Rao Vikram R (Orcid ID: 0000-0002-6389-2638)
Tcheng Thomas K (Orcid ID: 0000-0002-5491-5295)
Duun-Henriksen Jonas (Orcid ID: 0000-0003-1558-8225)
Baud Maxime Olivier (Orcid ID: 0000-0002-8297-7696)

Learning to generalize seizure forecasts

Marc G. Leguia[1*], Vikram R. Rao[2], Thomas K. Tcheng[3], Jonas Duun-Henriksen[4], Troels W. Kjær[5,6], Timothée Proix[7,**] & Maxime O. Baud[8,9**]

[1]Wyss Center fellow, Sleep-Wake-Epilepsy Center, Center for Experimental Neurology, NeuroTec, Department of Neurology, Inselspital Bern, University Hospital, University of Bern, Switzerland

[2]Department of Neurology and Weill Institute for Neurosciences, University of California, San Francisco, United States

[3]NeuroPace, Inc., Mountain View, California, United States

[4]UNEEG medical A/S, Allerød, Denmark

[5]Department of neurology, Zealand University Hospital, Roskilde, Denmark

[6]Department of clinical medicine, University of Copenhagen, Copenhagen, Denmark

[7]Department of Basic Neurosciences, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[8]Sleep-Wake-Epilepsy Center and Center for Experimental Neurology, Department of Neurology, Inselspital Bern, University Hospital, University of Bern, Switzerland

[9]Wyss Center for Bio and Neuroengineering, Geneva, Switzerland.

*correspondence: mgrauleg@gmail.com
**co last authors.

Acknowledgments : We thank the Wyss Center for Bio and Neuroengineering in Geneva for the support in this work.

Disclosure of Conflicts of Interest: MOB reports personal fees and grants from the Wyss Center for Bio and Neuroengineering in Geneva, and has a patent application pending under the Patent Cooperation Treaty (62665486). MOB holds shares with Epios, Ltd., a medical device company based in Geneva. VRR reports personal fees from NeuroPace, outside the submitted work. TKT is an employee of NeuroPace and receives salary and stock as compensation. JDH is an employee of UNEEG. TWK consults for UNEEG. All other authors declare no competing interests. We confirm that we have read the Journal's position on issues involved in ethical publication and affirm that this report is consistent with those guidelines.

Contribution: MGL, TP, VRR and MOB designed the study. TKT and JDH provided the data. MGL, TP and MOB performed the analysis, interpreted the data and drafted the manuscript. All authors edited the manuscript into its final version.

Word count

- Abstract: 295 words
- Main text: 5444 words

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/epi.17406

Abstract

Objective: Epilepsy is characterized by spontaneous seizures that recur at unexpected times. Yet, using years-long EEG recordings, we previously found that patient-reported seizures consistently occur when interictal epileptiform activity (IEA) cyclically builds up over days. This multidien (multi-day) interictal-ictal relationship, which is shared across patients, may bear phasic information for forecasting seizures, even if individual patterns of seizure timing are unknown. To test this rigorously in a large retrospective dataset, we pre-trained algorithms on data recorded from a group of patients, and forecasted seizures in other, previously unseen patients.

Methods: We used retrospective long-term data from participants (N=159) in the RNS System clinical trials, including intracranial EEG recordings (icEEG), and from two participants in the UNEEG clinical trial of a sub-scalp EEG system (sqEEG). Based on IEA detections, we extracted instantaneous multidien phases and trained generalized linear models (GLMs) and recurrent neural networks (RNNs) to forecast the probability of seizure occurrence at a 24-hour horizon.

Results: With GLMs and RNNs, seizures could be forecasted above chance in 79% and 81% of previously unseen subjects with a median discrimination of AUC=0.70 and 0.69 and median Brier skill score of BSS=0.07 and 0.08. In direct comparison, individualized models had similar median performance (AUC=0.67, BSS=0.08), but for fewer subjects (60%). Moreover, calibration of pre-trained models could be maintained to accommodate different seizure rates across subjects.

Significance: Our findings suggest that seizure forecasting based on multidien cycles of IEA can generalize across patients, and may drastically reduce the amount of data needed to issue forecasts for individuals who recently started collecting chronic EEG data. In addition, we show that this generalization is independent of the method used to record seizures (patient-reported vs. electrographic) or IEA (icEEG vs. sqEEG).

Key words: Multidien, Seizure forecasting, Intracranial EEG, Subscalp EEG, Transfer learning.

Key Points

1. Most prior seizure forecasting models were individualized for a given patient, based on large amounts of chronologically recorded data.
2. Pre-trained models that can transfer to previously unseen patients may drastically diminish the need for individual data
3. Models that are data-source agnostic can forecast seizures for individuals using intracranial or sub-scalp EEG devices
4. The generalizability of forecasting at long horizons enables pooling data across patients for training more sophisticated models
5. General forecasting methods may serve as a basis for later individualization, once data accumulates longitudinally

Introduction

The unpredictability of seizures creates stress and disability for many people with pharmacoresistant epilepsy, regardless of the seizure rate, focus or symptoms[1,2]. Patients and families participating in a recent survey stated that a reliable method to forecast seizure risk over a 12- to 24-hour horizon could help plan the day around seizures[3], addressing a central problem in epilepsy[4,5] and potentially improving psychological well-being for all. Whether such forecasting schemes, akin to those used for weather, can one day be deployed in widespread clinical practice hinges upon finding facile ways of tracking fluctuations in seizure risk and integrating shared and individual temporal patterns of seizure occurrence.

A decade ago, the prospective NeuroVista trial demonstrated in 9 out of 15 included participants that a certain degree of predictability minutes ahead of seizures was indeed possible with a warning system relying on real-time analysis of intracranial EEG[6]. Despite this technical feat, it remained unclear whether seizure warnings may be useful to users, in part due to the variability in seizure rates, and the time required to personalize the algorithms[6–8]. At the current stage of knowledge, launching much-needed further prospective seizure forecasting trials can be discouraging, given the scientific risk entailed and the amount of data (i.e. months to years) and resources needed (e.g. algorithm personalization and optimization). For example, post hoc studies on the NeuroVista dataset showed that forecasting performance crucially depended on tuning the right algorithm to individual factors, a resource-intensive optimization effort that required crowd-sourcing the problem within an international community of machine learning experts[9,10]. Ideally, a seizure forecasting algorithm would leverage fundamental determinants of seizure timing such that training in one group of individuals would yield a forecaster applicable to many others.

In the meantime, the accumulation of long calendar and EEG data (over months to years) in hundreds to thousands of patients has revealed that people with epilepsy have repeating seizure patterns that can be both individual and shared at the group level[11–15]. Crucially, studies in which central (EEG)[13,15–17] or peripheral (wrist sensor)[18] biomarkers were chronically monitored converged to the same conclusion: interictal and ictal epileptic activity is modulated cyclically over a range of circadian (about 24 hours)[13,16,17], multidien (many days)[13,15,17], and circannual (about yearly) timescales[13]. Based on these novel, but widely-accepted observations[19,20], a number of studies have shown that seizure cycles can be forecasted with invasive[15,21], minimally-invasive[22], and non-invasive methods[23–25]. In fact, leveraging the robustness of multidien cycles in epilepsy, we previously showed that models trained on one individual's prior recordings could forecast seizure risk days in advance for the same individual[21], an unprecedented forecast horizon that may enable a range of seizure mitigation strategies. In this probabilistic approach, dichotomy (predicted *presence* or *absence* of a seizure) is deemphasized to the benefit of forecasting cyclical fluctuations in the likelihood of upcoming seizures (forecasted *probability* between 0% and 100%), reminiscent of the approach taken to weather forecasting[26].

Because multidien cycles span several days to weeks[20], learning their association with relatively low and varying seizure rates within individuals necessitates acquiring months to years of longitudinal data. Yet, the phasic relationship between seizures and underlying multidien cycles of interictal activity is largely shared across patients, with seizure likelihood increasing when IEA rises over days[13],

suggesting the generalizability of a forecasting scheme based on multidien cycles. Thus, we hypothesized that models pre-trained on the multidien cycles of a subset of participants could be transferred to other, previously unseen subjects, thereby obviating the need for individual months-long training data[21].

