

1 Title:

2 Prospective validation of a seizure diary forecasting falls short

3

4 Running head:

5 Prospective validation of seizure forecasting

6

7 Daniel M. Goldenholz, MD PhD^{1,2}

daniel.goldenholz@bidmc.harvard.edu

8 Celena Eccleston, BS^{1,2}

ceccleston6celena@gmail.com

9 Robert Moss, BA³

rob@seizuretracker.com

10 M. Brandon Westover, MD, PhD^{1,2,4,5}

mwestover@mgh.harvard.edu

11

12 1 Dept. of Neurology, Beth Israel Deaconess Medical Center, Boston 02215 MA

13 2 Dept. of Neurology, Harvard Medical School, Boston 02215 MA

14 3 Seizure Tracker LLC, Springfield 22151 VA

15 4 Dept. of Neurology, Massachusetts General Hospital, Boston 02114 MA

16 5 McCance Center for Brain Health, Boston, 02114 MA

17

18 Corresponding author:

19 Daniel Goldenholz

20 330 Brookline Ave Baker 5

21 Boston MA 02215

22 617 632 8934

23 daniel.goldenholz@bidmc.harvard.edu

24

25 Title characters=62

26 Running head characters=45

27 Word count:

28 Abstract: 188

29 Body: *

30 Number of figures: 1

31 Number of color figures: 1 *(black and one figure for in-print version of paper)

32 Number of tables: 1

33

34

35 **Abstract (max 300 words)**

36

37 OBJECTIVE: Recently, a deep learning AI model forecasted seizure risk using retrospective
38 seizure diaries with higher accuracy than random forecasts. The present study sought to
39 prospectively evaluate the same algorithm.

40 METHODS: We recruited a prospective cohort of 46 people with epilepsy; 25 completed
41 sufficient data entry for analysis (median 5 months). We used the same AI method as in our
42 prior study. Group-level and individual-level Brier Skill Scores (BSS) compared random forecasts
43 and simple moving average forecasts to the AI.

44 RESULTS: The AI had an AUC of 0.82. At the group level, the AI outperformed random
45 forecasting (BSS=0.53). At the individual level, AI outperformed random in 28% of cases. At the
46 group and individual level, the moving average outperformed the AI. If pre-enrollment (non-
47 verified) diaries (with presumed under-reporting) were included, the AI significantly
48 outperformed both comparators. Surveys showed most did not mind poor quality LOW-RISK or
49 HIGH-RISK forecasts, yet 91% wanted access to these forecasts.

50 SIGNIFICANCE: The previously developed AI forecasting tool did not outperform a very simple
51 moving average forecasting this prospective cohort, suggesting that the AI model should be
52 replaced.

53

54 **Key points**

55 A previously developed e-diary based AI seizure forecasting tool was prospectively tested.

56 Although by some metrics the tool was successful, the overall AI performance was

57 unacceptably low.

58 It was much easier to outperform a random forecast; it was much harder to outperform a

59 moving average forecast.

60 Using unverified diaries can skew forecasting metrics in favor of underperforming tools.

61

62 **Introduction**

63 Not knowing when the next seizure will happen reduces quality of life for people living with
64 epilepsy. Roughly a decade ago, it was discovered that it is possible to provide seizure forecasts
65 using invasive technology¹. Since then, novel approaches involving highly invasive²⁻⁵ and less
66 invasive tools^{6,7} have been proposed. Using a retrospective study of 5,419 unverified self-
67 reported electronic diaries from Seizure Tracker, our group reported that 24-hour forecasts
68 from seizure diaries alone were possible using deep learning⁸. The present study aimed to
69 validate these findings prospectively.

