HHS Public Access

Author manuscript

*Epilepsia*. Author manuscript; available in PMC 2026 March 04.

Published in final edited form as:

*Epilepsia*. 2026 February ; 67(2): 753–761. doi:10.1111/epi.18677.

# Rigorous evaluation of five models for e-diary-only seizure forecasting–retrospective and prospective datasets do not outperform the Napkin method

**Chi-Yuan Chang, PhD**[1,2], **Robert Moss, BS**[3], **M. Brandon Westover, MD PhD**[1,2], **Daniel M. Goldenholz, MD, PhD**[1,2]

[1-]Harvard Medical School, Boston MA

[2-]Beth Israel Deaconess Medical Center, Boston, MA

[3-]Seizure Tracker LLC, Springfield, VA

## Abstract

**Objective:** Seizure forecasting using e-diaries may help patients with seizures to organize their daily life. Until now, most methods were not rigorously tested against a strict standard. This study aims to assess whether the performance of various models for seizure forecasting using e-diaries is better than the performance of a moving window average (aka, the Napkin method, due to simplicity of calculation).

**Methods:** We analyzed 3 cohorts from Seizure Tracker – a retrospective study and two prospective studies. E-diaries and the type of seizures were extracted from the datasets. We implemented 5 machine learning models (Perceptron, 1D-convolution, Multilayer Perceptron, Cycle, PPGLM) and compared their performance at seizure forecasting against the Napkin forecast. The models predicted the probability of having at least one seizure in the next 24-hour period based on a 90-day historical window. Model performances were evaluated by commonly used metrics (AUPRC, AUROC, and Brier Score). We considered a model to be clinically ineffective if it did not outperform the Napkin method across metrics and seizure frequencies.

**Results:** 5,501 retrospective patients (3,300 training, 1,100 validation, and 1,101 testing) and 36 prospective patients (21 from one cohort, 15 from the other) were included in the analysis. No model achieved significantly better performance than the Napkin method across metrics and frequencies.

**Significance:** Clinically effective seizure forecasting (i.e. beyond the Napkin method) for 24-hour risk using e-diaries alone may be infeasible with currently available techniques.

**Corresponding author**: Daniel M. Goldenholz, 330 Brookline Ave, Baker 5, Boston MA 02215, daniel.goldenholz@bidmc.harvard.edu.

Ethics Approval Statement

Three clinical cohorts were evaluated, all approved by BIDMC Institutional Review Board.

Ethical Publication Statement

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.

### Keywords

Seizure diary; forecasting; machine learning; biostatistics

## Introduction

Unpredictability of seizures can be very disabling. Seizure forecasts could warn patients to enact short-term preventive therapies. However, forecasting is unsolved. There are many proposed methods. They make use of e-diaries[1–6] and/or physiological signals (e.g. electroencephalogram (EEG)[7–15] or electrocardiogram[16,17]). It is unrealistic to record these signals without expensive equipment. On the other hand, e-diaries allow patients to record seizures unassisted with no cost using a mobile device or a website.

Previous studies focused on using e-diaries to forecast seizures[1,2,4,6]. However, those studies (with two exceptions[1,6]) typically evaluated models by comparing them to a permutation test (improvement over chance or IoC) instead of to a moving average (MA). Moreover, prior studies implicitly ignore the effect of seizure frequency (SF) on forecasting metrics. Our recent study[18] found that MA is a stronger clinical benchmark and SF dependency should be considered when comparing metrics to avoid mistakenly promoting ineffective models. Because it is simple enough that it can be calculated on the back of a napkin, we refer to MA forecasts as the Napkin method.

In contrast to IoC, the Napkin method is a causal and interpretable forecasting model applicable to real-world scenarios. The Napkin method is also among the simplest and most intuitive forecasting models, which patients may naturally use when guessing their seizure risk. However, as shown in our previous study[19], the method is not accurate enough for informed safety decisions. Therefore, the Napkin method isn't clinically effective.

This study aims to leverage e-diaries to compare 5 machine learning forecasters to the Napkin method: a Cycle model[2], point-process generalized linear model (PPGLM)[10], perceptron[20], multilayer perceptron (MLP)[21], and 1D convolutional neural network (Conv1D)[22]. The models we selected covered different structures and complexities. Any model considered clinically effective must surpass the efficacy of the Napkin method across metrics and seizure frequencies. We therefore refer to such models as "passing the napkin test".