In stark contrast to prior seizure prediction competitions, where hundreds of machine-learning models were generated for single patients[9,10], we here generalize a single explicit probabilistic approach (generalized linear models, GLMs) as well as a single recurrent neural network (RNN)[27] to hundreds of patients from a large retrospective dataset. The primary outcome of this post hoc study is the percentage of 'significant subjects' for whom forecasts performed above chance at a 24-hour horizon using and comparing pre-trained (trained on unrelated datasets from different subjects) and individualized models (trained on chronological data within subject). As the secondary study outcome, we quantify and compare models' performance in terms of discrimination, resolution, and calibration for subjects with forecasts above chance (see also review by Baud et al. in this issue for a technical introduction)[26].

Methods

Subjects and data.

Intracranial EEG (icEEG) cohort. As reported previously[21], we analyzed data from 159 out of the 256 adults with medically-refractory focal epilepsy who participated in the U.S. clinical trials of the RNS® System (NeuroPace, Inc., Mountain View, CA), a chronically implanted intracranial device that treats seizures with direct brain-responsive neurostimulation[28–30]. Subjects who did not report disabling seizures, who had more than 50% of disabling seizure days, or who had less than 6 months of continuous EEG data were excluded (Figure 2). We included two distinct types of data in our analysis: (1) interictal epileptiform activity (IEA), defined as hourly counts of detections of typically brief epileptiform discharges, such as spikes, sharp waves, and faster oscillations[17]; (2) diaries of self-reported seizures, recorded as counts per calendar day of patient-classified 'complex partial,' or 'generalized tonic-clonic' seizures. Seizures classified by patients as 'simple motor' and 'simple other' (auras) are often less accurate and clinically less relevant than disabling seizures and were not included in the analyses[31]. Retrospective analysis of the data was approved by the Institutional Review Boards of the participating centers.

Subscalp EEG (sqEEG) subjects. We forecasted seizures in two subjects out of the 9 adults with medically-refractory temporal lobe epilepsy who participated in the Danish trial of a chronically implanted subscalp device (24/7 EEG™ SubQ, UNEEG™ medical, Allerød, Denmark) that provides nearly-continuous EEG recordings via two bipolar electrodes inserted between the skull and the temporalis muscle[32,33]. Seven subjects who provided discontinuous or short data (headpiece in place <90% of the time for <60 days) and had less than 10 recorded electrographic seizures were excluded.

Data pre-processing. The RNS System provides a limited form of icEEG by continuously monitoring intracranial EEG signals from the seizure focus/foci and quantifying detections of epileptiform activity per hour (IEA) based on embedded algorithms relying on calculations of the line-length, area-under-the-curve, and band-pass amplitude of the EEG signal. To derive the IEA from continuous subscalp EEG, we developed an interictal epileptiform discharge detection algorithm suitable for extracranial EEG (see Supplementary materials, Figures S1 and S2).

Data processing. Extracted IEA data was processed as described previously[21]. We first calculated daily counts of IEA by summing the counts over a calendar day and interpolated gaps of short duration. We then computed the *broadband* Morlet wavelet transform of IEA daily counts for periods between 4 to 45 days and averaged the resulting complex coefficients over that period range to extract the instantaneous phase[13]. This is unlike our prior approach on the same cohort, where individualized forecasts relied on the phases of compound multidien cycles[20] (up to three distinct peak-periodicities) derived from up to three *narrow band* filters[21]. We favored the *broadband* method after noticing that it emphasizes the instantaneous dominant periodicity[13], while accommodating different periodicities in different patients in the form of a single input feature into period-agnostic models (Figure S3). For the sqEEG subjects, we could only evaluate periodicity from 4 to 25 days given the shorter duration of

recording. Note that the estimation of the phase using a wavelet transform is a non-causal estimation since it uses future information to compute the instantaneous phase at a given point.

Point process generalized linear models (PP-GLMs). To regress a number of input variables to seizure probabilities, we used PP-GLMs that are flexible and highly interpretable statistical models[17,27]. PP-GLMs can evaluate the association between a sequence of event (seizure) times, and temporal features upon which the event probability may depend. Here we used PP-GLMs with a log-link function and a (conditionally) Poisson distribution to forecast the probability of a seizure as a function of features extracted from the most recent seizure times, the most recent counts of IEA, and the instantaneous phase of the multidien cycle. This probability is related to the 'instantaneous' rate or conditional intensity function λ (t) of the point process, here modeled as:

$$log\ (\lambda \mid S_{\{t\text{-}1,\ldots t\text{-}p\}}, I_{\{t\text{-}1,\ldots t\text{-}p\}}, \theta_{\{t\text{-}1\}} = \beta_0 + \beta^{cos}\cos(\theta_{t\text{-}1}) + \beta^{sin}\sin(\theta_{t\text{-}1}) + \sum_{i=1}^{p}\left(\beta_i^S S_{t\text{-}i} + \beta_i^I I_{t\text{-}i}\right) \quad (1)$$

with θ_t the instantaneous non-causal phase estimation of the broadband IEA at time t, β_0 the intercept, β^{cos} and β^{sin} the model coefficients attributed to the broadband multidien phase, and β_i^S and β_i^I the model coefficients attributed to the most recent seizure timeseries S_t and recent IEA timeseries I_t with an identical number of time points p for both. For both the single feature individualized and pre-trained models, $p = 5$ days. Depending on the model used, the model coefficients are fitted for each subject separately (individualized models) or across subjects (pre-trained models, see below). In the results, we used models that included one or several of the features described above. Specifically, we build the following models: (i) only recent seizures ($\beta^{cos} = \beta^{sin} = \beta_i^I = 0$), (ii) only recent IEA ($\beta^{cos} = \beta^{sin} = \beta_i^S = 0$), (iii) only the instantaneous multidien phase ($\beta_i^S = \beta_j^I = 0$).

For the multifeature pre-trained models, we additionally included multiple time points of the most recent timeseries of the instantaneous phases to match the history of the recurrent neural network (see next section), extended to 50 days:

$$log\ (\lambda \mid S_{\{t-1,\ldots,t-p\}}, I_{\{t-1,\ldots,t-p\}}, \theta_{\{t-1,\ldots,t-p\}}) = \beta_0 + \sum_{i=1}^{p}\left(\beta_i^{cos}cos(\theta_{t-i}) + \beta_i^{sin}sin(\theta_{t-i}) + \beta_i^S S_{t-i} + \beta_i^I I_{t-i}\right) \quad (2)$$

using $p = 50$ days.

Recurrent neural network (RNN). RNN is a state-of-the art machine-learning technique for learning temporal relationships within time-series. We used gated recurrent units ('GRU'), a simple design of RNN[34], which can capture long-term effects such as underlying multidien cycles. We used 60 nodes with three layers of GRUs (Figure S4), followed by 2 dense layers of 30 and 15 nodes, respectively. Finally we added a dropout layer (dropout rate of 0.2) and an output dense layer of one unit with a sigmoid activation. The RNN used binary cross-entropy as a loss function and was trained with enough epochs to ensure performance. As in the multifeature pre-trained GLM, input data included seizures, IEA and multidien phases over the 50 previous days in order to account for the longest multidien cycles considered in this study and for comparison with prior studies[27].

Training and testing

We used these models to learn from data in two radically different schemes, using the same proportion of training and testing data for better compatibility:

Individualized training. As in our prior study, individualized forecasts were obtained by training a PP-GLM for each subject independently[21]. We used the minimum between 60% of the data or 480 days for the training set, and the remaining data for the testing set.

Pre-training. In this study, pre-trained models designate models (GLMs or RNNs) trained on a subset of the subjects (60%) and tested on remaining subjects that were not part of the training set (40%). For the subjects in the training set, we used all the data available. For the subjects in the test set, we used the same test data as would have been used for the individualized model to allow for direct comparability (Figure 3). Pre-trained models were thus used to output forecasts for unseen subjects, in a transfer learning design. Train and test sets were chosen by randomly partitioning the pool of subjects (40 times) such that all subjects were at least five times in the test set (cross-validation). For example, subject 10 could be in the train set for the first fold of the cross-validation, and in the test set for the second one. For each iteration, data across subjects were concatenated separately for the training and test sets. Issued forecasts for a given subject were deemed significant only when passing statistical testing five times (see below), and their performance was averaged across the five times to ensure the robustness of our results to random selection of training sets. For comparison between more complex models (i.e. multifeature GLM vs. multifeature RNN in Figure 4) all available data from test subjects was used since the direct comparison to single feature models was no longer the focus.