70

71 **Methods**

72 Patients

73 The protocol was deemed Exempt by the BIDMC Institutional Review Board. Participants were
74 recruited by Seizure Tracker⁹ via email. Participants with 1) epilepsy, 2) age 18 or older, 3) an
75 active Seizure Tracker e-diary account, 4) at least 3 seizures recorded in their account, and 5) at
76 least 3 months of previous e-diary data were eligible. Verified participants linked their e-diary
77 and a RedCap^{10,11} survey account to the study. They completed an initial survey and then
78 weekly surveys (verifying diary completion) for 5 months. They also maintained seizure e-
79 diaries. For safety, only retrospective forecasts were provided monthly.

80

81 The AI forecaster

82 Using our pre-trained deep learning algorithm⁸ (hereafter: AI), seizure forecasts were calculated
83 for every day possible. The AI uses a recurrent neural network connected to a multilayer

84 perceptron trained on 3806 users (Appendix A). All model parameters and hyperparameters
85 remained unchanged from the original model.

86 The AI computes a probability of *any seizures* occurring within a 24-hour period. The AI uses the
87 84-day trailing history of daily seizure counts leading up to that forecasted day as input. The
88 tool was applied sequentially with a sliding window that moves forward one day at a time. Each
89 patient could have up to 57 daily forecasts (8 weeks and one day), representing the prospective
90 observation period. In some patients, this number was lower due to incomplete diary
91 information (Appendix B). The 3-month pre-enrollment diaries were retained for additional
92 analysis.

93

94 The random forecaster

95 The daily AI forecast was compared with a permuted forecaster as a benchmark (hereafter
96 “random”). The random forecaster is generated by permuting forecasts from the AI at the
97 subject level. This can be thought of as shuffling a deck of cards, where each card is the AI
98 forecast for a given day, and there is a different deck for each patient. A useful forecast should
99 (at minimum) outperform a permuted forecaster¹². Where appropriate, the average outcome
100 metric from 1000 such permutations was used, such as for computing the Brier Score.

101

102 The moving average forecaster

103 The daily AI forecast was also compared with a moving average forecaster which accounted for
104 the typical seizure rate from each patient. Moving average forecasts were computed by taking
105 the total number of seizure days in each trailing 84-day history and dividing by 84 to obtain a

106 simple estimate of daily risk of any seizures for the coming 24-hour forecast (Appendix A). Of
107 note, unlike a similar comparator used our prior study (there called the “rate matched random”
108 forecaster), this moving average forecaster uses total seizure days, not total seizure counts⁸.
109 This change was made to provide a more stringent comparator for the AI. Also of note, all
110 summary results were computed using only the verified post-enrollment period due to
111 concerns about possible under-reporting during the pre-enrollment period (see Discussion).

112

113

114 Outcome metrics

115 Performance of each model was measured using area under the receiver operating
116 characteristic curve (AUC), and the Brier Score. AUC values range between 0 and 1, with 0.5
117 representing a tool indistinguishable from coin flipping, and 1 representing a perfect
118 discriminator. Brier Scores range between 0 and 1, with values closer to 0 representing higher
119 accuracy. Our primary outcome (Appendix B) was comparing AI to the random forecasts using
120 Brier Skill Scores (BSS). Brier Skill Score of 1 represents the AI algorithm is perfect, 0 indicates
121 the AI is not better than the reference forecast, and -1 indicates the reference forecast is
122 perfect).

123

124 BSS was computed both at the group-level and at the individual participant level. When using as
125 reference test the random forecaster to calculate BSS, “group-level” means that random
126 forecasts were generated by randomly shuffling the AI predictions across all patients, and
127 randomly reassigning them. Note that this means that forecasts from one patient may be

128 randomly reassigned to other patients. By contrast, calculating BSS at the “individual level”
129 relative to random forecasting means that random forecasts are all from the same patients,
130 albeit in a randomly shuffled order. This means that the group and individual level BSS scores
131 are not directly comparable, and the median of the individual-level BSS scores need not match
132 the group-level BSS score. Additional BSS values were computed using the moving average as
133 an alternative reference.

