Seizure forecasting: bifurcations in the long and winding road

Maxime O. Baud[1,2], Timothée Proix[3], Nicholas M. Gregg[4], Benjamin H. Brinkmann[4], Ewan S. Nurse[5], Mark Cook[5], Philippa Karoly[5].

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

[2] Wyss Center for bio- and neuro-engineering, Geneva, Switzerland

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

[4] Bioelectronics Neurophysiology and Engineering Laboratory, Department of Neurology, Mayo Clinic, Rochester, MN, USA.

[5] Graeme Clark Institute, The University of Melbourne, Melbourne, VIC, Australia

Correspondence: maxime.baud.neuro@gmail.com

Acknowledgments : We thank Dr. Florian Mormann for authorizing the use of the “long and winding road” metaphor for this review, thus encouraging the continuation of a decades-long tradition.

Conflict of 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. MOB reports personal fees and grants from Wyss Center for Bio- and Neuro-engineering outside the submitted work. MOB has a pending patent under the Patent Cooperation Treaty (#62665486). MOB is a shareholder of Epios Ltd., a medical device company based in Geneva, Switzerland. ESN, PJK, and MJC report personal fees and financial interest in Seer Medical Pty. Ltd., a medical technology company based in Melbourne, Australia. NMG is an investigator for the Medtronic Deep Brain Stimulation Therapy for Epilepsy Post-Approval Study. BHB reports licensed IP to Cadence Neuroscience Inc, and a consulting agreement with Otsuka Pharmaceuticals. BHB has research support from the National Institutes of Health and Seer Medical Pty. Ltd. ESN and BHB are supported by the Epilepsy Foundation of America’s ‘My Seizure Gauge’ grant.

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.17311

Contribution: MOB and PK conceptualized the review. MOB drafted the manuscript and all authors contributed to its final version.

Abstract

To date the unpredictability of seizures remains a source of suffering for people with epilepsy, motivating decades of research into methods to forecast seizures. Originally, only few scientists and neurologists ventured into this niche endeavor, which, given the difficulty of the task, soon turned into a *long and winding road*. Over the past decade however, our narrow field has seen a major acceleration with trials of chronic EEG devices and the subsequent discovery of cyclical patterns in the occurrence of seizures. Now, a burgeoning science of seizure timing is emerging, which in turns informs best forecasting strategies for upcoming clinical trials. Although the finish line might be in view, many challenges remain to make seizure forecasting a reality. This review covers the most recent scientific, technical and medical developments, discusses methodology in detail and sets a number of goals for future studies.

Keywords:

- Seizure forecasting
- Seizure cycles
- Wearable devices
- Chronic EEG
- Circadian
- Multidien

Key points.

- The field of seizure forecasting has made four major advances in the last decade that together will benefit upcoming clinical trials.
- 1) the recent unraveling of mutlidien cycles of epileptic brain activity enables forecasting schemes at days-long horizons
- 2) the generalizability of the ictal-interictal relationship allows for transferring pre-trained forecasters to unseen participants
- 3) the probabilistic estimation of seizure risk, benefits from identified cyclical variables, increasing amounts of data and refined methods
- 4) the recent inclusion of minimally and non-invasive technology help circumvent the risks linked to intracranial forecasting systems

I) Introduction

To date, the unpredictable nature of seizures has dramatic consequences for people with epilepsy as seizures can occur in any situation possibly leading to injuries or even death. Consequently, deploying strategies to provide people with epilepsy and their families with any degree of reliable information about upcoming seizures, would undoubtedly be met with great enthusiasm. Different clinical gains are expected for different forecasting strategies. Issuing accurate last-minute alerts about imminent seizures could enable the adoption of rapid safety measures[1]. Forecasting the probability of one or more seizures over hours or even days[2], akin to weather forecasting, would allow patients and families to plan their lives around periods of high-risk and develop risk-mitigation strategies[3–5]. Over the past four decades, seizure prediction has been a niche endeavour for a few epileptologists and scientists acquainted with non-linear systems and equipped with the necessary mathematical background. Today, with the rapid development of wearable or implantable mobile devices, together with access to suitable computational power, the idea is gaining ground in the clinical community.

Anticipating upcoming clinical trials, the purpose of this review is to contextualize the current state of clinical, scientific and technical knowledge and to clarify what goals these trials should strive to achieve. Major reviews have already covered work done before 2017[3,6–9], which will only be briefly discussed here. Recent work has clearly demonstrated the superiority of studying continuous data over months, as opposed to data collected over days in the hospital, hence studies based on shorter-term data will not be covered.

In the past five years, four bifurcations have been reached on the long and winding road towards forecasting seizures. Different groups have focused on different forecasting strategies and this has diversified the work for better advances in the field. To highlight novelty and clarify concepts, we purposefully contrast alternative approaches when reporting themes discussed between October 29 and 30 2021 in Copenhagen, at the Congress on Mobile Devices and Seizure Detection in Epilepsy. An appendix covers the more technical aspects of this progress.

First, the horizon for forecasting seizures has lengthened from minutes to days since the inclusion of cyclical variables[2,10,11]. Second, methodology for probabilistic forecasting was borrowed from the field of meteorology to challenge the initial hope for fully deterministic predictions[2,12]. Third, non- and less-invasive technologies were developed[11,13–15] and challenge the exclusive use of intracranial devices for the sole purpose of forecasting seizures. Fourth, the notion that forecasting algorithms must be individualised is questioned by the generalizability of forecasting strategies based on cyclical variables[16]. These different ways have not split far apart though. We conclude this review by explaining why they must reunite into one lane to pass the ultimate test: real-life, prospective clinical trials.

II) The road traveled in retrospect

Starting point: Work in the 1980s-1990s

With theoretical advances in mathematics in the 1980s, including chaos and non-linear systems theory, predicting seizures seemed within reach[17–19]. Researchers applied these tools to scalp or intracranial EEG from inpatient epilepsy workups, aiming at understanding the chaotic dynamics of the brain switching from one state (normal) to another (seizure)[17].

First milestone: International Workshop on Seizure Prediction (IWSP)

The enthusiasm was such that a first international hands-on workshop took place in 2002 in Bonn, Germany to try and predict 51 seizures in a dataset taken from 5 patients[19]. However, none of the participants was able to predict seizures above chance. This set the goal of the second and third workshops that took place in Bethesda, USA, and Freiburg, Germany to develop consensual methods for comparing system performance across labs[20]. This is when the community realized it was walking on a long and winding road[7].

A group of epileptologists, neurosurgeons, neuroscientists, computer scientists, engineers, physicists, and mathematicians continued to regularly convene at the workshop every other year, alternatively in Europe[20–22], the USA[23,24] or Australia[25]. But after the two first workshops trying to crack the problem as another hackathon, attention shifted to the need for a deeper understanding of the mechanisms of ictal transitions, and for identifying states promoting or impairing their occurrence. A major concern about earlier work was that seizure forecasting may be hampered by the non-stationarity in recordings[26,27] from patients brought into the hospital for diagnostic work-ups that necessitate the cessation of anti-epileptic drugs[28] and often sleep deprivation[29]. Thus, by the 2010s, the benefits in attempting seizure prediction in ambulatory patients, on stable medication, in their natural environment was evident. The field had turned to multivariate measures and machine-learning as the default approach to personalizing predictive algorithms on longitudinal datasets, a number of standards were established for the statistical testing of algorithm performance[7,30], but sufficiently large amounts of data were still lacking to prove the point. As a consequence, a keen interest in developing a device for ambulatory EEG monitoring arose. A forced detour: commercializing devices for monitoring EEG

To this day, five devices capable of chronically recording ambulatory EEG in epilepsy exist, or have existed, from which two are at the commercial stage. All come with advantages and drawbacks :

1) Ten years ago, the Neurovista device was developed as a subclavicular recorder connected to subdural leads with a total of 16 electrode contacts, capable of continuously recording and storing intracranial EEG. As the first and, to-date, only prospective trial of seizure prediction, the landmark Neurovista trial[1] between 2010-2012 incarnated previously-developed concepts into a seizure advisory system which successfully collected up to two years of continuous intracranial EEG data in 15 participants. Although above-chance warnings could be issued in 9 of these participants, a clinical benefit could not be formally established. Nevertheless, the

results, first presented at the 6th IWSP in San Diego in November 2013, finally raised hope that seizure forecasting is achievable.

Once the study concluded, the device was never commercialised due to a lack of investment beyond 2013, and data collection stopped in humans (but continued in dogs). The recorded data has proven invaluable and has fueled the decade of work[12,31–39] that followed the first publication[1].

2) The commercially-available RNS® System (NeuroPace, Inc, USA), received FDA approval in 2013 as a therapeutic intracranial cortical neurostimulator with 4, now 8 electrodes and is implanted to date in more than 3000 American people with epilepsy. Unlike Neurovista, the RNS System monitors, but does not store raw continuous intracranial EEG[40], which sets boundaries to EEG analysis. Nevertheless, the device comes with customizable algorithms based on line-length, area-under-the-curve and band-pass filtering that can be tuned by the clinicians to capture patient-specific epileptiform patterns and count them on an hourly basis (above threshold detections). As the company actively engages in collaborations with academia, longitudinal data collected in more than 200 participants[41–43] over the 12 years (2006-2018) of the clinical trials has yielded invaluable insights into focal epilepsy and its treatment by cortical stimulation[2,10,44–49] while data continues to accumulate.

3) The research RC+S device from Medtronic, with a trial limited to 5 human and 15 canine patients that is now closed, provided raw EEG data (4 channels) selected from 16 electrode contacts and has started to yield results on the recorded data and the effect of stimulation[50,51]. Its sister device, the commercially-available Percept™ PC deep brain stimulatior (DBS) has very limited recording capability in the ambulatory setting, storing one bandpass power epoch every 10 minutes (increased sampling rate possible only with firmware changes)[52] and is not further discussed here.