Forecasting performance. Performance of the forecast was assessed with three distinct scores, drawn from the same total testing data, regardless of the training scheme, ensuring direct comparability:

Percentage of significant subjects (% sig). The percentage of subjects for whom forecasts can be issued above-chance, i.e. for whom a significant improvement of the AUC is obtained compared to the AUC obtained when using randomized surrogate data (see below). As seizure forecasting is not expected to work for all patients, performance is further quantified only in those with above-chance forecasts. Beyond a simple improvement over chance, which may not be clinically meaningful, the AUC and BSS offer complementary insights in order to assess the goodness of the forecasting model in subjects with above-chance forecasts.

Area under the sensitivity vs. time in warning curve (AUC): The AUC is a deterministic score that assesses *discrimination*: how well-separated are the forecasts associated with seizures compared to forecasts associated with the absence of seizures. Instead of the classical receiver operating characteristic of the sensitivity versus specificity, we computed the sensitivity versus (corrected) time in warning at different threshold forecasted probabilities (see Supplementary materials), as customary in the field of seizure forecasting[35]. As a deterministic score, we used the AUC under this curve to assess the discrimination performance of our forecasting models. The AUC is 1 when perfect discrimination between

seizure and non-seizure instances is obtained for all thresholds, and close to 0.5 when the forecast is no better than a random classification.

Brier skill score (BSS). As a complement to the AUC, the BSS[36–38] is a probabilistic score that assesses *resolution* and *calibration*: how well do different forecasted probabilities (*a priori*) correspond to different observed probabilities (*a posteriori*) of having a seizure. We refer the reader to a review by Baud et al. in this issue for an in-depth discussion on this technical topic[26]. The Brier Score (BS) is defined as the mean squared distance between a forecasted probability (f_i) and an observation o_i, (1 for a seizure and 0 for no seizure):

$$BS = \frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$$

where n is the number of forecasted time points. The Brier Skill Score (BSS) calculates the improvement of the Brier score (BS) over a reference forecast as:

$$BSS = 1 - \frac{BS}{BS_{ref}}$$

where BS_{ref} is an uninformative reference forecast. In our case, we generated 1000 randomly shuffled versions of the forecasted probabilities on which we calculated a mean reference BS. The BSS tends to 1 when issued forecasts are sharp and perfectly represent the true probability of a seizure to occur (i.e. they tend to be deterministic), and to 0 when there is no improvement over the reference (negative values are possible for forecasts worse than the reference). The BSS is a compound score that captures a range of probabilistic performance between 0 and 1 and can be decomposed in terms of resolution and calibration for in-depth diagnostics (see Baud et al., Epilepsia, 2022).[37]

Calibration

Calibration measures how well observed probability and forecasted probabilities agree with each others and is defined as a calibration loss (distance to the diagonal line in the reliability diagram):

$$\mathrm{Cal} = \frac{1}{n}\sum_{k=1}^{m} n_k(\overline{f_k} - \overline{o_k})$$

where $\overline{f_k}$ and $\overline{o_k}$ are respectively the forecasted probabilities and the observed probabilities averaged over k bins (here 10 bins equally spaced, 0-10%, 10-20%, etc.). A lower calibration loss indicates better performance.

Resolution

As a complement, resolution measures how much the forecasted probabilities depart from the long-term prevalence of events (distance to the horizontal 'no resolution' line in the reliability diagram):

$$\mathrm{Res} = \frac{1}{n}\sum_{k=1}^{m} n_k(\overline{o_k} - \overline{o})$$

Where $\overline{o_k}$ is the observed probabilities averaged in k bins and $\bar{o}$ is the long-term average event rate. A higher resolution indicates better performance.

Surrogates and statistical analysis. Surrogate time series[39] for models relying on occurrence times of recent seizures were obtained by randomly shuffling the seizure test time series, under the null hypothesis that the seizure process is memoryless (i.e., events are independent of one another). To obtain surrogates for models relying on IEA or the underlying multidien phase, the daily IEA test time series were shuffled, under the null hypothesis that seizure timing does not depend on specific values or trends in IEA[31]. Just like for individualized models, the surrogates were also obtained for each subject independently to test the performance of the pre-trained models. This computationally-effective design allowed for testing all temporal features in all models, using 200 chance-level surrogate datasets. Additionally, we also performed the computationally-intense surrogate design used in Proix et al., 2020[21] for training and testing a number of models proposed here, which led to very similar statistical results (Table S1). For each tested subject, we assessed the significance by correcting the p-values for multiple comparisons across subjects and models (one for the individualized model, five for the pre-trained model) by using the false discovery rate (FDR) with a target α=0.05.

Results

We retrospectively analyzed data collected from 159 (74 females) out of 256 participants in the RNS System clinical trials with median (range) age 35 (25-43) and mesiotemporal epilepsy for the majority (67%, the rest neocortical, see[21] for full description). Intracranial EEG was monitored over a median of 1722 (321-3509) days and continuous records of hourly IEA detections were stored by the device, with a median loss of data of 21 days due to infrequent downloads. Over this duration, subjects reported a median of 143 (13-1233) days with at least one disabling seizure. Additionally, we included two female patients aged 33 and 38 with temporal lobe epilepsy who participated in the trial of a sub-scalp EEG system during 75 and 95 days and met our preset criterion of >10 recorded seizures.

To illustrate the main concept of this study, we show how a model pre-trained on data from one person with epilepsy due to periventricular temporo-occipital heterotopias (Figure 1A) can be transferred to forecast seizures for another, previously unseen individual with bitemporal epilepsy (Figure 1B). Even when two subjects exhibit different periodicities of IEA cycles (around 11 and 26 days in Figure 1C), their seizures tend to occur at similar phases of these cycles (Figure 1D). Knowing the instantaneous multidien phase allowed for transferring a GLM pre-trained to forecast seizures in one subject to another subject with good performance (Figure 1E-F; AUC = 0.73), illustrating the generalizability of seizure forecasting at the timescale of days.

Following the process outlined in Figure 1, we systematically pre-trained single-feature GLMs with a subset of the subjects (60%, median of 167,363 training days across subjects) and tested them on the remaining, unseen subjects (40%, median of 86,086 test days across subjects), repeating this operation to test each subject with five different pre-trained models (cross-validation, see Methods, Figure 2). For comparison, we repeated our previously published individualized forecasting scheme[17], by chronologically training GLMs on early data (median of 475 training days per subject) and testing on the remaining later data (≥ 40% of data, median of 1242 testing days per subject). As a primary outcome, we found that the proportion of significant subjects with forecasts above chance was 62% vs. 60% for pre-trained and individualized models, respectively, when relying on multidien phase as the sole input feature (Figure 3A, % Sig in Table 1). Models trained on recent IEA or recent seizure timeseries alone had lower performance (Table 1 and S1, Figure S5).

As a secondary outcome, we quantified forecast performance only for the significant subjects for whom forecasts were above chance. For a complete interpretation of Table 1 and Figure 3, it is useful to clarify the difference between *deterministic* and *probabilistic* scores. The AUC is deterministic, as it evaluates the ability of the forecast to discriminate two categorical observations: seizure versus no-seizure. The BSS is probabilistic, as it evaluates the correspondence between forecasted seizure probabilities (*a priori*), and observed seizure frequencies (*a posteriori*) within probability strata (here: 0-10%, 11-20%, etc.) in terms of calibration and resolution[26]. For completeness and to facilitate comparability with other studies, we report in Table 1 both (1) individual *scores per subject* that are solely sensitive to time-varying probabilities[6,8–10,21,25,40], and (2) aggregate *scores across subjects*[27] obtained by pooling daily forecasts and observations from all subjects, which are strongly influenced by the large difference in seizure rates across included subjects (here, from <1% to 49% seizure days). First, we evaluated discrimination per subject of our pre-trained models using the multidien phase and found a median AUC = 0.68 (interquartile range (IQR): 0.65-0.76, Table 1, Figure 3A), comparable to the individualized models with a median AUC = 0.67 (IQR 0.63-0.72) per subject. Sensitivity and time-in-warning at a chosen threshold was also similar (Figure S6). Second, the per-subject distribution of probabilistic scores for the pre-trained models was below that of individualized models, resulting in a lower median BSS[37,41] of 0.05 (IQR 0.03-0.07) versus 0.08 (IQR 0.03-0.14, Table 1, Figure S5B). This shows that despite a similar degree of discrimination, simple pre-trained models lack probabilistic performance.

