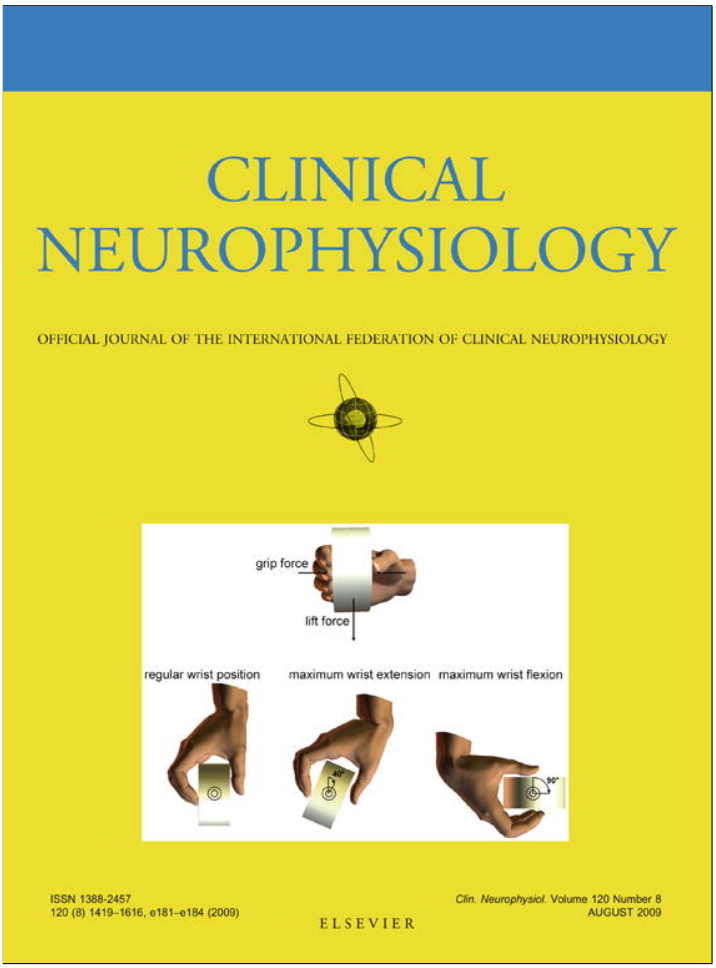

Clinical Neurophysiology 120 (2009) 1465–1478

Contents lists available at ScienceDirect

Clinical Neurophysiology

journal homepage: www.elsevier.com/locate/clinph

Seizure prediction: Any better than chance?

Ralph G. Andrzejak [a,*], Daniel Chicharro [a], Christian E. Elger [b], Florian Mormann [b,c]

[a] Department of Information and Communication Technologies, Universitat Pompeu Fabra, Carrer Roc Boronat 138, 08018 Barcelona, Spain
[b] Department of Epileptology, University of Bonn, Bonn, Germany
[c] Division of Biology, California Institute of Technology, Pasadena, USA

ARTICLE INFO

Article history:
Accepted 23 May 2009
Available online 2 July 2009

Keywords:
Nonlinear dynamical EEG analysis
Epilepsy
Seizure prediction
Surrogates
Monte Carlo simulation

ABSTRACT

Objective: To test whether epileptic seizure prediction algorithms have true predictive power, their performance must be compared with the one expected under well-defined null hypotheses. For this purpose, analytical performance estimates and seizure predictor surrogates were introduced. We here extend the Monte Carlo framework of seizure predictor surrogates by introducing alarm times surrogates.
Methods: We construct artificial seizure time sequences and artificial seizure predictors to be consistent or inconsistent with various null hypotheses to determine the frequency of null hypothesis rejections obtained from analytical performance estimates and alarm times surrogates under controlled conditions.
Results: Compared to analytical performance estimates, alarm times surrogates are more flexible with regard to the testable null hypotheses. Both approaches have similar, high statistical power to indicate true predictive power. For Poisson predictors that fulfill the null hypothesis of analytical performance estimates, the frequency of false positive null hypothesis rejections can exceed the significance level for long mean inter-alarm intervals, revealing an intrinsic bias of these analytical estimates.
Conclusions: Alarm times surrogates offer important advantages over analytical performance estimates.
Significance: The key question in the field of seizure prediction is whether seizures can in principle be predicted or whether algorithms which have been presumed to perform better than chance actually are unable to predict seizures and simply have not yet been tested against the appropriate null hypotheses. Alarm times surrogates can help to answer this question.

1. Introduction

For patients with medically intractable epilepsy, it is the sudden, unforeseen way in which seizures occur that represents one of the most disabling aspects of the disease. Apart from the risk of serious injury, there is often an intense feeling of helplessness that has a strong impact on the everyday life of a patient. In principle, there are two different scenarios of how a spontaneous seizure could evolve (Lopes da Silva et al., 2003). It could be caused by a sudden and abrupt transition, in which case it would not be preceded by detectable dynamical changes in the EEG. Such a scenario would be conceivable for the initiation of seizures in primary generalized epilepsy. Alternatively, this transition could be a gradual change in dynamics, which could in theory be detected. This type of transition could be more likely in focal epilepsies.

Around the turn of the millennium, a number of studies reported that nonlinear and linear signal analysis techniques applied to electroencephalographic (EEG) recordings from epilepsy patients allowed to reliably predict impending epileptic seizures (for a review see Mormann et al. (2007)). If these promising results could have been substantiated, the impact on the therapeutic possibilities for epilepsy patients would have been enormous. Apart from simple warning devices, one could envision automated implantable closed-loop systems that would prevent seizures by applying fast-acting anticonvulsant drugs or by electrical or other stimulation (Mormann et al., 2007).

Until now, however, these early promising results on the predictability of seizures could not be substantiated. Instead recent studies using more rigorous methodological concepts revealed that the performance of these seizure prediction algorithms is far too low to be considered for clinical application (De Clerq and Lemmerling, 2003; Winterhalder et al., 2003; Aschenbrenner-Scheibe et al., 2003; Maiwald et al., 2004; Lai et al., 2004; Harrison et al., 2005a,b; Mormann et al., 2005). Moreover, it became clear that sensitivity, specificity, and overall performance derived from these predictors are difficult to assess. The sensitivity is commonly defined as the number of true positive predictions normalized by the total number of seizures. However, the standard definition of specificity as the number of true negatives normalized by the sum of true negatives and false positives is not applicable because it is not straightforward to define the number of true negative

* Corresponding author.
E-mail address: ralphandrzejak@yahoo.de (R.G. Andrzejak).

1388-2457/$36.00
doi:10.1016/j.clinph.2009.05.019

predictions for seizure prediction algorithms (Mormann et al., 2007). To overcome this problem, specificity is routinely assessed using the average number of false positive predictions per time. The interpretation of such false positive rates however depends not only on the definition of false positives but also on the exact definition of true positives. Moreover, different normalizations are used to convert counts of false positive predictions into false positive rates (Mormann et al., 2007). Furthermore, even if the original predictor lacks any true predictive power, non-zero sensitivity values and low false positive rates can be obtained just by chance (Andrzejak et al., 2003). To test whether certain performance values of the original predictor are indeed indicative of a true predictive power, it is therefore indispensable to compare these values against the performance expected under various well-defined null hypotheses. Besides the central assumption that the seizure predictor lacks any true predictive power, these null hypotheses will generally include further assumptions. Two approaches for such null hypotheses tests have been suggested: analytical performance estimates (Winterhalder et al., 2003; Schelter et al., 2006a; Wong et al., 2007; Snyder et al., 2008) and seizure predictor surrogates (Andrzejak et al., 2003; Kreuz et al., 2004).

The application of these different null hypothesis tests suggested that while the performance of current seizure prediction algorithms would not yet suffice for clinical application some of them at least perform better than chance (Winterhalder et al., 2003; Aschenbrenner-Scheibe et al., 2003; Maiwald et al., 2004; Mormann et al., 2003, 2005; Chaovalitwongse et al., 2005; Schelter et al., 2006a,b; Winterhalder et al., 2006; Sackellares et al., 2006; Schelter et al., 2007; Wong et al., 2007; Schad et al., 2008; Snyder et al., 2008). One interpretation of these findings is that seizures are not completely unpredictable and that the goal to reliably predict them can ultimately be reached. However, an alternative explanation is that these algorithms cannot at all predict seizures and their apparent better-than-random-performance arises merely because they were not tested against an appropriate null hypothesis.

2. Definition of the problem and existing approaches

To determine whether a prediction algorithm performs better than chance, is inevitable to rigorously formulate the tested null hypotheses and the underlying assumptions. We therefore describe the various null hypotheses considered here in Section 2.1. We then discuss two fundamental approaches to estimate the performance expected under such null hypotheses. Section 2.2 outlines analytical approaches which provide formulae to determine the expected performance from the false positive rate of the seizure prediction algorithm. Section 2.3 describes the numerical approach of seizure predictor surrogates, which are based on Monte Carlo simulations.

2.1. Seizure predictors and composition of null hypotheses

As seizure predictor we denote a combination of algorithms used to extract alarms before impending seizures from multi-channel long-term EEG recordings from epilepsy patients which share the following characteristics. First some characterizing measure, e.g. based on the correlation dimension, the largest Lyapunov exponent or some measure derived from the power spectrum is extracted from the EEG using a moving window technique (for reviews see Stam (2005) and Mormann et al. (2007)). The characterizing measure can be extracted for single time series, pairs, or groups of time series. In general, the resulting temporal *measure profile* will be multivariate. This profile or some derivative thereof is then further evaluated for signatures that are considered predictive of impending seizures, resulting in a temporal sequence of seizure prediction alarm times. The crossing of a pre-defined threshold by the characterizing measure is a typical example for such a signature triggering an alarm. These alarm times are then evaluated with regard to their sensitivity and specificity for impending seizures. Given a certain specificity value of a seizure predictor one should test whether its sensitivity is indeed any better than what would be expected by chance. Only in this case is there evidence that the seizure predictor can have true predictive power. To address this issue, one has to first advance from the imprecise expression 'expected by chance' to a precisely formulated null hypothesis. To arrive at such a precise null hypothesis we shall begin by considering different assumptions.

The fundamental assumption $\mathscr{N}$ is that the alarms are raised by a naïve and unspecific predictor, i.e. by a predictor which has no true predictive power. Such a predictor has no access to any information that would be indicative of an impending seizure. Moreover, it uses no information about intervals between previous seizures. Accordingly, the alarm times are in no way related to the times of upcoming seizures. One can make the further assumption $\mathscr{S}_1$ that the predictor is stationary during the entire recording, implying that the alarms are raised at a time-independent mean rate. Alternatively, this assumption can be modified to account for the strong impact post-seizure EEG changes can have on characterizing measures and seizure predictors derived from them (cf. Mormann et al., 2005). A weakened stationarity assumption $\mathscr{S}_2$ is that $\mathscr{S}_1$ is wrong but the predictor exhibits the same time-dependence for all inter-seizures intervals where, importantly, time is measured relative to the preceding seizure. In addition, it can be assumed that the event of the actual seizure does not influence the predictor state ($\mathscr{R}_1$). Conversely, one can assume that an actual seizure resets the predictor to a state which is generally different from the one it assumes after raising an alarm ($\mathscr{R}_2$), or that an actual seizure resets the seizure predictor to the same state it assumes after raising an alarm ($\mathscr{R}_3$). Note that the assumptions $\mathscr{R}_1, \mathscr{R}_2$, and $\mathscr{R}_3$ are mutually exclusive. The intervals between alarms can, for example, be assumed to be uncorrelated and exponentially distributed ($\mathscr{D}$).

These assumptions can be combined in different ways to constitute different null hypotheses (Tables 1 and 2). A general null hypothesis is directly given by the first assumption ($\mathscr{H}_0^{\mathrm{I}} = \mathscr{N}$). However, a direct test of $\mathscr{H}_0^{\mathrm{I}}$ is impossible since in concrete implementations of null hypotheses tests further assumptions are inevitable. Hence, by making further assumptions one arrives at more specific null hypotheses. For example, $\mathscr{H}_0^{\mathrm{V}} = \mathscr{N}$ & $\mathscr{S}_1$& $\mathscr{D}$ is that the alarms arise from a homogenous Poisson process, i.e. from an uncorrelated random process with a time-independent mean alarm rate and an exponential inter-alarm interval distribution. The Tables 1 and 2 cannot represent a lists of all possible assumptions and null hypotheses. Rather they comprise assumptions and null hypotheses that have been used in previous studies as well as further plausible assumptions and null hypotheses that we address

Table 1
Overview of the different assumptions about the original seizure predictor.