## Materials and Methods

### Datasets and data preprocessing

Three clinical cohorts were evaluated, all approved by BIDMC Institutional Review Board. We received access to the data through a data use agreement with Seizure Tracker LLC, facilitated by the International Seizure Diary Consortium. Seizure Tracker[23] provided de-identified self-reported diaries. Seizure Tracker is a free tool available on web and mobile devices. It gives users the option to document their seizure date, time, and type. The purpose of the tool is to empower people living with epilepsy to play an active role in their own healthcare. The "retrospective" data comprised an export from Seizure Tracker

of any patient in their database that did not opt-out of research. Because the retrospective patients were evaluated post-hoc on user data, the completeness of these diaries is unknown. The “prospective” datasets came from two separate prospective remote data collection studies at BIDMC where patients were recruited explicitly for seizure forecasting. In both prospective studies, Seizure Tracker was used for documenting seizures; however, study staff independently checked in with patients to ensure complete diaries on a weekly basis. The first study had a 5-month duration, and the second had a 10-month duration. There were no overlapping patients between the two prospective studies. For safety reasons, all forecasting analysis was done after the conclusion of these two studies, meaning patients were not provided with “live” forecasts.

Patients in all datasets were screened to have a diary length longer than 90 days and an average SF in the range of 0.5 to 9.5 seizures per month. These limits were chosen based on the SF distribution from Seizure Tracker[23]. We used a 90-day historical window, during which most SFs would be empirically expected to be approximately steady[1]. Seizure types were also recorded in the datasets. We additionally categorized seizure types into different severity levels (Supplement Table 1). In the retrospective dataset, we split 5,501 patients into training (3,300 patients), validation (1,100 patients), and testing (1,101 patients) sets. In the prospective dataset, we assigned all patients to the testing set (36 patients). We transformed all diaries into binary arrays indicating if there was at least one seizure in a 24-hour period (00:00–24:00).

## Forecasting models

We implemented a series of models, based on the literature as well as commonly used tools for time-series analysis. In all cases, we attempted to use state-of-the-art model libraries to prevent errors and use industry standard default values wherever possible. With the exception of the Cycle model, all models used a 90-day lookback window as input for forecasting. Additional detail about each model was provided in the Supplement.

We implemented the Cycle model using unaltered code provided by the Karoly Lab. The model aims to find a combination of cycle periods which fit previous history best and uses this combination to calculate the likelihoods of having seizures in the future[2,11,17,24]. We trained Cycle models for each forecast day based only on seizure events that happened before the forecasted day. The temporal resolution of the Cycle model is hourly by default. To align the forecasting results with other models, we calculated the daily likelihoods by taking an average of hourly likelihoods within the same day. Unlike the other models, the Cycle model did not use a 90-day lookback, but rather used all available seizure data from the past in order to make forecasts for the next day, to be aligned with prior publications using this method[2,11,17,24].

The point-process generalized linear model (PPGLM) was described in Proix et al[10,25] for seizure forecasting. It assumes seizures to be a process that depends on a conditional Poisson distribution based on patient-specific time series.

Perceptron and multilayer perceptron (MLP, also known as deep learning networks) are two non-linear models commonly used in classification and regression. The MLP is a

multiple-layer fully connected extension of the basic perceptron concept. The MLP had 8 hidden layers.

To investigate the performance of a more complicated model, we trained a 1D convolutional network (Conv1D). There were two convolutional layers, followed by two fully connected layers for classification purposes.[26]

To enhance each model's sensitivity to SF, we trained a collection of models (also known as mixture of experts[26]) instead of a single model for each type of selected model. In training, we categorized each training diary into SF bins ranging from 1 seizure day/month to 9 seizure days/month with a 1-seizure day/month bin size. The SF was estimated using the entire diary in training. Next, we trained 9 SF-dependent models using the training diary in each SF bin respectively. In testing and validation, we estimated the SF of the incoming 90-day testing sample. This was then forecasted by the model whose SF was closest to the estimated SF. In other words, the SF for testing was always estimated prior to forecast and only used the 90-day history.

Models were tested using a simulated dataset to ensure that all metrics and testing code would perform as expected. See Supplement for details.

### Benchmark model and metrics of interest

Our previous study demonstrated that performance metrics are expected to have SF dependency and MA appears to be a stronger benchmark model compared to a permutation test when evaluating model performance[18]. Moreover, unlike permutation tests, MA is a causal forecasting model and can be simply calculated on the back of a napkin (we also call MA the Napkin method). We considered a model to have potential clinical utility only if it has better metrics than MA.

In this study, we selected MA as the benchmark model. MA was calculated using a 90-day lookback window. We measured the Brier score for evaluating calibration, and area under the receiver operating characteristic (AUROC) and area under precision recall curve (AUPRC) to evaluate discrimination[27–30]. The calibration curves of models are available in the Supplement. The cross-diary results were summarized by SF bins and reported as the median within each bin. The number of diaries in each bin was normalized by the total number of diaries for visualization purposes.

