Statistical validation of event predictors: A comparative study based on the field of seizure prediction

Hinnerk Feldwisch-Drentrup,[1,2,3,4,5,*] Andreas Schulze-Bonhage,[1,6] Jens Timmer,[1,2,4,5,7] and Björn Schelter[2,5,8]

[1]*Bernstein Center Freiburg (BCF), University of Freiburg, Freiburg, Germany*

[2]*Freiburg Center for Data Analysis and Modeling (FDM), University of Freiburg, Freiburg, Germany*

[3]*Department of Neurobiology and Biophysics, Faculty of Biology, University of Freiburg, Freiburg, Germany*

[4]*Freiburg Institute for Advanced Studies, University of Freiburg, Freiburg, Germany*

[5]*Department of Physics, University of Freiburg, Freiburg, Germany*

[6]*Epilepsy Center, University Hospital of Freiburg, Freiburg, Germany*

[7]*Department of Clinical and Experimental Medicine, Linköping University, Linköping, Sweden*

[8]*Institute for Complex Systems and Mathematical Biology, SUPA, University of Aberdeen, Aberdeen, United Kingdom*

(Received 23 January 2011; revised manuscript received 9 April 2011; published 20 June 2011)

The prediction of events is of substantial interest in many research areas. To evaluate the performance of prediction methods, the statistical validation of these methods is of utmost importance. Here, we compare an analytical validation method to numerical approaches that are based on Monte Carlo simulations. The comparison is performed in the field of the prediction of epileptic seizures. In contrast to the analytical validation method, we found that for numerical validation methods insufficient but realistic sample sizes can lead to invalid high rates of false positive conclusions. Hence we outline necessary preconditions for sound statistical tests on *above chance* predictions.

DOI: 10.1103/PhysRevE.83.066704 PACS number(s): 05.10.−a, 07.05.Tp, 87.19.xm, 91.30.pd

I. INTRODUCTION

The occurrence of rare but severe events in fields like seismology [1], epileptology [2], or meteorology [3] poses considerable risk to human life. If the time of occurrence of earthquakes, epileptic seizures, or windstorms could be reliably predicted, warnings could be issued in order to enable appropriate preparations. Hence considerable research efforts have been invested aiming at developing appropriate prediction methods, which analyze continuous measurements of the underlying system to identify precursors of an upcoming event.

In order to elicit possibly predictive information from the measurements, linear as well as nonlinear time series analysis techniques are applied. This leads to derived time series that measure specific quantities—so-called features. These feature time series are supposed to contain characteristic changes that render raising alarms possible. Alarms, which are triggered, e.g., when the features cross certain thresholds, are assumed to be related to the occurrence of the events. However, it is *a priori* not known whether the alarms indeed reflect predictive information. In principle, it is conceivable that alarms are raised at random times, which then also could—randomly—predict events. The application of such unspecific "prediction" methods would certainly not be reasonable. Unjustifiable stress would be put on the persons involved, who in turn might disregard the predictions [4]. Thus, statistically speaking, it must be tested whether the null hypothesis H_0, that the achieved performances based on the triggered alarms do not differ from performances obtained by chance, can be rejected.

Several methodologies have been suggested for the statistical validation of event prediction performances. Basically, they could be grouped into two classes. First, there exist analytical approaches that allow an immediate calculation of critical values for judging on the significance of prediction performances. Only if the actual performance of a prediction method is above this critical value can the predictor be considered to show above-chance performance. Second, the class of Monte Carlo based validation methodologies provides critical values by applying constrained randomizations to either the series of events or to the prediction methods under consideration. Again, to be considered significant the performance of the actual prediction method has to exceed the critical values.

In this study, we compare these methods in order to identify their advantages and limitations. We exemplify our analysis on the field of the prediction of epileptic seizures. Here, methods were developed analyzing electroencephalographic (EEG) recordings in order to identify seizure precursors. For a review see [2]. Although we base our study on evaluation methods developed for this specific application, respective implications also apply to other fields. Problem-specific adaptations might be necessary, such as the inclusion of spatial information. Discussing these in detail would go beyond the scope of this paper. However, the conclusions drawn about the characteristics of validation methodologies can be applied in an analogous manner to related fields.