$\mathscr{N}$	Unspecific, naïve predictor
$\mathscr{S}_1$	Stationary during the entire recording
$\mathscr{S}_2$	Non-stationary, but same time-dependence for all inter-seizure intervals with time measured relative to previous seizure
$\mathscr{R}_1$	A seizure does not reset the predictor state
$\mathscr{R}_2$	A seizure resets the predictor to a state different from the one it assumes after raising an alarm
$\mathscr{R}_3$	A seizure resets the predictor to the same state it assumes after raising an alarm
$\mathscr{D}$	The intervals between alarms are uncorrelated and exponentially distributed

Table 2
Overview of the different null hypotheses. The letters T and F, respectively, indicate whether the null hypothesis is fulfilled or violated in a particular Section.

Null hypothesis	Assumptions	4.1	4.2	4.3	4.4
$\mathscr{H}_0^{\mathrm{I}}$	$\mathscr{N}$	T	T	T	F
$\mathscr{H}_0^{\mathrm{II}}$	$\mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{R}_3$	F	T	F	F
$\mathscr{H}_0^{\mathrm{III}}$	$\mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{R}_1$	F	F	T	F
$\mathscr{H}_0^{\mathrm{IV}}$	$\mathscr{N}$ & $\mathscr{S}_2$ & $\mathscr{R}_2$	T	F	F	F
$\mathscr{H}_0^{\mathrm{V}}$	$\mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{D}$	F	F	T	F

in the present study. These lists can readily be extended by further assumptions and null hypothesis in future work.

2.2. Analytical sensitivity and performance estimates

An analytical approach to test null hypotheses for epileptic seizure predictors was proposed by Winterhalder et al. (2003), extended by Schelter et al. (2006a), Wong et al. (2007), and Snyder et al. (2008), and applied e.g. in Aschenbrenner-Scheibe et al. (2003), Maiwald et al. (2004), Winterhalder et al. (2006), Schelter et al. (2006b), Schelter et al. (2007) and Schad et al. (2008)). Winterhalder et al. (2003) derived an analytical expression for the maximal sensitivity expected under the null hypothesis $\mathscr{H}_0^{\mathrm{V}}$ for a given specificity of the original seizure predictor. For any performance measure defined as a function of the sensitivity and specificity, the analytical sensitivity estimate can readily be transformed into an analytical performance estimate. If the performance of the original predictor is higher than this analytical estimate, then $\mathscr{H}_0^{\mathrm{V}}$ can be rejected. Care has to be taken, however, in the interpretation of such a rejection as it only provides a necessary but not sufficient condition for a true predictive power of the original predictor. $\mathscr{H}_0^{\mathrm{V}}$ represent the conjunction of assumptions $\mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{D}$ and, accordingly, the violation of any of these assumptions is sufficient for its rejection. If $\mathscr{H}_0^{\mathrm{V}}$ was rejected, it would be informative to test further distinct null hypotheses to collect further evidence. This evidence could either provide further support for a true predictive power of the original predictor or help to rule it out. Unfortunately, apart from the special case of $\mathscr{H}_0^{\mathrm{V}}$, the derivation of analytical performance estimates for arbitrary null hypotheses seems hardly possible. In particular, like also pointed out by Wong et al. (2007), it is not possible to account for time-dependent mean alarm rates.

2.3. Seizure predictor surrogates

The framework of seizure predictor surrogates constructed from constrained randomizations of the original seizure predictor offers greater flexibility than analytical performance estimates. Seizure predictor surrogates can be designed to test a variety of distinct null hypotheses about the original predictor. Any assumption made for the null hypothesis has to be represented by a corresponding property that the surrogates share with the original predictor. Constrained randomization schemes are used to generate surrogates that exhibit these specified properties, but are otherwise random. For example, if assumption $\mathscr{S}_1$ is made, the surrogates must be constrained to be time-independent, regardless of potential time-dependencies of the original predictor. If assumption $\mathscr{D}$ is made, the surrogate must have an exponential inter-alarm interval distribution, regardless of the original distribution. If in contrast $\mathscr{S}_1$ and $\mathscr{D}$ are not assumed, the surrogate should be constrained to share potential time-dependencies and the inter-alarm interval distribution with the original predictor. In particular, any feature of the original predictor which is evidently not related to a true predictive power, but which might influence its predictive performance, should be translated into an assumption and be imposed as a constraint on the surrogates. For example, $\mathscr{S}_2$ can be motivated if post-seizure EEG changes manifest themselves in stereotypical time-dependent features of the predictor, and in this case the seizure predictor surrogates should be constrained to exhibit these same features.

After constructing an ensemble of independent realizations of the surrogates, the performance should be calculated for the original predictor and all surrogates. If the performance of the original predictor is significantly higher than the distribution of performance values obtained for the surrogates, the corresponding underlying null hypothesis can be rejected. As we will illustrate in this article, it is key to the flexibility of this Monte Carlo approach that various distinct null hypotheses can be tested by composing appropriate sets of assumptions and constraints.

The first Monte Carlo approach proposed for the evaluation of seizure prediction statistics was the technique of seizure times surrogates (Andrzejak et al., 2003). For this approach, the original seizure times are replaced with random surrogate seizure times while the original measure profile is kept unchanged. Andrzejak et al. (2003) generated random seizure times by shuffling the original inter-seizure intervals. Thereby the inter-seizure interval distribution was maintained while possible correlations in the seizure time sequence were destroyed. Recently, it has been proposed to constrain seizure times surrogates to preserve possible temporal correlations in the sequence of seizure times and severities (Sunderam et al., 2007). Apart from assumption $\mathscr{N}$, the null hypothesis of seizure times surrogates always includes $\mathscr{R}_1$ because real seizure times are deleted and surrogate seizure times are inserted. Hence, the actual seizures are assumed to have no influence on the seizure predictor. Accordingly, the interval from the last alarm prior to a certain seizure to the first alarm after this seizure is regarded as one continuous inter-alarm interval. The original inter-alarm interval distribution is therefore maintained by construction, and no explicit assumption concerning the inter-alarm interval distribution such as $\mathscr{D}$ can be tested. Furthermore, assumption $\mathscr{S}_1$ is inevitable for seizure times surrogates: the original predictor is assumed to be stationary for the entire recording, thus assumption $\mathscr{S}_2$ cannot be tested. Hence, while seizure times surrogates offer a straightforward and computationally inexpensive way to test the null hypothesis $\mathscr{H}_0^{\mathrm{III}}$, this first type of seizure predictor surrogates does not offer a high flexibility regarding the assumptions which can be included in the null hypothesis. Moreover, some recordings may contain only a few seizures, or recordings can be interrupted by gaps. Both problems can make the generation of a sufficient number of independent realizations of seizure times surrogates impossible, if one wants to preserve the inter-seizure interval distribution. Nevertheless, seizure times surrogates have been applied in a number of different studies (Andrzejak et al., 2003; Mormann et al., 2005; Chaovalitwongse et al., 2005), see also (Mormann et al., 2003).

To overcome problems related to seizure times surrogates, Kreuz et al. (2004) proposed measure profile surrogates. This type of seizure predictor surrogates is generated by randomizing the original measure profiles while keeping the original seizure times unchanged. In Kreuz et al. (2004) this technique was illustrated using the preservation of the autocorrelation function and amplitude distribution of the original measure profile as constraints for the measure profile surrogates. The autocorrelation was calculated for the entire recording rather than in a moving window and also across actual seizure times. Therefore, assumptions $\mathscr{S}_1$ and $\mathscr{R}_1$ were implicitly made. Despite the preservation of the original measure profile's amplitude distribution the original inter-alarm distribution is not preserved by this particular form of measure profile surrogates. In general, not even the total number of alarms will be preserved. However, using the technique of simulated annealing, measure profile surrogates can in principle be constrained to

preserve any feature of the original seizure predictor (Kreuz et al., 2004). For example, the measure profile surrogate can be constrained to result approximately in the original inter-alarm distribution or in some pre-defined distribution such as in assumption $\mathscr{D}$. Similarly, constraints can be implemented in order to test $\mathscr{S}_2$. However, a drawback of this approach is that it can be complicated to implement and computationally expensive. Especially when several constraints are combined, the generation of measure profile surrogates can become prohibitively time consuming. Hence, while measure profile surrogates in principle offer flexibility with regard to the testable null hypotheses, practical issues can render this approach unfeasible.

3. Methods

We here propose the concept of alarm times surrogates as a novel Monte Carlo approach for the evaluation of seizure prediction statistics. As in the case of measure profile surrogates, the original seizure times are kept unchanged, but the original multivariate measure profile is not directly manipulated. Instead the randomization of the original predictor is carried out at the level of the univariate temporal sequences of alarm times. Alarm times surrogates thus have the advantage of being as easy to implement and computationally inexpensive as seizure times surrogates while offering the same flexibility as measure profile surrogates with regard to the imposed constraints and corresponding null hypotheses.

We introduce different types of alarm times surrogates (Section 3.1) and compare them to analytical performance estimates (Section 3.4) on a variety of examples. To relate to earlier work, we use previously published definitions of sensitivity, specificity, and performance (Section 3.2). We deliberately refrain from analyzing seizure predictors extracted from real EEG recordings of epilepsy patients. Rather we generate artificial seizure time sequences and artificial measure profiles (Section 3.3), thereby creating controlled conditions under which we can specifically design artificial original seizure predictors to be consistent or inconsistent with the different assumptions and null hypotheses specified in Tables 1 and 2. Importantly, this also allows us to generate large ensembles of data from which we can estimate the frequency of null hypothesis rejections under well-defined conditions. While the entire analysis presented here is thus based on artificial data, we will use terms such as 'patients', 'EEG recording' or 'seizure times' to make the analogy clear.

3.1. Alarm times surrogates

In the following we describe algorithms to generate different types of alarm times surrogates. Throughout this study we use $q = 19$ alarm times surrogates. Hence, if the null tested hypothesis is correct, there should be a 5% probability that the original performance is higher than the maximal performance of the surrogates as well as a 5% probability that the performance of the original predictor is lower than the minimal performance of the surrogates.

3.1.1. $\mathscr{H}_0^{\mathrm{II}}$ alarm times surrogates

As a first example of alarm times surrogates we consider the null hypothesis $\mathscr{H}_0^{\mathrm{II}}$. That means the null hypothesis of a stationary ($\mathscr{S}_1$), unspecific, naïve ($\mathscr{N}$) predictor which is reset by an actual seizure to the same state it takes after an alarm ($\mathscr{R}_3$). No assumptions are made about the distribution of the inter-alarm intervals. Accordingly, the inter-alarm intervals of the surrogates should be resampled from the inter-alarm interval distribution of the original predictor. Furthermore, to meet assumption $\mathscr{S}_1$ the surrogates should be constructed to be time-independent regardless of possible time-dependencies of the original predictor. To design $\mathscr{H}_0^{\mathrm{II}}$ alarm times surrogates, let us assume that a recording from a patient includes a total of Q seizures and denote their onset times by $t_j^{\mathrm{s}}\,(j = 1,\ldots,Q)$. We shall further assume that the recording was started immediately after a seizure and index this seizure with $j = 0$ to facilitate the notation. Finally, the recording is assumed to end directly after the last seizure with index $j = Q$. For simplicity, the duration of the seizure is neglected and set to zero. We denote by d_j^{s} the interval between the seizures $j-1$ and j. The times of alarms that are raised during d_j^{s} are denoted by $t_{j,l}^{\mathrm{a}}\,(l = 1,\ldots,A_j)$. The interval between two consecutive alarms $l-1$ and l is denoted by $d_{j,l}^{\mathrm{a}}$ for $l > 1$. By $d_{j,1}^{\mathrm{a}}$ we denote the interval from the seizure $j-1$ to the first alarm. The distribution of $d_{j,l}^{\mathrm{a}}$ across all seizures j and alarms l is denoted by $\hat{\rho}^*$. Depending on the length of the recording and mean inter-alarm interval, the distribution $\hat{\rho}^*$ can be a very limited sample and poor estimate of the true inter-alarm interval distribution of the original predictor. In particular, this $\hat{\rho}^*$ will be biased towards shorter inter-alarm intervals, as explained below. Fortunately, this bias can be reduced substantially using the correction scheme described hereafter.