To characterize the mechanisms by which our models issued time-varying forecasts with above-chance AUCs, we inspected the coefficients (β_0, β^{cos}, and β^{sin} in equation 1) learned by each model pre-trained with the multidien phase as a feature. On one hand, pretrained models had stable parameters regardless of the sub-sample of subjects used for training, capturing the central tendency in the included population (white dots in Figure 3C). On the other hand, the coefficients of the individualized models had a wider range of values (half-violin in Figure 3C), but the median of those distributions (black horizontal lines in Figure 3C) coincided with the stable parameters found in the pre-trained models (white dots in Figure 3C). Additionally, in individualized models based on the multidien phase, the intercept (coefficient β_0) learned from the data was almost perfectly correlated with the individual observed seizure rates across subjects (Figure 3D $R^2 = 0.99$, $p < 10^{-120}$, Pearson's correlation), whereas, by design, the phase was captured by β^{cos} and β^{sin}. Based on this observation, we recalibrated pre-trained models for each test subject by individualizing the intercept β_0 using the average number of

seizures per day of the corresponding test subject (i.e. introducing a random-effect), while β^{cos} and β^{sin} retained their values (i.e. a fixed-effect). This post-training reparameterization that simply requires estimation of an individual's long-term expected seizure rate (e.g. one seizure per week) improved aggregate resolution of the model across subjects nearly to the level of individualized forecasts (overlap of dashed-red and gray curves in Figure 3B), resulting in a BSS across subjects of 0.21. This reparameterization did not improve the BSS per subject (median 0.04 (IQR 0.03-0.07), Table 1) but resulted in an increase in calibration per subject (Table S2). To summarize, single-feature pre-trained multidien GLMs were able to learn the shared tendency of seizures to cluster within rising multidien phases across subjects (Figure 3F). This learning can be transferred to forecast seizures for new subjects yielding similar discrimination independently of their individual seizure rate (Figure 3E), with the possibility to re-calibrate the output probabilities (Figure 3B).

Using pre-trained models across subjects increases the amount of data available for training, thus allowing the use of more complex models with more trainable parameters while avoiding overfitting. We first developed a multifeature GLM that combined all features previously used in isolation: recent seizures and IEA counts, as well as multidien phases over the past 50 days. These additional temporal features (or 'covariates'), may yield information about seizure risk that is not entirely captured by multidien cycles alone. Indeed, compared to the pre-trained multidien GLM (Figure 3A), the pre-trained multifeature GLM (Figure 4A) increased the proportion of subjects with forecasts above chance from 62% to 79% and discrimination from AUC = 0.68 to AUC = 0.70 (Table 1). Additionally, records of past seizures could inform the model about the long-term expected seizure rates, resulting in more calibrated forecasts without the need for re-calibration. Indeed, using seizure occurrence over 50 previous days as an initialization, the pre-trained multifeature GLM yielded better calibrated outputs (Figure 4B) than the pre-trained multidien GLM (Figure 3B), as shown by a median BSS of 0.07 (IQR 0.05 – 0.1, Table 1). This result indicates that discriminative and calibrative information is present in the recent IEA and seizure counts, that is independent of the multidien phase (Figure S7). Using all available information, we showed that pre-trained GLMs could forecast risk above chance two to five days in advance in some subjects (Figure S8).

Asking whether more complex models may better capture seizure risk, we then trained a multifeature recurrent neural network (RNN) across subjects with the same input features and training-testing design as before. Compared to the multifeature GLMs, multifeature RNNs pre-trained on 50 days of past seizures, IEA, and multidien phases increased the proportion of significant subjects slightly to 81% and improved the BSS per subject to a median of 0.08 (IQR 0.06-0.11, Table 1, Figure 4). However, the AUC per subject did not increase (Figure 4A, Table 1), indicating that simpler models can have good discrimination, but adding trainable parameters can further improve calibration. As a side note, the AUC and BSS across subjects seemed to mostly be influenced by varying observed seizure rates across subjects, as these scores did not improve with increasing model complexity as opposed to the per-subject scores (Table 1, see also Figure S9).

To characterize the relationship between electrographic and self-reported seizures, an important clinical issue, we evaluated the transferability of pre-trained models across seizure types. To that end, we used a cohort of N=18 subjects with focal epilepsy in which electrographic seizures were identified (see Supplementary materials and Proix et al., 2020)[21]. We found that training on electrographic seizures and comparing forecasted risk to seizures documented in diaries held by other subjects

yielded above-chance discrimination with a median AUC per subject 0.68 (IQR 0.64-0.74, Figure S9). Conversely, training with seizures taken from subject-held diaries and comparing forecasted risk to electrographic seizures recorded from other subjects resulted in a median AUC per subject of 0.68 (IQR 0.66-0.77, Figure S10).

Finally, given the growing use of less invasive technology such as sub-scalp EEG, we asked whether pre-trained models could transfer across recording modalities. First, we showed that multidien cycles of IEA can be detected with subscalp EEG, reflected in fluctuations in epileptic spike rates[40] and, to a lesser extent, in the variance of sqEEG (Figure 5A-C), although the discontinuity of data in many patients represents a technical challenge[32]. Second, using sqEEG only for testing, we showed that a pre-trained model relying on multidien phases of IEA fluctuations recorded with icEEG (Figure 5D) was able to forecast seizures based on multidien phases from IEA or variance fluctuations recorded with sqEEG in one subject (Figure 5A) with a discrimination of AUC = 0.69 (p-value <0.05 for both, Figure 5E). A second subject also had multidien cycles recorded with sqEEG (Figure S11) but pre-trained models could not forecast seizures that occurred at a different (falling) phase in this particular case, as seen in a few subjects in Figure 3.

Discussion

Aiming at transfer learning to forecast seizures, we used statistical (GLMs) and deep learning models (RNNs) to show a number of novel and robust results, drawn from the longest existing dataset of combined chronic EEG recordings and self-reported seizure diaries. In the framework of generalized linear models, we characterized the transferability of shared and individual ictal and interictal temporal patterns across patients. First, we showed that models pre-trained on the phase of underlying multidien cycles from a subset of subjects can be transferred to other, previously unseen subjects to discriminate periods of relatively high and low risk. We then turned to multifeature models that could forecast time-varying risk while maintaining the calibration of output probabilities, by capturing the individual long-term expected seizure rate. Finally, in a limited proof-of-principle analysis, we showed that our models depend neither on the seizure counting methodology (electrographic vs. self-reported) nor on EEG acquisition modality, including a minimally-invasive device[33,40]. In essence, learning from shared ictal-interictal phasic relationship across subjects and individual seizure rates, our models forecasted varying seizure risk over days, highlighting the generalizability of extrapolating seizure cycles in a large majority of people (here ~80%) with different focal epilepsies and seizure periodicities[13].

This study has limitations. Participants in the RNS System clinical trials may not be representative of all people with epilepsy, though inclusion of two patients with sqEEG may help address this concern. To emphasize the clinical relevance of our approach, we used self-reported seizures to construct most of our models, knowing that seizure diaries themselves can be inaccurate[42]. Future studies based on electrographic seizures will likely surpass model performances presented here. On the other hand,

our estimations of multidien phases were accurate but non-causal. In a prospective trial, a causal filter will be necessary, which may reduce the accuracy of the forecasts. Finally, we focused on the horizon of one to a few days, and have not included circadian cyclicity in this study[8,21], as it is known that the circadian time at which seizures tend to occur is variable across individuals[13,16,17,43]. In contrast, the phasic relationship at the multidien timescale is more generalizable[20], unless an outlier individual has seizures on a different phase (e.g. cycle trough, as shown for a few in Fig. 3), in which case a personalized model would be necessary.

In direct comparison to individualized models, and when combining all available input features (here, recent seizures and IEA, and multidien phase), our best pre-trained models generalized to ~80% vs. ~60% of the cohort while maintaining deterministic (AUC~0.7) and probabilistic (BSS~0.08) performance, but requiring only 50 vs. 480 days for training (i.e. ~10 times less). More sophisticated RNNs, which typically require larger datasets for training, further improved calibration while maintaining discrimination, illustrating how distributing learning across subjects enables state-of-the-art machine learning approaches in epilepsy. Thus, pre-trained models have key advantages over individualized models: (i) they are trained on varied data from different patients, thus avoiding overfitting; (ii) by pooling larger amounts of data, models of increasing sophistication can be used; and (iii) they allow for drawing general conclusions about seizure forecasting across patients, highlighting interpretability and the robustness of the results.