134

135 Calibration curves were generated for the AI, random, and moving average forecasters using
136 equally spaced bins. Confidence intervals for AUC and BSS values were obtained by 1000
137 bootstrapped samples, selecting patients with replacement.

138 Code is available here: <https://github.com/GoldenholzLab/deepManCode>.

139

140 **Results**

141 Of 46 recruited participants, 1 was ineligible, 3 were seizure-free, and 11 provided insufficient
142 diary data. Within the remaining 31, there were 3 dropouts, and 8 who missed some of the
143 weekly diary completeness responses. Only 25 patients had sufficient contiguous data to
144 perform forecasts based on 3 months of prospectively collected history. Forecastable diary days
145 (Appendix C) ranged 15-57 (median 57) days. Total seizures per patient ranged from 1-56,
146 (median 13). Participant characteristics are summarized in Table 1.

147

148 Group level results:

149 The following represent group level metrics (Figure 1). Confidence intervals were obtained via
150 1000 bootstrapped samples with replacement at the patient level. The AUC for AI was: 0.82
151 [95% CI 0.72-0.90], and for the permuted AI (i.e. random forecast) was 0.50 [95% CI 0.46-0.54].
152 The Brier Score for AI was 0.14. The AI performed significantly better than the random
153 forecaster at the group level, with a Brier Skill Score (AI vs. random) of 0.53 [95% CI 0.27-0.70].
154 However, the AUC of the moving average forecaster was also 0.82 [95% CI 0.72-0.89], which
155 was not significantly different from the AI (Mann Whitney U, $p=0.13$); and the Brier Skill Score
156 of the AI relative to the moving average forecaster was -0.01 [95% CI -0.04 - 0.02], suggesting
157 minimal difference in performance.

158

159 Individual level results:

160 In 7 patients (28%) the AI was superior (i.e., individual Brier Skill Score >0) to the random
161 forecaster whereas for 9 patients (36%) the AI was superior to the moving average. The
162 individual Brier Skill Scores (mean permuted AI forecasts¹² as comparator) were median 0.00
163 (95% CI: -0.03 – 0.20). These values were notably lower than the group level BSS values (see
164 Appendix I). Individual Brier Skill Scores with moving average as comparator were median -0.01
165 with 95% confidence range (-0.08-0.17). Individual level AI AUC values were very poor quality
166 0.43 +/- 0.21, as were individual level moving average values AUC 0.43 +/- 0.13.

167

168 Complete diaries with AI and moving average forecasts were plotted (Appendix D and E). There
169 were 25 patients reporting less than 3 seizures in the pre-enrollment period (see Appendix D).
170 Time-in-warning analysis was conducted (Appendix G).

171

172 The above analyses were also re-computed using the full set of 31 patients using the 3-month
173 pre-enrollment diaries (Appendix F). This showed the AI was superior to random and moving
174 average at the group level, and superior to the moving average at the individual level in 14
175 patients (45%). However, pre-enrollment data seizure rate was dramatically lower than the
176 enrollment seizure rates, suggesting severe under-reporting.

177

178 The initial surveys (n=46), filled out prior to any forecasting, included questions related to
179 seizure forecasting (Appendix G). Many (52%) patients stated they would not mind poor quality
180 HIGH-RISK forecasts, and many (52%) did not mind poor quality LOW-RISK forecasts, yet almost
181 all (91%) wanted access to forecasts. In the setting of LOW-RISK forecasts, 80% said they would
182 not change their behavior, yet in HIGH RISK only 28% would not change – many stated that they
183 would avoid risk-taking behavior (54%).