Let ρ denote the inter-alarm interval distribution of an original naïve and unspecific seizure predictor and let us assume that during d_j^{s} a number of $l_0 \leq A_j$ alarms have already been raised by the predictor. Let us further assume that the remaining time from the l_0-th alarm to the subsequent seizure is short, i.e. of the order of a typical inter-alarm interval. The predictor would raise alarm $l_0 + 1$ after an interval according to a further random sample from ρ. Evidently, the shorter this $l_0 + 1$-th interval, the higher the probability that it fits into the remaining time before the subsequent seizure. Long intervals are simply more likely to be interrupted by the event of a seizure which is assumed to reset the predictor according to $\mathscr{R}_3$. Therefore, $\hat{\rho}$ overestimates the probability of short intervals in ρ. This problem can be treated by taking into account also those intervals that are interrupted by a seizure and are thereby cut short: suppose once again that A_j alarms were raised during d_j^{s}. This means that alarm $A_j + 1$ was not raised because seizure $j + 1$ took place. We know that the interrupted interval between the alarms A_j and $A_j + 1$ would have been at least as long as the interval from the alarm A_j to this seizure. Instead of ignoring this unfinished interval, which would lead to the bias just described, one can estimate its length from a random sample drawn from the subset of all intervals in $\hat{\rho}^*$ that are longer than the interval from alarm A_j to seizure $j + 1$. Denoting this random sample by d_{j,A_j+1}^{a}, an improved estimate $\hat{\rho}$ of ρ can be obtained from the set of all $d_{j,l}^{\mathrm{a}}$ across $j = 1,\ldots,Q$ and $l = 1,\ldots,A_j + 1$. It is important to note that this improved estimate can still have a remaining bias towards shorter inter-alarm intervals. Once $\hat{\rho}$ has been determined, the generation of an alarm times surrogate is straightforward. Surrogate intervals $\tilde{d}_{1,l}^{\mathrm{a}}$ for the first interval d_1^{s} are drawn with replacement from $\hat{\rho}$ until the alarm time $\tilde{t}_{1,\tilde{A}_1+1}^{\mathrm{a}}$ falls after the time of the first seizure and this alarm is discarded. The same procedure is then carried out for the remaining inter-seizure intervals. Note that under the given constraints $\tilde{A}_j$ can differ from A_j. Furthermore, $A_j \geqslant 0$, and $\tilde{A}_j \geqslant 0$.

3.1.2. $\mathscr{H}_0^{\mathrm{III}}$ alarm times surrogates

To generate $\mathscr{H}_0^{\mathrm{III}}$ alarm times surrogates, only slight modifications of the scheme described in Section 3.1.1 are necessary. To account for assumption $\mathscr{R}_1$, the interval $d_{j,1}^{\mathrm{a}}$ is not defined from the beginning of the inter-seizure interval j to the first alarm in this interval but rather by the time from the last alarm of the preceding

inter-seizure interval to this first alarm. Furthermore, $\mathscr{R}_1$ implies that the only unfinished inter-alarm interval is the one interrupted by the end of the recording. Accordingly, the correction scheme for the estimation of $\hat{\rho}$ from $\hat{\rho}^*$ should only be applied for this last interval. Surrogate intervals are generated starting at $\tilde{d}^{\mathrm{a}}_{1,l}$ by drawing with replacement from $\hat{\rho}$ until the alarm time $\tilde{t}^{\mathrm{a}}_{Q,\widetilde{A}_Q+1}$ falls after the end of the recording.

Since the assumption that the intervals between alarms are uncorrelated and exponentially distributed ($\mathscr{D}$), implies that the predictor is memoryless, there are no correlations between subsequent predictor states regardless of whether or not a seizure took place between them. Therefore, an actual seizure does not influence the predictor ($\mathscr{R}_1$). Hence, assumption $\mathscr{D}$ implies $\mathscr{R}_1$, and accordingly $\mathscr{H}_0^{\mathrm{V}}$ implies $\mathscr{H}_0^{\mathrm{III}}$. In consequence, a test based on $\mathscr{H}_0^{\mathrm{III}}$ alarm times surrogates should not be rejected if $\mathscr{H}_0^{\mathrm{V}}$ is valid.

3.1.3. $\mathscr{H}_0^{\mathrm{IV}}$ alarm times surrogates

In contrast to the cases of $\mathscr{H}_0^{\mathrm{II}}$ and $\mathscr{H}_0^{\mathrm{III}}$ surrogates we now drop the stationarity assumption ($\mathscr{S}_1$) and assume instead that the seizure predictor shows a time-dependence which is time-locked to the previous seizure ($\mathscr{S}_2$), as e.g. in the case of a post-seizure state that influences a characterizing measure. Accordingly, the inter-alarm intervals of the surrogates must follow this time-dependence of the original predictor. We illustrate this approach using the following specific example. Suppose that the inspection of the original predictor revealed that the intervals from the preceding seizure to the first alarm are significantly longer than all subsequent intervals prior to the next seizure, which in turn all seem to originate from the same distribution. That is, the distribution of intervals $d^{\mathrm{a}}_{j,1}$ $(j = 1,\ldots,Q)$, which we denote by $\hat{\rho}^*_1$, is different from the one obtained from all $d^{\mathrm{a}}_{j,l}$ $(j = 1,\ldots,Q,\ l = 2,\ldots,A_j)$, which we denote by $\hat{\rho}^*_2$. Consequently, the alarm times of the original predictor exhibit some non-random structure that is evidently not reflecting any true predictive power because it is time-locked to the preceding rather than to the subsequent seizure. Therefore, assumption $\mathscr{S}_2$ is fulfilled, and one should constrain the alarm times surrogates to exhibit the same time-dependence by generating $\hat{\rho}_1$ and $\hat{\rho}_2$ separately: using the scheme described in Section 3.1.1, one should estimate $\hat{\rho}_2$ exclusively from $\hat{\rho}^*_2$ by drawing the additional samples d^{a}_{j,A_j+1} from $\hat{\rho}^*_2$. Importantly, this implies that these additional samples are only taken for inter-seizure intervals with $A_j \geqslant 2$. If during every inter-seizure interval the original predictor raised at least one alarm, one can directly use $\hat{\rho}_1 = \hat{\rho}^*_1$. However, if there are inter-seizure intervals without any alarms, i.e. some intervals to the first alarm were cut short by the seizure, then these unfinished intervals should be estimated from additional samples $d^{\mathrm{a}}_{j,1}$ drawn from $\hat{\rho}^*_1$ to derive $\hat{\rho}_1$. Once $\hat{\rho}_1$ and $\hat{\rho}_2$ have been determined, a surrogate is generated as described above for the $\mathscr{H}_0^{\mathrm{II}}$ alarm times surrogates. For the case of $\mathscr{H}_0^{\mathrm{IV}}$ alarm times surrogates, however, the first surrogate alarm time $\tilde{d}^{\mathrm{a}}_{j,1}$ after a seizure is always drawn from $\hat{\rho}_1$, and all subsequent intervals up to $\tilde{d}^{\mathrm{a}}_{j,\widetilde{A}_j+1}$ are drawn from $\hat{\rho}_2$. Like for $\mathscr{H}_0^{\mathrm{II}}$ surrogates, the last resulting alarm time $\tilde{t}^{\mathrm{a}}_{j,\widetilde{A}_j+1}$ falling after seizure j is discarded. Jointly these samples result in $\tilde{d}^{\mathrm{a}}_{j,1,\ldots,\widetilde{A}_j}$. Note that even the sample for the first surrogate alarm time interval $\tilde{d}^{\mathrm{a}}_{j,1}$ determined by the sample from $\hat{\rho}_1$ can result in a surrogate alarm time $\tilde{t}^{\mathrm{a}}_{j,1}$ falling after seizure j. In this case this first alarm is discarded, and we obtain $\widetilde{A}_j = 0$.

Note that this algorithm is adapted to the particular type of time-dependence studied here: the intervals to the first alarms are longer than all subsequent intervals. Indeed, our aim is not to propose a general-purpose algorithm to construct alarm times surrogates for original predictors with *arbitrary* time-dependencies. Rather we here suggest to carefully inspect the original predictor for potential time-dependencies and then adapt the algorithm described in Section 3.1.1 accordingly. The procedure described in this section is meant as one example for such an adaptation. Importantly, the concept of constrained randomization generally allows for such adaptations.

3.2. Quantification of sensitivity, specificity, and performance

Suppose that a multi-channel EEG recorded from an epilepsy patient has a total duration of d hours and includes Q seizures. Let us further assume that some characterizing measure was extracted from this recording by means of a moving window technique, and that the temporal profile of this measure is evaluated for signatures that are assumed to be predictive for impending seizures, resulting in a univariate temporal sequence of alarm times. To analyze whether these alarms have any true predictive power one first has to quantify their sensitivity and specificity. For this purpose we define the periods directly preceding the seizures as prediction horizons. The length of these prediction horizons in hours is denoted by h and assumed to be the same for all seizures. We assume that the minimal distance between two consecutive seizures is not shorter than this prediction horizon. Cases where at least one alarm is raised within the prediction horizon of a seizure are counted as true positive predictions. Seizures for which no alarm is raised during the prediction horizon count as false negative predictions. All alarms outside of any prediction horizon are counted as false positive predictions. The sensitivity is given by the ratio of the total number of true positive predictions (P^+) to the total number of seizures:

$$S = \frac{P^+}{Q} \tag{1}$$

Several ways exist to derive a specificity value from false positive predictions. To relate to other studies (e.g. Winterhalder et al., 2003; Aschenbrenner-Scheibe et al., 2003; Mormann et al., 2003; Maiwald et al., 2004; Chaovalitwongse et al., 2005; Iasemidis et al., 2003; Schelter et al., 2006a,b; Sackellares et al., 2006; Winterhalder et al., 2006; Schelter et al., 2007; Schad et al., 2008) we here use the false positive rate. This should be determined by dividing the total number of false positive alarms (P^-) by the total time in hours during which such false alarms can occur. This time is given by the duration of the recording minus the total time covered by the prediction horizons, since by construction no false alarms can occur during the prediction horizons:

$$F = \frac{P^-}{d - Qh} \tag{2}$$

As definition of the performance we use (cf. Chaovalitwongse et al., 2005):

$$P(S,F) = 1 - \sqrt{(1-S)^2 + \frac{F^2}{F_0^2}} \tag{3}$$

Here we use $F_0 = 1\ \mathrm{h}^{-1}$ to turn also the second summand in the square-root into a dimensionless quantity and since one per hour is commonly used as unit for false positive rates (Winterhalder et al., 2003; Aschenbrenner-Scheibe et al., 2003; Iasemidis et al., 2003; Maiwald et al., 2004; Mormann et al., 2003; Chaovalitwongse et al., 2005; Schelter et al., 2006a,b; Winterhalder et al., 2006; Sackellares et al., 2006; Schelter et al., 2007; Schad et al., 2008). For $S = 1$ and $F = 0$ we get $P = 1$, while lower values are obtained for deviations from this perfect predictor. However, the performance is not normalized. For poor predictors also negative values can be obtained.

3.3. Artificial seizure times, characterizing measures, and alarm times

We generated artificial seizure time sequences to represent the seizure times included in a continuous EEG recording from an individual patient. In total we generated artificial data for an arbitrary but high number of $K = 100,000$ patients. The purpose of using such a large patient ensemble is to reliably derive the frequencies of rejecting and accepting the different null hypotheses. In the following, we describe the procedure carried out for each individual patient using the notation and index ranges introduced in Section 3.1. For simplicity we use no additional index to identify individual patients.

