HHS Public Access

Author manuscript

Epilepsy Res. Author manuscript; available in PMC 2025 December 01.

Published in final edited form as:

Epilepsy Res. 2024 December ; 208: 107474. doi:10.1016/j.eplepsyres.2024.107474.

Necessary for seizure forecasting outcome metrics: seizure frequency and benchmark model

Chi-Yuan Chang, PhD[1,2], **Boyu Zhang, MS**[3,4,6], **Robert Moss, BS**[5], **Rosalind Picard, ScD**[3,4], **M. Brandon Westover, MD PhD**[1,2], **Daniel Goldenholz, MD, PhD**[1,2]

[1]-Harvard Medical School, Boston MA

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

[3]-Massachusetts Institute of Technology, Cambridge, MA

[4]-Empatica USA, Cambridge, MA

[5]-Seizure Tracker LLC, Springfield, VA

[6]-Brigham and Women's Hospital, Boston, MA

Abstract

Background: This study aims to illustrate the connection between seizure frequency (SF) and performance metrics in seizure forecasting, and to compare the effectiveness of a moving average (MA) model versus the commonly used permutation benchmark.

Methods: Metrics of calibration and discrimination were computed for each dataset, comparing MA and permutation performance across SF values. Three datasets were used: (1) self-reported seizure diaries from 3,994 Seizure Tracker patients, (2) automatically detected and sometimes manually reported or edited generalized tonic-clonic seizures from 2,350 Empatica Embrace 2 and Mate App users, and (3) simulated datasets with varying SFs.

Results: Most metrics were found to depend on SF. The MA model outperformed or matched the permutation model in all cases. These more advanced metrics show that comparison to permutation will falsely elevate poor forecasting models.

Conclusions: The findings highlight SF's role in seizure forecasting accuracy and the MA model's suitability as a benchmark. This study underscores the need for considering patient SF in forecasting studies and suggests the MA model may provide a better standard for evaluating future seizure forecasting models.

Keywords

forecasting; bioinformatics; epilepsy; seizure; statistics

Corresponding author: Daniel Goldenholz, 330 Brookline Ave, Baker 5, Boston MA 02215.

Introduction

Many studies attempt to forecast seizures.[1-12] However, patients' seizure frequency (SF) is usually ignored when reporting the model performance. It has been observed across studies that patients have vastly different SFs.[13] It is possible that model performance metrics calculated over a cohort might be influenced by SF and, thus, confound the evaluation of model performance.

Benchmark model selection is another important consideration. Often model performances are compared against a permutation testing.[1,14] Using permutation testing to assess a forecasting model is a very low bar to overcome, and probably does not have any clinical significance (see Appendix). Conversely, a moving average model ("what happened before is likely to happen again') may be a better litmus test for a successful forecasting tool. In our prior work[15], we found that a deep-learning-based forecasting model which surpasses the permutation testing does not perform better than a moving average model.

Recently, more forecasting studies use a probabilistic rather than a deterministic approach[2,8,12,15,16].Probability based models (like modern weather forecasts) can be used as-is, or a threshold cutoff can be applied to artificially create a binary "yes/no" output for users who choose to view simple outputs. In this study, we focused our attention on evaluating the probabilities given by forecasting models.

We hypothesized that (1) there is a SF dependence that affects the performance of some forecasting metrics, and (2) using a moving average model is a better benchmark model compared to permutation testing. This study aims to explore these two hypotheses with simulation and real-world data.

Materials and Methods

Datasets and data preprocessing

Simulated dataset—We produced a structured simulation of 9 seizure diaries with 9 different seizure frequencies respectively. Each diary was a 10000-days-long binary array where 0 indicates there is no seizures and 1 indicates there is at least 1 seizure in that day. The monthly SF is determined by the number of seizure days in a month ranging from 1 seizure day to 9 seizure days per month. Of note, most patients from both clinical datasets had SF values within 1-9/month. All the seizure days occurred consecutively at the beginning of each month (Appendix). This organization of when seizures occurred was arbitrary – the key was that each diary had a prespecified number of seizures per month. Our simulated dataset also served as a perfect recording which contained no biases from under-/over-reporting of seizure events.

Clinical datasets—Two clinical datasets were evaluated, both approved by BIDMC IRB with Exempt status. We received access to the e-diary data through a data use agreement with Seizure Tracker LLC, facilitated by the International Seizure Diary Consortium. Seizure Tracker[13] provided de-identified self-reported diaries. We selected patients based on the recording period and the length of diary (Appendix).

Another dataset was recorded by Empatica's FDA-cleared Embrace 2, a wearable device for generalized tonic-clonic seizure (GTCS) detection.[9] De-identified wearable-derived seizure diaries were provided by Empatica for the purposes of this statistical analysis. We selected patients based on recording duration and on the reliability of their e-diary interactions (Appendix).