Besides the statistical validation of prediction methods, another fundamental question regards their practical relevance, i.e., whether their performance is sufficiently high for a given application. The practical relevance needs to be related to the desired type of intervention. For example, if actual warnings are given to persons, a low rate of false predictions is usually required. In contrast, in fields where automatic interventions can be delivered, higher rates of false predictions are often possible if the interventions are accompanied by negligible side effects. In the following, we focus on studying the statistical significance, which allows an unequivocal evaluation of the validity of a prediction method.

*hinnerk.feldwisch@bcf.uni-freiburg.de

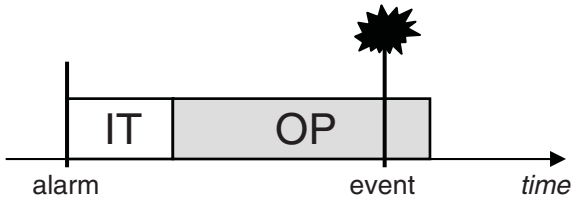

FIG. 1. Prediction scheme considered in this study. An alarm is followed by an intervention time (IT), during which no event should occur such that an intervention could be applied. During the occurrence period (OP), the event is expected to occur. If, indeed, an event occurs during OP, the alarm is regarded correct.

General notions and notations for the evaluation of event predictions are introduced in the next section. This is followed by a description of the validation approaches studied, motivated and introduced by the seizure prediction application. Their properties are compared in a simulation study, incorporating artificial event times and simulated predictors with adjustable predictive power. Subsequently, the results and their implications are discussed to determine necessary preconditions for statistically sound tests on above chance performances.

II. QUANTIFYING PREDICTION PERFORMANCES

When evaluating the predictive performance of a prediction method M, it can be regarded as a "black box" triggering alarms at specific points in time, which should predict the occurrence of upcoming events. For each alarm triggered by M, a well-defined occurrence period (OP) has to be specified, for which the subsequent event is expected to occur (Fig. 1). In order to predict an event, the alarms have to precede the event by a specified amount of time, which then renders a preparation possible—the so-called intervention time (IT). The prediction sensitivity S is defined as the fraction of events that were predicted correctly, i.e., that occurred during the OP of the previous alarm.

An alarm is considered to be false if no event occurs following IT during OP. To quantify specificity, we here follow the approach of a false prediction rate (FPR), i.e., the number of false alarms divided by the time during which false alarms could be triggered.

III. STATISTICAL VALIDATION IN THE FIELD OF SEIZURE PREDICTION

In seizure prediction, both analytical and Monte Carlo based validation approaches have been introduced [5–13].

A. Analytical random predictor

In order to generate unpredictive alarms, a Poisson process in time is assumed for the analytical approach, triggering alarms independently from each other at a constant rate over time. This rate is determined by the false prediction rate γ [6,11], leading to the probability $P_h = \gamma h$ for an alarm during a short time interval h, strictly speaking in the limit $h \to 0$. The probability to randomly predict a seizure correctly, i.e., the probability to trigger an alarm followed by a seizure in the corresponding occurrence period of length Ω, can be approximated by $P \approx 1 - e^{\gamma\Omega}$ [6]. In many studies, parameters of the prediction method under consideration are optimized in order to improve its performance [10,11]. This includes, e.g., the selection of an optimal duration of IT, or an optimal channel of the multichannel EEG recording. The random predictor (RP) can be corrected for this increased degree of freedom. If d independent optimizations of, for example, one optimal EEG channel out of d channels are performed, the probability to randomly predict at least n out of N seizures reads [11]

$$P_d(n,N,P) = 1-\left[1-\sum_{j\geqslant n}\binom{N}{j}P^j(1-P)^{N-j}\right]^d. \qquad (1)$$

For a significance level α, a critical sensitivity of

$$S_{\mathrm{RP}} = \mathrm{argmax}_n\{P_d(n,N,P) > \alpha\}/N \times 100\% \qquad (2)$$

can be achieved by chance performance. If the observed sensitivity S of the actual prediction method is higher than S_{RP}, it can be regarded statistically significant.