4) The commercially-available UNEEG™ device (UNEEG medical A/S, Denmark) that received CE-labeling in 2019 for a two-channel unihemispheric sub-scalp (i.e. electrodes and device are between scalp and skull) EEG system so-far implanted in 9 subjects with epilepsy

represents a less-invasive solution and offers recordings over months that are similar to scalp EEG, albeit with a limited number of electrodes.[53–56]

5) The research Epiminder device (Minder, Australia), a four electrode, two-channel bi-hemispheric sub-scalp EEG system, now in trial in 10 patients (aiming at 16) has yielded ongoing continuous recordings over months, up to one year[15].

Collectively, this 'long data' collected in a clinical ambulatory context in different formats in different groups of patients, amounts to years of data, and has enabled the recognition of previously uncharacterized interictal and ictal patterns among which the (re)[57]-discovery of cycles in epilepsy is central (see section III)[5,10,32–34,44,45].

Onboarding friends

Now with long data available, the development of learning algorithms became the new bottleneck. Given the complexity of the task, the community has turned to experts in Artificial Intelligence for help. Using rigorous designs, internet-based machine-learning competitions (www.kaggle.com) have pushed the performance of individual algorithms, tested on held-out data not available to the participants[38,39]. In these crowd-sourcing efforts, hundreds of researchers (654 and 646) attempted to solve the problem with a high number of trials (17,856 and > 10,000 algorithms entries) on two-year long datasets from two and three humans with epilepsy (some additional data from dogs) yielding top AUCs of 0.81 and 0.84, respectively[38,39]. Of course, crowd-sourcing efforts to develop personalized algorithms for a few patients cannot be implemented in real-life but the approach is nevertheless informative in regard to the amount of engineering effort needed for algorithm optimization to gain performance on datasets that were reputedly difficult. Yet, it is difficult to synthesise the many EEG input features on which these algorithms typically rely into one coherent explanation[4] and the AUC should not be the only assessor of performance (see below).

Lessons learned so far

From decades of work, a number of fundamental and practical points were clear:

- *The need for high specificity and sensitivity.* In a seizure alarm system, false-alarms (low specificity) can be stressful at first, but then decrease confidence in the system, lead to alarm habituation and ultimately defy the intended purpose. Related to this issue, patients with very frequent seizures (e.g. daily) likely will benefit less from seizure forecasting. On the other hand, a system capable of detecting low risk, is likely as valuable as an alarm, and may contribute to decreasing stress. However, this is only true granted the system has high sensitivity, as false negatives also decrease confidence in the system. To achieve reasonable sensitivity-specificity trade-offs, forecasting algorithms need to be patient-specific and require a large amount of training data. This commands the acquisition of long data within the same patient, and personalized algorithms although these imperatives are challenged in this review.
- T*The need for continuous raw data and better batteries.* In a field where much of the science remains to be done, raw data is of the essence. Outside of the temporal constraints of a prospective trial, retrospective studies on existing raw data likely yield more knowledge than immediately launching the next clinical trial. However, obtaining raw data immediately relates to the battery problem. Technically, devices that provide raw data must be rechargeable because streaming rapidly depletes batteries. Since high-performance rechargeable implantable batteries is an unresolved biomedical engineering problem, many have adopted battery-less implants, combined with rechargeable small external batteries[56].
- *Invasiveness is undesirable.* The invasiveness of neurotechnologies that do not directly provide treatment is difficult to accept from the patient side, calling for minimally- or non-invasive solutions.

These learned lessons were some of the motivation for a number of groups in academia and/or industry to bifurcate in slightly different ways. Crucially, scrupulous retrospective studies have

enabled the burgeoning of a small scientific revolution for our field and advanced technology for seizure forecasting by leaps and bounds.

III) An emerging science of seizure timing

To better reflect the need for a deeper understanding of seizures and abandon the sole focus on prediction the IWSP changed names in 2017 to become ICTALS: the International Conference on Technology and AnaLysis of Seizures. With this new impetus, the conference saw major knowledge advances over the past years, which together constitute the emergence of a true science of seizure timing. This knowledge has recently been covered in a major review[58] and will be here only succinctly summarized. Accounting for their prevalence and effect-size on relative seizure risk, a number of endogenous cyclical risk factors (some of which are paced by the environment) have now been recognized as critical for seizure timing, whereas a number of external and/or sporadic influences have seen their importance relativized (Table1).

Seizure cycles

Evidence indicates that the momentary likelihood of a seizure is co-modulated by cycles operating at various time scales: from the shorter ultradian (<<24 hours)[10,47] and circadian (~24 hours)[10,32,35,44,45,47] cycles to longer multidien (~1 week to ~1 month)[10,33,45] and circannual (~1 year)[45] influences. In any given individual, a combination of cyclical modulations at specific timescales and phases may give rise to a unique temporal pattern of seizure occurrence[9]. Shared characteristics at the level of groups of people with epilepsy led to the notion of seizure chronotypes[45], originally described in a handful of landmark reports at the turn of the 20th century[57,59–61]. Highlighting the importance of these clinical phenomena, modern chronic EEG data shows a ~90%, ~60%, and ~10% prevalence of circadian[35,45], multidien[45], and circannual[45] seizure cycles, respectively. Although the mechanisms by which seizures occur

cyclically at these different timescales are currently unknown, there has been long standing interest in the question and chronic EEG studies have recently advanced our understanding.

Sleep and seizures

Although the modulation of seizures by the sleep–wake cycle and the circadian cycle are often conflated in the epilepsy literature, they should be treated as distinct but intertwined modulators of epileptic brain activity[62,63]. Historical reports have documented prevalence of 20–30% for sleep-related seizures[57,59–61]. Although long-suspected, an additional role for sleep homeostasis (the need for regenerative sleep that grows with the length of wake) in triggering seizures has not been fully established yet[64]. One recent retrospective study on NeuroVista data found that 10/12 patients had a slight decrease in seizure risk over 48h when they slept more than 11.2 hours, but no heightened seizure risk when the usual sleep time was curtailed[65]. However, the sleep-protective effect was not confirmed in a pseudo-prospective study on the same data[36]. The issue remains unresolved as these and other studies were not designed as experiments to assess the effect of shorter sleep (i.e. experimental sleep deprivation)[64].

Catamenial epilepsy

Catamenial epilepsy has been discussed for decades, and progesterone and estrogens have been hypothesized to have variable degrees of anti-ictal and pro-ictal effects. However, these mechanisms cannot account for observed multidien rhythms in men and children, highlighting that catamenial epilepsy is merely a special case of multidien rhythmicity in epilepsy. Additionally, the effect of menses as a temporal cue reported in the literature is lower (~RR of 1.2-1.5)[66,67] than that of multidien cycles found in women with chronic EEG.

External modulation

Environmental factors such as weather changes clearly have weaker predictive value. In a pseudo-prospective study, four of eight included subjects had seizure risk weakly influenced by temperature (3 subjects) and humidity (2 subjects, see Table 1)[36]. Another larger study reporting only group statistics showed that seizure risk increased slightly with low atmospheric pressure and high humidity (Table 1)[68].

Endogenous cycles of ictal-interictal activity

Highlighting the power of monitoring epileptic discharges, the long-debated interictal-ictal relationship can be understood in more general terms of long dynamics in epilepsy by examining the relationship between seizure timing and fluctuations in IEA at long time-scales: cycle after cycle, seizures tend to occur when IEA raises over days[10]. This phasic relationship (technically phase-clustering of seizures within rising phases of IEA) holds true for cycles spanning about a week to a few months and is consistent across studies[10,33,45], individuals[10,33,45], and species[51,69], representing a general basis to forecast seizures over longer horizons. In a study of > 200 individuals, very few contradicted this rule[45]. Importantly this phase relationship did not depend on the period-length of the underlying cycle, pointing to a dynamical process that can accommodate several timescales.

Additionally, ictal and interictal activity also fluctuate rhythmically at shorter, circadian timescale, but with less consistent phase relationship across patients[10,32,33,45]. Further, circadian and multidien cycles of interictal and ictal activity can be detected with intracranial[10,33] or sub-scalp[16,70] EEG, and correlate with heart-rate variability in some patients[71] opening the way for less invasive methods. Related to these intricate cycles, some clinical evidence[33,72,73] supports the theoretical prediction[74,75] that critical slowing may signal approaching ictal transitions (a bifurcation) at timescales longer than the seizures themselves.

Relative importance of time varying risk factors

In medicine, effect sizes are often compared as a ratio of risk or odds for belonging to one category or another. In the circular domain, effect sizes are best evaluated on the continuum of phases, as the phase-locking value. Dichotomizing (thresholding) cycles into one critical phase versus low-risk phases can approximate circular data in terms of relative risk. Across different studies using chronic EEG, relative risks linked to nearly-ubiquitous endogenous cycles combined into 3-9 fold increases[2,10], far above relative risk found for categorical variables in other studies, which rarely exceeded 1.5 fold increases (Table 1 for comparison). Contrasting the relative effect-size of established time-varying risk factors informs best strategies to forecast seizures. Building on this, pseudo-prospective studies have shown that a promising strategy to improve the performance of seizure forecasting at longer horizons is to account for slow variables at circadian[12] and multidien timescales[2,10,33].

IV) First bifurcation: forecasting horizon of days versus minutes

One of the most striking advances in the past five years is the realisation that it is possible to forecast seizures over days[2], whereas previous attempts had focused on the minutes preceding seizures[1,12]. This advance, which has not yet been tested prospectively, was only possible with the discovery of multidien rhythms of epileptic brain activity[10,32,33]. In pseudo-prospective studies, forecasting algorithms were able to output above-chance forecasts at a 24-hour horizon on the unseen test-data in about two-thirds of subjects[2]. The inclusion of circadian and multidien rhythms was also demonstrated to be the best predictor in datasets that had been previously exhaustively investigated with machine-learning[32,33], illustrating how the most advanced algorithms cannot compensate for important features that are missing in the input data. Opening the way to less invasive studies, multidien cycles of IEA can be detected with sub-scalp EEG[15,16]. In addition, characterising and modelling seizure cycles using patient reported seizure calendars may be sufficient to forecast to some degree upcoming seizure risk over the next calendar day[11,76]. Multidien rhythms may not be straightforward to detect with wearable technology, although some data is promising in that

sense too[71]. Including a circadian influence can be approximated by simply characterizing the preferential hour of the day for a patient's seizure occurrence (as opposed to tracking a circadian biomarker) and this very simple approach should be included in any probabilistic seizure forecasting scheme. Indeed, the past is the best predictor of the future, and repeating patterns can be found in almost all patients with epilepsy. While extremely promising for future risk-modifying strategies, forecasting cycles of seizures comes with its own technical challenges that will need to be addressed to enable the design of clinical trials (see section IX).