184

185 **Discussion**

186 Our results prospectively attempted validation of a deep learning seizure forecasting system
187 that is based entirely on seizure diaries. At the group level (considering all forecasts from all
188 patients equally), one may mistakenly believe that the AI has strong potential. Using a random
189 permutation surrogate as our comparator, the AI forecasts better than chance. However, a
190 simple moving average forecaster turns out to perform just as well as the AI. Moreover, at the
191 individual level (summarizing each patient separately first, then aggregating results), the AI
192 outperforms the random permutation and the moving average in a small minority of cases,

193 showing very poor overall individual level performance in AUC and Brier scores. The present
194 work mirrors the previous retrospective study⁸, however it focuses on the individual patient
195 level with physician curated, verified complete diaries. By reporting multiple metrics in different
196 ways, this study highlights deficiencies of the present AI algorithm, and in certain outcome
197 metrics. Clearly, the AI is not better than moving average forecasts; however, when missing
198 data is present, the AI outperforms the moving average.

199 Qualitatively, the data (Appendices D, E, F) suggests that at least one driver of periods of better
200 forecasts relates to the AI being better able to forecast multi-day clusters of seizures compared
201 with the random permutation or the moving average. These clusters may reflect multi-day
202 seizure susceptibility periods, though they do not appear to be periodic^{3,13}, and they do not fit
203 the classical definition of seizure clusters^{14,15}.

204

205 Unlike our retrospective study⁸ that did not have verified complete diaries, the prospective
206 study utilized weekly verified diaries from patients with clinical data confirming their epilepsy
207 diagnosis. The misalignment of results between the former study and the present one may
208 reflect the difference between the self-report and closely monitored self-report. In the case of
209 the former, some events may be missed (under-reporting¹⁶), but in the case of the latter, some
210 dubious events may be included (over-reporting¹). There are no rigorous studies of over-
211 reporting, which is challenging to accurately quantify. Here, the verified diaries have
212 dramatically higher rates during the prospective phase compared to the pre-enrollment 3-
213 month periods (see Appendix D) – strongly suggesting under-reporting.

214 The apparent under-reporting from the pre-enrollment period appears to reflect that without
215 supervision, diaries might be incomplete. Our study required for enrollment the existence of a
216 Seizuretracker account with at least 3-months of data prior to enrollment, however we did not
217 verify or demand that such diaries were complete. This oversight is significant, because during
218 the observed portion of the study we asked the participants weekly if their diaries were
219 complete, and the seizure rates were consistently much higher (see Appendix D). Importantly,
220 multiple lines of evidence^{13,17-21} show that, contrary to what we observed in our cohort,
221 unverified seizure diaries often do reproduce patterns confirmed in verified systems, thus
222 unsupervised seizure diaries may not always suffer from underreporting bias. Nevertheless,
223 future studies will need to either confirm with participants that pre-enrollment diaries are
224 complete or obtain longer duration observation periods and use only data obtained during
225 confirmed timeframes.

226
227 Perhaps, one might suspect that patients with very high seizure rates would be unlikely to
228 benefit from seizure forecasts at all. On the other hand, our cohort included only patients who
229 wanted to be involved in a forecasting study (there was no compensation for this study), and
230 39% of them had very high seizure rates. Patient preferences (Appendix G) may even support
231 inaccurate forecasts rather than no forecasts. It is worthwhile to note that the preferences
232 reported were obtained prior to obtaining any forecasts from our team, therefore these can be
233 viewed as the opinion of optimistic patients who had just enrolled in a study. Nevertheless,
234 patients with less frequent seizure days are likely the most important to forecast (based on the
235 need to make temporary changes in behavior), and the present algorithm did not excel in this

236 area. More study is needed to better understand what the characteristics are of patients who
237 would be most interested in seizure forecasts, and who would benefit most. It should be
238 emphasized that in the absence of a nearly perfect forecast system, patients should never be
239 encouraged to engage in risky behavior during periods of forecasted low risk.