B. Numerical approaches

Alternatively, Monte Carlo simulations can be used to approximate the performance of random predictors. In the field of seizure prediction, it has been proposed to randomize either the seizure onset times [7], the alarms triggered by a predictor [5], or the time series of the features extracted from the EEG [8]. We emphasize that for other fields of event predictions, the same strategies can be followed when substituting seizures by events. For the randomization of the features, a large number of constraints for the randomization is possible. Yet it was not established which properties of the original features should be preserved in the randomization process. Also due to its high computational demand, it was not used frequently [5]. In the following, we concentrate on the other methods. For so-called "seizure times surrogate" (STS), the time intervals between seizures are randomly permuted by drawing without replacement to generate randomized seizure onset times [7,10]. A random offset is added to the first interval because otherwise all realizations would add up to the same total duration and the last seizure onset times would always coincide with the original ones [10]. As an alternative, the intervals can be drawn with replacement, which corresponds to the bootstrap resampling approach (BST). For both STS and BST, the prediction method has to be applied to the randomized seizure onset times as it was applied to the original ones. For "alarm time surrogate" (ATS), randomized alarm times are generated by drawing intervals with replacement from the pool of original inter-alarm intervals, which were triggered by the prediction method M under consideration. Here, the performance is assessed based on the original seizure times. If the internal state of M is reset after the occurrence of each seizure, special care has to be taken for the ATS in order to obtain a valid sample of the inter-alarm intervals [5].

For a given significance level α, the performance of M can be considered *above chance level* for a Monte Carlo based validation method if it exceeds the $(1-\alpha)$ quantile of the empirical distribution of the performances under H_0. A minimum

number of realizations is $\lceil 1/\alpha - 1 \rceil$, i.e., 19 realizations for $\alpha = 5\%$. With more realizations, the distributions of the performances of the validation methods can be approximated with higher accuracy.

IV. CHARACTERISTIC PROPERTIES OF VALIDATION METHODS

For all validation methods, it has to be shown that their empirical size, which is the empirical probability for α errors of the resulting test on above-chance performance [14], sticks correctly to the nominal size α. Otherwise, the test would not adhere to the chosen maximum probability for α errors; thus it would be invalid. Additionally, it should be tested whether the statistical power is sufficient, which is the probability that a null hypothesis is rejected correctly for various strengths of violations of the null hypothesis.

In order to analyze the empirical size, it has to be warranted that no predictive information is incorporated under H_0. For the analytical random predictor, this is ensured by assuming a Poisson process, which is characterized by exponentially distributed inter-alarm times. Hence the occurrence of alarms is neither correlated to other alarms nor to seizures, independent of the alarm distribution of M and independent of the original seizure distribution. For STS, BST, or ATS, the original seizure or alarm distributions are randomized as described above in order to eliminate possibly predictive information. In cases of degenerated distributions, this may be impossible. E.g., for periodically occurring seizures with an almost fixed duration of the period, the inter-seizure intervals have all approximately the same duration. Hence no independent realizations can be drawn—for each realization, the resulting event times are correlated to the original ones. In such cases, the randomization would fail and a true predictive performance of M would not be detected.

For the asymptotic theory of bootstrap tests [15], conditions are known under which the tests fail [16–18]. One necessary condition is a consistent estimation of the underlying sample distribution, i.e., the distribution of the intervals between seizures for STS and BST, and between alarms for ATS. It has to be ensured that both the sample size and the number of bootstrap realizations is sufficiently large. Otherwise, bootstrap does not necessarily provide consistent and unbiased results, which could eventually lead to invalid high rates of α errors and/or a decreased power.

V. SIMULATION STUDY

A. Design

To study empirical size and power of RP, STS, BST, and ATS, we analyze them based on simulated events and simulated "predictors" M_{sim}. Here, each simulation instance represents one patient. For a number of simulation instances, exponentially distributed inter-event intervals are drawn to generate artificial series of onset times. Each simulation lasts for a given total simulation duration T_{sim}. To be close to an actual application, a rate r_{sz} of 0.15 events per hour is used, which is a typical seizure rate occurring during clinical EEG recordings [19]. For an exemplary IT of 10 min and an OP of 30 min, which are again typical durations in seizure prediction, artificial prediction methods M_{sim} are simulated that trigger correct alarms prior to each seizure with probability P_{CA}. Additionally, alarms are triggered that are uncorrelated to the seizures, following a Poisson process with exponentially distributed inter-alarm intervals and based on a false alarm rate r_{FA}. These alarms thus carry no predictive information. If P_{CA} is set to zero, M_{sim} is a random predictor.