V) Second bifurcation: probabilistic versus deterministic forecasting strategies

A deterministic forecast seeks to provide a categorical answer to the question of whether an event will occur or not ('yes' or 'no' categories). In contrast, probabilistic forecasting strives to reproduce the probability of events. When accurate, a deterministic forecasting approach is adequate, because it provides spot forecasts for best-informed decisions. While this is most often conceivable at very short horizons, perfectly accurate deterministic forecasts do not exist for seizures, nor for weather. Deterministic forecasts are indeed as good as a combination of the accuracy of the model generating them, the accuracy of the collected data, the interpretability of the output, and the time horizon to take action before the realization of the forecast. In the 1960s, the acknowledgement that the accuracy of deterministic forecasts heavily depended on minute parameter changes or measurement inaccuracies (initial conditions) led the field of meteorology to promote a probabilistic approach[77]. Many other forecasting problems greatly benefit from a probabilistic approach, because they are too complex to be modeled accurately, or because the very nature of the events' timing is stochastic. For seizure forecasting, a combination of both approaches will be ultimately needed in clinical practice (probability and threshold), as a given seizure risk must be

translated into a practical decision (e.g. taking a medication above a certain risk), but we here purposefully emphasize the nuances between the two approaches for the sake of clarity.

Forecasting probabilities versus categories

The choice between a deterministic and a probabilistic approach has repercussions on (I) the goal to attain, (II) the choice of forecasting algorithm, (III) the information provided to users, (IV) the methods to evaluate performance, and (V) the amount of data needed to do so:

(I) Two different goals may be sought: (a) always forecast a category but accept that this may sometimes (or often) be wrong, or (b) forecast a probability of belonging to a category (continuum between 0% and 100%), striving to produce a reliable quantification of uncertainty.

(II) The types of supervised algorithms that best match these distinct goals are different. Deterministic algorithms act as classifiers (e.g. logistic regression) that dichotomize values in binary categories by learning how to assign a label to each data sample. Probabilistic algorithms act as regressions (e.g. linear regression and its generalisation) that predict continuous values, and seek to optimise the conditional probability of observing a label given a data sample. Although there are methods to convert outputs of deterministic algorithms into probabilistic ones, opting for probabilistic algorithms allows for using methods adapted to the probabilistic nature of the problem (e.g., likelihood optimization).

(III) Deterministic outputs to users require threshold optimization to achieve a given goal (e.g. high specificity, low sensitivity), which is not required if users directly access forecasted probabilities to make informed decisions based on a certain degree of uncertainty.

(IV) While benefiting from more observations, deterministic scores can already be evaluated on small sample sizes (e.g. 20)[78]. In contrast, evaluating a probabilistic forecast requires orders of magnitude larger sample size, because it is based on

comparing probability distributions that include some values which are rarely output by the algorithm for low-probability events. The critical lack of longitudinal data originally undermined burgeoning interest for a probabilistic approach in epilepsy[79], but this strategy has now been established with longer datasets[2,12].

(V) "How often are the forecasts correct?" Correctness is appealing to characterize deterministic forecasts but is generally considered inappropriate to evaluate probabilistic forecasts.[78,80] Probabilistic scoring metrics reward a forecaster for reporting risk when an event could have occurred, even in the absence of an observed event (absence of realization of the risk). A deterministic forecast on the other hand strives for *discrimination* between event and no-event datapoints; Deterministic scoring metrics punish reporting risk when no event took place. Indeed, higher deterministic scores reflect that lower and higher forecasted probabilities are associated with the non-events and events, respectively, but not how well this forecast captures the underlying event probabilities. Importantly, they are blind to the calibration of a forecast as they only rely on the relative —rather than absolute— probabilities. For example, an algorithm forecasting 0% when no event occurs and 0.1% on each day when an event occurs would lead to an area-under-the-curve (AUC, see Table 2) of 1 (perfect performance), even if the event occurred every other day on average (i.e. an expected probability of 50%). Probabilistic metrics on the other hand can quantify the calibration and resolution of the forecast[78]. The average relationship between forecasted probabilities (a priori) and observed probabilities (a posteriori) can be visualized in the form of a reliability diagram which allows for the evaluation of resolution and calibration (Fig. 1 and mathematical formulas in the appendix) as well as the calculation of the Brier skill score. The reference forecast for comparison can be drawn from simply shuffling the original individual forecasts[2,12], or by issuing a trivial forecast, such as the long-term expected event probability[81]. The best strategy has not been determined with certainty in the field of epilepsy.

Rationale for probabilistic seizure forecasting

One key advantage of a probabilistic framework is that it includes the possibility to issue fully committed predictions (i.e. 0% and 100%) at both ends of a continuum of intermediate degrees of confidence. Conversely and by design, a deterministic approach severs the connection to model outputs by irreversibly thresholding values into two mutually exclusive categories. The price to pay for a deterministic approach is to be wrong on some (or many) occasions. The price to pay for a probabilistic approach is to never (or rarely) be certain.

Opting for a probabilistic approach in epilepsy, like was the case in weather forecasting in the 1960s can be motivated by several factors: (1) the lack of infallible seizure precursors in the *pre-ictal* period (minutes preceding seizures)[9,58], (2) the existence of cycles of epileptic brain activity that determine *pro-ictal* states at certain phases, which constrain the timing of seizures in a probabilistic manner over different durations[9,58], (3) the possibility to use explicit Bayesian[12] or generalized linear model frameworks[2] that can fit any probability distribution; and (4) the potential for probabilistic forecasts to be interpreted by people with epilepsy in terms of quantified uncertainty about upcoming seizures with forecasted probabilities rendered as a continuum of increasing risk, a “seizure gauge”[83–85] (Fig. 1a).

In summary, from technical nuances that seem subtle at first, it is clear that the choice of the methodological strategy has a key impact on the scientific and clinical aspects of seizure forecasting. As opposed to black-box machine-learning approaches, an explicit combination of time-varying risk factors in probabilistic terms favors the scientific understanding of the relative importance of predictors within and across patients. Additionally, by offering a graded assessment, probabilistic forecasting circumvents the core issues of specificity and sensitivity as well as false positives and negatives, which may raise stress for patients and create medico-legal issues, respectively. In our opinion, to move the field forward safely, the goal of immediate deterministic forecasts must be extended to include progressively improved probabilistic forecasts by increasing their sharpness (higher prevalence of extreme forecasted

values, e.g.. close to 1% and 99%) and resolution (actual degree of certainty on the observed outcome). For a full technical description on these issues, we refer the reader to the appendix of this review.

VI) Third bifurcation : invasive versus non-invasive seizure forecasting

It is not surprising that many people living with epilepsy would much prefer to have seizure forecasts derived from noninvasive signals without requiring an invasive brain implant[79]. While seizure forecasting methods have now been established with invasive EEG devices, forecasting with noninvasive and minimally invasive signals has begun to emerge. It has now been demonstrated that signals from wearable devices, including heart rate, electrodermal activity, actigraphy, and temperature show circadian and multidien cyclical relationships with seizure risk, though often these relationships are less significant than those demonstrated with interictal discharges and will have to be confirmed with a larger cohort. These signals are measurable using wrist-worn research devices (Gregg et al. American Clinical Neurophysiology Society Annual Meeting, conference abstract, 2022) or commercially available fitness trackers[13,71] and represent form factors that many patients find acceptable and easy to use[86,87]. The ability to non-invasively track epileptic rhythms is a key development, given the focus on individual-specific multidien cycles to inform the next generation of seizure forecasting devices.

Recent research highlights the potential for mobile and wearable seizure forecasting[88]. Some studies have validated methods to track the likelihood of self-reported seizures using mobile diaries[11,76], and wearable biosensors[13]. In addition, wearable measurements can be used to detect the pre-ictal state of electrographic seizures. For instance, accelerometry, blood volume

pulse, electrodermal activity, and temperature were predictive of both focal and generalized electrographic seizures in 43% of people (n=69) with a short prediction horizon (on the order of minutes)[14], although findings were limited to inpatient recordings. Compared to simply using time-of-day, Nasseri et al. 2021[89] showed improvement of forecasting performance (AUC range of 0.72–0.92) of electrographic seizures using the same[14] wearable signals in a cohort of 6 participants with a minimum of 6 months ambulatory recording. Taken together, these studies show early promise that wearable signals may be useful to forecast the risk of seizures.

Minimally invasive subcutaneous EEG systems have also very recently demonstrated the ability to forecast seizures, using multiple devices. A proof-of-concept case was published using the EpiMinder device in two patients with epilepsy[15] employing a cyclical critical slowing approach, previously demonstrated using the NeuroVista dataset[33]. The UNEEG SubQ device has also demonstrated the ability to measure cycles of brain excitability over long periods (Viana et al., American Clinical Neurophysiology Society Annual Meeting, 2021), and generate long-term seizure forecasts[16]. Additionally, short term forecasts may also be possible using this device in patients from Denmark and the UK (Viana and Pal Attia, current issue).

In the end, it is likely a combination of physiological, environmental and behavioral signals measured from implantable, wearable and mobile systems will contribute to seizure forecasts, with less invasive approaches potentially useful to screen individuals who would benefit from higher precision forecasts using continuous brain recordings[90]. In the meantime, more work is needed to characterize the relationships between seizure risk, brain excitability, autonomic regulation, behavior, and other features of body homeostasis to realize the full potential of noninvasive multimodal seizure risk forecasting.