With few exceptions[27], previous seizure forecasting models were patient-specific and sought to maximize discrimination of minutes-long epochs preceding seizures[6–10,25], a classification problem which is blind to the model calibration. Using this approach, crowd-sourced seizure prediction competitions[9,10] benchmarked achievable AUC at ~ 0.8. In contrast, our study presents seizure forecasting as a generalizable regression problem involving capture of different degrees of time-varying seizure risk. Median discrimination obtained here at a time horizon of one to a few days was lower with AUCs of ~ 0.7. In addition to the obvious difference in forecasting horizon, we here emphasize the importance of assessing forecasts both deterministically and probabilistically for an in-depth understanding of their true value[26,37]. Forecasted probabilities, in our case, were calibrated (Figure 4B) and evolved smoothly over time, avoiding flickering predictions at short horizons, which can lead to alarm fatigue and defy the intended purpose of risk management. In the framework of temporal fluctuations in latent seizure risk[44], the BSS[8,21,38,40] score calculated *per* subject complements the AUC because it accounts for the fact that a true transient heightened seizure risk does not always result in the occurrence of a seizure, while a deterministic assessment would consider this case a 'false positive'. Whether a BSS of ~0.1 can be clinically useful is a question for prospective trials, but resolution of risk in time will undoubtedly increase further with incorporation of additional covariates (i.e. temporal risk factors). Related to this issue, aggregate scores across subjects (as presented here and elsewhere[27]) can be misleadingly high as they are largely influenced by variable seizure rates across subjects. Although they help the forecaster obtain reliable statistics, for example by pooling a sufficient number of observations in the reliability diagram (as done here), scores *per* subject are of the essence from a user's standpoint in a clinical context, as they are solely influenced by time-varying risk.

Increasing interpretability, inspection of our model's coefficients enabled insight into how GLMs are fit to statistically model seizure timing with different rates in different patients. First, including broadband multidien phases allowed for capturing time-varying risk within subjects, on a range of timescales (here 4-45 days) and defining periods of high and low *relative risk,* agnostic to the individual long-term expected seizure rate. In the framework of GLMs, this covariate leverages the generalizable ictal-interictal phasic relationship and can be seen as a *fixed* effect that provides discrimination. Second, given that seizure rates are highly variable across individuals with epilepsy (including in this cohort), forecasts had to be (re-)calibrated to take on a true probabilistic meaning as an *absolute risk* (a probability between 0 and 1 of having a seizure). This was achieved either via a simple re-calibration of the model output or by initializing pre-trained models with the timing of recent seizure occurrences. In the framework of GLMs, this represents a *random* effect, providing calibration specific to each individual by capturing their long-term expected seizure rate. The generalizability of seizure forecasting over days can be interpreted from a theoretical viewpoint. Seizure occurrences can be modeled as an inhomogeneous Poisson process, wherein each individual has a set expected seizure rate with superimposed multidien cycles and other factors governing *relative* fluctuations around this rate[44]. In this view, seizures remain stochastic events drawn from a conditional Poisson distribution, while seizure risk evolves according to latent (potentially stochastic) dynamics[44].

Once a fanciful goal, the ability to anticipate seizures in people with epilepsy is now a near-term reality. Patient-specific forecasting algorithms that leverage cyclical patterns of brain activity have shown much promise[8,15,21,24], but collecting training data and tailoring models on an individual basis limits scalability. Results presented here highlight the generalizability of seizure forecasting models across individuals, including those with minimally invasive devices for chronic recording of brain activity. One-size-fits-all approaches are scarce in epilepsy, but our findings indicate that knowledge of fundamental temporal relationships between ictal and interictal activity powerfully enables transfer learning and may yield economies of scale in seizure forecasting. While gradual personalization will likely bring further increases in performance as data accumulate, readily available generalized models will facilitate imminent prospective seizure forecasting trials.

References

1. Dumanis SB, French JA, Bernard C, et al. Seizure forecasting from idea to reality. Outcomes of the my seizure gauge epilepsy innovation institute workshop. eNeuro. 2017; 4(6):0–11.
2. Moss A, Moss E, Moss R, et al. A Patient Perspective on Seizure Detection and Forecasting. Front Neurol. 2022; 13(February):11–4.
3. Grzeskowiak CL, Dumanis SB. Seizure Forecasting: Patient and Caregiver Perspectives. Front Neurol. 2021; 12(September).
4. Arthurs S, Zaveri HP, Frei MG, et al. Patient and caregiver perspectives on seizure prediction. Epilepsy Behav. 2010; 19(3):474–7.
5. Baud MO, Rao VR. Gauging seizure risk. Neurology. 2018; 91(21):967–73.
6. Cook MJ, O'Brien TJ, Berkovic SF, et al. Prediction of seizure likelihood with a long-term, implanted seizure advisory system in patients with drug-resistant epilepsy: A first-in-man study. Lancet Neurol. 2013; 12(6):563–71.
7. Kuhlmann L, Lehnertz K, Richardson MP, et al. Seizure prediction — ready for a new era. Nat Rev Neurol. 2018; 14(10):618–30.
8. Karoly PJ, Ung H, Grayden DB, et al. The circadian profile of epilepsy improves seizure forecasting. Brain. 2017; 140(8):2169–82.
9. Kuhlmann L, Karoly P, Freestone DR, et al. Epilepsyecosystem.org: crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. Brain. 2018; :2619–30.
10. Brinkmann BH, Wagenaar J, Abbot D, et al. Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain. 2016; 139(6):1713–22.
11. Baud MO, Perneger T, Rácz A, et al. European trends in epilepsy surgery. Neurology. 2018; 91(2):e96–106.
12. Karoly PJ, Goldenholz DM, Freestone DR, et al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. Lancet Neurol. 2018; 17(11):977–85.
13. Leguia MG, Andrzejak RG, Rummel C, et al. Seizure Cycles in Focal Epilepsy. JAMA Neurol. 2021; 78(4):454.
14. Ferastraoaru V, Goldenholz DM, Chiang S, et al. Characteristics of large patient-reported outcomes: Where can one million seizures get us? Epilepsia Open. 2018; 3(3):364–73.
15. Maturana MI, Meisel C, Dell K, et al. Critical slowing down as a biomarker for seizure susceptibility. Nat Commun. 2020; 11(1):2172.
16. Karoly PJ, Freestone DR, Boston R, et al. Interictal spikes and epileptic seizures: Their relationship and underlying rhythmicity. Brain. 2016; 139(4):1066–78.
17. Baud MO, Kleen JK, Mirro EA, et al. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun. 2018; 9(1):1–10.
18. Karoly PJ, Stirling RE, Freestone DR, et al. Multiday cycles of heart rate modulate seizure likelihood at daily, weekly and monthly timescales: An observational cohort study. Medrxiv. 2020; .
19. Khan S, Nobili L, Khatami R, et al. Circadian rhythm and epilepsy. Lancet Neurol. 2018; 17(12):1098–108.
20. Karoly PJ, Rao VR, Gregg NM, et al. Cycles in epilepsy. Vol. 0123456789, Nature Reviews Neurology. Springer US; 2021.
21. Proix T, Truccolo W, Leguia MG, et al. Forecasting seizure risk in adults with focal epilepsy: a development and validation study. Lancet Neurol. 2020; 4422(20):6–9.
22. RE Stirling, PJ Karoly, MI Maturana, ES Nurse, K McCutcheon, DB Grayden, SG Ringo, J Heasman, TL Cameron, RJ Hoare, A Lai, W D'Souza, U Seneviratne, L Seiderer KM, Cook M. Seizure Forecasting Using a Novel Sub-Scalp Ultra-Long Term EEG Monitoring System. Medrxiv. 2021; .
23. Stirling RE, Grayden DB, D'Souza W, et al. Forecasting Seizure Likelihood With Wearable Technology. Front Neurol. 2021; 12(July):1–12.
24. Karoly PJ, Cook MJ, Maturana M, et al. Forecasting cycles of seizure likelihood.