The diaries in both datasets were further transformed into binary diaries, i.e. whether there is at least one seizure within that day.

Metrics of interest

We focused on 4 commonly used metrics for evaluating probabilistic forecasting tools: two for calibration (Brier Score and calibration curve) and two for discrimination (area under curve of receiver-operating characteristics (AUROC), and area under curve of precision-recall curve (AUPRC)).[1,2,4-8,14,17,18] There are some other commonly used metrics[19,20] designed to evaluate binary forecasting tools (e.g. tools that only provide a yes/no answer without a probability) not evaluated in our study. Any metric that might have an SF dependence (e.g. time-in-warning) would need to account for it.

A desirable metric profile for forecasting tools is: having a low Brier Score (closer to 0 is better), a high AUROC and a high AUPRC (closer to 1 is better for both). The calibration curve should be as close to the diagonal line as possible.

To summarize results across diaries, we categorized each diary into SF bins ranging from 1 seizure day/month to 9 seizure days/month with a 1-seizure day/month bin size and reported the average within each bin. For results on extremely high and low SF, please see Appendix. The number of diaries in each bin was normalized by the total number of diaries for visualization purposes.

Benchmark model: Moving average model vs. permutation testing

Moving average model (MA) is a simple causal forecasting model. It predicts the probability of having seizure events by calculating the rate of seizure-present intervals (here, 24-hour intervals) in the diary history using a lookback window. In this study, we used a 90-day window, during which most SFs would be empirically expected to be steady.[8,13] Since MA is intuitive and requires minimal computation (could even be computed manually by a patient/caregiver), we consider MA a candidate benchmark model.

Permutation testing is a widely used benchmark in forecasting tasks.[1,10,14] It permutes the model forecasts and calculates the metrics of interest. This process is then repeated (e.g., 1,000 times). The average metric across all permutations is typically reported.

Improvement over chance (IOC) is another way to quantify a model performance, as shown in Eq. (1).

$$IOC(model) = mean(Brier(permuted\ \ model)) - Brier(model) \quad (1)$$

It can be shown that IOC for a perfectly accurate forecasting model (“truth”) is maximal (Appendix). Therefore, we consider the average result of permutations of truth, denoted permuted truth, as another candidate benchmark test, because it would provide the largest possible IOC for a given SF.

One prominent model for seizure forecasting accounts for multidien cycles[4,21]. For parsimony, we focused on the simpler MA model. (See Appendix)

Results

After preprocessing, there were 3,994 patients from Seizure Tracker with diary durations of 91-5,337 (median 525) days, and 2,350 patients from Empatica with diary durations of 90-1,551 (median 280) days.

Figure 1 shows the results of comparing the MA and permuted truth, across the four metrics for each of the three datasets. Seizure Tracker and Empatica have more diaries with low SF, as expected [13,14]. For all three datasets, MA always has smaller Brier Scores and equal or higher AUPRC than the permuted truth across SF. Though all AUROC values fluctuate around 0.5-0.6 across SF, MA never has an AUROC smaller than the permuted truth across SF. The calibration curves of MA show slight overestimates in probability for low SF but improve as SF increases, indicating by aligning with the diagonal line. On the other hand, all the calibration curves of the permutation truth are nearly horizontal.

Among the metrics of interest, Brier Score, calibration curve, and AUPRC show SF dependencies. The calibration curve and AUPRC improve as SF increases while Brier Score decreases as SF increases. AUROC shows no dependency on SF.

Discussion

There are two main findings in our study. First, MA appears to be a better benchmark compared to permutation testing. Second, three of the metrics, the calibration curve, Brier score and AUPRC, show a dependence on SF while AUROC appears to be relatively SF-independent.

It is necessary to report individual patient SF with these metrics when comparing model performances across different studies. Bins of seizure frequencies can be used if narrowly defined (as done here). When comparing models evaluated on datasets with different SF ranges, we suggest imputing metric performance for a common SF range and including SF independent metrics, such as AUROC (Appendix). Critically, some SF values may not be very important to forecast (e.g., daily risk in patients who have a seizure per 2-days, or daily risk in patients with yearly seizures, etc.).

When comparing the performance of MA and permuted truth, we found MA always performs the same or better than permuted truth. Additionally, MA is preferable as a benchmark because it (1) is causal (i.e. does not require knowledge of the future), (2) is easily computed (“back of the envelope calculation”), and (3) is interpretable (“the previous seizure rate will recur”). These three characteristics are also a good standard for

any real-world seizure forecasting tool. Conversely, permutation is noncausal (knowledge of the future is required) and requires more computational resources. Anecdotally, our investigations have found MA to be surprisingly accurate in multiple seizure forecasting contexts and therefore a more challenging benchmark to overcome for a candidate model. It is important to emphasize that if a method was better than chance yet worse than MA, there would be no clinical utility for it.