For each patient we drew $Q = 15$ random inter-seizure intervals d_j^s, the lengths of which were, unless stated otherwise, uniformly distributed between 2 and 14 h. The concatenation of these intervals resulted in a random sequence of 15 seizure times t_j^s. Accordingly, across patients, these recordings had an average duration of 5 days = 120 h (cf. Fig. 1).

3.3.1. Non-stationary integrate-and-fire predictor

For the setting in Section 4.1, we used a simple non-stationary stochastic integrate-and-fire process to generate artificial measure profiles. These measure profiles are supposed to be extracted from the EEG using a moving window technique and are denoted by $m_{i,j}$ for $i = 1,\ldots,N_j$, where N_j is the number of analysis windows in the j-th inter-seizure interval. The measure profile was initialized to zero and held at this value during a post-seizure delay period D_j, i.e. $m_{i,j} = 0$ for $i = 1,\ldots,D_j$. Values for subsequent analysis windows were generated according to:

$$m_{i+1,j} = \begin{cases} 0 & \text{for } |m_{i,j}| = r \\ \begin{cases} m_{i,j} - 1 & \text{with probability } 0.5 + b \\ m_{i,j} + 1 & \text{else} \end{cases} & \text{else} \end{cases} \tag{4}$$

using $0 < b < 0.5$. In the first of these cases, in addition to resetting the predictor an alarm was raised at this window: $t_{j,l}^a = i$. Once the time index reached $i = N_j$, the process was stopped and re-initialized for the subsequent inter-seizure interval. We set the post-seizure delay period D_j to be uniformly and randomly distributed between 1250 and 1750 windows. Assuming the length of individual windows to be 20 s, this corresponds to 6.94 and 9.72 h, respectively. By construction no alarms could occur during this post-seizure delay period, which is meant to simulate the well-known post-seizure EEG alterations. The mean length and distribution of the intervals between subsequent alarms can be influenced by the parameters b and r. For the non-stationary integrate-and-fire predictor we used $b = 0.125$ and $r = 40$. An example for a temporal profile of this predictor is shown in Fig. 1.

Since the alarm times are random and have no true predictive power for the seizure times, the assumption $\mathscr{N}$ holds. The assumption $\mathscr{D}$ does not hold since the inter-alarm interval distribution is not exponential. Furthermore, due to the post-seizure delay period $\mathscr{S}_2$ as well as $\mathscr{R}_2$ hold. Hence, for this non-stationary integrate-and-fire predictor $\mathscr{H}_0^{\mathrm{I}}$ and $\mathscr{H}_0^{\mathrm{IV}}$ are valid.

3.3.2. Stationary integrate-and-fire predictor

For the setting in Section 4.2, we used a stationary stochastic integrate-and-fire predictor. This was constructed identical to the non-stationary integrate-and-fire predictor but with the post-seizure delay period D_j set to zero. The parameters were changed to $b = 0.055$ and $r = 70$ to obtain a similar false positive rate as for the non-stationary case. As opposed to the non-stationary case, for the stationary integrate-and-fire predictor $\mathscr{S}_1$ holds instead of $\mathscr{S}_2$, and $\mathscr{R}_3$ holds instead of $\mathscr{R}_2$. On the other hand, $\mathscr{N}$ still holds, and $\mathscr{D}$ still does not hold. Hence, for this stationary integrate-and-fire predictor $\mathscr{H}_0^{\mathrm{I}}$ and $\mathscr{H}_0^{\mathrm{II}}$ are valid.

3.3.3. Poisson predictor

For the setting in Section 4.3, we did not use artificial measure profiles but rather generated alarm times directly from a homogenous Poisson process. Such a Poisson process has an exponential inter-alarm interval distribution, and its only parameter is the mean inter-alarm interval $F^{\mathscr{P}}$ which we specify in units h^{-1}. In addition to $\mathscr{D}$ the Poisson predictor also fulfills $\mathscr{N}, \mathscr{S}_1$, and $\mathscr{R}_1$. In consequence $\mathscr{H}_0^{\mathrm{I}}$ and $\mathscr{H}_0^{\mathrm{V}}$ are true.

3.3.4. Non-naïve predictors

To determine not only the statistical size but also the power of the different null hypotheses tests, we used two different non-naïve predictors with variable degrees of true predictive power for the setting in Section 4.4. For this purpose, we first constructed a non-naïve predictor. This predictor raised alarms in the prediction horizons of s randomly selected seizures, with $s = 0,\ldots,15$.

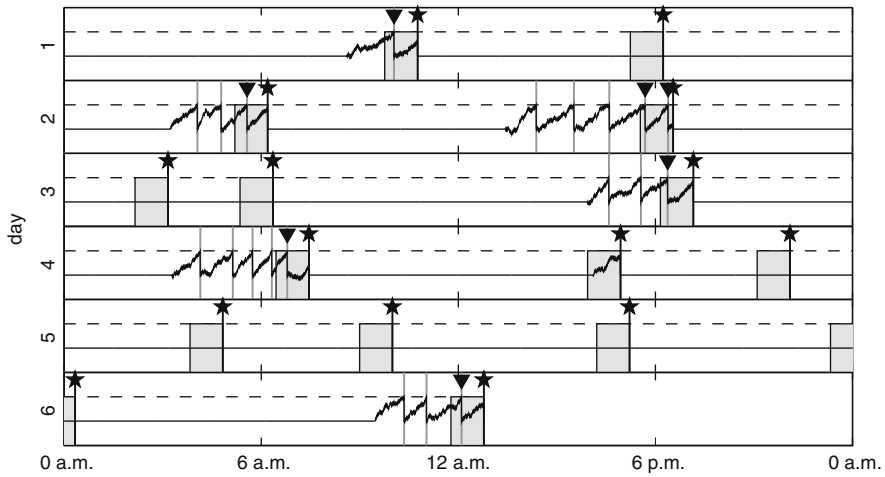

Fig. 1. Scheme of events during an EEG recording with a duration of approximately 5.5 days. Seizure times are shown by vertical lines with stars. The preceding prediction horizons are depicted as gray frames. The measure profile of the non-stationary integrate-and-fire predictor (Section 3.3.1) is depicted as black temporal profile. Alarms, which are triggered whenever the measure profile reaches a pre-defined threshold (depicted as dashed horizontal lines), are shown by gray vertical lines. Whenever an alarm falls into a prediction horizon, this is indicated by a triangle. Cases in which more than one alarm falls into the same prediction horizon (e.g. for the fourth seizure) contribute only one count to the total number of true positive predictions (see Section 3.2). Every alarm outside of any prediction horizon is counted as a false positive alarm.

The times of these alarms were set randomly and uniformly distributed in the prediction horizons.

For a first example we combined this non-naïve predictor with the naïve non-stationary integrate-and-fire predictor (Section 3.3.1). The value of s determines the average sensitivity of the non-naïve part of the predictor. To relate this sensitivity to the influence of the naïve part, we use the percentage of alarms raised by the non-naïve part. Pre-analysis showed that for the given values of b, r, and D_j and averaged across the entire patient group, the naïve integrate-and-fire predictor raises 19.85 alarms during the recording. Hence, this percentage is given by $s^* = 100\frac{s}{s+19.85}$. For a second example we combined the non-naïve predictor with the naïve Poisson predictor (Section 3.3.3, $F^{\mathscr{P}} = 0.33\ \mathrm{h}^{-1}$). Given that the average duration of the recordings is 120 h, the percentage of alarms raised by the non-naïve part of the predictor is given by $s^* = 100\frac{s}{s+40}$. The first and second example will be referred to as hybrid non-naïve integrate-and-fire predictor and hybrid non-naïve Poisson predictor, respectively. For $s > 0$ both predictors violate the assumption $\mathscr{N}$, and in consequence all null hypotheses in Table 2 are false.

3.4. Analytical estimates

Suppose an original seizure predictor fulfills $\mathscr{H}_0^{\mathrm{V}}$ and has a false positive rate of F per time unit. The probability that this predictor raises at least one alarm in a prediction horizon of h time units is approximately:

$$p_1 = p(\mathrm{alarm}|F,h,\mathscr{H}_0^{\mathrm{V}}) \approx 1 - \mathrm{e}^{-Fh} \tag{5}$$

(see Winterhalder et al., 2003; Schelter et al., 2006a). The probability that the predictor reaches a sensitivity of at least S_0 thus is:

$$p(S \geqslant S_0|F,h,\mathscr{H}_0^{\mathrm{V}}) = \sum_{j \geqslant q = QS_0}^{Q} \binom{Q}{j} p_1^j (1-p_1)^{Q-j} \tag{6}$$

For a designated significance level α the analytical sensitivity estimate is defined as:

$$S_{\mathrm{A}}(F,h) = \max\{S_0 \mid p(S \geqslant S_0|F,h,\mathscr{H}_0^{\mathrm{V}}) > \alpha\} \tag{7}$$

(Winterhalder et al., 2003; Schelter et al., 2006a). Hence, the probability that the original Poisson predictor exceeds S_A is less than α. Throughout the study we use $\alpha = 5\%$ as significance level. The analytical performance estimate is defined by using Eq. (3):

$$P_{\mathrm{A}}(F,h) = 1 - \sqrt{(1-S_{\mathrm{A}}(F,h))^2 + \frac{F^2}{F_0^2}} \tag{8}$$

Winterhalder et al. (2003) gave an analytical sensitivity estimate for a periodic predictor with an alarm rate of F. The probability that this periodic predictor raises at least one alarm in a prediction horizon of h is: $p_1^* = \min\{1, Fh\}$. For the typical case of Fh considerably smaller than 1 one can approximate Eq. (5) by p_1^* (Schelter et al., 2006a), the Poisson and periodic predictor become equivalent. We therefore restrict our evaluation of analytical estimates in this study to the Poisson predictor.

4. Results

4.1. Non-stationary integrate-and-fire predictor

For the first setting we apply the naïve non-stationary integrate-and-fire predictor for which $\mathscr{H}_0^{\mathrm{I}}$ and $\mathscr{H}_0^{\mathrm{IV}}$ are valid (Section 3.3.1). The prediction horizon was set to $h = 1$h. Fig. 2 shows detailed results obtained under this setting for 10 exemplary patients, and Fig. 3 summarizes the rejection frequencies of the different null hypotheses for the entire ensemble of $K = 100,000$ patients. In almost 80% of the patients the performance of the original predictor is higher than the analytical performance estimate. Given that we used a significance level of 5% for the analytical performance estimate, the underlying null hypotheses $\mathscr{H}_0^{\mathrm{V}}$ is clearly rejected. This is primarily due to the non-stationarity caused by the post-seizure delay period: consider a certain inter-seizure interval j and suppose that the first alarm was raised before the onset of the prediction horizon of the subsequent seizure ($t_{j,1}^{\mathrm{a}} < t_j^{\mathrm{s}} - h$, see e.g. the third seizure in Fig. 1). Due to the non-zero post-seizure delay period D_j, the interval $d_{j,1}^{\mathrm{a}}$ spanning the time between the preceding seizure at t_{j-1}^{s} and this first alarm at $t_{j,1}^{\mathrm{a}}$ is long while the period from this first alarm to the subsequent seizure at t_j^{s} is covered by shorter inter-alarm intervals: $d_{j,2}^{a}, \ldots, d_{j,A_j}^{\mathrm{a}}$, resulting in a time-dependent false positive rate. What is used to calculate the analytical performance estimate is an average false positive rate obtained from all intervals between any previous seizure at t_{j-1}^{s} and the onsets of the prediction horizon of the subsequent one at $t_j^{\mathrm{s}} - h$. Thereby, the false positive rate reflects the mean alarm rate covering one long and a number of short inter-alarm intervals. What is decisive for the actual sensitivity of the naïve random seizure predictor, however, are the short inter-alarm intervals $d_{j,2}^{\mathrm{a}}, \ldots, d_{j,A_j}^{\mathrm{a}}$ after $t_{j,1}^{\mathrm{a}}$. Furthermore, given the distribution of inter-seizure intervals and the length of the post-seizure delay period, there are inter-seizure intervals for which $t_j^{\mathrm{s}} - h < t_{j,1}^{\mathrm{a}} < t_j^{\mathrm{s}}$ or for which no alarm is raised at all (see e.g. first and second seizure, respectively, in Fig. 1). The false positive rate is zero in both cases, and the sensitivity is one and zero, respectively. In consequence, these two cases contribute further to a deviation of the actual performance of the predictor from the analytical estimate. Overall, due to the time-dependence of the original predictor, its performance is higher than the analytical performance estimate. This deviation in no way reflects any inherent incorrectness of the analytical performance estimate. Instead it simply reflects that one of the underlying assumptions, namely the stationarity assumption $\mathscr{S}_1$, is violated. Accordingly, the tested null hypothesis