Epilepsia. 2020; (October 2019):2019.12.19.19015453.
25. Meisel C, El Atrache R, Jackson M, et al. Machine learning from wristband sensor data for wearable, noninvasive seizure forecasting. Epilepsia. 2020; 61(12):2653–66.
26. Baud MO, Proix T, Gregg NM, et al. Seizure forecasting: bifurcations in the long and winding road. Epilepsia. 2022; .
27. Goldenholz D, Goldenholz S, Romero J, et al. Development and validation of forecasting next reported seizure using e-diaries. Ann Neurol. 2020; :1–8.
28. Morrell MJ. Responsive cortical stimulation for the treatment of medically intractable partial epilepsy. Neurology. 2011; 77(13):1295–304.
29. Nair DR, Laxer KD, Weber PB, et al. Nine-year prospective efficacy and safety of brain-responsive neurostimulation for focal epilepsy. Neurology. 2020. 10.1212/WNL.0000000000010154.
30. Bergey GK, Morrell MJ, Mizrahi EM, et al. Long-term treatment with responsive brain stimulation in adults with refractory partial seizures. Neurology. 2015; 84(8):810–7.
31. Fertig E, Fureman BE, Bergey GK, et al. Inclusion and exclusion criteria for epilepsy clinical trials-Recommendations from the April 30, 2011 NINDS workshop. Epilepsy Res. 2014; 108(5):825–32.
32. Weisdorf S, Duun-Henriksen J, Kjeldsen MJ, et al. Ultra-long-term subcutaneous home monitoring of epilepsy—490 days of EEG from nine patients. Epilepsia. 2019; 60(11):2204–14.
33. Weisdorf S, Gangstad SW, Duun-Henriksen J, et al. High similarity between EEG from subcutaneous and proximate scalp electrodes in patients with temporal lobe epilepsy. J Neurophysiol. 2018; 120(3):1451–60.
34. Cho K, van Merrienboer B, Bahdanau D, et al. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. In: Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. Stroudsburg, PA, USA: Association for Computational Linguistics; 2014. p. 103–11.
35. Snyder DE, Echauz J, Grimes DB, et al. The statistics of a practical seizure warning system. J Neural Eng. 2008; 5(4):392–401.
36. Brier GW. Verification of Forecasts Expressed in Terms of Probability. Mon Weather Rev. 1950; 78(1):1–3.
37. Mason SJ. On using “climatology” as a reference strategy in the Brier and the ranked probability skill scores. Mon Weather Rev. 2004; 132(7):1891–5.
38. Jachan M, Feldwisch Genannt Drentrup H, Posdziech F, et al. Probabilistic forecasts of epileptic seizures and evaluation by the brier score. IFMBE Proc. 2008; 22(1):1701–5.
39. Schreiber T, Schmitz A. Surrogate time series. Phys D Nonlinear Phenom. 2000; 142(3–4):346–82.
40. Stirling RE, Maturana MI, Karoly PJ, et al. Seizure Forecasting Using a Novel Sub-Scalp Ultra-Long Term EEG Monitoring System. Front Neurol. 2021; 12(August):1–11.
41. Bröcker J, Smith LA. Increasing the reliability of reliability diagrams. Weather Forecast. 2007; 22(3):651–61.
42. Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. Lancet Neurol. 2018; 17(3):279–88.
43. Gowers WR. Epilepsy and Other Chronic Convulsive Diseases; Their Causes, Symptoms and Treatment. Churchill J& A, editor. London; 1881.
44. Baud MO, Proix T, Rao VR, et al. Chance and risk in epilepsy. Curr Opin Neurol. 2020; 33(epub).

			Individualized models					Pre-trained models				
			Scores per subject			Scores across subjects		Scores per subject			Scores across subjects	
			% Sig.	AUC	BSS	AUC	BSS	% Sig.	AUC	BSS	AUC	BSS
GLM	Single temporal feature	Recent seizures over 5 d	71 45%	0.58 (0.52-0.63)	0.05 (0.03-0.1)	0.74	0.21	47 30%	0.63 (0.60-0.68)	0.07 (0.04-0.1)	0.70	0.12
		Recent IEA over 5 d	50 31%	0.64 (0.61-0.68)	0.05 (0.03-0.09)	0.74	0.22	40 25%	0.63 (0.60-0.68)	0.01 (0.0-0.02)	0.58	0.01
		Multidien IEA phase	96 60%	0.67 (0.63-0.72)	0.08 (0.03-0.14)	0.75	0.22	98 62%	0.68 (0.65-0.76)	0.05 (0.03-0.07)	0.65	0.04
		Multidien IEA phase with recalibration	-	-	-	-	-	98 62%	0.68 (0.65-0.76)	0.04 (0.03-0.07)	0.75	0.20
	Multiple temporal features over 50 days		-	-	-	-	-	126 79%	0.70 (0.65-0.74)	0.07 (0.05-0.1)	0.77	0.22
RNN			-	-	-	-	-	128 81%	0.69 (0.64-0.74)	0.08 (0.06-0.11)	0.77	0.20

Table 1. Performance of the individualized and pre-trained models. Number and proportion of significant subjects (% Sig.) with forecasts above chance (i.e. tested against 200 surrogates) out of 159 included subjects. Median (IQR) area under the sensitivity versus time-in-warning curve (AUC) and Brier skill score (BSS) for significant subjects. Scores *per* subject are calculated for each significant subject and then averaged. Scores *across* subjects are calculated by aggregating the forecasts and the observations across all significant subjects. See also Table S1 for similar results obtained with a different surrogate design. See Table S2 for corresponding calibration loss and resolution.

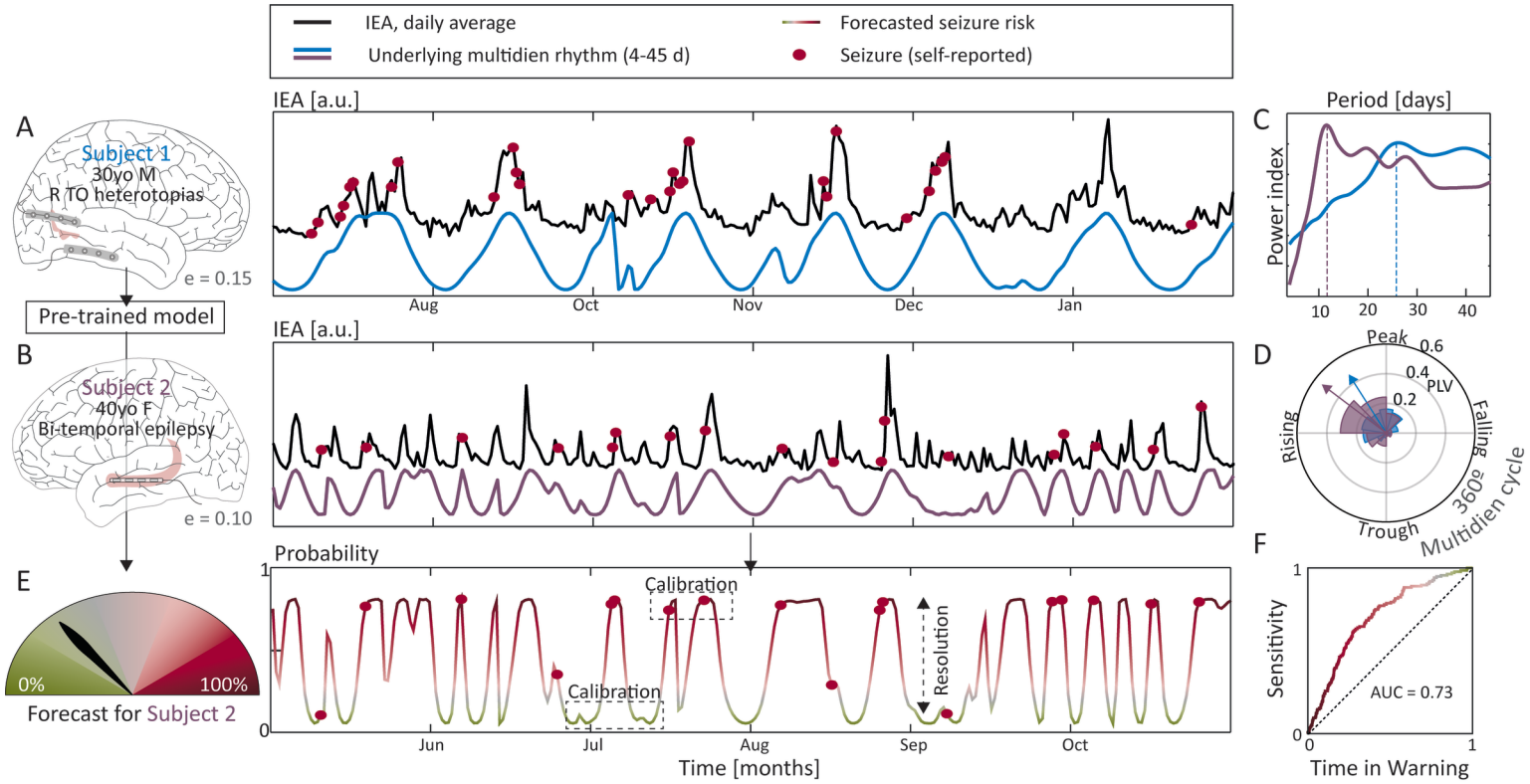
Figure 1. Illustrative example of pre-trained models. (**A**) Example of a six-month recording from a 30 year-old male with seizures stemming in clusters from right temporo-occipital heterotopias at a rate of once per week on long-term average (expected rate, e = 15% seizure days). From these recordings, three temporal features used as inputs for training the model included: (1) daily z-scored IEA over months recorded with one latero-temporal and one occipital electrode (black trace) along with (2) the cosine of the broadband (4-45 d) wavelet-derived phase (blue trace), and (3) self-reported seizures (red dots). (**B**) Same as (A) but from a 40 year-old female with bi-temporal epilepsy and seizures stemming from both hippocampi at a somewhat regular weekly to bi-weekly rate (e = 10% seizure days). IEA recorded from bi-hippocampal electrodes and self-reported seizures (red dots) were used for testing the pre-trained model. (**C**) Corresponding periodogram derived from the IEA in the training (A, blue) and test datasets (B, magenta). Note that the subjects used for training and testing the model have distinct peak-periodicities at 26 and 11 days, respectively. (**D**) Corresponding circular distribution and resultant vector showing that, in both subjects, seizures preferentially occur at similar rising phases of the underlying IEA cycle, regardless of its period-length. (**E**) Seizure probabilities forecasted for subject 2 (B), by the model pre-trained in subject 1 (A) showing that most observed seizures (red dots) fall in times of increased probability. (**F**) Receiver operating characteristic curve of the forecast in (E) with AUC = 0.73. Complementary probabilistic assessment of the forecast is captured under the notions of *resolution*, i.e. *how much* the forecast separates observed probabilities from the long-term expected seizure rate (dotted double arrow), and *calibration*, i.e. *how well* the forecasted probability agrees with the observed probability (dotted rectangles for low and high probabilities).