To derive the performance of the predictor M_{sim} and the corresponding performances of the validation methods, sensitivities and specificities of the triggered alarms of M_{sim} were determined as described in Sec. II. If further alarms follow each other within the ongoing IT or OP, it would be possible to prolong the first OP [20]. Yet, this could result in excessively long prediction windows. Instead, in such scenarios we consider only the first alarm and discard all further alarms during IT and OP after an alarm. These intervals do not enter the calculation of the false prediction rate (FPR) γ. If two seizures follow each other too closely, i.e., within the duration of IT plus OP, the later seizure is regarded unpredictable and is excluded from the analysis.

In order to compare the performances of prediction methods to the critical values of validation methods, a scalar performance measure is required that quantifies both the sensitivity S and the FPR γ. In the following we use [5,9]

$$\Pi_M(S,\gamma) = 1 - \sqrt{(1-S)^2 + \gamma^2/\gamma_0^2}. \tag{3}$$

The performance measure Π_M equals 1 for the perfect case of $S = 1$ and $\gamma = 0$. It decreases with decreasing S and increasing γ, and has no lower bound. The parameter γ_0 adjusts whether Π_M depends more on S or γ and can be chosen depending on the priorities of the application. For seizure prediction, the correct prediction of seizures is considered more important than the reduction of false alarms [21]. Thus the normalization parameter γ_0 was set to a rather high value of 1/h in order to down-weight the influence of γ. Typical values for γ of 0 to 0.5/h result in a maximum contribution of γ^2/γ_0^2 of 0.25, while S contributes a value between 0 and 1. Thereby, changes in S lead to multiple higher changes of Π_M than changes in γ.

B. Results

The simulations were based on 1000 instances, for which the ratio of instances with performances higher than the different validations methods was determined. The STS, BST, and ATS were based on 100 realizations each. For varying rates of false alarms r_{FA}, Fig. 2 shows the empirical size, i.e., the empirical probability that above-chance prediction performances are detected for the case that H_0 is correct ($P_{\text{CA}} = 0$). In order to yield statistically sound results, the empirical size is required to be smaller than or equal to the chosen significance level $\alpha = 5\%$ (marked by black horizontal lines), since otherwise excessive α errors occur. In (a)–(c), the simulation durations T_{sim} were varied as an additional parameter; in (d) and (e), the seizure rates r_{sz}. For small $r_{\text{FA}} \leqslant 0.05$/h, the empirical size decreases for all methods because only rarely false alarms are issued that could "predict" an event correctly by chance.

In general, it can be observed that for the numerical validation methods the empirical size exceeds α for short