### Multivariate analysis of variance (MANOVA) and Permutation test

The difference between models and the benchmark model was evaluated by multivariate analysis of variance (MANOVA) on Brier, AUROC, and AUPRC. Each of the 9 metrics from the SF bins was treated as independent variables in MANOVA. There are 5 forecasting tools evaluated, with 1 MANOVA for each of 3 metrics (Brier, AUROC, AUPRC), resulting in 15 MANOVA results. To correct for multiple comparisons, we implemented false discovery rate (FDR) correction over the 15 MANOVA results.

Permutation tests are conventionally reported in previous studies[2,17,31]. We performed permutation tests for each model (without accounting for SF in the result) including

benchmark MA model. For each testing set diary, we performed 1000 permutations of each model's outcomes and evaluated if the model metrics were within the top 5% of the distribution of permuted metrics. We then calculated the percentage of the diaries which passed the top 5% threshold in the testing diaries.

**Seizure Severity Analysis:** To maximize the models' performance, we retrained our models using seizure severity information. We focused on predicting the most severe seizures (generalized tonic-clonic), as these pose a greater danger to patients. We designed three kinds of training input: all seizure events as our original method, severity levels, and only severe events. The same testing as above was used; however, Cycle was excluded because it was not developed for inputs other than seizure timing. Further methodological details for the severity analysis are available in the Supplement.

## Results

The retrospective data comprised 5,501 patients with ages ranging from 0 to 91. The prospective datasets (prospective study 1=24 patients, study 2=12 patients, total 36 patients) had ages 6 to 77. See Supplement for further detail.

After training and validating our models on the retrospective dataset, we tested the models on 1,101 held-out patients in retrospective dataset and 36 patients in prospective dataset and calculated our metrics of interest. Training loss, calibration curves, and summary plots are available in the Supplement.

Fig. 1 shows the comparison between Brier scores, AUROC, and AUPRC of different models across SF. There are more patients at SF=1 and 2, indicated by the sizes of the markers. In both retrospective and prospective datasets, Brier scores and AUPRC showed SF dependencies but not AUROC, as reported in our previous study[18]. We observed the differences of Brier scores, AUROC, and AUPRC between models and MA were small. Also, there was substantial overlap among the 95% confidence intervals (CIs) of those metrics. The 95% confidence intervals were wider in the prospective dataset due to the small number of patients. There was only one patient in the SF=6 bin in the prospective dataset, resulting in no CI available at SF=6. Moreover, there was no patient in the SF=7 bin in the prospective dataset. The AUROCs of all models were slightly above 0.5 across SF in retrospective dataset. However, all models, except for the Cycle model had an AUROC lower than 0.5 at SF=2 in the prospective dataset. In both retrospective and prospective datasets, AUPRC was lower than 0.5.

Table 1 shows the results of MANOVA against MA. In the MANOVA results, no model was significantly different from MA, except the Cycle model in the retrospective dataset. Visual inspection in Fig. 1 reveals that the Cycle model performed worse than the MA across all metrics, thus explaining why MANOVA found a difference. At higher frequencies Cycle and PPGLM dramatically underperform on Brier Score retrospectively (higher scores are undesirable for Brier). At low frequencies, MLP and PPGLM beat the MA only on AUPRC retrospectively, however they did not beat MA on AUROC nor Brier. In Table 2, the permutation tests show all models, including MA, exceeded the 5% threshold for all

metrics to more than 20% of the patients in the retrospective dataset. On the other hand, in prospective datasets, all models, except Cycle, exceeded the 5% threshold for Brier scores on all patients while only around 10% of patients exceeded the 5% threshold for AUROC and AUPRC.

All models, including MA, had a very low Brier score due to the low probability of having seizures at typical seizure rates. Hence, the Brier scores' difference between models and MA were small especially when SF was low. The only two models that had an observable Brier score difference from MA were the Cycle model and PPGLM (Fig. 1 and Table 1). Cycle model had higher Brier scores than MA because Cycle uses sinusoidal functions to mimic seizure period. Thus, Cycle gives higher likelihoods for those non-seizure days near the day when sinusoidal functions reach maximum (suggesting poor calibration). The Brier score difference between Cycle and MA increases as SF increases because the periods of Cycle's sinusoidal functions decrease, resulting in having more high-likelihood non-seizure days. On the other hand, we observed that PPGLM appeared overconfident and pushed most of the estimated probabilities toward 0 and 1, resulting in underestimation when the estimated probabilities were low and overestimation when the estimated probabilities were high (Supplement Fig. 3).

Using all seizures to forecast only severe seizures resulted in no model outperforming MA across the 3 metrics. When using all seizures in combination with severity scores to predict severe seizures, no model outperformed MA across the three metrics. When using only severe seizures to predict severe seizures, again no model outperformed MA across metrics (see Supplement).

## Discussion

Our results show that there is no model that has overall better metrics than MA using our rigorous evaluation approach.