VII) Fourth bifurcation : personalized versus generalized forecasting

With the advent of artificial intelligence, machine- and deep-learning algorithms were applied to EEG to try and forecast seizures relying on large amounts of individual training data.. However, we previously saw that, at the multidien timescale, a majority of patients have an alternance of pro-ictal and low-risk states with smooth transitions over days[2]. Building on the generalizability of the multidien interictal-ictal phasic relationship (see above), a recent study proposed algorithms that are transferable across patients to forecast the risk of seizures over days[16]. Indeed, once the multidien phasic relationship is captured by a statistical model, it is sufficient to know the past seizure (emission) rate and the recent IEA trends of an individual to forecast absolute seizure probabilities over coming days. Simply put, as general principles of seizure timing are being discovered, forecasters can learn from cohorts of patients and forecast seizure risk for previously unseen patients. Along the same line, others have generated generalized forecasts at shorter time horizons using machine learning algorithms trained on data from different subjects (Pal Attia and Viana, current issue). Being able to learn and generalize across cohorts of patients has a number of practical consequences: 1) Cohort-based statistical models will undoubtedly be more robust as they are based on more data from different patients, thus avoiding overfitting 2) more advanced methods that require such large datasets can be used, including deep-learning, and 3) for individual patients who freshly engage into trials of seizure forecasting, less preliminary data will be needed before producing accurate seizure risk estimates. Cohort-based models also have drawbacks, as by design, they lack personalization and tend to capture average effects.

VIII) Merging into one lane

This review purposefully contrasted approaches for the sake of clarity of the concepts. Overall, it is worthy to note that most recent progress has been achieved by methods different from complex EEG analysis. Studying counts of seizures[11,35] and interictal discharges[10,32,33,45] has been instrumental in delineating a probabilistic approach to seizure forecasting relying on repeating patterns. The addition of sub-scalp EEG and peripheral monitoring may further enable less-invasive approaches or complementary measures. In clinical trials, where forecast performance should be optimised for each individual participant, the strengths of different approaches should be combined. Given recent advances, future holistic solutions will likely be:

- *multi-timescale*, using sleep-wake, circadian and multidien cyclical influences that coexist at different degrees in most patients to best inform momentary risk estimates.
- *multimodal*, combining a range of peripheral, central, and behavioural measurements to best capture coexisting dynamics.
- *probabilistic*, accounting for mixed time-varying risk factors, but also *deterministic* when a categorical decision must be made (e.g. take an add-on medication or not) or for closed-loop paradigms.
- *generalised*, learning robust forecasting models from cohorts of patients, but also *personalised*, optimising the final forecast output for each individual.

IX) The finish line ?

To date only the Neurovista trial has taken the ultimate test, setting high standards for future trials1. With the advances presented here, new trials with less invasive methodology and forecasts at extended horizons are within reach. However, a number of problems remain.

Unsolved problems

Additional chronorisk factors. Additional time-varying risk factors have been under-researched. It is well-known that certain patients have seizures in preferential brain states, regardless of the circadian time, for example, patients with seizures occurring exclusively during specific sleep states, either at night or during a nap. Simply tracking brain states and using this information as another covariate will help narrow windows of risk. As mentioned above, an additional role of sleep homeostasis remains unresolved. Additional chronorisk factors may remain to be identified, as uncovering the previously unrecognized importance of circadian and multidien cycles has been a humbling lesson for the field. Different time-varying risk factors likely incorporate into one latent variable – the momentary seizure risk – which may be more directly tested (for example with stimulation) in the future.

Real-time phase estimation. The causal estimation of the instantaneous phase of a cycle in real-time (i.e. without knowledge of future signal fluctuations) represents a major challenge. While this may be relatively easy for cycles with well-characterized states (sleep-wake) or with a set and externally-paced period-length (e.g. circadian), it is particularly difficult for free-running quasi-rhythms defined by their non-stationarity leading to changes in period-length and phase-shifts (e.g. multidien)[91]. So far, the issue of causal instantaneous phase estimations of multidien rhythms of brain excitability and heart-beat has not been solved in the forecasting studies using this feature[2,15,33]. Solving this issue is necessary to be able to implement seizure forecasting on the scale of days.

Usability. Keeping the user engaged with the forecasts will be another key issue in future trials. Once reliable forecasts are achieved, incorporating psychological aspects of individual users may reveal another *long and winding road* for the field. While prior probability estimation at daily and hourly timescales are undoubtedly going to improve seizure alarms, whether users prefer last-minute alarms, hourly or daily risk estimates is unknown and should be one outcome of future trials. Although useful to prepare trials, existing surveys[83,85,92] cannot really inform the final choice of a forecasting horizon, and this will have to be determined prospectively. The human mind varies across individuals in its anticipatory plans, but typically a 24-hours advanced notice seems a reasonable horizon, granted risk estimation does not

change in the interval. If the forecast is re-issued at a later point, say 2 hours later, it should not be radically different from its predecessor. This desirable feature, which we here term *forecast stability,* is not accounted for by the deterministic or probabilistic performance assessments presented in this review (they are insensitive to the sequence of issued forecasts) and represents another open issue. As an increasing number of risk factors are included at finer temporal scales, the sharpness of forecasts will undoubtedly increase (more forecasted probabilities at the extremes), but flickering between low and high risk will likely remain undesirable for users. Separating daily forecasts into finer hourly forecasts may be helpful for some users. Additionally, seizure alarms can be issued at minutes-long horizon in real-time. Defining set forecast horizons (here daily, hourly and last-minute alarms) can help comparing future results of clinical trials, but also has practical consequences; if 24-hours are most helpful to users, then the use of devices only capable of batch-transfers (as opposed to continuous data streaming) is possible, lowering the needs of constant connectivity and high-performance batteries.

Design. Potential users have expressed a strong interest in using forecasting devices[92], but even a perfectly performing forecast must be interpreted by users and their care-givers. For instance, it must be made clear to the user that on days with 50% seizure probability means that seizure will occur one day out of two with the same forecast. It is therefore important for investigators to invest in high-quality design, making it as simple as possible to understand the information being presented[85,92]. It is important the systems are cosmetically acceptable, and not stigmatizing in themselves. This should have input from experts such as industrial designers, user interface and experienced product/service designers. There have already been some efforts to explore possible designs beyond the stand-alone Neurovista system of 3 lights (low, unsure, or high chance of seizure)[85] and whether the forecasts should be integrated with smart-devices already used by patients.

Design of upcoming prospective trials

Trial objectives. It is important as well to recognize that the performance measures and time scales in seizure forecasting will necessarily be driven by the particular application of the

forecast, requiring a clear use case definition. For example, an approach focused on providing a patient with an advisory to guide activities may practically incur a high penalty for incorrectly declaring a low-risk state and a low penalty for incorrectly distinguishing high from medium risk states. In contrast, a forecast changing neuromodulation settings or prompting a supplemental medication may have a low penalty practically for occasional false alarms, but a high penalty for missed seizures. Further it may be most appropriate to evaluate some applications using a deterministic measure if the forecasting application must result in a deterministic action (e.g. taking a medication).

Patient selection. Seizure forecasting will not work for all patients with epilepsy and given the clinical complexity of implementing seizure forecasting, trials should likely focus on the most "forecastable" epilepsies at first and broaden to more difficult cases later on. Therefore, studies may screen for people who may most benefit from forecast as an inclusion criterion, acknowledging an explicit selection bias.

Generalization to the epilepsies. Most of what has been learnt about forecasting comes from the study of focal epilepsies. In particular, the Neurovista cohort was relatively small, and focused on patients with temporal foci. While larger studies also had a majority of temporal-lobe epilepsies, the method worked just as well for a number of extra-temporal focal epilepsies[2]. It is to be seen how well electrographic forecasting technologies work for generalized epilepsies, as well as the more complex developmental and epileptic encephalopathies. Pediatric patients and their families may benefit from seizure forecasting technologies in the future.

Self-forecasting. one comparator that trials ought to include is the capacity of certain patients to achieve seizure self-forecasting based on subjective perception or feeling that a seizure is likely or unlikely to occur over future horizons[37,38], although self-forecasting in itself represents a field of ongoing research. Nevertheless, complex and costly technological developments should at least be better than introspection to be of any use.

Collection of *continuous raw EEG data*. Continuous subscalp EEG data is trickling in, as patients are being recruited into trials[15,56,93], but is currently limited to only two to three

channels. To continue to advance the field at the pace set over the past decade, raw EEG data from multiple recording electrodes will be needed. Despite the recent push for sub-scalp EEG methods[56], intracranial EEG remains valuable: many insights could still be gained from multisite high-density intractranial EEG, as critical information resides in the fine characterization of functional connectivity, brain states, seizure propagation and so on. Most importantly EEG datasets must be long (years) and nearly continuous, even if from a limited number of patients. In the field of seizure forecasting *long* rather than *big* data prevails.

Covert forecasts. For robust evaluation of the forecasting performance, forecasts need to be tested covertly and extensively before opening to patients. Indeed, once forecasts are communicated to the participants, behaviour may change, which in turn may change seizure risk. The risk that trials become uninterpretable if forecasts are communicated to participants too soon is a concern. Covert forecasting may have to go on for a year or more, depending on the seizure rate, especially for trials adopting a probabilistic evaluation of forecasting performance.