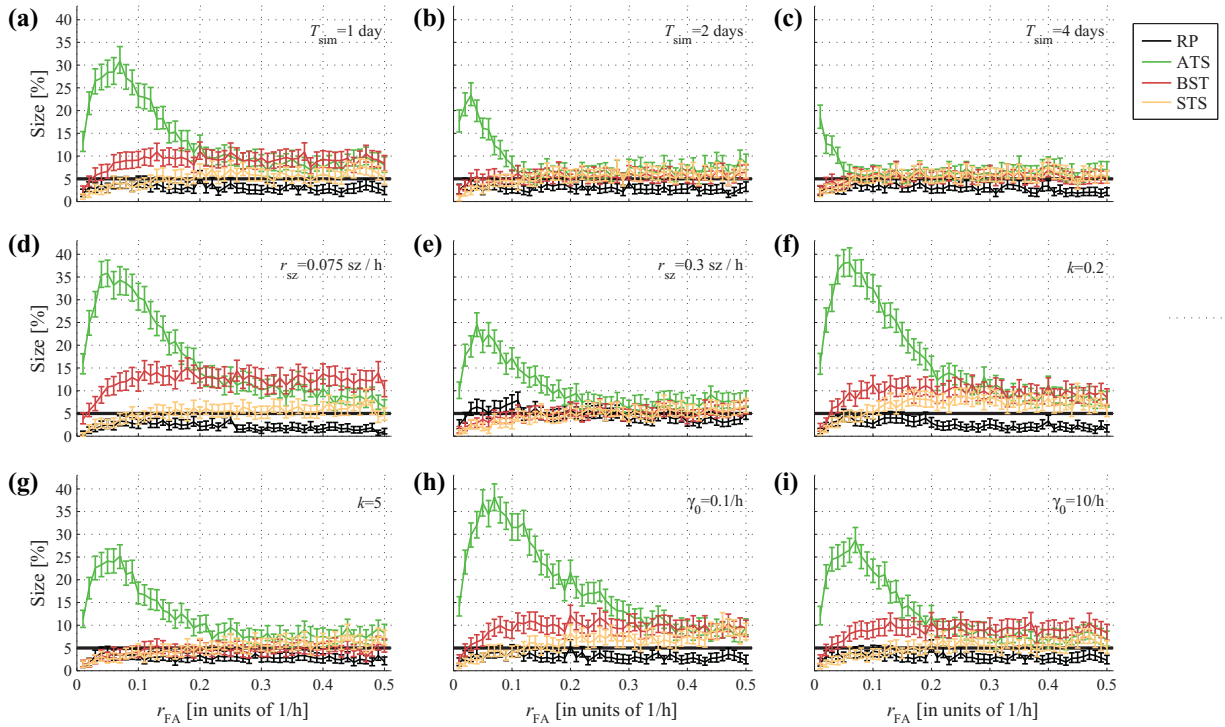

FIG. 2. (Color online) Empirical size of the analytical random predictor (RP, black), the alarm time surrogates (ATS, green or medium gray), and seizure time surrogates with random offset (STS, orange or light gray) and with bootstrap resampling (BST, red or dark gray) depending on the rate of false alarms r_{FA}, for $\alpha = 5\%$ (black horizontal lines). For exponentially distributed inter-seizure intervals, the results are shown for varying simulation durations and fixed seizure rate $r_{sz} = 0.15/\text{h}$ and $\gamma_0 = 1/\text{h}$ in (a)–(c), and for varying r_{sz} and fixed simulation duration $T_{sim} = 1$ day and $\gamma_0 = 1/\text{h}$ in (d) and (e). In (f) and (g), the empirical size is shown for gamma distributed inter-seizure intervals with shape parameter $k = 0.2$ and $k = 5$ for $r_{sz} = 0.15/\text{h}$ and $T_{sim} = 1$ day. In (h) and (i), the normalization factor γ_0 of the performance measure Π_M was varied, again for $r_{sz} = 0.15/\text{h}$ and $T_{sim} = 1$ day, for $k = 1$. Based on 1000 simulation instances, 95% confidence intervals are given.

simulation durations T_{sim} (a) or small seizure rates r_{sz} (d). Hence they could lead to invalid conclusions for small numbers of events. For the BST, the empirical size clearly exceeds α for $T_{sim} = 1$ day and $r_{sz} = 0.15/\text{h}$. Since for increasing numbers of events the empirical size adheres correctly to the value of 5%, this discrepancy can be explained by an insufficient estimation of the underlying sample distributions. This is especially prominent for the ATS, for which the results depend strongly also on r_{FA}. For an average of five alarms or less, i.e., a rate of $r_{FA} \leqslant 0.2/\text{h}$ for $T_{sim} = 1$ day (a) or $r_{FA} \leqslant 0.1/\text{h}$ for $T_{sim} = 2$ days (b), α is considerably exceeded. These deviations decrease for increasing numbers of events. For the STS, in comparison to BST and ATS, α is exceeded only slightly for $T_{sim} = 1$ day (a) and $r_{sz} \leqslant 0.15/\text{h}$ (d), and the empirical size adheres to α for increasing T_{sim} or r_{sz}. For the RP, all approximations are designed to be conservative, like the calculation of the maximum number of seizures predicted by chance in Eq. (2). Hence the RP is statistically conservative by definition (cf. [11]), which is reflected in an empirical size smaller than or equal to α.