240
241 The present study has several limitations. First, some people with epilepsy have very low (e.g.,
242 1-2 seizures per year) or very high (i.e., \geq daily) seizure rates²². Such patients would not be likely
243 to benefit from the current generation of daily forecasting tools. Second, it can be challenging
244 for patients to maintain a seizure diary²³, thus limiting tools of this nature to patients and
245 caregivers willing to maintain a diary. Third, our prior⁸ and present study did not have available
246 EEG data to augment forecasts. Although speculative, including EEG data may enhance the
247 performance of these models. Fourth, the 5-month prospective duration of the present study
248 may be too short to make definitive conclusions about the utility of the AI algorithm. To
249 address this deficiency, our group will be conducting a larger study soon with a longer
250 observation period to allow for sufficiently large windows of investigator-verified seizure
251 diaries. Sixth, there was a presumed dramatic under-reporting in the pre-enrollment period. In
252 our future study, we will not include a pre-enrollment period due to the challenges in verifying
253 that they are complete. Finally, the choice of reference standard comes at a cost. Our average
254 permutation (a.k.a. random) forecaster standard could not be realistically provided to patients
255 in real-time. Conversely, our second reference standard was the moving average forecaster.
256 This can be implemented in a real-time system, making it a realistic comparator. A comparison
257 of the calibration curve (Figure 1) shows very poor calibration of the permuted AI, but decent

258 calibration of moving average and AI. In using both, we highlight the advantages and
259 disadvantages of each.

260

261 We hope that future advances in wearables⁶ and minimally invasive tools^{7,24} can synergistically
262 be applied to diary-based forecasting tools to achieve higher accuracy and wider patient
263 appeal.

264

265 **Data availability:**

266 Data is available on reasonable request.

267

268 **Epilepsia ethical publication statement:**

269 We confirm that we have read the Journal's position on issues involved in ethical publication
270 and affirm that this report is consistent with those guidelines.

271

272 **Acknowledgements and Funding**

273 DMG was supported by NINDS KL2TR002542 and K23NS124656. MBW received funding support
274 from the American Academy of Sleep Medicine through an AASM Foundation Strategic
275 Research Award; the NIH (R01NS102190, R01NS102574, R01NS107291, RF1AG064312,
276 RF1NS120947, R01AG073410), and NSF (2014431). Dr. Westover is a co-founder of Beacon
277 Biosignals, and Director for Data Science for the McCance Center for Brain Health.

278

279 **Author contributions**

280 DG, MW, and RM contributed to conception and design of the study; CE, DG and RM
281 contributed to acquisition and analysis of data; DG drafted a significant portion of the
282 manuscript and figures.

283

284 **Potential Conflicts of interest**

285 There are no conflicts of interest for any of the authors.

286

287 **References**

- 288 1. Cook MJ, O'Brien TJ, Berkovic SF, Murphy M, Morokoff A, Fabinyi G, et al.
289 Prediction of seizure likelihood with a long-term, implanted seizure advisory
290 system in patients with drug-resistant epilepsy: A first-in-man study. *Lancet*
291 *Neurol*. 2013; 12(6):563–71.
- 292 2. Leguia MG, Andrzejak RG, Rummel C, Fan JM, Mirro EA, Tcheng TK, et al. Seizure
293 Cycles in Focal Epilepsy. *JAMA Neurol* [Internet]. 2021; 78(4):454–63. Available
294 from: <http://www.ncbi.nlm.nih.gov/pubmed/33555292>
- 295 3. Baud MO, Kleen JK, Mirro EA, Andrechak JC, King-Stephens D, Chang EF, et al.
296 Multi-day rhythms modulate seizure risk in epilepsy. *Nat Commun* [Internet].
297 2018; 9(1):1–10. Available from: <http://dx.doi.org/10.1038/s41467-017-02577-y>
- 298 4. Proix T, Truccolo W, Leguia MG, Tcheng TK, King-Stephens D, Rao VR, et al.
299 Forecasting seizure risk in adults with focal epilepsy: a development and validation
300 study. *Lancet Neurol* [Internet]. 2021; 20(2):127–35. Available from:
301 <http://www.ncbi.nlm.nih.gov/pubmed/33341149>
- 302 5. Nasser M, Pal Attia T, Joseph B, Gregg NM, Nurse ES, Viana PF, et al. Ambulatory
303 seizure forecasting with a wrist-worn device using long-short term memory deep
304 learning. *Sci Rep* [Internet]. 2021; 11(1):21935. Available from:
305 <http://www.ncbi.nlm.nih.gov/pubmed/34754043>
- 306 6. Karoly PJ, Stirling RE, Freestone DR, Nurse ES, Maturana MI, Halliday AJ, et al.
307 Multiday cycles of heart rate are associated with seizure likelihood: An
308 observational cohort study. *EBioMedicine* [Internet]. 2021 [cited 2021];
309 72:103619. Available from: <https://pubmed.ncbi.nlm.nih.gov/34649079/>
- 310 7. Viana PF, Pal Attia T, Nasser M, Duun-Henriksen J, Biondi A, Winston JS, et al.
311 Seizure forecasting using minimally invasive, ultra-long-term subcutaneous
312 electroencephalography: Individualized inpatient models. *Epilepsia* [Internet].
313 2022; . Available from: <http://www.ncbi.nlm.nih.gov/pubmed/35395101>
- 314 8. Goldenholz DM, Goldenholz SR, Romero J, Moss R, Sun H, Westover B.
315 Development and Validation of Forecasting Next Reported Seizure Using e-Diaries.