The AUROC of all models were slightly above 0.5 in the retrospective dataset, indicating the models have very limited prediction capabilities. It is worth noticing that Cycle has AUROC worse than MA at most of the SF bins as shown in Fig. 1 and Table 1. This result indicates that the pattern of seizures may be more complicated than periodic functions (the primary assumption of Cycle). Conversely, the result may reflect other factors such as averaging hourly forecasts from Cycle, potentially diluting underlying cyclical patterns. In the prospective dataset, the AUROC below 0.5 of PPGLM, Perceptron, MLP, and Conv1D indicate that the seizure patterns across cohorts are different. Our findings show that seizure forecasting using e-diaries only is not easily transferable across cohorts of e-diary users. It is important to note that the prospective cohorts had physician curation and weekly supervision, whereas the retrospective cohort had neither. Thus, one might suspect the retrospective data had higher levels of noise.

Our benchmark requirement that a model must overcome the Napkin method on 3 metrics across SF is based on our prior study[18] showing that without this consistent result, one may erroneously expect clinically helpful results from underperforming models. Using Brier

score addresses calibration, and AUROC/AUPRC addresses discrimination. If a model is successful with either calibration or discrimination alone, it will still have problems in deployment.

The Napkin method (a.k.a. MA), Perceptron, 1D-Convolution, Multilayer Perceptron and PPGLM used a 90-day history window. The Cycle method, consistent with prior work, used all available history from the diary. It is perhaps possible that other durations of history would be more suitable (e.g. 180 days, 360, etc.). Our choice of 90 days reflects the tradeoff between greater historical knowledge, the challenge of expecting longer accurate records, and the unknown possibility of brain networks shifting over time. It is worthwhile to reflect that prior studies find that 3 months is a very fruitful choice across modalities[32], and that the median seizure rate in a very large population of outpatients was 2.7 seizures/month[23], which means that roughly 50% of such a population would have 8 or more seizures recorded in 3 months. Nevertheless, future studies may discover robust reasons for unique window sizes in specific cases.

The permutation test results (Table 2) show that all models had a fraction of patients that are improved over chance. However, considering SF dependency and all aspects of metrics, none of the models meet our rigorous standard for clinical utility. A less demanding evaluation method may promote inaccurate or even harmful forecasting models. For example, if one trains a forecasting model without considering SF dependency, the model might perform well only around the majority SF of the patients tested. This would result in overestimation for SFs lower than the majority and underestimation for SFs higher than the majority.

We expected better performance when accounting for severity levels. However, no models performed better than MA in all three training scenarios. Moreover, incorporating severity levels into models did not enhance model performance compared to the original method. This result suggests that the information provided by adding severity levels is limited for the seizure forecasting task.

In this study, we explored 5 different flexible forecasting models as well as several input forms (diaries with and without different severity markers). None of these techniques pass “the napkin test” for clinical utility, meaning none of them exceed the performance of MA (the “Napkin method”). These results imply that e-diaries do not provide enough information for clinically effective seizure forecasting. There are many potential covariates associated with seizures which are not included in our model, e.g. medication adherence[33], external stimuli[34], and sleep patterns[35]. Also, auxiliary physiologic inputs may be more directly associated with underlying seizure mechanisms[5,10,16,17,31,36].

A limitation of the present study is that we cannot evaluate all possible models. However, the machine learning models we selected encompassed a wide range of architectures and varied in their number of trainable parameters. When considering deep learning models such as MLP and Conv1D, it is possible that the retrospective dataset did not include a sample size sufficiently large to fully take advantage of those flexible models. However, the retrospective dataset we used is one of the largest seizure diary datasets available in the world. Thus,

until new datasets become available, this may be the best chance possible to explore such models. Moreover, as shown the Supplement, there were no signs of overfitting with the deep learning models, suggesting that the sample size from the retrospective dataset may have been acceptable.

Another limitation of this study is the perceived reliability of self-reported seizure diaries. Under-reported or over-reported self-reported data in the retrospective dataset and the limited amount of prospective data may cause the poor performance of all models. A previous study of 23 patients showed that only 26% of the patients with epilepsy were always aware of their seizures[37]. Another study of 76 patients found that 57% of generalized or focal seizures seem to be not recognized by witnesses within proximity of the patient[38]. It is unknown how much over-reporting truly exists, but under-reporting has been estimated to vary widely from patient to patient[39]. However, numerous studies find that results found in self-reported seizure diaries match findings in other validated forms of data[40–49]. Another concern is the small cohort of prospectively collected data. It is possible that larger prospective datasets may reveal new findings, however it is encouraging that a highly curated prospective dataset matches the findings in the uncontrolled retrospective data.

It is important to recognize that the current approach[18] has not yet been applied to other forms of seizure forecasting, including scalp EEG[50], intracranial EEG[51], or other physiologic signals[36,52]. It remains unknown if adding e-diaries to those additional measures would facilitate forecasting that would overcome the limitations seen here. The typical benchmark that has been tested is permutation testing (also known as chance forecasting), and as seen here, methods can deceptively look promising if compared to permutation. We encourage other research groups to adopt our more rigorous benchmark[18], which is much harder to overcome, but also has a clinically practical interpretation.