Performance evaluation. Although the field has increased in its rigor over the decades, variations of the definition of chance-level still exist, sometimes defying comparisons between studies, an issue reported over the years in our field[3,7,30,94]. To be realistic, the performance of a forecasting system must be assessed over the entire time the forecast might be used in a clinical setting (as opposed to retrospectively selected pre-ictal/interictal periods). Indeed, selection of interictal data, that is always biased, may not cover all brain states (e.g. sleep stages)[95] and overestimates specificity in an unbalanced problem such as seizure forecasting. Chance level should be derived empirically from randomized time series used for training and testing models that generate chance-level outputs on out-of-sample, unseen test datasets. Informative approaches include generating a given amount of artificial seizure-onset times using naïve forecasting schemes, such as random[7,94,96] or periodic (e.g. circadian) forecasters[7,96] for deterministic schemes, as well as issuing the long-term expected seizure rate for probabilistic schemes[81]. However, since seizure risk is inhomogenous in time and seizures have interdependencies, a favored approach that increases confidence in the result[3] consists in

randomly shuffling the original inter-seizure intervals to generate surrogate seizure timeseries that share essential statistics (mean, variance, distribution) with the original, patient-specific data[30]. Further confusion arises from different definitions of forecasting horizons, the definition of what constitutes a seizure, and the exclusion of variably defined clustered seizures, historically motivated as a fair concern to not overestimate forecast performance in a deterministic framework (if one seizure is known, the second is easier to predict). However, in the more recent probabilistic view, this approach is likely contra-productive, as clustered seizures indeed confirm the existence of states of heightened likelihood. As seizure forecasting may not work for all patients, future studies ought to report the proportion of subjects in the included cohort with forecasts showing statistical improvement over chance-forecasting, and, for those, also quantify complementary attributes of goodness of forecasts. Indeed, that a forecast is better than chance does not yet mean that it is of any potential use in clinical practice. To facilitate comparison between future studies, we here propose the systematic report of a set of pre-defined metrics covering deterministic and probabilistic evaluation of forecast performance. For deterministic metrics, the field has favored[3] the report of the area-under the curve (AUC) of the sensitivity versus time-in-warning over the entire dataset (as opposed to selected pre-ictal and interictal clips), as well as the actual sensitivity and the time-in-warning for an optimized threshold. For probabilistic metrics, a number of studies[2,12,15,76] have reported the reliability curve, calibration, resolution and associated Brier skill score. These metrics can be calculated for individual[1,2,13] or aggregated[2,76] data, that is by pooling daily forecasts and observations across subjects. While the former is most informative from a patient's standpoint, the latter is strongly influenced by the large differences in expected absolute risk across patients (F_4 & F_5 in Fig. 1) and be misleading. Indeed, an algorithm that is merely able to capture the difference in seizure rates across patients in a cohort would lead to high scores. However, aggregating data can assess whether the algorithm accommodates for a wide range of individual seizure rates and can pool a sufficient number of observations to obtain well-defined curves. We recommend reporting individual scores, and report aggregate data only when data is insufficient at the individual level to populate the bins of the

calibration curve. If some metrics cannot be reliably derived from the data (e.g. lack of large test dataset for probabilistic evaluation), this should be discussed as a limitation, because the true forecasting performance of a given algorithm cannot be known with a single metric[78].

Patient outcomes. Studies of forecasting systems largely focus on the performance of the algorithms. However, in order to show true clinical benefit to people with epilepsy and their carers, we must demonstrate that forecasting systems can provide benefit in the management of epilepsy, which has not been achieved by any study so far. Tracking outcomes such as quality of life, stress levels, depression/anxiety scores, and changes in seizure rates will be important in demonstrating benefit to users, regulators, and payors.

After the trials

Deployment of neurotechonologies. One essential point learned since the first chronic EEG device in a human head, is that developing and commercializing neurotechnologies is a complex endeavor. From a novel idea to the first device sold, one should roughly count 10-20 years and a minimum of 50-100 million dollars. Even when the idea is good, the technology performs, and the clinical trials succeed, practical aspects may come in the way of commercialization. Many neurologists are not familiar with neurotechnologies and prefer to read the EEG visually, a heritage that will remain for years until the majority of them are convinced that machines do just as well. For patients and investors alike, invasiveness is unattractive, which leads to hesitations to accept such technology, especially those that are not lifesaving, unlike implanted defibrillators. Thus, minimal- or non-invasiveness is a way to promote and has enabled cardiology to adopt diagnostic and treatment devices at a large scale. As a community, raising awareness early on, highlighting how close-monitoring of biomarkers has potential to transform epilepsy management (just like in diabetes), while openly discussing the upcoming issues as attempted here is probably the way to make such deployment more likely.

Learning from people with epilepsy. Very few people have ever received a forecast of their seizures. As the symptoms which a forecast would potentially alleviate are largely subjective in nature (that is, the uncertainty of seizure timing), it is essential that those who have exposure

to forecasts are able to express their experiences. For instance, in the relatively limited NeuroVista cohort, the experience of participants varied dramatically[97]. Furthermore, there is still much to learn about which patient cohorts receive the most benefit from seizure forecasting, for example, those with high versus low seizure counts, those with single or multiple seizure types, those who have had epilepsy for many years or those with relatively new diagnoses.

Conclusion

The field of seizure forecasting has made steady progress over the past 20 years but has definitely accelerated in the last decade, with a more profound understanding of what the important predictive factors for seizures are. The (re-)discovery of cycles in epilepsy and the generalizability of the notion of pro-ictal states unlocked a number of aspects in seizure forecasting that were previously thought impossible, mainly: 1) the possibility of forecasting seizure risk over days as opposed to minutes, and 2) the possibility of forecasting seizure risk non-invasively with simple wristbands. With these scientific and engineering leaps forward, the time has come to accelerate the resolution of unsolved problems, take the ultimate test of *prospective* trials and learn how seizure forecasting may best help people with epilepsy. While the road is still long, the path ahead is clear.

Appendix

Methodology for evaluating probabilistic forecasts

Refining measurement of the value of probabilistic forecasts has a decades-long history in meterology[77,92]. More recently, the field of machine-learning has also adopted similar definitions and concepts[93]. Nomenclature differs somewhat from one community to the other. Our own definitions inspired from these fields are given in Table 3, and their geometric meaning can be visualized in reliability diagrams (Fig. 1) which evaluates how well the forecasted probabilities of an event correspond to their observed probabilities. Excellent online sources can be found at:

- https://www.cawcr.gov.au/projects/verification
- http://checkmyai.com/index.php?get=methods

Metrics to evaluate probabilistic forecast performance do not rely on classical definitions of false/true positives/negatives and the related deterministic scores (e.g., Sensitivity, Specificity), as probabilities are not thresholded. Rather, probabilistic scores determine how well a group of probabilistic predictions correspond to reality.

The reliability diagram

The reliability diagram, also called calibration diagram, compares forecasted, *a priori* probabilities to observed (i.e. empirical), *a posteriori* probabilities and provides an excellent diagnostic tool. These assessments require a large number of trials, as the measure is based on a histogram method with multiple observations per bin. For example, consider a given forecasted value of 5% probability of an event with a 24-hour horizon. If we assume that such a forecast is issued one day out of 10, verifying, with some degree of confidence, that observations match this forecast would require observing at least 2 events out of 40 forecasts with 5% probability, i.e. it would necessitate ~400 days of observation for daily forecasts. The

calculation must be repeated for higher and lower forecasted values associated with more and less frequent events, respectively. This leads to a rapid expansion of the number of observations required to correctly assess the performance of a model at different forecasted values. The reliability diagram, when sufficiently populated, allows for a geometric understanding of all characteristics of a good (or bad) forecast, which compose the Brier score (below)[90]. The construction of the reliability curve is in itself the focus of a body of litterature in meteorology. For example, to visualise the behavior at extreme probabilities, one strategy is to uniformly define 10 bin boundaries at each decile (i.e. 0-10%, 10-20%, etc.). In contrast, to better understand the average behavior of the forecast, one can define bin boundaries, such that each bin is populated by an equal number of forecasts[94]. Additionally, to aid visual interpretation of a good forecast, it is useful to plot 5-95% bootstrap limits on the expected behavior (the *diagonal*)[94].

The Brier score

The *Brier score*[98] developed for meteorological forecasts was first proposed for seizure forecasting a decade ago[79]. It measures the magnitude of the probabilistic forecast errors, and squares it to put weight on larger errors. The Brier score can be partitioned into three attributes for better interpretability[90,96]. First, a forecast is evaluated accounting for the *uncertainty* intrinsic to the problem, as not all forecasting problems are equally difficult (i.e. some are imbalanced). In the case of epilepsy, clinicians and patients typically calculate the proportion of days (or hours) with seizures, which represents the expected individual daily seizure rate that can be known *a priori*, before issuing any forecast. This rate may take extreme values from <1% (e.g. one seizure per year) to 100% (i.e. one or more seizure per day) representing simpler forecasting problems (less uncertainty) because a trivial solution is to forecast 0% seizure and 100% seizure every day and be accurate a vast majority of times[78]. Patients with intermediary seizure rates, say 30-50%, meaning one seizure every other day or two pose a greater problem, as uncertainty as to which days is greater[78]. Second, a good probabilistic forecast is *calibrated* (or *reliable*). A *loss of calibration* is defined as datapoints deviating from

the diagonal and can result in over or under-forecasting *bias*. Calibration can be adjusted after algorithm training, in a subsequent step of "re-calibration," which can compensate for systematic biases. Third, a good probabilistic forecast has *resolution*, i.e. it is able to predict different outcomes, when it takes on different values[78]. If the outcome is independent of the forecast, the forecast has no resolution and is useless. For example, a trivial solution of always forecasting the expected daily seizure rate (e.g. 30-50%) will be accurate but brings no information. When *resolution* is absent, *discrimination* is also absent for each chosen threshold (ie AUC = 0.5). *Resolution* is conditioned on the forecasts: are different outcomes obtained given different forecasted values? Conversely, *discrimination* is conditioned on the observations: are different forecasted values obtained given different outcome categories? *Resolution* cannot be adjusted after algorithm training. To develop an intuitive sense for resolution, it is useful to clarify the difference between absolute risk (a calibrated probability between 0 and 1) and relative risk (a probability difference or ratio). An accurately forecasted probability of say 10% on a given day is a low risk in absolute (closer to zero than one) and has the probabilistic meaning of seeing one event among 10 days with such accurate forecasts, independently of the expectation of the patient for whom it is issued. However, this forecast may have drastically different meanings to different individuals. For a patient with a long-term expected risk of 2% per day (i.e. one seizure every two months, as established by observation over a long period), a daily forecast of 10% would mean a 5x higher risk on that day relative to the average risk. For another patient with a long-term expected risk of 20%, a daily forecast of 10% would mean a halving of the risk on that day relative to the average. Discrimination (AUC) and resolution emphasize the benefits of the forecast for the user in terms of time-varying relative risk, because they ignore (AUC) or account (BSS) for the expected seizure rate. Discrimination interprets this relative risk as a category, whereas resolution offers a graded view on how high or low is the risk, relative to the long-term expected probability (i.e. a risk ratio or absolute risk difference). Additionally, a *sharp* forecast tends to issue mostly higher and/or lower probabilities, i.e. it offers greater confidence in the outcome and approaches deterministic strategies, provided it is calibrated.