We elected to focus on the parsimonious computation of MA for this study. To date, it has not been shown that the prominent cycles model[4,22] outperforms MA. (See Appendix)

Based on our findings, we encourage investigators to consider the impact of SF on metrics and replace permutation test by MA as a benchmark model when evaluating their forecasting tools. This rigorous evaluation method can avoid errors caused by SF variance. It is also more practical to compare a causal forecasting tool with a causal benchmark aligned with realistic assumptions.

One thing to emphasize is that the potential SF dependence seen across metrics would be expected for any candidate forecasting model. This can be understood from a simple example: low SF patients would have good metrics for a model that often undercalls seizures and high SF patients would have poor metrics for the same model.

Data from self-report or devices can suffer from varying degrees of under-reporting or noise. Because the SF dependence issue is fundamental to the performance metrics themselves, we assert that the conclusions here transcend specific data gathering techniques. To confirm this, we presented perfectreporting noise-free synthetic data as well as two independent real-world datasets.

All datasets come with biases. We selected 3 distinct forms of data. The encouraging finding is convergence of results. While we recognize that only certain types of patients would use Seizure Tracker or Empatica, we believe the principles we explored here have a firm mathematical foundation, and are being illustrated by the imperfect datasets presents as real examples of that foundation at work.

There were some differences between the simulation and the clinical datasets. Many of these differences reflect the simplistic assumptions used for the simulation, as well as methodological choices made for our study (Appendix).

In this study, we illustrated our ideas a 90-day historical window and 24-hour forecasts. Other choices for both time windows would not change our two key recommendations (see Appendix).

The emphasis of this paper is identifying mathematical guideposts for testing algorithms. In contrast, the *value* of seizure forecasting tools is beyond the scope of this study; patient attitudes, beliefs, desires, and behaviors all need to be accounted for prior to deploying a forecasting tool.

In summary, this study provides insight into the importance of including patients' seizure frequency in seizure forecasting tasks and demonstrates that MA represents a valuable benchmark with minimal computational complexity.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

FUNDING:

DG and CC are supported by NINDS K23NS124656.

BZ is supported by T32 HL007901-25.

Dr. Westover was supported by grants from the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312, RF1NS120947, R01AG073410, R01HL161253, R01NS126282, R01AG073598), and NSF (2014431).

Bibliography

1. Leguia MG, Rao VR, Tcheng TK, Duun-Henriksen J, Kjær TW, Proix T, et al. Learning to generalize seizure forecasts. Epilepsia. 2023; 64(S4).
2. 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]
3. 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.
4. 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]
5. 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.
6. 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).
7. Onorati F, Regalia G, Caborni C, LaFrance WC, Blum AS, Bidwell J, et al. Prospective Study of a Multimodal Convulsive Seizure Detection Wearable System on Pediatric and Adult Patients in the Epilepsy Monitoring Unit. Front Neurol. 2021; 12.
8. 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]
9. Onorati F, Regalia G, Caborni C, Migliorini M, Bender D, Poh MZ, et al. Multicenter clinical assessment of improved wearable multimodal convulsive seizure detectors. Epilepsia. 2017; 58(11):1870–9. [PubMed: 28980315]
10. Mormann F, Andrzejak RG, Elger CE, Lehnertz K. Seizure prediction: the long and winding road. Brain. 2007; 130(Pt 2):314–33. [PubMed: 17008335]
11. Mormann F, Andrzejak RG. Seizure prediction: making mileage on the long and winding road. Brain. 2016; 139(Pt 6):1625–7. [PubMed: 27234060]
12. Baud MO, Proix T, Gregg NM, Brinkmann BH, Nurse ES, Cook MJ, et al. Seizure forecasting: Bifurcations in the long and winding road. Epilepsia. 2023; 64(S4).
13. 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]

14. Snyder DE, Echauz J, Grimes DB, Litt B. The statistics of a practical seizure warning system. J Neural Eng. 2008; 5(4):392–401. [PubMed: 18827312]
15. 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]
16. Andrzejak RG, Zaveri HP, Schulze-Bonhage A, Leguia MG, Stacey WC, Richardson MP, et al. Seizure forecasting: Where do we stand? Epilepsia. 2023; 64(S3).
17. Brier GW. Verification of forecasts expressed in terms of probability. Mon Weather Rev. 1950; 78(1):1–3.
18. 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).
19. 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]
20. 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).
21. 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]
22. 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).