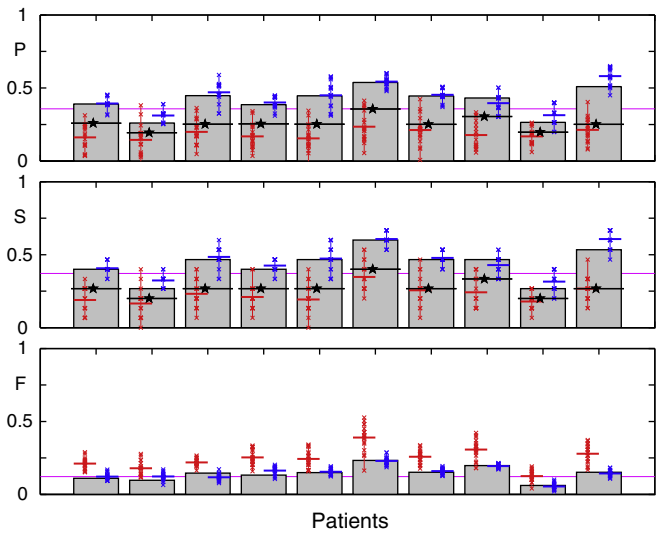

Fig. 2. Results for the non-stationary integrate-and-fire predictor (Section 4.1) obtained for 10 exemplary patients. Top, performance P; middle, sensitivity S; bottom, false positive rate F. Gray bars represent results obtained for the original predictor. Magenta horizontal lines depict the corresponding mean value taken across the entire group of $K = 100,000$ patients. Red and blue crosses represent values obtained for 19 $\mathscr{H}_0^{\mathrm{II}}$ and $\mathscr{H}_0^{\mathrm{IV}}$ surrogates, respectively. Blue and red horizontal bars indicate the corresponding mean values of the surrogates. Black bars with asterisks indicate analytical sensitivity and performance estimates. The leftmost case corresponds to the example shown in Fig. 1.

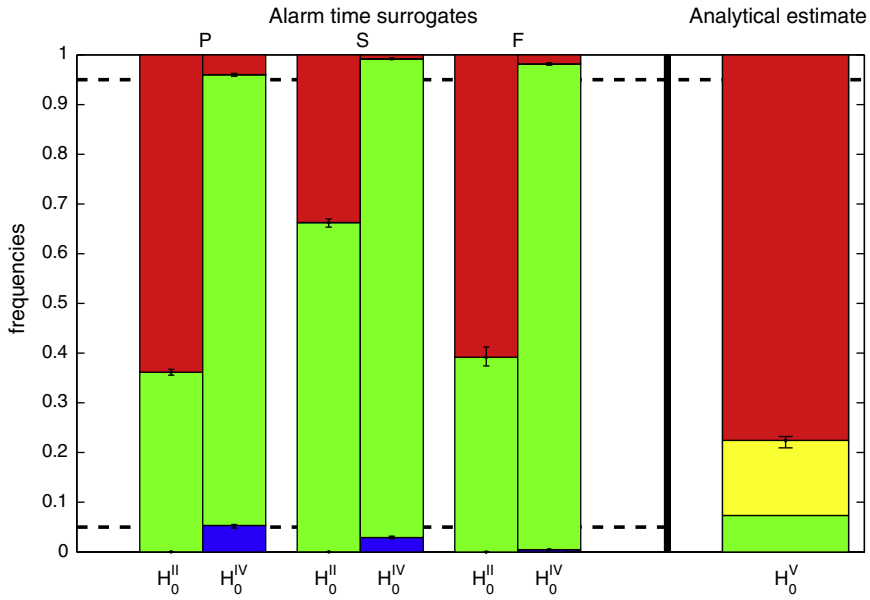

Fig. 3. Results for the non-stationary integrate-and-fire predictor (Section 4.1) obtained for the entire patient group for both $\mathscr{H}_0^{II}$ and $\mathscr{H}_0^{IV}$ alarm times surrogates as well as for the $\mathscr{H}_0^{V}$ analytical performance estimate. For the performance P and the sensitivity S, the blue, green, and red fractions represent the percentages of the 100,000 patients for which the performance of the original predictor was lower than the minimal value for all 19 surrogates, within the surrogate distribution, and higher than the maximal value for all surrogates, respectively. For the false positive rate F bars in blue, green, and red indicate the fractions for which the value for the original predictor was higher than the maximal value for all surrogates, within the surrogate distribution, lower than the minimal value for all surrogates, respectively. The rightmost stacked bars depict in green, yellow, and red the percentages for which the performance of the original predictor was lower, equal, and higher, respectively, than the analytical performance estimate. To derive confidence intervals all values were determined from 10 non-overlapping sub-divisions of 10,000 patients each. Bars depict the mean values across these sub-divisions, error bars depict the corresponding ranges. The dashed lines indicate the 5% significance levels.

$\mathscr{H}_0^{V}$ is correctly rejected. In fact, the assumption of uncorrelated and exponentially distributed inter-alarm intervals ($\mathscr{D}$) is also violated, and we illustrate the relevance of this assumption in Section 4.2.

The performance values of the $\mathscr{H}_0^{II}$ alarm times surrogates do not reach the original predictor either (Figs. 2 and 3). For this surrogate type, the inter-alarm intervals are randomized without constraints. Therefore, the long intervals which were located at the beginning of the inter-seizure intervals for the original predictor ($d^a_{1,1},\ldots,d^a_{Q,1}$) are randomly distributed within the inter-seizure intervals for the alarm times surrogate. In particular, they can overlap with, or even completely cover the prediction horizons. The latter case reduces the sensitivity, and in both cases, more of the short inter-alarm intervals fall into periods not covered by prediction horizons, thereby increasing the false positive rate. Overall, the surrogates' performance is lower than that of the original predictor, the null hypothesis $\mathscr{H}_0^{II}$ is thus rejected. As in the case of the analytical performance estimate, this does not reflect a shortcoming of the applied surrogate technique but a violation of the underlying stationarity assumption $\mathscr{S}_1$ for the original predictor. In consequence, the rejection of $\mathscr{H}_0^{II}$ is correct.

So far we have established that the corresponding null hypotheses tested by the $\mathscr{H}_0^{V}$ analytical estimate and $\mathscr{H}_0^{II}$ alarm times surrogates were rejected correctly, because some of the underlying assumptions were not met by the original predictor. Importantly, while the analytical approach cannot be adjusted to account for deviations from these assumptions, alarm times surrogates can be constructed to test a different null hypothesis. Before we proceed, however, we should recall that we made use of controlled conditions. We used a post-seizure delay period to construct a non-stationary predictor. What if we had no access to this *a priori* knowledge but rather our only information were the alarm times and the rejections of $\mathscr{H}_0^{II}$ and $\mathscr{H}_0^{V}$? In that case, one should scrutinize the original predictor for possible non-stationary features. Indeed plots of the intervals to the first alarm in each inter-seizure interval pooled across all seizures of a given patient, measured in one instance with respect to the preceding seizure and in another instance with respect to the subsequent seizure, reveal an evident non-stationarity (Fig. 4).

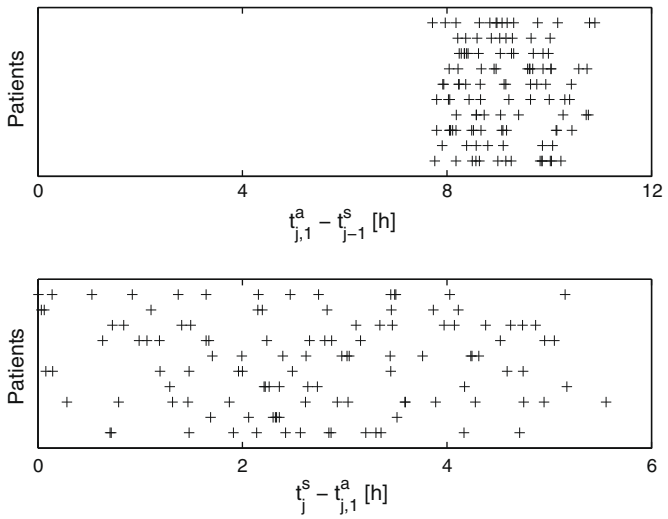

Fig. 4. (upper) Lengths of intervals from seizures with index $j-1$ to the first alarms in the subsequent inter-seizure intervals ($d^a_{j,1} = t^a_{j,1} - t^s_{j-1}$). Rows contain sets of intervals obtained for 10 exemplary patients (the same patients as in Fig. 3). Each marker represents an interval obtained for an individual inter-seizure interval. (lower) Analogous to upper panel but here showing intervals from the first alarms in inter-seizure intervals j to the subsequent seizures ($t^s_j - t^a_{j,1}$). The variability relative to the mean interval length is evidently higher in the lower panel, providing evidence that the times of the first alarms are related to the preceding but not to the subsequent seizures.

This non-stationarity is time-locked to each previous but not to the subsequent seizure and thereby does not indicate any true predictive power of the predictor. Accordingly, the reason for the rejections of $\mathcal{H}_0^{\mathrm{II}}$ and $\mathcal{H}_0^{\mathrm{V}}$ can be deduced empirically from the data without the use of any *a priori* knowledge. These plots furthermore provide evidence that while the stationarity assumption $\mathcal{S}_1$ is clearly violated, the assumption of the same time-dependence for all inter-seizure intervals with time measured relative to the previous seizure ($\mathcal{S}_2$) seems plausible. This assumption is backed by Fig. 1 which shows that only the first intervals $d_{j,1}^{\mathrm{a}}$ stand out from the remaining intervals $d_{j,2}^{\mathrm{a}}, \ldots, d_{j,A_j}^{\mathrm{a}}$. Hence, one should drop $\mathcal{S}_1$ and instead use $\mathcal{S}_2$, which corresponds to testing $\mathcal{H}_0^{\mathrm{IV}}$ instead of $\mathcal{H}_0^{\mathrm{II}}$ (Table 2). As stated in Section 3.1.2, the construction of $\mathcal{H}_0^{\mathrm{IV}}$ surrogates always needs to be adapted to the particular time-dependence found for the original predictor. The corresponding algorithm for the case considered here, i.e. the interval to the first alarm stands out from all subsequent inter-alarm intervals, has been described in Section 3.1.2.

Results for exemplary patients shown in Fig. 2 indicate that $\mathcal{H}_0^{\mathrm{IV}}$ alarm times surrogates indeed match the performance, sensitivity, and false positive rate of the original predictor. This is further substantiated by statistics derived from the entire patient ensemble (Fig. 3) which show that the performance of the original seizure predictor exceeds the maximal performance of all surrogates in 4.0% of the cases (range for ten sub-divisions of $\frac{K}{10}$ patients each: 3.8–4.4%). The performance of the original seizure predictor is smaller than the minimal performance of all surrogates in 5.3% of the cases (range: 4.8–5.5%). Hence, the empirical size of the $\mathcal{H}_0^{\mathrm{IV}}$ surrogates is close to the significance level of 5%. This match is not perfect, and somewhat larger deviations are found for the sensitivity and false positive rate in this example. However, as already indicated in Section 3.1.1 and further discussed in Section 5, the surrogates cannot be expected to perfectly match the original predictor. We here regard the empirical size as consistent with the significance level of the test and interpret these results to provide strong evidence that $\mathcal{H}_0^{\mathrm{IV}}$ is true, which is in fact correct. At first sight the direction of the remaining deviations might suggest that the test is too conservative against accepting a true predictive performance of the original predictor. However, results of this first setting cannot be conclusive for this issue. The power of a test can only be studied when the null hypothesis is false, and we return to this point in Section 4.4.