The *Brier skill score* incorporates all of these attributes[99] and assesses the improvement in performance of the output forecast (its skill) relative to a random reference (for example, the random shuffling of the original forecast)[81] that has the same average probability. Indeed, as the Brier score also depends on uncertainty, the Brier skill score is typically used instead for better comparability across forecasts with different uncertainties. Nevertheless, the Brier skill score is a composite score, and reporting calibration, sharpness and resolution separately (e.g. in the form of the reliability diagram) is useful.

In this probabilistic context, the *AUC* represents a complementary metric to the *Brier skill score*, influenced by the problem's *uncertainty* as well as by forecast *resolution* and *sharpness* but not *calibration*. In seizure forecasting, the imbalanced nature of the problem leads most researchers to use the *time-in-warning* as opposed to specificity for the calculation of the *AUC*, so as to avoid an over-weighting of true negatives[3,94]. While the *Brier score* and the *AUC* are influenced by the degree of *uncertainty* (imbalance), the reliability diagram is not and therefore represents an excellent visual tool to assess performance and even point out potential problems/biases with the forecasting scheme (Fig. 1).

References

1. 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.
2. 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.
3. Kuhlmann L, Lehnertz K, Richardson MP, et al. Seizure prediction — ready for a new era. Nat Rev Neurol. 2018; 14(10):618–30.
4. Baud MO, Rao VR. Gauging seizure risk. Neurology. 2018; 91(21):967–73.
5. Freestone DR, Karoly PJ, Cook MJ. A forward-looking review of seizure prediction. Curr Opin Neurol. 2017; 30(2):167–73.
6. Freestone DR, Karoly PJ, Peterson ADH, et al. Seizure Prediction: Science Fiction or Soon to Become Reality? Curr Neurol Neurosci Rep. 2015; 15(11).
7. Mormann F, Andrzejak RG, Elger CE, et al. Seizure prediction: The long and winding road. Brain. 2007; 130(2):314–33.

8. Mormann F, Andrzejak RG. Seizure prediction: making mileage on the long and winding road. Brain. 2016; 139(6):1625–6.
9. Baud MO, Proix T, Rao VR, et al. Chance and risk in epilepsy. Curr Opin Neurol. 2020; 33(epub).
10. Baud MO, Kleen JK, Mirro EA, et al. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun. 2018; 9(1):1–10.
11. Karoly PJ, Cook MJ, Maturana M, et al. Forecasting cycles of seizure likelihood. Epilepsia. 2020; (October 2019):2019.12.19.19015453.
12. Karoly PJ, Ung H, Grayden DB, et al. The circadian profile of epilepsy improves seizure forecasting. Brain. 2017; 140(8):2169–82.
13. Stirling RE, Grayden DB, D'Souza W, et al. Forecasting Seizure Likelihood With Wearable Technology. Front Neurol. 2021; 12(July):1–12.
14. 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.
15. 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.
16. Leguia MG, Proix T, Tcheng TK, et al. Learning to generalize seizure forecasting. 2021.
17. Iasemidis LD, Chris Sackellares J, Zaveri HP, et al. Phase space topography and the Lyapunov exponent of electrocorticograms in partial seizures. Brain Topogr. 1990; 2(3):187–201.
18. Litt B, Echauz J. Prediction of epileptic seizures. Lancet Neurol. 2002; 1(1):22–30.
19. Lehnertz K, Litt B. The First International Collaborative Workshop on Seizure Prediction: Summary and data description. Clin Neurophysiol. 2005; 116(3):493–505.
20. Schelter B, Timmer J, Schulze-Bonhage A. Seizure Prediction in Epilepsy: From Basic Mechanisms to Clinical Applications. Seizure Predict Epilepsy From Basic Mech to Clin Appl. 2008; :1–334.
21. Tetzlaff R, Elger CE, Lehnertz K. Recent Advances in Predicting and Preventing Epileptic Seizures. 2013. 304 p.
22. Zaveri HP, Schelter B, Schevon CA, et al. Controversies on the network theory of epilepsy: Debates held during the ICTALS 2019 conference. Seizure. 2020; 78(March):78–85.
23. Zaveri HP, Frei MG, Arthurs S, et al. Seizure prediction: The fourth international workshop. Epilepsy Behav. 2010; 19(1):1–3.
24. Frei MG, Zaveri HP, Arthurs S, et al. Controversies in epilepsy: Debates held during the Fourth International Workshop on Seizure Prediction. Epilepsy Behav. 2010; 19(1):4–16.

25. Kuhlmann L, Grayden DB, Cook MJ. Proceedings of the 7th international workshop on seizure prediction. Int J Neural Syst. 2017; 27(1):1–2.
26. Ung H, Baldassano SN, Bink H, et al. Intracranial EEG fluctuates over months after implanting electrodes in human brain. J Neural Eng. 2017; 14(5).
27. Sillay KA, Rutecki P, Cicora K, et al. Long-term measurement of impedance in chronically implanted depth and subdural electrodes during responsive neurostimulation in humans. Brain Stimul. 2013; 6(5):718–26.
28. Meisel C, Schulze-Bonhage A, Freestone D, et al. Intrinsic excitability measures track antiepileptic drug action and uncover increasing/decreasing excitability over the wake/sleep cycle. Proc Natl Acad Sci U S A. 2015; 112(47):14694–9.
29. Díaz-Negrillo A. Influence of Sleep and Sleep Deprivation on Ictal and Interictal Epileptiform Activity. Epilepsy Res Treat. 2013; 2013:1–7.
30. Andrzejak RG, Chicharro D, Elger CE, et al. Seizure prediction: Any better than chance? Clin Neurophysiol. 2009; 120(8):1465–78.
31. Cook MJ, Varsavsky A, Himes D, et al. The dynamics of the epileptic brain reveal long memory processes. Front Neurol. 2014; 5(OCT):1–8.
32. Karoly PJ, Freestone DR, Boston R, et al. Interictal spikes and epileptic seizures: Their relationship and underlying rhythmicity. Brain. 2016; 139(4):1066–78.
33. Maturana MI, Meisel C, Dell K, et al. Critical slowing down as a biomarker for seizure susceptibility. Nat Commun. 2020; 11(1):2172.
34. Chen Z, Grayden DB, Burkitt AN, et al. Spatiotemporal Patterns of High-Frequency Activity (80-170 Hz) in Long-Term Intracranial EEG. Neurology. 2021; 96(7):e1070–81.
35. 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.
36. Payne DE, Dell KL, Karoly PJ, et al. Identifying seizure risk factors: A comparison of sleep, weather, and temporal features using a Bayesian forecast. Epilepsia. 2021; 62(2):371–82.
37. Payne DE, Karoly PJ, Freestone DR, et al. Postictal suppression and seizure durations: A patient-specific, long-term iEEG analysis. Epilepsia. 2018; 59(5):1027–36.
38. Brinkmann BH, Wagenaar J, Abbot D, et al. Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain. 2016; 139(6):1713–22.
39. Kuhlmann L, Karoly P, Freestone DR, et al. Epilepsyecosystem.org: Crowd-sourcing reproducible seizure prediction with long-term human intracranial EEG. Brain. 2018; 141(9):2619–30.
40. Sun FT, Morrell MJ. The RNS System: Responsive cortical stimulation for the

