of drops, peaks, and quasiplateaus. The variety among the surrogates clearly demonstrates that the imposed constraints leave enough degrees of freedom for the randomization and do not overspecify the surrogates.

The remaining and most crucial question is whether the original measure profile stands out from the surrogates with

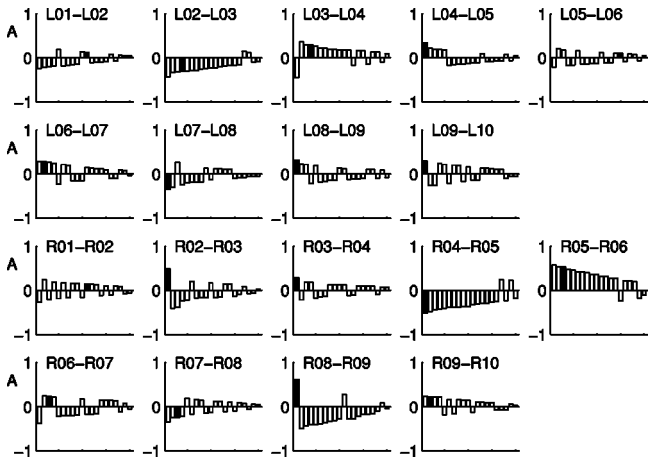

FIG. 5. Performance values for the first evaluation scheme (parameter optimization is performed for each channel combination separately) of the mean phase coherence R for the original measure profiles (highlighted by solid bars) and the surrogates. For each channel combination of the right and left depth electrode signed ROC values A are depicted, sorted by their absolute value. Asterisks mark channel combinations yielding maximum performance for the original measure profile.

respect to its correspondence of the seizure times. To answer this question, the seizure prediction statistics is applied to the original measure profiles as well as to their surrogates. In the first evaluation scheme each channel combination is regarded separately, performing exactly the same optimization for the original measure profile as well as for each of the 19 surrogates. The resulting performance values are shown in Fig. 5 for the mean phase coherence R and in Fig. 6 for the event synchronization Q. Signed ROC values A are depicted to indicate whether a pre-ictal decrease or increase of synchronization is observed for the respective profiles. In order to show the rank of the original performance inside the distribution of the values obtained for the surrogate measure profiles, all performances are sorted by their absolute value.

When considering the performances obtained for the original measure profiles only, highly nonuniform results can be observed. For most channel combinations ROC values

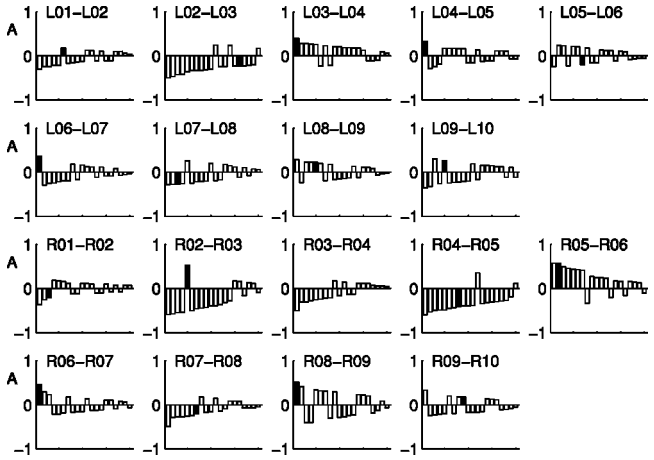

FIG. 6. Same as Fig. 5, but for event synchronization Q.

close to zero are obtained, reflecting the fact that pre-ictal and inter-ictal amplitude distributions are almost indistinguishable. But for some channel combinations (e.g., R02-R03, R05-R06, and R08-R09) high ROC values indicating a considerable degree of discrimination between these distributions can be observed, no matter which of the two measures is used. This might correctly reflect the existence of a pre-ictal state which can be detected using either measure, but it could also be the spurious result of statistical fluctuations.

This ambiguity can be resolved by the method of measure profile surrogates. First of all, the information gathered by the surrogates is nonredundant to the information of the original performance values. This can be seen, e.g., when turning our attention to the results of event synchronization in channel combinations L04-L05 and R04-R05 (cf. Fig. 6). In the channel combination from the left hemisphere the absolute performance value obtained for the original measure profile is quite low, but still larger than all values yielded by the surrogates, whereas in the right channel combination a higher absolute performance value is observed, which, however, does not prove to be significant.