Figure 2: Forecasting schemes and performance. (A) Subjects included had more than 6 months of continuous data and fewer than 50% seizure days. (B) Each subject's seizures were forecasted according to radically different schemes: 1) using individualized models on chronological training and testing datasets within each subject; 2) transferring models trained on some subjects to test them on other unseen subjects in a five-fold (x5) cross-validation design. For better comparability between 1 and 2, we used the same testing data from each patient (40% of later recordings, black rectangles). (C) Output forecast with daily probability [0-1] for two individuals with low (purple) and high (orange) long-term expected seizure rate, respectively. The assessment of these forecasts can be done for each individual and then averaged, or aggregated by pooling all forecasts and observations together (vertical double arrows on the right). (D) Three characteristic metrics of forecast performance were calculated: First, using within-subject surrogate testing, we calculated the percentage of subjects with forecasts showing improvement-over-chance. Second, for significant subjects with above-chance forecasts, we evaluated the probabilistic performance of individual forecasts by the Brier Skill Score (BSS; forecast calibration and resolution. Third, by thresholding output probabilities, we evaluated the deterministic performance of individual forecasts by the area under the sensitivity versus time-in-warning curve (AUC).

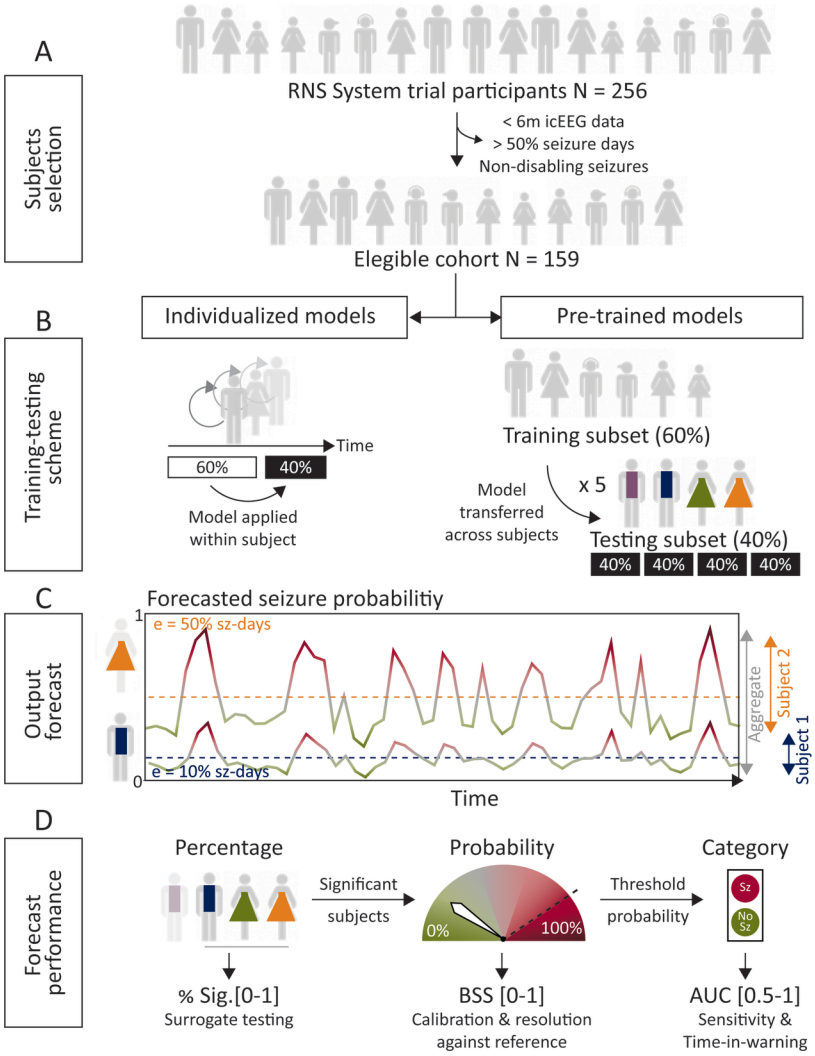
Figure 3. Performance of the individualized and pre-trained GLMs. (**A**) Violin plots comparing the distributions of per subject AUCs obtained with the individualized or the pre-trained models that incorporate the multidien phase as single input temporal features into GLMs. Filled dots denote significant subjects with forecasts above chance; empty dots denote subjects for whom forecasts were not significantly different from chance (not included in the violin plots). White horizontal lines are the distribution's median. (**B**) Corresponding reliability diagram showing observed (*a posteriori*) seizure probabilities versus forecasted (*a priori*) probabilities for individualized (grey) and pre-trained (red) models aggregating all significant forecasts across subjects. Note how post-hoc recalibration of forecasts issued by pre-trained models markedly improves resolution across subjects (full red to dashed red line). Shaded histograms show for each model the proportion of datapoints in each probability bin. Error bars are the 95% confidence interval based on the binomial fit at the median observed probability. (**C**) Coefficients of the GLMs trained with instantaneous multidien phases as input features (see Eq. 1, β_0 intercept, β^{Sin} and β^{cos} coefficients on the sine and the cosine of the phase, respectively). Half-violins (left) represents the distribution of coefficients obtained with individualized models, whereas white dots represent the different realizations of the pre-trained models across the cross-validation folds. (**D**) Tight relationship between the intercepts of the individualized models and the observed seizure rate of each subject. (**E**) Lack of relationship between the AUC and the observed seizure rate for the individualized (grey) and pre-trained (red) models. Filled dots represent significant subjects. (**F**) Tight relationship between the AUC and the preferential multidien phase for seizure occurrence (as shown in Fig. 1D). Seizures from subjects with seizure clustering between the rising and peak-phase can be forecasted with individualized or generalized models (filled dots: significant subjects). Arrowheads point to the few outlier subjects, whose seizures can only be forecasted with individualized and not generalized models.

Figure 4. Performance of multifeature models. (**A**) Distribution of per-subject AUCs using pre-trained multifeature GLM (red) or RNN (blue). Filled dots denote significant subjects with forecasts above chance; empty dots denote subjects for whom forecasts were not significantly different from chance (not included in the violin plots). (**B**) Corresponding reliability diagram as in Fig 3B. Forecasts issued by the RNN are closer to the perfect calibration line (dotted).