treatment of refractory partial epilepsy. Expert Rev Med Devices. 2014; 11(6):563–72.
41. Jobst BC, Kapur R, Barkley GL, et al. Brain-responsive neurostimulation in patients with medically intractable seizures arising from eloquent and other neocortical areas. Epilepsia. 2017; 58(6):1005–14.
42. 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.
43. 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.
44. Rao VR, Leguia M, Tcheng TK, et al. Cues for seizure timing. Epilepsia. 2020; (April):epi.16611.
45. Leguia MG, Andrzejak RG, Rummel C, et al. Seizure Cycles in Focal Epilepsy. JAMA Neurol. 2021; 78(4):454.
46. Duckrow RB, Tcheng TK. Daily variation in an intracranial EEG feature in humans detected by a responsive neurostimulator system. Epilepsia. 2007; 48(8):1614–20.
47. Spencer DC, Sun FT, Brown SN, et al. Circadian and ultradian patterns of epileptiform discharges differ by seizure-onset location during long-term ambulatory intracranial monitoring. Epilepsia. 2016; 57(9):1495–502.
48. Sisterson ND, Wozny TA, Kokkinos V, et al. Closed-Loop Brain Stimulation for Drug-Resistant Epilepsy: Towards an Evidence-Based Approach to Personalized Medicine. Neurotherapeutics. 2019; 16(1):119–27.
49. Khambhati AN, Shafi A, Rao VR, et al. Long-term brain network reorganization predicts responsive neurostimulation outcomes for focal epilepsy. Sci Transl Med. 2021; (in press).
50. Gregg NM, Sladky V, Nejedly P, et al. Thalamic deep brain stimulation modulates cycles of seizure risk in epilepsy. Sci Rep. 2021; 11(1):1–12.
51. Gregg NM, Nasseri M, Kremen V, et al. Circadian and multiday seizure periodicities, and seizure clusters in canine epilepsy. Brain Commun. 2020; 2(1):1–13.
52. Gregg NM, Marks VS, Sladky V, et al. Anterior nucleus of the thalamus seizure detection in ambulatory humans. Epilepsia. 2021; (August):1–7.
53. 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.
54. Gangstad SW, Mikkelsen KB, Kidmose P, et al. Automatic sleep stage classification based on subcutaneous EEG in patients with epilepsy. Biomed Eng Online. 2019; 18(1):106.
55. 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.
56. Duun-Henriksen J, Baud M, Richardson MP, et al. A new era in electroencephalographic monitoring? Subscalp devices for ultra–long-term recordings. Epilepsia. 2020; 61(9):1805–17.
57. Griffiths G, Fox JT. Rhythm in Epilepsy. Lancet. 1938; 232(5999):409–16.
58. Karoly PJ, Rao VR, Gregg NM, et al. Cycles in epilepsy. Vol. 0123456789, Nature Reviews Neurology. Springer US; 2021.
59. Gowers WR. Epilepsy and Other Chronic Convulsive Diseases; Their Causes, Symptoms and Treatment. Churchill J& A, editor. London; 1881.
60. Langdon-Down M, Brain WR. Time of Day in Relation to Convulsions in Epilepsy. Lancet. 1929; :1029–32.
61. Magnussen G. 18 cases of epilepsiy with fits in relation to sleep. Acta Psychiatr Scand. 1936; 11(1):289–321.
62. Janz D. The grand mal epilepsies and the sleep-waking cycle. Epilepsia. 1962; 3:69–109.
63. Khan S, Nobili L, Khatami R, et al. Circadian rhythm and epilepsy. Lancet Neurol. 2018; 17(12):1098–108.
64. Rossi KC, Jalyoung J, Makhija M, et al. Insufficient sleep, EEG activation, and seizure risk: Re-evaluating the evidence. Ann Neurol. 2020; :617–32.
65. Dell KL, Payne DE, Kremen V, et al. Seizure likelihood varies with day-to-day variations in sleep duration in patients with refractory focal epilepsy: A longitudinal electroencephalography investigation. EClinicalMedicine. 2021; 37:100934.
66. Herzog AG, Harden CL, Liporace J, et al. Frequency of catamenial seizure exacerbation in women with localization-related epilepsy. Ann Neurol. 2004; 56(3):431–4.
67. Haut SR, Hall CB, Masur J, et al. Seizure occurrence: Precipitants and prediction. Neurology. 2007; 69(20):1905–10.
68. Rakers F, Walther M, Schiffner R, et al. Weather as a risk factor for epileptic seizures: A case-crossover study. Epilepsia. 2017; 58(7):1287–95.
69. Baud MO, Ghestem A, Benoliel JJ, et al. Endogenous multidien rhythm of epilepsy in rats. Exp Neurol. 2019; 315(February):82–7.
70. 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; .
71. Karoly PJ, Stirling RE, Freestone DR, et al. Multiday cycles of heart rate are

associated with seizure likelihood: An observational cohort study. EBioMedicine. 2021; 72:103619.
72. Chang WC, Kudlacek J, Hlinka J, et al. Loss of neuronal network resilience precedes seizures and determines the ictogenic nature of interictal synaptic perturbations. Nat Neurosci. 2018; 21(12):1742–52.
73. Xiong W, Nurse ES, Lambert E, et al. Seizure forecasting using long-term electroencephalography and electrocardiogram data. Int J Neural Syst. 2021; 31(9):1–14.
74. Scheffer M, Bascompte J, Brock WA, et al. Early-warning signals for critical transitions. Nature. 2009; 461(7260):53–9.
75. Scheffer M, Carpenter SR, Lenton TM, et al. Anticipating critical transitions. Science (80-). 2012; 338(6105):344–8.
76. Goldenholz D, Goldenholz S, Romero J, et al. Development and validation of forecasting next reported seizure using e-diaries. Ann Neurol. 2020; :1–8.
77. Lorenz E. Deterministic Nonperiodic Flow. J Atmos Sci. 1963; :130–41.
78. Mason S. Guidance on Verification of Operational Seasonal Climate Forecasts. Seevccc.Rs. 2013.
79. 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.
80. Winkler RL., Murphy AH. " Good " Probability Assessors. J Appl Meteorol. 1968; 7(May 2020):751–8.
81. 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.
82. Bröcker J, Smith LA. Increasing the reliability of reliability diagrams. Weather Forecast. 2007; 22(3):651–61.
83. Janse SA, Dumanis SB, Huwig T, et al. Patient and caregiver preferences for the potential benefits and risks of a seizure forecasting device: A best–worst scaling. Epilepsy Behav. 2019; 96:183–91.
84. Chiang S, Moss R, Patel AD, et al. Seizure detection devices and health-related quality of life: A patient- and caregiver-centered evaluation. Epilepsy Behav. 2020; 105:106963.
85. Chiang S, Moss R, Black AP, et al. Evaluation and recommendations for effective data visualization for seizure forecasting algorithms. JAMIA Open. 2021; 4(1):1–14.
86. Nasseri M, Nurse E, Glasstetter M, et al. Signal quality and patient experience with wearable devices for epilepsy management. Epilepsia. 2020; (January):1–11.
87. Bruno E, Biondi A, Richardson MP. Pre-ictal heart rate changes : A systematic review

and meta-analysis. 2018; 55:48–56.
88. Tang J, El Atrache R, Yu S, et al. Seizure detection using wearable sensors and machine learning: Setting a benchmark. Epilepsia. 2021; 62(8):1807–19.
89. Nasseri M, Pal Attia T, Joseph B, et al. Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning. Sci Rep. 2021; 11(1):1–9.
90. 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.
91. Leguia MG, Rao VR, Kleen JK, et al. Measuring synchrony in bio-medical timeseries. Chaos. 2021; 31(1).
92. Grzeskowiak CL, Dumanis SB. Seizure Forecasting: Patient and Caregiver Perspectives. Front Neurol. 2021; 12(September).
93. Viana PF, Duun-Henriksen J, Glasstëter M, et al. 230 days of ultra long-term subcutaneous EEG: seizure cycle analysis and comparison to patient diary. Ann Clin Transl Neurol. 2020; :acn3.51261.
94. Snyder DE, Echauz J, Grimes DB, et al. The statistics of a practical seizure warning system. J Neural Eng. 2008; 5(4):392–401.
95. Schelter B, Winterhalder M, Maiwald T, et al. Do false predictions of seizures depend on the state of vigilance? A report from two seizure-prediction methods and proposed remedies. Epilepsia. 2006; 47(12):2058–70.
96. Mormann F. Seizure prediction. Scholarpedia. 2008; 3(10):5770.
97. Gilbert F, O'Brien T, Cook M. The Effects of Closed-Loop Brain Implants on Autonomy and Deliberation: What are the Risks of Being Kept in the Loop? Cambridge Q Healthc Ethics. 2018; 27(2):316–25.
98. Brier GW. Verification of Forecasts Expressed in Terms of Probability. Mon Weather Rev. 1950; 78(1):1–3.
99. Murphy AH. A New Vector Partition of the Probability Score. J Appl Meteorol. 1973; 12(4):595–600.

Figure 1. Deterministic and probabilistic performance of five illustrative seizure forecasts. **a**: Visual representation of a seizure gauge conveying low to high daily seizure risk, as compared to the expected seizure rate expressed in seizures per day (e, blue dotted line). **b**: seizure probabilities issued over the first 60 days. **c**: histogram of the 1000 generated daily forecasts in 10 bins of probability deciles and corresponding sharpness (S, range 0-0.25). Of note, bins of forecasted probabilities can be equally populated or spaced to show the central or extreme tendency, respectively[82]. **d**: Reliability diagrams depicting the calibration of each forecasted probability decile to the observed seizure frequency. Graphically, calibration loss (CL, range 0-1.0, lower is better) is the average distance to the diagonal (perfect calibration line) and resolution (R, range 0-0.25, higher is better) the average distance to the horizontal 'no resolution' line, which corresponds to the expected seizure rate. **e**: Area under the curve (AUC) of the proportion of seizures found (sensitivity) versus proportion of time spent (time-in-warning) above a given probability threshold (gradient-color). Each of the five daily forecasts were generated over 1000 days in an arbitrary, but realistic manner to illustrate how the deterministic and probabilistic metrics can be used conjointly to interpret the result. $\mathbf{F_1}$ is the best of the five forecasts, because it is calibrated and sharp, with most output probabilities either low (close to zero) or high (close to one), leading to higher resolution and discrimination (AUC). Although as well calibrated (same CL), $\mathbf{F_2}$ is slightly less sharp than F_1, with values concentrated around the expected seizure rate, which decreases resolution and discrimination on the same set of observed seizures. This means that intermediate output probabilities (e.g. 0.3) are accurate, but less discriminative than more extreme, accurate output probabilities in F_1 (e.g. 0.1), which is reflected in almost halving the BSS. $\mathbf{F_3}$ is the same forecast as F_2 but applied to another set of observed seizures, where additional seizures (red dots) resulted from stochastic noise, representing a case where the forecaster does not capture difficult-to-measure seizure triggers. As a result, the forecast is biased, systematically underestimating the event probability, resulting in calibration loss. Discrimination is also decreased because the observation is saturated with frequent seizures that are not accounted for by the forecast (top right corner in d). Both $\mathbf{F_4}$ and $\mathbf{F_5}$ have excellent calibration but mediocre resolution, forecasting the expected seizure rate ±10%. F_4 remains nevertheless more useful than F_5 as the forecasted relative fluctuations between low (e.g. 0.1) and low-intermediate probabilities (e.g. 0.3) bears more discrimination than relative fluctuations between intermediate-high (e.g. 0.5) and high (e.g. 0.7) probabilities. This example illustrates that the AUC brings complementary information to resolution and calibration. Conversely, evaluating either F_5 or F_4 against the observations O_4 made for F_4 yields the same AUC = 0.62 because the relative fluctuation around the expected seizure rate is the same, highlighting that the AUC is blind to calibration. Finally, the joint evaluation of F_4 and F_5 and their corresponding observations in aggregate increases the AUC to 0.77 and the BSS to 0.2, better than any of the two individual forecasts, simply because F_4 and F_5 are both well calibrated for low and high seizure rates, respectively. This highlights that high scores for aggregated forecasts can be misleading and give a false impression of excellent forecasting performance, when in fact, they merely reveal the ability to distinguish patients with low and high seizure rates. This figure was created using computations available at https://checkmyai.com/.