4.2. Stationary integrate-and-fire predictor

For this setting, we use the stationary integrate-and-fire predictor for which $\mathcal{H}_0^{\mathrm{I}}$ and $\mathcal{H}_0^{\mathrm{II}}$ are valid (Section 3.3.1). The prediction horizon was again set to $h = 1$h. Comparing Figs. 2 and 5 we find that the performance of the non-stationary integrate-and-fire predictor in Section 4.1 is substantially higher than the performance of the stationary predictor studied in the current setting. This further illustrates the effect of the non-stationarity induced by the non-zero post-seizure delay period in the former setting. In the current setting there is no such non-stationarity, and in consequence $\mathcal{H}_0^{\mathrm{II}}$ alarm times surrogates match the original predictor within the expected accuracy (Figs. 5 and 6). Hence, this null hypothesis is correctly accepted. In contrast, the performance of the original predictor is still higher than the analytical performance estimate in 15.4% of the cases, and thus $\mathcal{H}_0^{\mathrm{V}}$ is correctly rejected, although the predictor has no true predictive power. This result illustrates once again that $\mathcal{H}_0^{\mathrm{V}}$ represents the conjunction of three assumptions and that the violation of any of these assumptions, in the current setting the assumption of uncorrelated and exponentially distributed inter-alarm intervals ($\mathcal{D}$), is sufficient for a rejection of this null hypothesis.

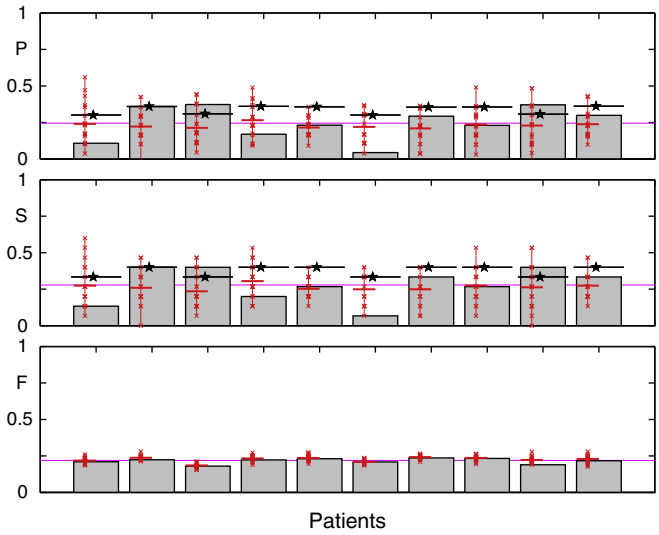

Fig. 5. Same as Fig. 2, but for the stationary integrate-and-fire predictor (Section 4.2) and only for $\mathcal{H}_0^{\mathrm{II}}$ surrogates.

4.3. Poisson predictor with variable mean alarm rate

In this setting we use the Poisson predictor (Section 3.3.3) for a range of mean inter-alarm intervals $F^{\mathscr{P}}$ and a prediction horizon of 2 h. Accordingly, $\mathcal{H}_0^{\mathrm{V}}$ is true and all assumptions underlying the analytical performance estimate are fulfilled. Nonetheless, for lower $F^{\mathscr{P}}$, the rejection frequency for the analytical performance estimate exceeds the significance level of 5%. For $F_{\mathscr{P}} = 1/26\,h = 0.0385\ \mathrm{h}^{-1}$ the null hypothesis is rejected in 7.8% of the cases (range: 7.2–8.2%, Fig. 7a). The reason for this mismatch are variations in F estimated for individual patients. Averaged across the entire patient ensemble, the mean F always matches the actual $F^{\mathscr{P}}$, regardless of the actual value of $F^{\mathscr{P}}$. For decreasing $F^{\mathscr{P}}$ values, however, the uncertainty in the estimated F for individual patients increases. In consequence, for individual patients F can be substantially lower or higher than the actual $F^{\mathscr{P}}$. For those patients for which $F^{\mathscr{P}}$ is underestimated or overestimated by F, and the rejection probability of the analytical performance estimate is increased or decreased, respectively. Importantly, while these cases balance when F is averaged across patients, they cannot balance with regard to the rejection frequency. Overall the empirical size is higher than the applied significance level.

This problem further aggravates when F has to be estimated from shorter inter-seizure intervals. To illustrate this point we show results for which we used a different set of $Q = 15$ shorter artificial inter-seizure intervals d_j^{s} with lengths uniformly distributed between 2 and 8 h (Fig. 7b). For this example the rejection frequency of the analytical performance estimate reaches values of up to 10.5% (range: 10.2–11.2%) at $F^{\mathscr{P}} = 0.385\ \mathrm{h}^{-1}$. Hence, even though $\mathcal{H}_0^{\mathrm{V}}$ is valid, there is a substantial probability of false positive rejections of this null hypothesis tested by the analytical performance estimate. This reveals an intrinsic bias of the analytical performance estimate that can cause its empirical size to be substantially higher than the significance level.

Fig. 7 further shows that $\mathcal{H}_0^{\mathrm{II}}$ surrogates exhibit a significantly higher performance than the original predictor. This becomes apparent in particular for lower $F^{\mathscr{P}}$ and for shorter inter-seizure intervals. The reason why $\mathcal{H}_0^{\mathrm{II}}$ surrogates do not match the original is that despite the correction scheme to account for unfinished intervals, the surrogates can have a remaining bias towards too short inter-alarm intervals. This bias becomes substantial if the original predictor exhibits an inter-alarm interval distribution ρ with

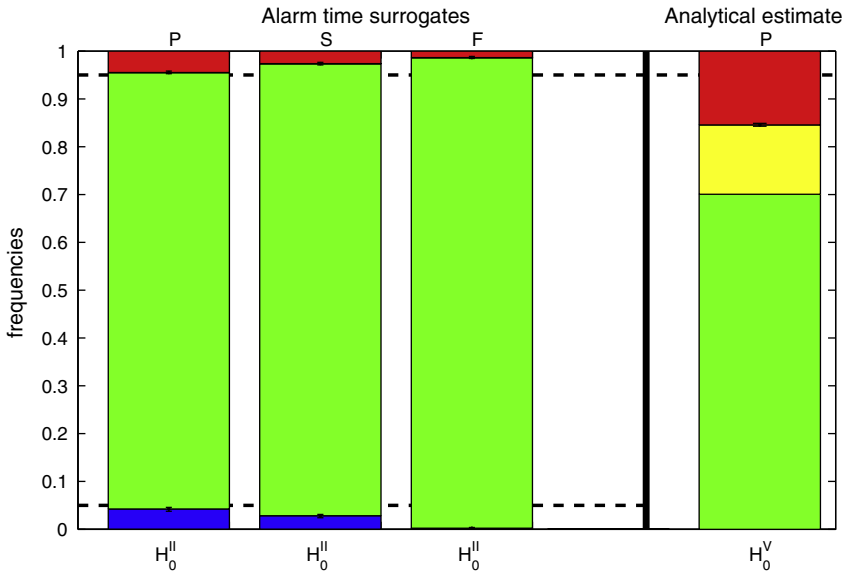

Fig. 6. Same as Fig. 3, but for the stationary integrate-and-fire predictor (Section 4.2). Here only $\mathscr{H}_0^{II}$ surrogates are used.

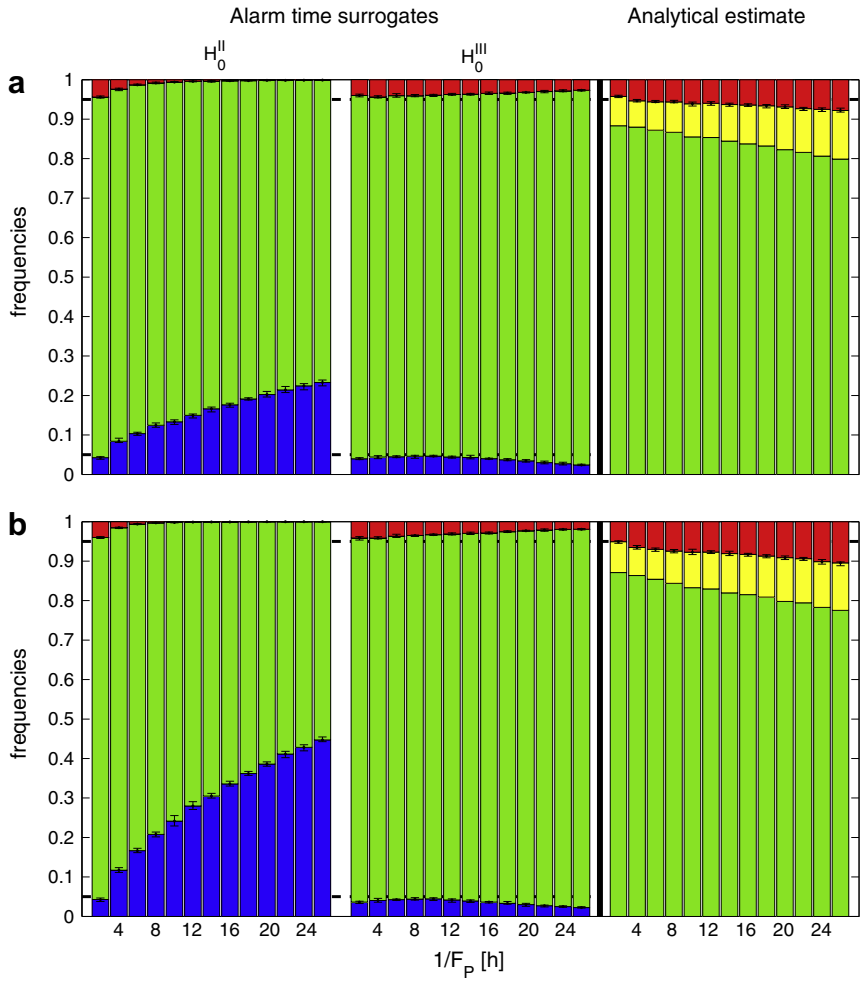

Fig. 7. Same as Fig. 3, but here for the Poisson predictor in dependence on $(F^{\mathscr{P}})^{-1}$ (Section 4.3) and restricted to rejection frequencies derived from the performance using $\mathscr{H}_0^{II}$ and $\mathscr{H}_0^{III}$ surrogates. (a) Results for the standard patient ensemble. (b) Results for a patient ensemble with shorter inter-seizure intervals.

a pronounced tail towards long intervals, as is the case for the exponential ρ of the Poisson predictor studied here. Evidently, the heavier the tail of inter-alarm intervals with regard to the length of inter-seizure intervals, the more difficult it becomes to reliably estimate ρ from $\hat{\rho}$. This explains why the probability that the original performance is smaller than the minimal performance

across an ensemble of $\mathscr{H}_0^{\mathrm{II}}$ surrogates rises with decreasing $F^{\mathscr{P}}$ and for shorter inter-seizure intervals.

While this mismatch would not lead to the erroneous conclusion of a true predictive power of the original predictor since the original predictor is actually outperformed by the surrogates, it still is sufficient for a rejection of $\mathscr{H}_0^{\mathrm{II}}$. In fact, this $\mathscr{H}_0^{\mathrm{II}}$ rejection is correct. For the Poisson predictor, inter-alarm intervals are not interrupted by seizures. Therefore, $\mathscr{R}_3$ is violated, and thereby $\mathscr{H}_0^{\mathrm{II}}$ is valid. However, like stated in Section 3.1.2 the null hypothesis $\mathscr{H}_0^{\mathrm{V}}$ implies $\mathscr{H}_0^{\mathrm{III}}$. The Poisson predictor is memoryless and uncorrelated, it always has the same state and is not reset by a seizure. $\mathscr{R}_1$ is valid. Accordingly, the interval from the last alarm in the inter-seizure interval j to the first alarm in interval $j+1$ does not represent two distinct inter-alarm intervals but rather one continuous interval. Indeed, results for $\mathscr{H}_0^{\mathrm{III}}$ surrogates show that for all values of $F^{\mathscr{P}}$ and regardless of the length of the inter-seizure intervals, the underlying null hypothesis is always correctly accepted (Fig. 7). For high values of $F^{\mathscr{P}}$, we find that the rejection frequency for these surrogates falls below the significance level. The reason is that due to the substantial tail of the inter-alarm distribution of the Poisson predictor, there is a growing probability that no alarm at all is raised for the entire recording for decreasing $F^{\mathscr{P}}$. These cases are counted as acceptances of the null hypothesis both for surrogates and the analytical performance estimate.