In order to test the robustness of the validation methods for varying distributions of the inter-seizure intervals, we also simulated these intervals following a gamma distribution, which resembles the actual inter-seizure intervals of epilepsy patients [22]. For again $P_{CA} = 0$, i.e., unpredictive alarms, intervals were generated with mean $k\theta$ and variance $k\theta^2$. The scale parameter θ of the gamma distribution was set to $\theta = 1/(kr_{sz})$ for a given shape parameter k such that the expected seizure rate is r_{sz} as well. For $k = 1$ the intervals are exponentially distributed, corresponding to Fig. 2(a). In Fig. 2(f), the empirical size is shown for $k = 0.2$, which reflects variances of the inter-seizure intervals larger than the ones for the exponential distribution. For this scenario, an invalid empirical size is observed for all numerical validation methods. Since more inter-seizure intervals with short durations occur in this case, leading to so-called "clusters,' the number of seizures with sufficient distance to the previous seizure decreases. We also performed the simulations for $k = 5$, shown in Fig. 2(g), for which the empirical size adheres to the chosen significance level even for small simulation durations for a wide range of r_{FA} for the BST and STS. This can be explained by the decreased variance and improved estimation of the underlying distribution of the inter-seizure intervals for large k, and the larger number of seizures available. However, for most epilepsy patients both long and very short intervals occur [22,23], such that in these cases an adequate number of seizures has to be ensured.

When varying the rates of γ_0 of the performance measure Π_M [cf. Eq. (3)], as shown in Figs. 2(h) and 2(i), the empirical size decreases slightly for numerical validation methods if γ_0 is chosen to be larger than $1/\text{h}$ [Fig. 2(i)]. In this case, false predictions have a smaller influence on the values of

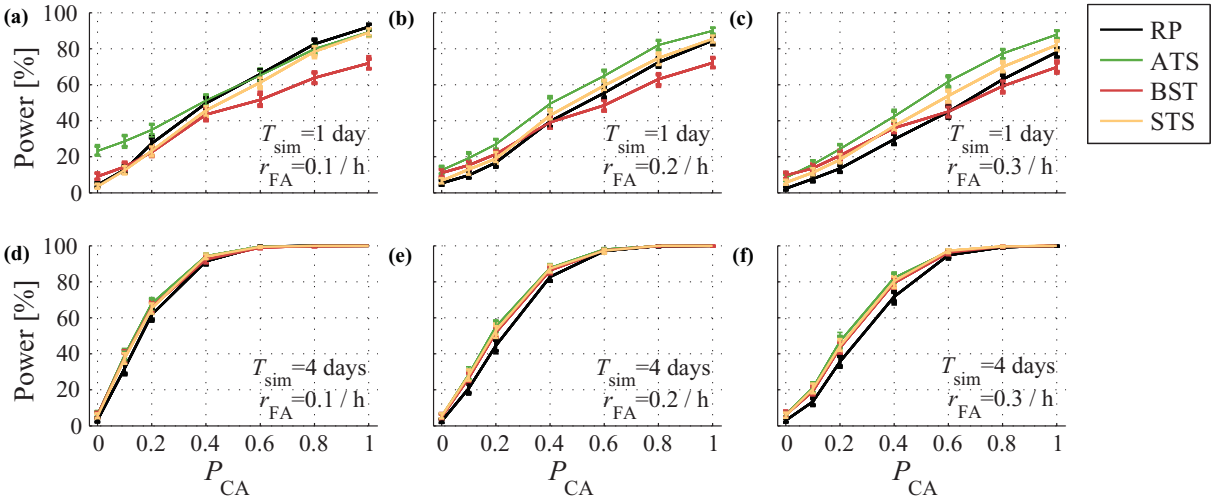

FIG. 3. (Color online) Statistical power of the analytical random predictor (RP, black), the alarm time surrogates (ATS, green or medium gray), seizure time surrogates with random offset (STS, orange or light gray), and with bootstrap resampling (BST, red or dark gray), depending on the probability of predictive alarms P_{CA}. Simulation durations of 1 day in (a)-(c) and 4 days in (d)-(f) are shown for exemplary rates of false alarms r_{FA} and $\gamma_0 = 1/\text{h}$. For the case $P_{CA} = 0$, the power is equivalent to the empirical size (cf. Fig. 2). 95 % confidence intervals are based on 1000 simulation instances.