	Risk Factor		Data	N	Prevalence if available	Categorical Effect size	Circular effect size	Reference
Sporadic	Stormy weather		Admission	604	Group-level	**1.1-1.5 (OR)**	-	Rakers F et al., Epilepsia (2017)
			NV	8	4/8	-	-	Payne et al., Epilepsia (2021)
	Missed medication		Diary	71	Group-level	1.2		Haut SR et al., Neurology (2007)
	Alcohol intake		Diary	71	Group-level	1.5		Haut SR et al., Neurology (2007)
	Mood	Favorable change in mood	Diary	**19**	**9/19**	**0.8** (OR)		Haut SR et al., Epilepsia (2013)
	Stress - anxiety		Diary	71	Group-level	**1.1 (OR)**		Haut SR et al., Neurology (2007)
	Sleep duration	decreased	Diary	71	Group-level	1.1 (OR)		Haut SR et al., Neurology (2007)
			NV	12	0/12	-		Dell et al., EClinicalMedicine, 2021
		increased	Diary	71	Group-level	**0.9 (OR)**		Haut SR et al., Neurology (2007)
			NV	8	**0/8**	-		Payne et al., Epilepsia (2021)
			NV	12	10/12	0.7 (OR)		Dell et al., EClinicalMedicine (2021)
Cyclical	Circannual		NP	194	12%	-	0.17	Leguia MG et al., JAMA Neurology (2021)
	Moon		Admission	859	Group-level	1.8-1.9 (OR)	-	Polychronopoulos P et al., Neurology (2006)
			NV	8	1/8	-	-	Payne et al., Epilepsia (2021)
			NP	186	0/186	-	-	Leguia MG et al., JAMA Neurology (2021), Epilepsia (2020)
	Days of the week		ST diary	**9849**	Group-level	**1.08 (max IRR)**		Ferastraoaru, Epilepsia open, 2018
			ST diary	**1118**	**7-21%**	**-**	-	Karoly, Lancet Neurology, 2018
			NV	**8**	0/8	**-**	-	Payne et al., Epilepsia (2021)
			NP	**186**	**5%**	**1.05 (max RR)**		Leguia et al., JAMA Neurology (2021), Epilepsia (2020)
	Menstrual		Diary		Group-level	~2 (max RR)	-	Herzog AG et al., Annals of Neurology (2004)
			Diary	71	Group-level	1.2	-	Haut SR et al., Neurology (2007)
			Diary	184	42%	-	-	Herzog, Neurology, 2012
			Diary	100	-	-	-	Herzog A, Epilepsia, 2015
	Multidien		NP	**14**	**13/14**	7	0.32	Baud M, Nature comm, 2018

		NV	**15**	**15**	-	0.67	Maturana M, Nature comm 2020
		NP	**186**	**60%**	**~7** (RR)	0.34	Leguia MG et al., JAMA Neurology (2021)
		Heart rate	**19**	**10/19**		0.37	Karoly, EBioMedicine, 2021
	Circadian	NP	**14**	**12/14**	5	0.32	Baud M, Nature comm, 2018
		ST diary	**1118**	**80-82%**	-	-	Karoly, Lancet Neurology, 2018
		ST diary	**9849**	Group-level	5.5 (IRR)	-	Ferastraoaru, Epilepsia open, 2018
		NV	**15**	**15**	-	0.77	Maturana M, Nature comm 2020
		NP	**85**	**76/85**	**~5** (RR)	0.34	Leguia MG et al., JAMA Neurology (2021)
		Heart rate	**19**	**14/19**		0.42	Karoly, EBioMedicine, 2021
	Sleep-wake	Night EEG		-	~50 (RR) REMS		Ng et al., Epilepsy Res Treat (2013)
	Self prediction	Diary Diary	**71** **19**	**-** **9/19**	**3.7 (OR)** **~5-9** (OR)		Haut SR et al., Neurology (2007) Haut, Epilepsia, 2013

Table 1: Time-varying risk factors, prevalence and effect-sizes. For simplicity, 95% confidence intervals are not shown, but values that are statistically significant are in bold. Group-level: no prevalence reported, effect-size calculated at the group level. Values in italics were recalculated, either from the raw data, or from processed values reported in the charts. OR: odds-ratio. RR: relative risk or risk ratio. IRR: incidence rate ratio. NV: NeuroVista. NP: NeuroPace. ST diary: SeizureTracker diary. Dash: value not available in original paper.

University of Freiburg EEG database	https://epilepsy.uni-freiburg.de/freiburg-seizure-prediction-project/eeg-database
SWEC-ETHZ iEEG Database	http://ieeg-swez.ethz.ch/
Temple University Hospital	https://isip.piconepress.com/projects/tuh_eeg/html/downloads.shtml
Epilepsy Ecosystem	Epilepsyecosystem.org
UCSF – UniBE – UniGE Dataset	https://zenodo.org/record/5094447#.YgEd8vXMLOQ

Table 2: Currently available intracranial EEG databases and datasets.

Terminology		Definition	Formula
General definitions	**Forecast horizon**	The future period of time for which a forecast is generated.	
	Uninformative forecasts	Forecasts that do not help decision-making. Trivial solutions, such as perpetually issuing 0% probability for rare events, have good performance but are uninformative (unskilled) and can be used as a reference.	
	Discrimination	Discrimination measures whether forecasts differ when their corresponding observations differ; for example, if forecasts for days that are wet indicate more rain than for days that are dry, the forecasts can discriminate wetter from drier days.	
Deterministic metrics	**Accuracy**	Measure of discrimination or how well a forecast correctly identifies or excludes a certain outcome.	$\frac{TP + TN}{All}$
	Sensitivity (Se)	How often the forecast correctly identifies an event.	$\frac{TP}{TP + FN}$
	Specificity (Sp)	How often the forecast avoids misidentification.	$\frac{TN}{TN + FP}$
	Time in warning (Tiw)	Duration of time a forecast indicates an event is likely.	$\frac{TP + FP}{All}$
	Area under the curve (AUC)	Typically assessed as the tradeoff between sensitivity and specificity (or time in warning) by systematically thresholding the algorithm output at all forecasted values.	*Se vs. 1-Sp* *or* *Se vs. Tiw*
	Relative risk	The ratio between the probability of an event in a category or state and the probability of this event in another category.	$\frac{TP / TP + FP}{FN / FN + TN}$
Probabilistic metrics	**Observed probability**	Frequency of events per unit of time observed in the data, ie their empirical probability.	$\frac{\sum_{i=1}^{n} o_i}{n}$
	Expected probability	Based on all previous observations, the frequency (probability) of events expected over long duration in the future.	$lim_{n\to\infty} \frac{\sum_{i=1}^{n} o_i}{n}$
	Forecasted probability	Probability of event forecasted for one time interval in the future	f_i
	Calibration (or reliability)	Agreement between forecasted probability and observed probability. Typically calculated by averaging n forecasts datapoints in m ranked bins (f_k , e.g. average forecast between 0 and 10%) and calculating the corresponding observed event probability, $\bar{o}_k$. For a calibrated forecast, the binned forecasted probability and observed probability match and therefore align on a diagonal in a reliability diagram. Graphically, distance to the diagonal (Fig. S1).	$\frac{1}{n}\sum_{k=1}^{m} n_k(\bar{f}_k - \bar{o}_k)^2$
	Resolution	Ability of the forecast to separate observed probabilities from the average observed probability. Resolution is zero for a flat line intersecting the y-axis at the expected probability, this corresponds to alignment of the ROC curve with the diagonal. Graphically, separation of the reliability curve from the horizontal line of no resolution (Fig. S1).	$\frac{1}{n}\sum_{k=1}^{m} n_k(\bar{o}_k - \bar{o})^2$
	Sharpness	Tendency to forecast probabilities, f_i, near 0 or 1, as opposed to uniformly distributed forecasts. Sharpness is an attribute belonging only to the forecast and is not influenced by the observations. Graphically, variance of the distribution of the forecasts.	$\frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})$
	Uncertainty	Uncertainty only depends of the frequency of events $\bar{o}$ and is not influenced by the forecast. Uncertainty tends to 0 with very rare (or frequent) observations (ie with increased imbalance) and is greatest (=0.25) when an event is observed 50% of the time, making forecasts more difficult.	$\bar{o}(1 - \bar{o})$
	Skill	Accuracy of a forecast relative to some reference forecast. The reference forecast is generally an unskilled forecast such as random chance, shuffled forecasts, or uninformative forecasts. A forecast may be better simply because it is easier to make, which is taken into account when calculating Skill.	$1 - \frac{Score}{Score_{ref}}$

	Bias	Mismatch between the mean forecast value, $\bar{f}$, and mean observed probability, $\bar{o}$.	$\bar{f} - \bar{o}$
	Brier score (BS)	Mean squared distance between the forecasted value, f_i, and the observation, o_i (set to 1 or 0), calculated at each ith timepoint for n forecasts. Better Brier scores are lower (ie tend to zero).	$\frac{1}{n}\sum_{i=1}^{n}(f_i - o_i)^2$
	Brier skill score (BSS)	Improvement of Brier score over a reference forecast. Brier skill scores tend to 1 when better, 0 when no improvement over reference, and $-\infty$ when worse than reference.	$1 - \frac{BS}{BS_{ref}}$

Table 3: Metrics for forecast performance. TP: true positive, TN: true negative, FP: false positive, FN: false negative, All: TP + TN + FP + FN. m, number of bins in the reliability diagram; n, number of data points (observed or forecasted); f_i , forecast probability for the *i*th forecast; o_i the *i*th observed probability; and the average observed probability.