HIGHLIGHTS

- Seizure forecasting metrics depend on the seizure frequency (SF)
- It is important to report forecasting metrics as a function of SF
- Comparing moving average (MA) to permutation reveals that MA is always as good or better
- Comparing a candidate forecasting model against the moving average model (MA) is a better benchmark than permutation

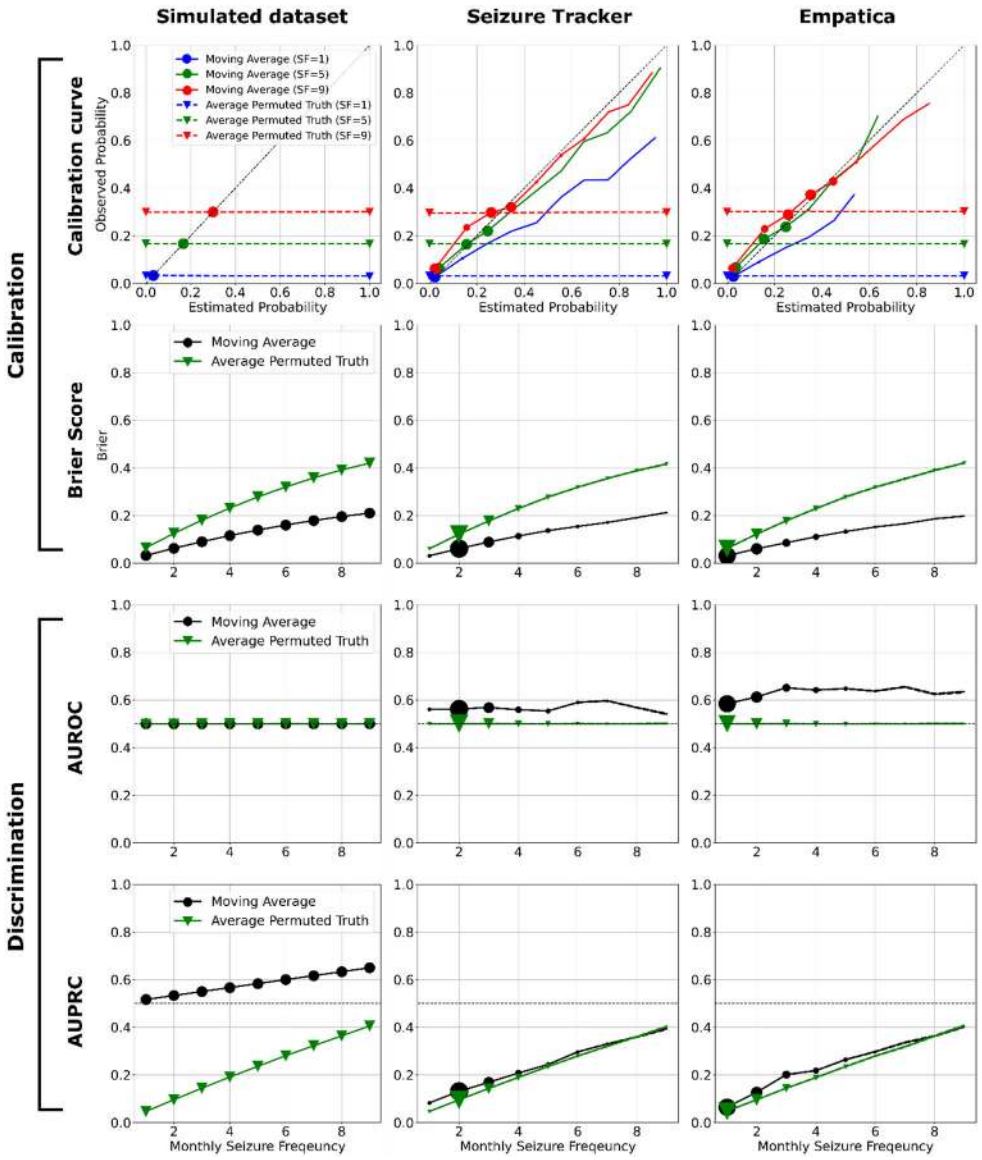

Figure 1.

Twelve scenarios comparing the performance of MA vs. permuted truth. The calibration curve, brier score, AUROC, and AUPRC are shown in rows. The results of simulated, Seizure Tracker, and Empatica datasets are shown in columns. In the calibration curves (first row), the monthly seizure frequencies 1, 5, and 9 are shown in blue, green, and red. The results of MA and permuted truth are indicated by solid line and dash line. The marker size indicates the normalized number of diaries within each estimated probability bin. Since the MA outcomes for the simulated dataset are constant, there is only one estimated probability in the calibration curve, resulting in a single marker instead of a solid line. The brier score, AUROC, and AUPRC of MA and permuted truth are indicated by black and green solid lines respectively in the second, third, and fourth rows. The marker size indicates the normalized number of diaries within each SF bin. Note that in all twelve comparisons, MA performs as well or better than permuted truth. Additionally, within this range of SF, all

metrics except AUROC vary monotonically with seizure frequency. Higher SF values were explored in simulation (Appendix).