Π_M than correct predictions. Thus it can be concluded that for small T_{sim} the false prediction rates are overestimated by the numerical methods, leading to an invalid high empirical size. Additionally, we also analyzed the empirical size for varying values of the occurrence period and intervention times (not shown). Since the number of seizures with sufficient inter-seizure duration depends on both OP and IT, it has to be ensured that a sufficient number is available for the chosen durations. For a disproportionately large OP of several hours, the estimation of the FPR fails for short T_{sim} if no false predictions are triggered, leading to an incorrect empirical size for the RP. In this case, a conservative estimate for FPR is given by $1/T_{sim}$ as an upper bound. Otherwise, we emphasize that we did not find considerable differences to the results presented. Hence for all prevailing scenarios and independent of the seizure distribution, the analytical random predictor was found to keep to α for all simulation durations and all settings studied.

In Fig. 3, the statistical power of the validation methods is shown for varying P_{CA} and simulation durations of one day in (a)–(c) and four days in (d)–(f) for three exemplary r_{FA}. As expected, the power is higher for longer simulation durations for all methods. Larger sample sizes allow a better detection of above-chance performances. For $T_{sim} = 4$ days, for which BST and STS adhere to the given significance level, and for false alarm rates $r_{FA} = 0.2/\text{h}$ [Fig. 3(e)] and $r_{FA} = 0.3/\text{h}$ [Fig. 3(f)], the numerical methods exhibit a slightly higher power than the RP.

As described in Sec. III, for the STS a random offset was used, which was uniformly distributed in $[0 \ldots \tau]$ with $\tau = 4\,\text{h}$, as proposed in the literature [10]. To test its influence, we also varied τ. For values near to zero, a decrease in power was observed for small T_{sim}, because in this case the last seizure time almost coincides with the original time. Hence the probability for a correct prediction is higher. For reasonable maximum offsets τ, i.e., in the order of the duration of the OP or larger, no considerable changes in power and empirical size of the STS were found.

VI. DISCUSSION

Both the analytical and numerical validation concepts are characterized by specific advantages. For the analytical random predictor, the assumptions made are explicitly stated and the dependency of its sensitivity on its parameters is known analytically. Hence it allows the design of studies. For example, the minimum number of events needed for a rejection of H_0 could be calculated in advance. The main advantage of Monte Carlo based methods, given that a sufficient number of events is available for valid numerical estimates, is their apparent flexibility. By applying constraints to the randomized data, it is possible to test prediction methods for specific properties beyond randomness of the underlying event generating process [24]. Such properties could be rhythms in event occurrences or the observation that events are followed by a refractory period, as done in a recent study [5]. However, since potentially predictive information is included into the "random" predictor here, a true prediction performance may not be detected. Hence it is of importance to clearly distinguish these tests on specific characteristics from the test on statistical significance of prediction performance. While the former allows insights in the event generating process, the latter is designed to detect performances above chance.

As shown by the simulation study performed, we summarize that numerical methods proposed for the validation of prediction performances can lead to invalid high rates of false positive conclusions for small but realistic numbers of events, which were used in several studies; for a review see [10]. We found that the absolute number of events with sufficiently long inter-event intervals is a decisive factor. Hence also for short durations of one day, a large number of about five such events would be required for the STS and BST. This is especially

important if the inter-event distribution deviates from the exponential distribution toward a long-tailed distribution. For the particular case of the ATS, the validity also depends critically on the total number of alarms. Here, an average of at least 5 alarms is required, or more than 0.2 alarms per hour for a duration of 1 day. This conflicts with the goal to trigger as few false alarms as possible in order to gain optimal performances. Overall, STS were found to be applicable in a higher number of prediction settings than BST and ATS. The analytical random predictor is statistically conservative. It might have a slightly lower power in some cases, but in general it complies to the given significance level and hence constitutes a valid test for statistical significance.

For other fields of event predictions, the situation is similar, like in the field of earthquake prediction. Here, the place of occurrence is another parameter that has to be taken into account. However, when restricting to a confined area, the considerations presented above apply analogously. Similar to epileptic seizures, earthquakes with significant impact are rare events. While the absolute occurrence rates are much smaller for high-magnitude earthquakes than the seizure rates considered in this study, they also can be modeled by a gamma distribution [25]. If, instead of a duration of 1 day for the seizure prediction example, an exemplary time frame of 2 years is considered with the same average number of 3.6 events per year, the situation is directly transferable. An OP of 30 min for $T_{\mathrm{sim}} = 1$ day would correspond to an OP of about 10 days here. Hence it can likewise be concluded that for numerical validation methods comparable numbers of events are required.