4.4. Non-naïve predictors

In this last setting, we use the two different non-naïve predictors described in Section 3.3.4 with variable degrees of true predictive power. Recall that this predictive power is quantified by s^*, i.e. the percentage of all alarms raised by the non-naïve part of the predictor. To test the hybrid non-naïve integrate-and-fire predictor, we use $\mathscr{H}_0^{\mathrm{IV}}$ alarm times surrogates. Here the case of $s^*=0$ is identical to the setting of Section 4.1 (compare Figs. 3 and 8a). Accordingly, for $s^*=0$ the rejection frequency of $\mathscr{H}_0^{\mathrm{IV}}$ surrogates is again close to the significance level, and $\mathscr{H}_0^{\mathrm{IV}}$ is correctly accepted. For $s^*=0$ this rejection frequency corresponds to the statistical size of the test, i.e. the probability of rejecting the null hypotheses although it is correct. For $s^*>0$, this rejection frequency represents the statistical power of the test, i.e. the probability that a false null hypotheses is indeed rejected. Evidently, for $s^*\to 0$ the power converges towards the size, but substantial deviations from the significance level are already established for the first non-zero $s^*\approx 5\%(s=1)$. For increasing s^* values, the power of the test rises further. Hence, at the level of the patient ensemble the null hypothesis $\mathscr{H}_0^{\mathrm{IV}}$ is correctly rejected for $s^*>0$. For individual patients on the other hand, a substantial probability of accepting the null hypothesis despite its incorrectness remains for small s^*. However, for $s^*\approx 25\%$, a statistical power as high as approximately 80% is achieved.

Results in Fig. 8a cannot be conclusive with regard to the power of the $\mathscr{H}_0^{\mathrm{V}}$ analytical performance estimate, since already for $s^*=0$ its null hypothesis is violated. This is different for the hybrid non-naïve Poisson predictor for which both $\mathscr{H}_0^{\mathrm{III}}$ and $\mathscr{H}_0^{\mathrm{V}}$ are true for $s^*=0$ (Fig. 8b). We here use this predictor for $F^{\mathscr{P}}=0.25\ \mathrm{h}^{-1}$ which for $s=0$ corresponds to one of the lower false positive rates considered in Section 4.3 (Fig. 7). Comparing the results for $\mathscr{H}_0^{\mathrm{IV}}$ surrogates obtained for the hybrid non-naïve integrate-and-fire with results for $\mathscr{H}_0^{\mathrm{III}}$ surrogates obtained for the hybrid non-naïve Poisson predictor (Fig. 8a versus b), we find that for the latter the statistical power rises more slowly as a function of s. However, here the random part of the predictor raises more alarms, and with regard to s^* the power of $\mathscr{H}_0^{\mathrm{III}}$ surrogates rises faster. A power of around 80% is reached already for $s^*\approx 15\%$. Comparing the $\mathscr{H}_0^{\mathrm{III}}$ alarm times surrogates and the analytical performance estimate for the hybrid non-naïve Poisson predictor, we find that the power of the analytical performance estimate rises slightly faster with increasing s^*. Overall these results show that alarm times surrogates and analytical performance estimates have a similar, high statistical power to result in true positive rejections if the seizure predictor has any true predictive power.

5. Discussion

In this study, we have extended the framework of seizure predictor surrogates by introducing the concept of alarm times surrogates. We compared this concept against an analytical performance estimates under controlled conditions. Our results provide ample evidence that alarm times surrogates offer a variety of important advantages over analytical performance estimates. The key advantage is the surrogates' higher flexibility with regard to the different null hypotheses that can be tested. While analytical sensitivity and performance estimates have been described only for periodic predictors and Poisson predictors ($\mathscr{H}_0^{\mathrm{V}}$) (Winterhalder et al., 2003; Schelter et al., 2006a; Wong et al., 2007; Snyder et al., 2008), seizure time surrogates allow one to test various distinct null hypotheses.

Furthermore, analytical performance estimates have been derived only as a function of false positive rates (Winterhalder et al., 2003; Schelter et al., 2006a) or the fraction of time under warning (Wong et al., 2007;Snyder et al., 2008). An advantage of seizure predictor surrogates is that they are in no way restricted to any particular definition of sensitivity, specificity, and performance. Moreover, seizure predictor surrogates are even robust against potential biases in these definitions, because any procedure applied to the original seizure predictor is applied in the same way to the seizure predictor surrogates. Accordingly, any potential bias included in these procedures affects the original and the surrogate predictor in the same way. Analytical performance estimates, in contrast, must rely on the correctness of the specificity estimate. If this is not warranted, false positive rejections of $\mathscr{H}_0^{\mathrm{V}}$ are to be expected. Seizure predictor surrogates can also be derived for schemes based on seizure warnings of variable duration (cf. Snyder et al., 2008) or based on the notion of permissive pre-ictal states that not always lead to seizures (cf. Wong et al., 2007).

At first sight it might seem that the flexibility of seizure predictor surrogates might allow one to just try different null hypotheses until one finds surrogates that match the original seizure predictor. Regardless of what other assumptions are made, however, the null hypotheses of seizure predictor surrogates must always include $\mathscr{N}$. Accordingly, the violation of $\mathscr{N}$, as of any other assumptions included in the null hypothesis, is sufficient for a rejection. Therefore, a predictor with true predictive power will generally outperform surrogates for any null hypothesis that includes $\mathscr{N}$. Certainly, the statistical power of seizure predictor surrogates cannot always be 100%. For seizure predictors with some weak but true predictive power, false acceptances of the null hypothesis for individual patients cannot be ruled out. However, on the level of patient ensembles, the null hypothesis is likely to be rejected even for weak sensitivity and limited specificity (see Section 4.4). When in doubt, it is recommended to extend the data set by adding more recordings to obtain robust results.

Results derived from seizure predictor surrogates always need to be interpreted with care. First of all, it is never warranted that all features which do not reflect a true predictive power of the original predictor but may influence its performance have been detected and imposed as constraints to the surrogates. Hence, it

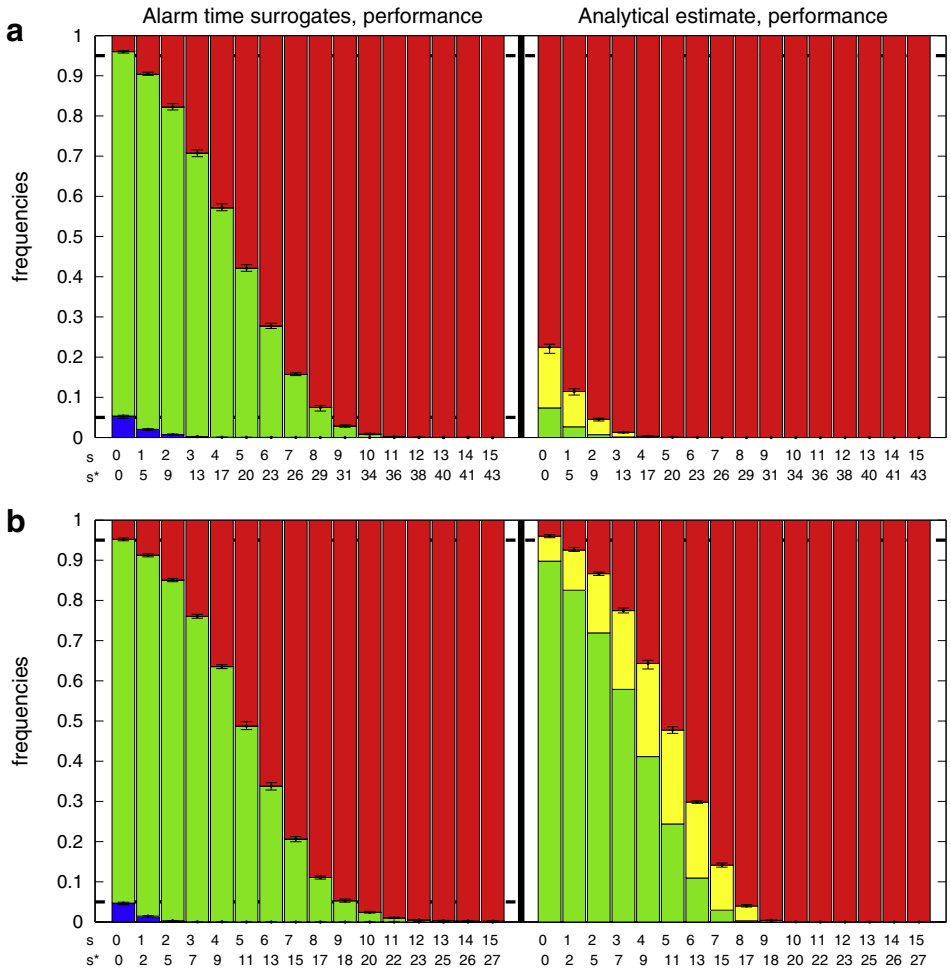

Fig. 8. Same as Fig. 3, but for the non-na predictors of Section 4.4. (a) Hybrid non-na integrate-and-fire predictor for variable s and s^* tested against $\mathscr{H}_0^{\mathrm{IV}}$ surrogates. (b) Same as (a) but for the hybrid non-na Poisson predictor tested against $\mathscr{H}_0^{\mathrm{III}}$ surrogates.

does not follow from the rejection of different null hypotheses that the original predictor is not consistent with some other null hypothesis that includes $\mathscr{N}$. We can never rule out that the original predictor is consistent with some appropriate stochastic model that we have not found yet. Furthermore, even if the appropriate null hypothesis is used, a perfect statistical match between predictor surrogates and the original predictor is not guaranteed. Such a perfect match can only be expected if the surrogates match the original in all constrained properties with a sufficient precision. The often limited sample size given by the original predictor makes it difficult to eliminate the possibility of a systematic bias in the surrogate properties. Specialized correction schemes, such as the one described in Section 3.1.1, can reduce such biases but might not eliminate them completely. Hence the frequency of false positive null hypothesis rejections may deviate from the significance level of the surrogate test. The deviations encountered in the present study were typically such that the empirical size was below the nominal size. This type of deviation would not provide any evidence for the erroneous conclusion that the original predictor has true predictive power. However, a bias in the surrogates' properties could also cause too many false positive rejections where the original performance is higher than the surrogate performance. In consequence, seizure predictor surrogate should not be overrated as a fail-safe tool for evaluating seizure prediction algorithms. Keeping these as well as general limitations intrinsic to the logic of null hypothesis testing (see e.g. Cohen (1994) and references therein) in mind, alarm times surrogates offer a straightforward way to test well-defined null hypotheses about seizure predictors derived from the EEG of epilepsy patients.