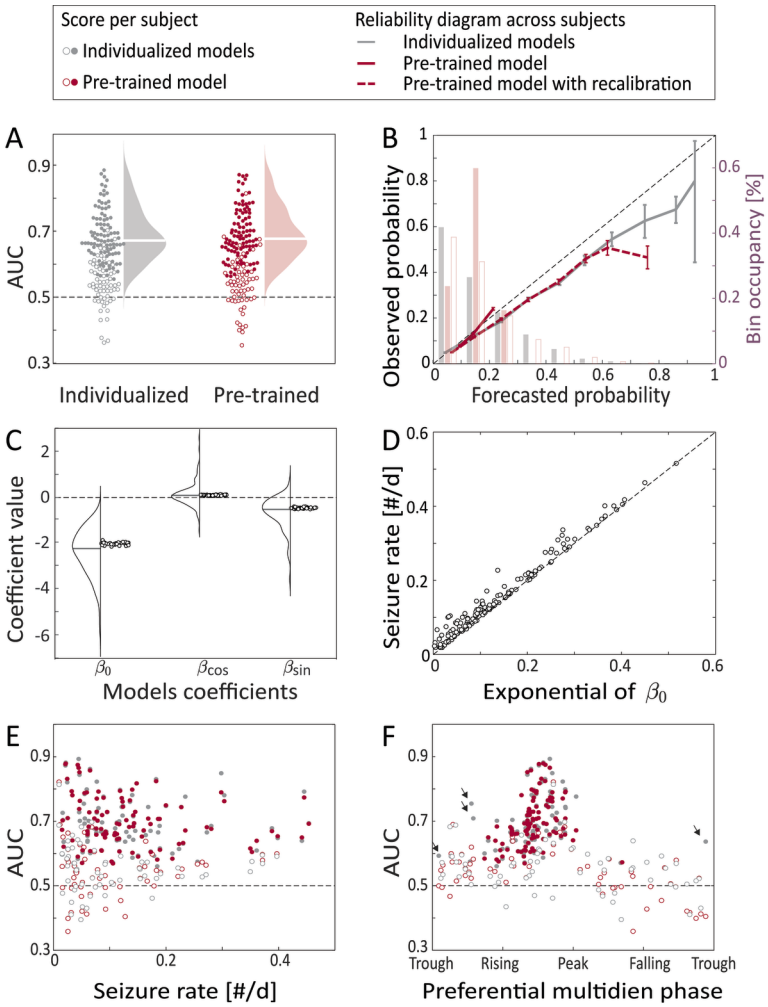
Figure 5. Forecasting seizures recorded with sub-scalp EEG, using pre-trained models. (**A**) Example of three parallel time-series derived from sub-scalp EEG data in one subject, consisting of detected IEA, EEG variance, and EEG autocorrelation (from top to bottom), along with cosine of the underlying multidien phase and seizure instances (red points) as in Fig. 1A. Grayed-out trace indicate periods where data was interpolated as described in[17]. (**B**) Corresponding periodogram derived from the detected IEA (black), EEG variance (blue), and EEG autocorrelation (purple). (**C**) Corresponding mean resultant vector indicating moderate-to-strong seizure clustering within cycles of IEA and variance, but not autocorrelation. (**D**) Transfer learning scheme where the model was learned from the cohort with icEEG and applied to this subject with sqEEG. (**E**) Receiver-operator curves for forecasts generated by models pre-trained on the RNS cohort applied to phases extracted from cycles of IEA, variance, or autocorrelation in sqEEG.



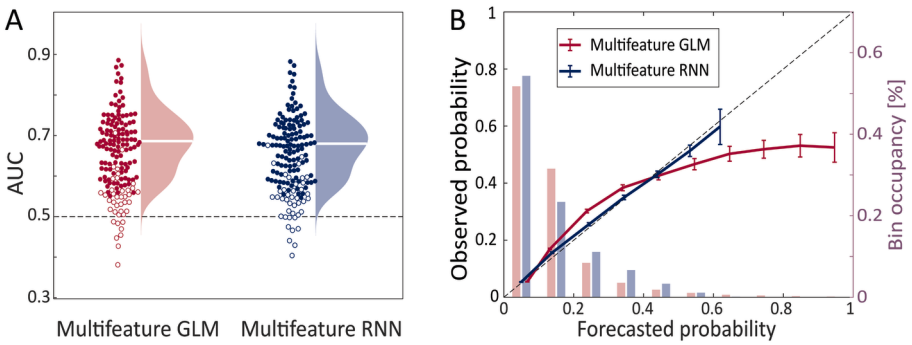
EPI_17406_Leguia-fig1.tif



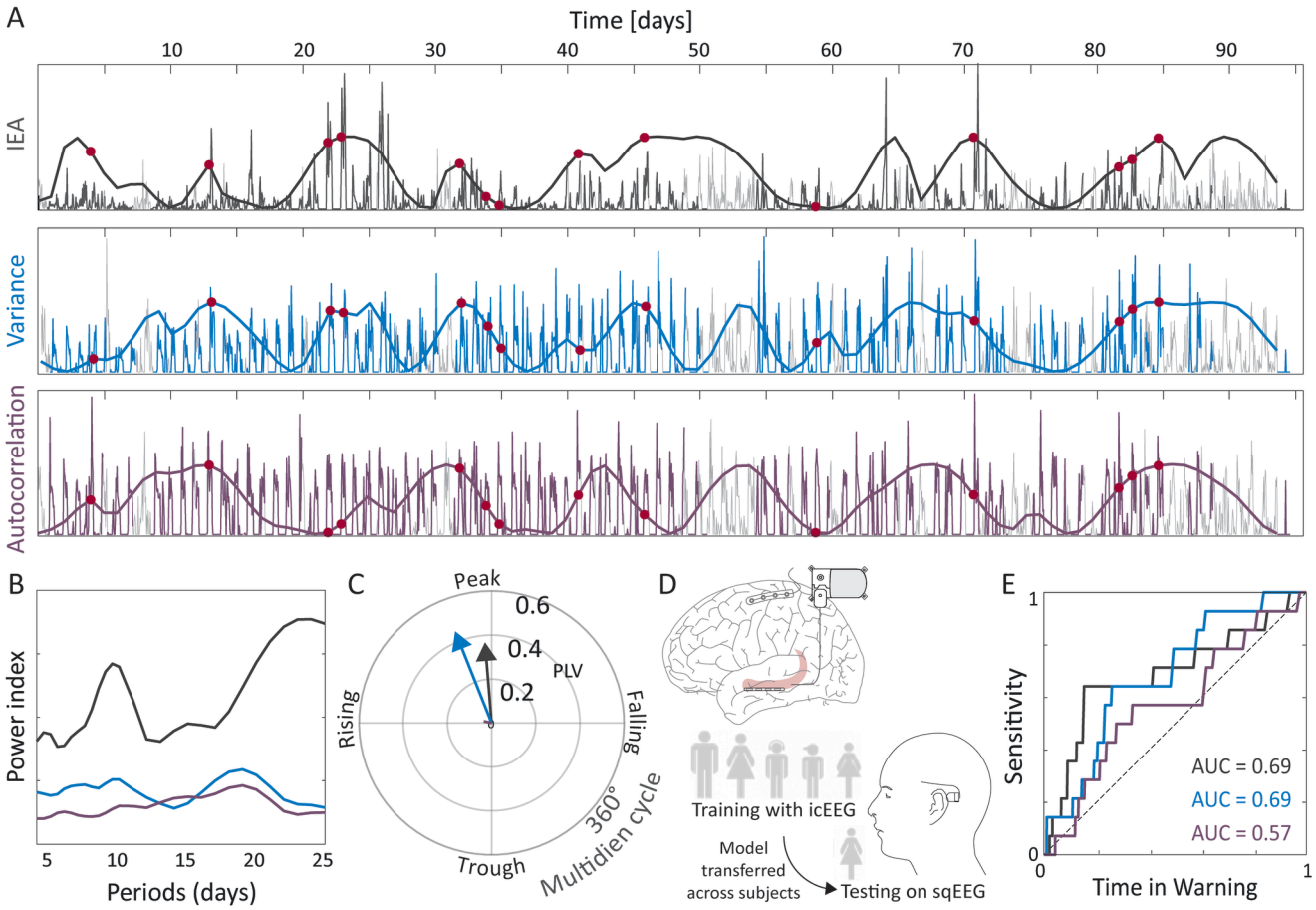
EPI_17406_Leguia-fig2.tif



EPI_17406_Leguia-fig3.tif



EPI_17406_Leguia-fig4.tif



EPI_17406_Leguia-fig5.tif