## Conclusion

In this study, we compared the performance of various types of machine learning models against MA for seizure forecasting tasks, considering SF dependency. None of the models outperformed the Napkin method, a standard clinical benchmark. We suggest that achieving clinically effective seizure forecasting within the next 24 hours using e-diaries alone is infeasible. Because all the input data derives from Seizure Tracker, there remains a possibility that alternative e-diary systems could overcome present obstacles, though we suspect that such systems would require additional data from the patient. Further study is needed with other systems, and other forms of physiologic inputs. It remains our hope that clinically effective forecasting will be discovered soon.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

RM is a co-founder and owner of Seizure Tracker LLC. Seizure diaries from Seizure Tracker were provided via the International Seizure Diary Consortium. We would like to thank Dr. Philippa Karoly and Dr. Rachel Stirling for generously assisting us with the Cycle methodology and source code.

**Conflicts of Interest Disclosure:**

Dr. Chang has no conflict. Mr. Moss is the cofounder and owner of Seizure Tracker, LLC, and has received personal fees from Courtagen Life Sciences, Engage Therapeutics, Epitel, LivaNova, Marinus Pharmaceuticals, Neurelis, Neuropace, UCB, and grants from the Tuberous Sclerosis Complex Alliance. Seizure Tracker was paid for the effort to participate in this project via NIH funding. Dr. Goldenholz is an unpaid advisor for Epilepsy AI and Eysz. He has been provided speaker fees from AAN, AES, ACNS, and NNS. He also previously has been a paid consultant for Neuro Event Labs, IDR, LivaNova and Health Advances. Dr. Westover is a co-founder, scientific advisor, and consultant to Beacon Biosignals and has a personal equity interest in the company. He also receives royalties for authoring Pocket Neurology from Wolters Kluwer and Atlas of Intensive Care Quantitative EEG by Demos Medical.

**Funding:**

DG and CC were supported by NINDS K23NS124656. MB was supported by grants from the NIH (RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119), and NSF (2014431).

## Data Availability

The retrospective data that support the findings of this study are available from Seizure Tracker LLC. Restrictions apply to the availability of these data, which were used under license for this study. The prospective data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

## References

1. Goldenholz DM, Goldenholz SR, Romero J, Moss R, Sun H, Westover B. Development and Validation of Forecasting Next Reported Seizure Using e-Diaries. Ann Neurol. 2020; 88(3):588–95. [PubMed: 32567720]
2. Karoly PJ, Cook MJ, Maturana M, Nurse ES, Payne D, Brinkmann BH, et al. Forecasting cycles of seizure likelihood. Epilepsia. 2020; 61(4):776–86. [PubMed: 32219856]
3. Privitera M, Haut SR, Lipton RB, McGinley JS, Cornes S. Seizure self-prediction in a randomized controlled trial of stress management. Neurology. 2019; 93(22):E2021–31. [PubMed: 31645468]
4. Gleichgerrcht E, Dumitru M, Hartmann DA, Munsell BC, Kuzniecky R, Bonilha L, et al. Seizure forecasting using machine learning models trained by seizure diaries. Physiol Meas. 2022; 43(12).
5. Brinkmann BH, Karoly PJ, Nurse ES, Dumanis SB, Nasseri M, Viana PF, et al. Seizure Diaries and Forecasting With Wearables: Epilepsy Monitoring Outside the Clinic. Vol. 12, Frontiers in Neurology. Frontiers Media S.A.; 2021.
6. Goldenholz DM, Eccleston C, Moss R, Westover MB. Prospective validation of a seizure diary forecasting falls short. Epilepsia. 2024; 65(6):1730–6. [PubMed: 38606580]
7. Cook MJ, O'Brien TJ, Berkovic SF, Murphy M, Morokoff A, Fabinyi G, 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. [PubMed: 23642342]
8. Cousyn L, Messaoud R Ben, Lehongre K, Frazzini V, Lambrecq V, Adam C, et al. Daily resting-state intracranial EEG connectivity for seizure risk forecasts. Epilepsia. 2023; 64(2):e23–9. [PubMed: 36481871]
9. Stirling RE, Maturana MI, Karoly PJ, Nurse ES, McCutcheon K, Grayden DB, et al. Seizure Forecasting Using a Novel Sub-Scalp Ultra-Long Term EEG Monitoring System. Front Neurol. 2021; 12.
10. Proix T, Truccolo W, Leguia MG, Tcheng TK, King-Stephens D, Rao VR, et al. Forecasting seizure risk in adults with focal epilepsy: a development and validation study. Lancet Neurol. 2021; 20(2):127–35. [PubMed: 33341149]
11. Baud MO, Kleen JK, Mirro EA, Andrechak JC, King-Stephens D, Chang EF, et al. Multi-day rhythms modulate seizure risk in epilepsy. Nat Commun. 2018; 9(1).

12. Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. Brain. 2007; 130(Pt 2):314–33. [PubMed: 17008335]
13. Mormann F, Andrzejak RG. Seizure prediction: making mileage on the long and winding road. Brain. 2016; 139(Pt 6):1625–7. [PubMed: 27234060]
14. Nejedly P, Kremen V, Sladky V, Nasseri M, Guragain H, Klimes P, et al. Deep-learning for seizure forecasting in canines with epilepsy. J Neural Eng. 2019; 16(3).
15. Brinkmann BH, Wagenaar J, Abbot D, Adkins P, Bosshard SC, Chen M, et al. Crowdsourcing reproducible seizure forecasting in human and canine epilepsy. Brain. 2016; 139(6).
16. Cousyn L, Dono F, Navarro V, Chavez M. Can heart rate variability identify a high-risk state of upcoming seizure? Epilepsy Res. 2023; 197.
17. Xiong W, Stirling RE, Payne DE, Nurse ES, Kameneva T, Cook MJ, et al. Forecasting seizure likelihood from cycles of self-reported events and heart rate: a prospective pilot study. EBioMedicine. 2023; 93:104656. [PubMed: 37331164]
18. Chang C-Y, Zhang B, Moss R, Picard R, Westover MB, Goldenholz D. Necessary for seizure forecasting outcome metrics: Seizure frequency and benchmark model. Epilepsy Res. 2024; 208:107474. [PubMed: 39522392]
19. Chang C-Y, Zhang B, Moss R, Picard R, Westover MB, Goldenholz D. Necessary for seizure forecasting outcome metrics: Seizure frequency and benchmark model. Epilepsy Res. 2024; 208:107474. [PubMed: 39522392]
20. Rosenblatt F The perceptron: A probabilistic model for information storage and organization in the brain. Psychol Rev. 1958; 65(6):386–408. [PubMed: 13602029]
21. Svozil D, Kvasnicka V, Pospichal J. Introduction to multi-layer feed-forward neural networks. Chemometrics and Intelligent Laboratory Systems. 1997; 39(1):43–62.
22. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: A survey. Mech Syst Signal Process. 2021; 151:107398.
23. Ferastraoaru V, Goldenholz DM, Chiang S, Moss R, Theodore WH, Haut SR. Characteristics of large patient-reported outcomes: Where can one million seizures get us? Epilepsia Open [Internet]. 2018; 3(3):364–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30187007 [PubMed: 30187007]
24. Karoly PJ, Goldenholz DM, Freestone DR, Moss RE, Grayden DB, Theodore WH, et al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. Lancet Neurol. 2018; 17(11):977–85. [PubMed: 30219655]
25. Truccolo W, Eden UT, Fellows MR, Donoghue JP, Brown EN. A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. J Neurophysiol. 2005; 93(2).
26. Yuksel SE, Wilson JN, Gader PD. Twenty years of mixture of experts. IEEE Trans Neural Netw Learn Syst. 2012; 23:1177–93. [PubMed: 24807516]
27. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950; 78(1):1–3.
28. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143(1):29–36. [PubMed: 7063747]
29. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning - ICML '05. New York, New York, USA: ACM Press; 2005. p. 625–32.
30. Boyd K, Eng KH, Page CD. Area under the Precision-Recall Curve: Point Estimates and Confidence Intervals. In 2013. p. 451–66.
31. Leguia MG, Rao VR, Tcheng TK, Duun-Henriksen J, Kjær TW, Proix T, et al. Learning to generalize seizure forecasts. Epilepsia. 2023; 64(S4).
32. Goldenholz DM, Goldenholz SR, Romero J, Moss R, Sun H, Westover B. Development and Validation of Forecasting Next Reported Seizure Using e-Diaries. Ann Neurol [Internet]. 2020; 88(3):588–95. Available from: http://www.ncbi.nlm.nih.gov/pubmed/32567720 [PubMed: 32567720]