In contrast to the high consistency in the two measures' ROC values regarding the original measure profiles only, qualitatively different results are obtained in a comparison of the performances yielded for the original measure profiles with the ones observed for the surrogates. For the mean phase coherence results appear to be significant for 9 out of 18 channel combinations (L04-L05, L06-L07, L08-L09, L09-L10, R02-R03, R03-R04, R04-R05, R08-R09, and R09-R10). For event synchronization in 5 channel combinations (L04-L05, L05-L06, R05-R06, R06-R07, and R08-R09) highest absolute ROC values are obtained for the original measure profiles.

If a hypothesis test with a nominal size p is performed q times, the likelihood P to get at least r rejections merely by chance is given by

$$P = \sum_{k=r}^{q} \binom{q}{k} p^k (1-p)^{q-k}. \tag{10}$$

Here a one-sided test with 19 surrogates (hence $p=0.05$) is performed for $q=18$ different channel combinations. This yields probabilities $P(r \geqslant 9) < 10^{-7}$ for the mean phase coherence and $P(r \geqslant 6) \approx 10^{-3}$ for event synchronization. The calculation of these values of significance is based on the implicit assumption that measure profiles from different channel combinations can be regarded as independent. To verify this assumption empirically, the correlation between all combinations of measure profiles is estimated. Most values are close to zero and only rarely is a distinct dependence observed. Furthermore, as can be seen from Figs. 5 and 6, also the performance values obtained for the original measure profiles do not show any clustering for values from neighboring channel combinations. But even when a slight reduction in the number of independent channel combinations is taken into account, the effect remains that the number of channel combinations to show significant ROC values by itself is significant. Thus the corresponding null hypothesis H_0^I—that the measure is not suitable to find a significant number of local effects predictive of epileptic seizures—can be rejected for both measures.

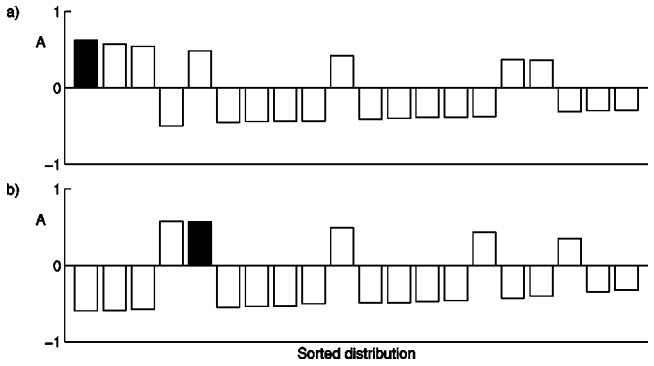

FIG. 7. ROC values A of the second evaluation scheme (best channel selected) for the original measure profiles (highlighted by a solid bar) and the surrogates, again sorted by their absolute value: (a) mean phase coherence R, (b) event synchronization Q.

When the surrogate test is performed for each channel combination separately, the mean phase coherence already seemed to show a slightly higher level of statistical validity. This difference becomes more striking and even leads to a principal distinction in significance in the second evaluation scheme. Here for each measure and for the original as well as for the 19 surrogates the channel combination with the highest performance is chosen. The resulting distributions of the overall performance values are shown in Fig. 7. While for the mean phase coherence results prove to be significant, rendering the highest overall performance value for an original measure profile, this time the corresponding null hypothesis H_0^{II} (already stated at the end of Sec. II D) cannot be rejected for event synchronization. Here the performance value of the best original measure profile falls into the distribution obtained for the ensembles of surrogates.

A closer look on the results obtained for event synchronization in Figs. 6 and 7(b) reveals that in the second evaluation scheme the best performance yielded by the original measure profile of channel combination R05-R06 is surpassed by performances obtained from surrogate measure profiles from other channel combinations—namely, R04-R05 once and R02-R03 twice. This effect is due to the fact that here an ensemble surrogate test is performed. For each measure the best performance yielded by the entirety of the 18 different original measure profiles is compared to the maximum performance values of 19 surrogate ensembles. These surrogate ensembles preserve the properties of the ensemble of original measure profiles as a whole, since they consist of 18 surrogate measure profiles each of which individually substitutes one of the original measure profiles. When the overall optimization from the second evaluation scheme is now applied to the original as well as to the surrogate ensembles, it thus can happen that the channel combination yielding the best performance is not the same for the original measure profiles and the surrogate ensembles. This effect is required to investigate the statistical validity of the optimization procedure performed, in this case the best channel selection.