- 316 Ann Neurol [Internet]. 2020; 88(3):588–95. Available from:
317 <http://www.ncbi.nlm.nih.gov/pubmed/32567720>
- 318 9. Casassa C, Rathbun Levit E, Goldenholz DM. Opinion and Special Articles: Self-
319 management in epilepsy: Web-based seizure tracking applications. *Neurology*.
320 2018; 91(21).
- 321 10. Harris PA, Taylor R, Minor BL, Elliott V, Fernandez M, O’Neal L, et al. The REDCap
322 consortium: Building an international community of software platform partners. *J*
323 *Biomed Inform*. 2019; 95:103208.
- 324 11. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic
325 data capture (REDCap)--a metadata-driven methodology and workflow process for
326 providing translational research informatics support. *J Biomed Inform*. 2009;
327 42(2):377–81.
- 328 12. Karoly PJ, Ung H, Grayden DB, Kuhlmann L, Leyde K, Cook MJ, et al. The circadian
329 profile of epilepsy improves seizure forecasting. *Brain*. 2017; 140(8):2169–82.
- 330 13. Karoly PJ, Goldenholz DM, Freestone DR, Moss RE, Grayden DB, Theodore WH, et
331 al. Circadian and circaseptan rhythms in human epilepsy: a retrospective cohort
332 study. *Lancet Neurol* [Internet]. 2018; 17(11):977–85. Available from:
333 <http://www.ncbi.nlm.nih.gov/pubmed/30219655>
- 334 14. Haut SR. Seizure clusters: characteristics and treatment. *Curr Opin Neurol*
335 [Internet]. 2015; 28(2):143–50. Available from:
336 <http://www.ncbi.nlm.nih.gov/pubmed/25695133>
- 337 15. Chiang S, Haut SR, Ferastraoaru V, Rao VR, Baud MO, Theodore WH, et al.
338 Individualizing the definition of seizure clusters based on temporal clustering
339 analysis. *Epilepsy Res*. 2020; 163.
- 340 16. Elger CE, Hoppe C. Diagnostic challenges in epilepsy: seizure under-reporting and
341 seizure detection. *Lancet Neurol* [Internet]. 2018; 17(3):279–88. Available from:
342 <http://www.ncbi.nlm.nih.gov/pubmed/29452687>
- 343 17. Goldenholz DM, Westover MB. Flexible realistic simulation of seizure occurrence
344 recapitulating statistical properties of seizure diaries. *Epilepsia* [Internet]. 2023;
345 64(2):396–405. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/36401798>
- 346 18. Goldenholz DM, Goldenholz SR, Moss R, French J, Lowenstein D, Kuzniecky R, et al.
347 Is seizure frequency variance a predictable quantity? *Ann Clin Transl Neurol*. 2018;
348 5(2):201–7.
- 349 19. Goldenholz DM, Goldenholz SR, Moss R, French J, Lowenstein D, Kuzniecky R, et al.
350 Does accounting for seizure frequency variability increase clinical trial power?
351 *Epilepsy Res* [Internet]. 2017; 137(June):145–51. Available from:
352 <http://dx.doi.org/10.1016/j.epilepsyres.2017.07.013>
- 353 20. LaGrant B, Goldenholz DM, Braun M, Moss RE, Grinspan ZM. Patterns of Recording
354 Epileptic Spasms in an Electronic Seizure Diary Compared With Video-EEG and
355 Historical Cohorts. *Pediatr Neurol* [Internet]. 2021; 122:27–34. Available from:
356 <https://doi.org/10.1016/j.pediatrneurol.2021.04.008>
- 357 21. Goldenholz DM, Tharayil J, Moss R, Myers E, Theodore WH. Monte Carlo
358 simulations of randomized clinical trials in epilepsy. *Ann Clin Transl Neurol*. 2017;
359 4(8):544–52.