In all of these fields the application of analytical validation methods constitutes a robust and valid approach in order to test whether prediction methods are indeed *better than chance*.

ACKNOWLEDGMENTS

We thank Jan Beyersmann and Klaus Lehnertz for fruitful discussions. This work was supported by the German Federal Ministry of Education and Research (Grant No. 01GQ0420), the German Science Foundation (Grant No. Ti315/4-2), the European Union (Grant No. 211713), and the Excellence Initiative of the German Federal and State Governments. B.S. is indebted to the Baden-Württemberg Stiftung for the financial support of this research project.

[1] R. Abercrombie, Nature (London) **438**, 171 (2005).
[2] F. Mormann, R. G. Andrzejak, C. E. Elger, and K. Lehnertz, Brain **130**, 314 (2007).
[3] R. A. Kerr, Science **304**, 946 (2004).
[4] A. Sol and H. Turan, Sci. Eng. Ethics. **10**, 655 (2004).
[5] R. G. Andrzejak, D. Chicharro, C. E. Elger, and F. Mormann, Clin. Neurophysiol. **120**, 1465 (2009).
[6] M. Winterhalder, T. Maiwald, H. U. Voss, R. Aschenbrenner-Scheibe, J. Timmer, and A. Schulze-Bonhage, Epilepsy Behav. **4**, 318 (2003).
[7] R. G. Andrzejak, F. Mormann, T. Kreuz, C. Rieke, A. Kraskov, C. E. Elger, and K. Lehnertz, Phys. Rev. E **67**, 010901 (2003).
[8] T. Kreuz, R. G. Andrzejak, F. Mormann, A. Kraskov, H. Stogbauer, C. E. Elger, K. Lehnertz, and P. Grassberger, Phys. Rev. E **69**, 061915 (2004).
[9] W. Chaovalitwongse, L. Iasemidis, P. Pardalos, P. Carney, D.-S. Shiau, and J. Sackellares, Epilepsy. Res. **64**, 93 (2005).
[10] F. Mormann, T. Kreuz, C. Rieke, R. G. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz, Clin. Neurophysiol. **116**, 569 (2005).
[11] B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, A. Schulze-Bonhage, and J. Timmer, Chaos **16**, 013108 (2006).
[12] S. Wong, A. B. Gardner, A. M. Krieger, and B. Litt, J. Neurophysiol. **97**, 2525 (2007).
[13] B. Schelter, R. G. Andrzejak, and F. Mormann, in *Seizure Prediction in Epilepsy—From Basic Mechanisms to Clinical Applications*, edited by B. Schelter, J. Timmer, and A. Schulze-Bonhage (Wiley-VCH, Weinheim, 2008), p. 237.
[14] E. L. Lehmann, *Testing Statistical Hypotheses* (Chapman & Hall, New York, 1986).
[15] B. Efron, Ann. Stat. **7**, 1 (1979).
[16] E. Mammen, *When Does Bootstrap Work? Asymptotic Results and Simulations*, Lecture Notes in Statistics Vol. 77 (Springer, New York, 1992).
[17] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap* (Chapman & Hall, New York, 1998).
[18] D. W. K. Andrews, Econometrica **68**, 399 (2000).
[19] S. R. Haut, C. Swick, K. Freeman, and S. Spencer, Epilepsia **43**, 711 (2002).
[20] D. E. Snyder, J. Echauz, D. B. Grimes, and B. Litt, J. Neural. Eng. **5**, 392 (2008).
[21] A. Schulze-Bonhage, F. Sales, K. Wagner, R. Teotonio, A. Carius, A. Schelle, and M. Ihle, Epilepsy Behav. **18**, 388 (2010).
[22] P. Suffczynski *et al.*, IEEE Trans. Biomed. Eng. **53**, 524 (2006).
[23] I. Osorio, M. G. Frei, D. Sornette, J. Milton, and Y.-C. Lai, Phys. Rev. E **82**, 021919 (2010).
[24] T. Schreiber, Phys. Rev. Lett. **80**, 2105 (1998).
[25] A. Corral, Phys. Rev. Lett. **92**, 108501 (2004).