IV. DISCUSSION

Within the method of measure profile surrogates, results yielded by a seizure prediction algorithm are tested against the fundamental null hypothesis H_0: The measure under investigation is not suited for the prediction of epileptic seizures. To demonstrate our approach we have used two different evaluation schemes to investigate the predictive performance of two measures of synchronization—namely, mean phase coherence and event synchronization—by means of a straightforward seizure prediction statistics. Measure profile surrogates have been generated by a constrained randomization of the original measure profiles. In the first evaluation scheme the significance of the measures' original performance values has been tested for each channel combination separately, resulting in a higher number of significant values for the mean phase coherence. Finally after choosing the best channel combination for each measure in the second scheme an ensemble surrogate test has been performed. Here only the mean phase coherence has reached a significant performance value. Thus for event synchronization only null hypothesis H_0^I is rejected. For the mean phase coherence both null hypotheses H_0^I and H_0^{II} could have been rejected. Note that positive results in a statistical approach like the one used in this study only constitute a necessary but not yet a sufficient condition for the suitability of the measures for seizure prediction. Whether these measures allow one to reliably predict epileptic seizures with both values of sensitivity and specificity sufficient for a clinical application remains to be shown in an algorithmic and prospective setting.

A method to statistically validate the performance of epileptic seizure prediction algorithms (such as the proposed method of measure profile surrogates or, if computationally infeasible, alternatively the method of seizure time surrogates [26]) should be applied whenever there is the slightest chance of any "in-sample" overoptimization. This is the general case since so far rarely a sufficient amount of data are available to perform a proper "out-of-sample" study, where the recordings are divided into a training set on which all algorithm parameters are adjusted and a test set on which later on the performance of the algorithm is evaluated.

In our opinion the method of measure profile surrogates is suited to serve the need for statistical validation of seizure prediction results. On the other hand, also in the application of this method there might be some caveats and pitfalls (e.g., a hidden bias between the original profiles and the surrogates). Thus we would like to stress that also the results obtained with this method should be interpreted with care and jumping to conclusions too quickly should thoroughly be avoided. In particular, the additional degree of freedom introduced in the choice of a suitable null hypothesis should always be considered. Furthermore, whenever a null hypothesis is rejected, it is always very important to keep in mind that the complementary hypothesis is very comprehensive and might include many different reasons that are possibly responsible for this rejection.

Concerning the practical implementation of our method, in some cases the computational cost can be lowered by simplifying the randomization scheme. Some characterizing measures from time series analysis (e.g., the effective correlation dimension evaluated for seizure prediction in Refs. [10,22] or the degree of nonlinear determinism applied in Ref. [26]) show measure profiles with a distinct ceiling effect. For these measures most values lie at the upper or lower end of the definition range, and only rarely can sparse deviations (i.e., drops or peaks) be found. In such cases the method of simulated annealing does not seem to be appropriate. A suitable randomization of the original measure profile could be achieved by performing a random shuffle of these deviations instead.

The application of the proposed method of measure profile surrogates is not restricted to the problem of seizure prediction. In principle it is rather universal and can be used for the statistical validation of the performance of time-resolved measures in many other detection and prediction problems. The only requirement is that a finite number of observables is measured and from their analysis certain circumscribed events are to be detected or predicted. Thus many other applications are also conceivable.

Regarding the particular application considered in this paper we would like to emphasize that it was not the aim of this study to prove or disprove the existence of a pre-ictal state, but rather to supply a general means to reliably evaluate the statistical validity of the performance of a seizure prediction algorithm. In future applications, we expect measure profile surrogates to be a powerful tool to distinguish between measures and algorithms unsuited for the prediction of epileptic seizures and more promising approaches.

ACKNOWLEDGMENT

This work was supported by the Deutsche Forschungsgemeinschaft (Grant No. SFB TR3).

[1] *Epilepsy: A Comprehensive Textbook*, edited by J. Engel, Jr. and T. A. Pedley (Lippincott-Raven, Philadelphia, 1997).
[2] C. E. Elger, Curr. Opin. Neurol. **14**, 185 (2001).
[3] S. S. Viglione and G. O. Walsh, Electroencephalogr. Clin. Neurophysiol. **39**, 435 (1975).
[4] J. Gotman, J. Ives, P. Gloor, A. Olivier, and L. Quesney, Epilepsia **23**, 432 (1982).
[5] H. H. Lange, J. P. Lieb, J. Engel, Jr., and P. H. Crandall, Electroencephalogr. Clin. Neurophysiol. **56**, 543 (1983).
[6] Z. Rogowski, I. Gath, and E. Bental, Biol. Cybern. **42**, 9 (1981).
[7] R. B. Duckrow and S. S. Spencer, Electroencephalogr. Clin. Neurophysiol. **82**, 415 (1992).
[8] L. D. Iasemidis, J. C. Sackellares, H. P. Zaveri, and W. J. Williams, Brain Topogr **2**, 187 (1990).
[9] C. E. Elger and K. Lehnertz, Eur. J. Neurosci. **10**, 786 (1998).