- 360 22. Ferastraoar V, Goldenholz DM, Chiang S, Moss R, Theodore WH, Haut SR.
361 Characteristics of large patient-reported outcomes: Where can one million
362 seizures get us? *Epilepsia Open* [Internet]. 2018; 3(3):364–73. Available from:
363 <http://www.ncbi.nlm.nih.gov/pubmed/30187007>
- 364 23. Fisher RS, Blum DE, DiVentura B, Vannest J, Hixson JD, Moss R, et al. Seizure
365 diaries for clinical research and practice: limitations and future prospects. *Epilepsy*
366 *Behav.* 2012; 24(3):304–10.
- 367 24. Stirling RE, Grayden DB, D’Souza W, Cook MJ, Nurse E, Freestone DR, et al.
368 Forecasting Seizure Likelihood With Wearable Technology. *Front Neurol.* 2021;
369 12(July):1–12.

370

371

372 **Figure 1: Calibration curves.** The prospective seizure forecasts (pooled across all patients) are
373 compared to the actual observed seizures for (1) the artificial intelligence (AI), (2) the rate
374 matched random forecast (RMR), and (3) random permutations of the AI. Confidence intervals
375 are shown by bootstrapping 1000 times (choosing patients with replacement). A perfectly
376 calibrated (dashed line) forecast would always forecast the correct percentage of observed
377 seizures. In this figure, the AI and random forecast deviate from the ideal somewhat, whereas
378 the permuted reference is very poorly calibrated (as expected).

379

380

	Number	%
Number of patients	31	
Females (%)	14	45%
Physician confirmed epilepsy	31	100%
EEG confirmed epilepsy		
Yes	27	87%
Unsure	4	13%
Handedness (right / left / mixed)		
Right	23	74%
Left	6	19%
Mixed	2	6%
Epilepsy type		
Generalized	8	26%
Focal	11	35%
Focal + Generalized	8	26%
Unknown	4	13%
Epilepsy location		
Frontal	1	3%
Temporal	6	19%
Parietal	0	0%
Occipital	1	3%
Multifocal	2	6%
Unknown	21	68%
Epilepsy cause		

Structural	9	29%
Genetic	6	19%
Infectious	1	3%
Metabolic	0	0%
Immune	0	0%
Unknown	15	48%
Prior epilepsy surgery (%)	16	52%

381

382 **TABLE 1:** Baseline characteristics of participants in the prospective study. Note, 31 patients had
383 sufficient information to proceed to analysis, however 6 did not have sufficient data for analysis
384 involving forecasts made only from 3 months of prospectively collected history.

Calibration Curve