33. Manjunath R, Davis KL, Candrilli SD, Ettinger AB. Association of antiepileptic drug nonadherence with risk of seizures in adults with epilepsy. Epilepsy and Behavior. 2009; 14(2):372–8. [PubMed: 19126436]
34. Beghi E, Carpio A, Forsgren L, Hesdorffer DC, Malmgren K, Sander JW, et al. Recommendation for a definition of acute symptomatic seizure. Epilepsia. 2010; 51(4):671–5. [PubMed: 19732133]
35. Derry CP, Duncan S. Sleep and epilepsy. Vol. 26, Epilepsy and Behavior. 2013. p. 394–404. [PubMed: 23465654]
36. Nasseri M, Pal Attia T, Joseph B, Gregg NM, Nurse ES, Viana PF, et al. Ambulatory seizure forecasting with a wrist-worn device using long-short term memory deep learning. Sci Rep. 2021; 11(1).
37. Blum DE, Eskola J, Bortz JJ, Fisher RS. Patient awareness of seizures. Neurology. 1996; 47(1).
38. Moraes J, Cook M, Nurse E. The silent witness: The unseen gaps in eyewitness recognition of seizures. Epilepsia [Internet]. 2025;. Available from: http://www.ncbi.nlm.nih.gov/pubmed/40492933
39. Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and seizure detection. Lancet Neurol. 2018; 17(3):279–88. [PubMed: 29452687]
40. Goldenholz DM, Goldenholz SR, Moss R, French J, Lowenstein D, Kuzniecky R, et al. Is seizure frequency variance a predictable quantity? Ann Clin Transl Neurol. 2018; 5(2):201–7. [PubMed: 29468180]
41. Ferastraoaru V, Goldenholz DM, Chiang S, Moss R, Theodore WH, Haut SR. Characteristics of large patient-reported outcomes: Where can one million seizures get us? Epilepsia Open. 2018; 3(3):364–73. [PubMed: 30187007]
42. Karoly PJ, Goldenholz DM, Freestone DR, Moss RE, Grayden DB, Theodore WH, et al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort study. Lancet Neurol. 2018; 17(11):977–85. [PubMed: 30219655]
43. Karoly PJ, Romero J, Cook MJ, Freestone DR, Goldenholz DM. When can we trust responders? Serious concerns when using 50% response rate to assess clinical trials. Epilepsia. 2019; 60(9):e99–103. [PubMed: 31471901]
44. Oliveira A, Romero JM, Goldenholz DM. Comparing the efficacy, exposure, and cost of clinical trial analysis methods. Epilepsia. 2019; 60(12):e128–32. [PubMed: 31724165]
45. LaGrant B, Goldenholz DM, Braun M, Moss RE, Grinspan ZM. Patterns of Recording Epileptic Spasms in an Electronic Seizure Diary Compared With Video-EEG and Historical Cohorts. Pediatr Neurol. 2021; 122:27–34. [PubMed: 34293636]
46. Goldenholz DM, Westover MB. Flexible realistic simulation of seizure occurrence recapitulating statistical properties of seizure diaries. Epilepsia. 2023; 64(2):396–405. [PubMed: 36401798]
47. Goldenholz DM, Goldenholz EB, Kaptchuk TJ. Quantifying and controlling the impact of regression to the mean on randomized controlled trials in epilepsy. Epilepsia. 2023; 64(10):2635–43. [PubMed: 37505116]
48. Goldenholz D, Brinkmann BH, Westover MB. How accurate do self-reported seizures need to be for effective medication management in epilepsy? Epilepsia. 2024; 65(7):e104–12. [PubMed: 38776216]
49. Zhang B, Chen WV, Regalia G, Goldenholz DM, Picard RW. Statistical characteristics of large-scale objective tonic–clonic seizure records from medical smartwatches used in daily life. Epilepsia. 2024; 65(11):3255–64. [PubMed: 39287615]
50. Ozcan AR, Erturk S. Seizure Prediction in Scalp EEG Using 3D Convolutional Neural Networks with an Image-Based Approach. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2019; 27(11):2284–93. [PubMed: 31562096]
51. Truong ND, Nguyen AD, Kuhlmann L, Bonyadi MR, Yang J, Ippolito S, et al. Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Networks. 2018; 105:104–11. [PubMed: 29793128]
52. Ding TY, Gagliano L, Jahani A, Toffa DH, Nguyen DK, Bou Assi E. Epileptic seizure forecasting with wearable-based nocturnal sleep features. Epilepsia Open. 2024; 9(5):1793–805. [PubMed: 38980984]

**Key Points**

- Five machine learning forecasts were compared to moving average ("Napkin method") across seizure frequencies and metrics for discrimination and calibration.
- None of the models, including Cycle, outperformed the Napkin method.
- Incorporating seizure severity into models did not improve forecasts over the Napkin method.
- Seizure forecasting with e-diaries alone was not clinically effective.