5.1. Implications for seizure predictors extracted from actual EEG recordings

While our paper focusses on the use of artificial seizure predictors under controlled conditions, our results have strong implications for applications to seizure predictors extracted from actual EEG recordings. We discuss these implications in the remainder of this paper. In Section 4.1, we used an artificial alarm-free delay period after each seizure to construct a non-stationary integrate-and-fire predictor. It is well-known that post-seizure EEG alterations can cause pronounced changes in characterizing measures which in turn result in time-dependencies of the seizure predictor. Furthermore such post-seizure time-dependencies can be intrinsic to the prediction algorithm. In several studies (e.g. Iasemidis et al. (2003), Chaovalitwongse et al. (2005), Sackellares et al. (2006), Iasemidis et al. (2005)) which used the largest Lyapunov exponent as characterizing measure, channel groups used for the predictor were re-selected after each seizure and were specifically composed of those measure profile channels which were most converged prior to seizure j and diverged after this seizure. A re-convergence of the measure profile in these channel groups was then regarded as predictive of

seizure $j+1$ and used to trigger an alarm. Choosing specifically those channels that are diverged after a seizure (out of equilibrium) and awaiting their re-convergence (equilibrium) may have the same effect as the artificial post-seizure delay period we used for the construction of the non-stationary integrate-and-fire predictor. As illustrated in Fig. 4, such time-dependencies can easily be detected since they are time-locked to each preceding seizure. For seizure predictor surrogates, one can account for this type of time-dependence by replacing $\mathscr{S}_1$ by $\mathscr{S}_2$. The derivation of analytical performance estimates for these time-dependencies, however, seems hardly possible. In the analytical framework $\mathscr{S}_1$ cannot simply be replaced by $\mathscr{S}_2$. Indeed we have shown that $\mathscr{H}_0^{\mathrm{IV}} = \mathscr{N}$ & $\mathscr{S}_2$ & $\mathscr{R}_2$ surrogates matched the performance of the original non-stationary integrate-and-fire predictor studied in Section 4.1, whereas $\mathscr{H}_0^{\mathrm{II}} = \mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{R}_3$ surrogates as well as the $\mathscr{H}_0^{\mathrm{V}} = \mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{D}$ analytical performance estimate fell short of this predictor's performance.

The distinction of the different assumptions and null hypotheses composed from them might seem too meticulous. However, as further substantiated by the present study, it is important to realize exactly what assumptions underlie the particular null hypothesis being tested. Aschenbrenner-Scheibe et al. (2003) reported that although they found no significant difference between results for the inter-seizure and pre-seizure data obtained from an original seizure predictor based on the correlation dimension, this predictor nonetheless outperformed the $\mathscr{H}_0^{\mathrm{V}}$ analytical performance estimate. The authors concluded that the pre-seizure EEG carried information about the forthcoming seizure and that the seizure predictor indeed captured this information to some degree. However, if we recall that $\mathscr{H}_0^{\mathrm{V}}$ represents the conjunction of three assumptions: $\mathscr{H}_0^{\mathrm{V}} = \mathscr{N}$ & $\mathscr{S}_1$ & $\mathscr{D}$, it is clear that the violation of any of these assumptions is sufficient for a rejection of this null hypothesis. Apart from a true predictive power of the original predictor (violating $\mathscr{N}$, cf. Section 4.4), a time-dependence of the mean alarm rate (violating $\mathscr{S}_1$, cf. Section 4.1), or a non-exponential distribution of the inter-alarm intervals (violating $\mathscr{D}$, cf. Section 4.2) can cause a rejection of $\mathscr{H}_0^{\mathrm{V}}$. Hence, this rejection provides merely a necessary yet not sufficient condition for a true predictive power of the original predictor. In fact, Aschenbrenner-Scheibe et al. (2003) deactivated the predictor after an alarm for the duration of a prediction horizon. Thereby, assumption $\mathscr{D}$ was violated by construction. In addition, an in-sample parameter optimization was carried out which by itself is sufficient to falsely reject null hypotheses of analytical performance estimates. Lastly, we have shown here that the analytical performance estimate has an intrinsic bias to underestimate the performance values expected under $\mathscr{H}_0^{\mathrm{V}}$ for low false positive rates such as those considered by Aschenbrenner-Scheibe et al. (2003) (Section 4.3). Therefore, rather than assuming true predictive power of the original predictor, many factors could explain why in (Aschenbrenner-Scheibe et al., 2003) the original predictor outperformed the $\mathscr{H}_0^{\mathrm{V}}$ analytical performance estimate.

Mormann et al. (2005) applied a total of 30 different characterizing measures to EEG recordings from five patients using different evaluation schemes. Since the available amount of data was insufficient for a division into training and testing data, a number of evaluation parameters were optimized in-sample. To account for this optimization, results were validated using seizure times surrogates, and the surrogates were given the same degrees of freedom for optimization. Periods of 30 minutes after each seizure were excluded from the analysis to diminish the influence of post-seizure time-dependencies that would violate the stationarity assumption $\mathscr{S}_1$ included in the null hypothesis of seizure times surrogates ($\mathscr{H}_0^{\mathrm{III}}$). Rejections of this null hypothesis were found for several measures, particularly for bivariate measures characterizing the interaction between different brain regions. However, although Mormann et al. (2005) discussed the problem of testing multiple characterizing measures, no formal correction for multiple tests was applied.

Chaovalitwongse et al. (2005) divided EEG recordings from 10 patients into training and testing data for each individual patient. The authors employed a seizure predictor based on the criterion of a pre-seizure convergence and post-seizure divergence of the largest Lyapunov exponent profile as described above and seizure times surrogates to validate the performance values. As free parameters the number of channel groups and the number of channels per group were optimized using the training data. These optimized parameters were then used on the testing data for both the original predictor and the surrogates. However, the authors depicted the full distribution of the surrogate testing data performance for only one of the 10 patients. Notably they picked the patient for whom the highest original predictor performance was found for the testing data. Only for this one patient this testing data performance of the original predictor was compared to the surrogate distribution, and the resulting p-value was reported. For the remaining nine patients, only the mean values but not the ranges of the surrogates' performance distributions were reported. Overall, eight out of 10 patients showed a performance of the original predictor that was better than the mean performance of the surrogates. However, since the original performance was not statistically tested against the full distribution of the surrogates, it is problematic to asses the significance of these performance differences for individual patients.

Sackellares et al. (2006) re-analyzed the same EEG recordings studied by Chaovalitwongse et al. (2005) using the same seizure predictor already used in (Chaovalitwongse et al., 2005). This time, however, the authors used a different quantification of the performance and studied the influence of the length of the prediction horizon. Furthermore, the data was not divided into training and testing data, and fixed values were used for the two free parameters of the prediction algorithm, i.e. the number of channel groups and the number of channels per group. The results were tested against the null hypotheses of a periodic predictor and Poisson predictor. Rather than using analytical sensitivity and performance estimates, Sackellares et al. (2006) generated actual alarms from a periodic and from a Poisson process. Hence, this procedure, which Snyder et al. (2008) extended to gamma distributed inter-alarm intervals, can be regarded as a variant of seizure predictor surrogates. Here surrogate alarms are not generated by constrained randomizations of the original predictor, but by using a concrete model, namely, a periodic or Poisson process. This form of *typical realization*-surrogates (cf. Schreiber and Schmitz (2000)) shares a number of the advantages of seizure predictor surrogates derived from constrained randomizations of the original predictor. However, just as Chaovalitwongse et al. (2005) and Sackellares et al. (2006) only provided the mean values and not the full surrogate performance distributions. Furthermore, no information was provided about the choice of the two parameters of the algorithm. In particular, it is not clear whether the values used by Sackellares et al. (2006) were indeed chosen independent and different from the optimal values of these parameters determined in Chaovalitwongse et al. (2005) from the training data portion of the recordings.

We conclude that previous applications of analytical performance estimates or seizure predictor surrogates to real seizure predictors extracted from actual EEG recordings have suffered from various problems or shortcomings. Our results presented here for artificial data under controlled conditions emphasize the relevance of these problems in that they can cause false positive null hypothesis rejections. These rejections can be erroneously

interpreted as indicative of true predictive power of the seizure prediction algorithm.

Acknowledgements

The authors would like to thank the three anonymous referees for their helpful comments and suggestions. RGA acknowledges grant BFU2007-61710 of the Spanish Ministry of Education and Science. DC was supported by the grant 2008FI-B 00460 of the 'Generalitat de Catalunya' and European Social Funds. FM was supported by the 6th Framework Programme of the European Commission (Marie Curie OIF 040445).

References

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

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 2003;126:2616–26.

Chaovalitwongse WA, Iasemidis LD, Pardalos PA, Carney PR, Shiau DS, Sackellares JC. Performance of a seizure warning algorithm based on the dynamics of intracranial eeg. Epilepsy Res 2005;64:93–113.

Cohen J. The earth is round (p less than 0.05). Am Psychol 1994;49:997–1003.

De Clercq W, Lemmerling P, Van Huffel S, Van Paesschen W. Anticipation of epileptic seizures from standard EEG recordings. Lancet 2003;361:970.

Harrison MAF, Frei MG, Osorio I. Accumulated energy revisited. Clin Neurophysiol 2005a;116:527–31.

Harrison MAF, Osorio I, Frei MG, Asuri S, Lai YC. Correlation dimension and integral do not predict epileptic seizures. Chaos 2005b;15:15 [article ID 033106].

Iasemidis L, Shiau D, Chaovalitwongse W, Sackellares J, Pardalos P, Principe J, et al. Adaptive epileptic seizure prediction system. IEEE Trans Biomed Eng 2003;50:616–27.

Iasemidis LD, Shiau DS, Pardalos PM, Chaovalitwongse W, Narayanan K, Prasad A, et al. Long-term prospective on-line real-time seizure prediction. Clin Neurophysiol 2005;116:532–44.

Kreuz T, Andrzejak RG, Mormann F, Kraskov A, Stogbauer H, Elger CE, et al. Measure profile surrogates: a method to validate the performance of epileptic seizure prediction algorithms. Phys Rev E 2004;69:061915.

Lai YC, Harrison MAF, Frei MG, Osorio I. Controlled test for predictive power of lyapunov exponents: their inability to predict epileptic seizures. Chaos 2004;14:630–42.

Lopes da Silva F, Blanes W, Kalitzin S, Parra J, Suffczynski P, Velis D. Dynamical diseases of brain systems: different routes of epileptic seizures. IEEE Trans Biomed Eng 2003;50:540–8.

Maiwald T, Winterhalder M, Aschenbrenner-Scheibe R, Voss HU, Schulze-Bonhage A, Timmer J. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. Phys D 2004;194:357–68.

Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. Brain 2007;130:314–33.

Mormann F, Andrzejak RG, Kreuz T, Rieke C, David P, Elger CE, et al. Automated detection of a preseizure state based on a decrease in synchronization in intracranial electroencephalogram recordings from epilepsy patients. Phys Rev E 2003;67:021912.

Mormann F, Kreuz T, Rieke C, Andrzejak RG, Kraskov A, David P, et al. On the predictability of epileptic seizures. Clin Neurophysiol 2005;116:569–87.

Sackellares JC, Shiau DS, Principe JC, Yang MCK, Dance LK, Suharitdamrong W, et al. Predictability analysis for an automated seizure prediction algorithm. J Clin Neurophysiol 2006;23:509–20.

Schad A, Schindler K, Schelter B, Maiwald T, Brandt A, Timmer J, et al. Spatio-temporal patient-individual assessment of synchronization changes for epileptic seizure prediction. Clin Neurophysiol 2008;119:197–211.

Schelter B, Winterhalder M, Drentrup HFG, Wohlmuth J, Nawrath J, Brandt A, et al. Seizure prediction: the impact of long prediction horizons. Epilepsy Res 2007;73:213–7.

Schelter B, Winterhalder M, Maiwald T, Brandt A, Schad A, Schulze-Bonhage A, et al. Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction. Chaos 2006a;16:10 [article ID 013108].

Schelter B, Winterhalder M, Maiwald T, Brandt A, Schad A, Timmer J, et al. Do false predictions of seizures depend on the state of vigilance? A report from two seizure-prediction methods and proposed remedies. Epilepsia 2006b;47:2058–70.

Schreiber T, Schmitz A. Surrogate time series. Phys D 2000;142:346–82.

Snyder D, Echauz J, Grimes DB, Litt B. The statistics of a practical seizure warning system. J Neural Eng 2008;5:392–401.

Stam CJ. Nonlinear dynamical analysis of eeg and meg: review of an emerging field. Clin Neurophysiol 2005;116:2266–301.

Sunderam S, Osorio I, Frei MG. Epileptic seizures are temporally interdependent under certain conditions. Epilepsy Res 2007;76:77–84.

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 2003;4:318–25.

Winterhalder M, Schelter B, Maiwald T, Brandt A, Schad A, Schulze-Bonhage A, et al. Spatio-temporal patient-individual assessment of synchronization changes for epileptic seizure prediction. Clin Neurophysiol 2006;117:2399–413.

Wong S, Gardner AB, Krieger AM, Litt B. A stochastic framework for evaluating seizure prediction algorithms using hidden markov models. J Neurophysiol 2007;97:2525–32.