[10] K. Lehnertz and C. E. Elger, Phys. Rev. Lett. **80**, 5019 (1998).
[11] J. Martinerie, C. Adam, M. Le Van Quyen, M. Baulac, S. Clemenceau, B. Renault, and F. J. Varela, Nat. Med. (N.Y.) **4**, 1173 (1998).
[12] B. Litt *et al.*, Neuron **30**, 51 (2001).
[13] M. Le Van Quyen, J. Martinerie, V. Navarro, P. Boon, M. D'Havé, C. Adam, B. Renault, M. Baulac, and F. J. Varela, Lancet **357**, 183 (2001).
[14] V. Navarro, J. Martinerie, M. Le Van Quyen, S. Clemenceau, C. Adam, M. Baulac, and F. Varela, Brain **125**, 640 (2002).
[15] F. Mormann, K. Lehnertz, P. David, and C. E. Elger, Physica D **144**, 358 (2000).
[16] L. D. Iasemidis, P. Pardalos, J. C. Sackellares, and D. S. Shiau, J. Comb. Optim. **5**, 9 (2001).
[17] F. Mormann, T. Kreuz, R. G. Andrzejak, P. David, K. Lehnertz, and C. E. Elger, Epilepsy Res. **53**, 173 (2003).
[18] K. Schindler, R. Wiest, M. Kollar, and F. Donati, J. Clin. Neurophysiol. **113**, 604 (2002).
[19] P. E. McSharry, L. E. Smith, and L. Tarassenko, Nat. Med. (N.Y.) **9**, 241 (2003).
[20] F. Mormann, R. G. Andrzejak, T. Kreuz, C. Rieke, P. David, C. E. Elger, and K. Lehnertz, Phys. Rev. E **67**, 021912 (2003).
[21] W. De Clerq, P. Lemmerling, S. van Huffel, and W. van Paesschen, Lancet **361**, 970 (2003).
[22] R. Aschenbrenner-Scheibe, T. Maiwald, M. Winterhalder, H. U. Voss, J. Timmer, and A. Schulze-Bonhage, Brain **126**, 2616 (2003).
[23] B. Litt and J. Echaux, Lancet Neurol. **1**, 22 (2002).
[24] B. Litt and K. Lehnertz, Curr. Opin. Neurol. **15**, 173 (2002).
[25] K. Lehnertz, F. Mormann, T. Kreuz, R. G. Andrzejak, C. Rieke, P. David, and C. E. Elger, IEEE Eng. Med. Biol. Mag. **22**, 57 (2003).
[26] R. G. Andrzejak, F. Mormann, T. Kreuz, C. Rieke, A. Kraskov, C. E. Elger, and K. Lehnertz, Phys. Rev. E **67**, 010901 (2003).
[27] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz Clin. Neurophysiol. (to be published).
[28] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, Physica D **58**, 77 (1992).
[29] T. Schreiber and A. Schmitz, Physica D **142**, 346 (2000).
[30] R. Quian Quiroga, T. Kreuz, and P. Grassberger, Phys. Rev. E **66**, 041904 (2002).
[31] *Surgical Treatment of the Epilepsies*, edited by J. Engel, Jr. (Raven Press, New York, 1993).
[32] D. Gabor, Proc. IEEE London **93**, 429 (1946).
[33] P. Panter, *Modulation, Noise, and Spectral Analysis*, (McGraw-Hill, New York, 1965).
[34] K. Mardia, *Probability and Mathematical Statistics: Statistics of directional data* (Academy Press, London, 1972).
[35] F. Mormann, Ph.D. thesis, University of Bonn, Germany, 2003.
[36] J. A. Hanley and B. J. McNeil, Radiology **143**, 29 (1982).
[37] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, UK, 1997).
[38] D. Prichard and J. Theiler, Phys. Rev. Lett. **73**, 951 (1994).
[39] T. Schreiber and A. Schmitz, Phys. Rev. Lett. **77**, 635 (1996).
[40] T. Schreiber, Phys. Rev. Lett. **80**, 2105 (1998).
[41] *Applied Simulated Annealing*, edited by R. V. V. Vidal (Springer-Verlag, Berlin, 1993).
[42] S. Kirkpatrik, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).