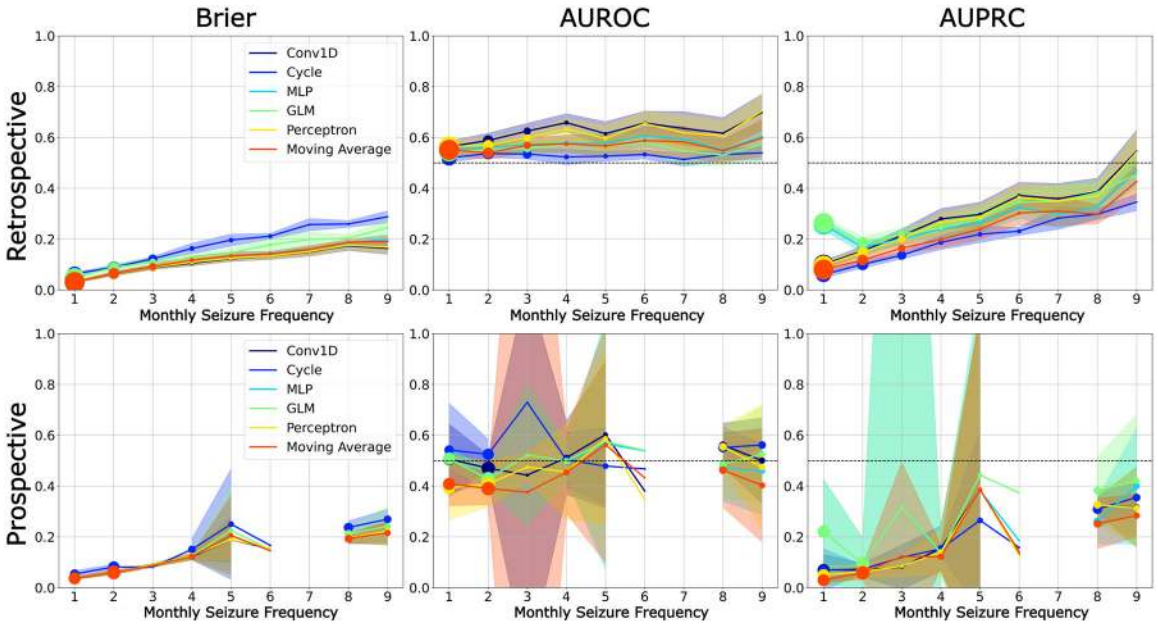


**Figure 1.**

Comparison of Brier score, AUROC, and AUPRC across models in retrospective dataset and prospective dataset. The solid line indicates the median across diaries within the SF bin. The shaded area indicates the 95% CI of the metrics. The sizes of the markers indicate the percentage of patients in the SF bin, showing that some bins have larger or smaller representative samples. Of note, unlike AUROC and AUPRC, in Brier scores, lower values indicate better performance. In all cases, there is no forecasting method that is consistently better than MA.

**Table 1.**

The MANOVA significance against the benchmark model (Moving Average) for each metric. The results of retrospective and prospective datasets are shown as Retro and Pro respectively. The p-values shown were corrected by FDR correction. The p-values which were considered significant are bolded. Overall requires statistical significance for all 3 metrics. Improved requires the forecasts to be better than MA visually.

| | Dataset | Brier | AUROC | AUPRC | Overall significance | Improvement over MA |
|---|---|---|---|---|---|---|
| **Cycle** | Retro | **0.000** | **0.000** | **0.000** | **True** | No |
| | Pro | 0.102 | 0.099 | 1.000 | False | No |
| **PPGLM** | Retro | **0.000** | 1.000 | **0.000** | False | No |
| | Pro | 1.000 | 0.581 | 0.484 | False | No |
| **Perceptron** | Retro | 0.447 | **0.001** | **0.000** | False | No |
| | Pro | 1.000 | 1.000 | 1.000 | False | No |
| **MLP** | Retro | 1.000 | 1.000 | **0.000** | False | No |
| | Pro | 1.000 | 0.653 | 1.000 | False | No |
| **Conv1D** | Retro | 0.125 | **0.000** | **0.000** | False | No |
| | Pro | 1.000 | 0.531 | 1.000 | False | No |

**Table 2.**

The percentage of patients whose metrics were significantly better than permutation tests. The results of retrospective and prospective datasets are shown as Retro and Pro respectively. Nearly all these fractions are greater than 5% which is the expected value of chance level.

| | Dataset | % of better Brier | % of better AUROC | % of better AUPRC |
|---|---|---|---|---|
| **Cycle** | Retro | 24.8 | 26.9 | 24.2 |
| | Pro | 48.2 | 13.8 | 10.3 |
| **PPGLM** | Retro | 86.6 | 46.9 | 38.1 |
| | Pro | 100 | 12.5 | 6.3 |
| **Perceptron** | Retro | 94.9 | 50.6 | 49.3 |
| | Pro | 100 | 11.1 | 13.9 |
| **MLP** | Retro | 89.1 | 51.7 | 42.5 |
| | Pro | 100 | 9.4 | 3.1 |
| **Conv1D** | Retro | 90.5 | 52.7 | 50.0 |
| | Pro | 100 | 13.9 | 13.9 |
| **Moving Average** | Retro | 86.1 | 44.4 | 36.6 |
| | Pro | 100 | 11.1 | 8.3 |