



Deep Thoughts—Predicting Initial Treatment Response in Newly Diagnosed Epilepsy

Epilepsy Currents

2023, Vol. 23(2) 90-92

© The Author(s) 2022

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/15357597221139365

journals.sagepub.com/home/epi



Development and Validation of a Deep Learning Model for Predicting Treatment Response in Patients With Newly Diagnosed Epilepsy

Hakeem H, Feng W, Chen Z, Choong J, Brodie MJ, Fong SL, Lim KS, Wu J, Wang X, Lawn N, Ni G. *JAMA Neurol.* 2022;79(10):986-996. doi:10.1001/jamaneurol.2022.2514

Importance: Selection of antiseizure medications (ASMs) for epilepsy remains largely a trial-and-error approach. Under this approach, many patients have to endure sequential trials of ineffective treatments until the “right drugs” are prescribed. **Objective:** To develop and validate a deep learning model using readily available clinical information to predict treatment success with the first ASM for individual patients. **Design, Setting, and Participants:** This cohort study developed and validated a prognostic model. Patients were treated between 1982 and 2020. All patients were followed up for a minimum of 1 year or until failure of the first ASM. A total of 2404 adults with epilepsy newly treated at specialist clinics in Scotland, Malaysia, Australia, and China between 1982 and 2020 were considered for inclusion, of whom 606 (25.2%) were excluded from the final cohort because of missing information in 1 or more variables. **Exposures:** One of 7 antiseizure medications. **Main Outcomes and Measures:** With the use of the transformer model architecture on 16 clinical factors and ASM information, this cohort study first pooled all cohorts for model training and testing. The model was trained again using the largest cohort and externally validated on the other 4 cohorts. The area under the receiver operating characteristic curve (AUROC), weighted balanced accuracy, sensitivity, and specificity of the model were all assessed for predicting treatment success based on the optimal probability cutoff. Treatment success was defined as complete seizure freedom for the first year of treatment while taking the first ASM. Performance of the transformer model was compared with other machine learning models. **Results:** The final pooled cohort included 1798 adults (54.5% female; median age, 34 years [IQR, 24-50 years]). The transformer model that was trained using the pooled cohort had an AUROC of 0.65 (95% CI, 0.63-0.67) and a weighted balanced accuracy of 0.62 (95% CI, 0.60-0.64) on the test set. The model that was trained using the largest cohort only had AUROCs ranging from 0.52 to 0.60 and a weighted balanced accuracy ranging from 0.51 to 0.62 in the external validation cohorts. Number of pretreatment seizures, presence of psychiatric disorders, electroencephalography, and brain imaging findings were the most important clinical variables for predicted outcomes in both models. The transformer model that was developed using the pooled cohort outperformed 2 of the 5 other models tested in terms of AUROC. **Conclusions and Relevance:** In this cohort study, a deep learning model showed the feasibility of personalized prediction of response to ASMs based on clinical information. With improvement of performance, such as by incorporating genetic and imaging data, this model may potentially assist clinicians in selecting the right drug at the first trial.

Commentary


We know the classic numbers: two-thirds of patients with epilepsy experience seizure-freedom on anti-seizure medications (ASMs). But consider this striking statistic stated differently: 71% of initial treatment is unsuccessful.¹ Now, “unsuccessful” is a grab-bag term conflating inefficacy, non-adherence, and intolerability. Regardless, that’s a problem. The question is—can we predict who will sink versus swim? Given 30+ available ASMs, currently trial-and-error represents the standard of care. Understanding what factors determine success

could better target our drug selection and inform who to counsel most aggressively.

Hakeem et al tackled this question.¹ In background, they point to a precursor model² developed on claims data predicting optimal first ASM selection, with modest discrimination. While claims provide sheer size, they imperfectly measure some key predictors and outcomes (e.g., seizures, electroencephalogram (EEG)/imaging test results, drug discontinuation). Thus, the authors pooled 5 institutional datasets containing 1800 patients mostly from Glasgow plus smaller samples from Malaysia, Australia, and China.



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).



The goal was to predict “success,” defined as 1-year seizure-freedom only on initial monotherapy. They unleashed 6 models, including deep learning algorithms (e.g., a “transformer” model, and a “multilayered perceptron” model). Discrimination was compared via the area under the curve (AUC; probability that a patient who “succeeded” had a higher predicted probability of “succeeding”; $0.5 = \text{chance}$, $>0.7 = \text{modest}$; $>0.8 = \text{strong}$).

Results were modest, sometimes no better than chance. The transformer model had an AUC 0.72 in the 80% training dataset, and 0.65 and maximal accuracy of $\sim 65\%$ in the 20% testing dataset. The other models fared no better, with AUCs ranging 0.58 to 0.64 and accuracy ranging 57% to 60%. When models were trained on Glasgow data and tested on the remaining 4 datasets, results were even less rosy. The AUCs ranged 0.44 to 0.58, and accuracy was similarly about coin flip.

Trial-and-error lives on, and the future remains hard to predict. The paper’s goal (predicting treatment success with the first ASM) implies a 2-fold question. 1. Should we treat? And 2. What drug? These data don’t quite answer either question. A low chance of success does not rule out benefit; the chance of seizure-freedom was still likely lower than had they not been treated. Likewise, a high chance of success does not guarantee benefit. A patient could have been destined for seizure-freedom regardless of treatment, or downsides of treatment could outweigh a small absolute risk reduction. Thus, a single-armed (everyone was treated) dichotomous prediction probably hides what matters most—absolute treatment effects. Still, documenting a low success rate remains useful. Perhaps this means that we commonly aren’t hitting the bullseye with the first ASM, which begets the second question—which drug? Interestingly, though, the actual ASM mattered little. The specific ASM was about 10 times less important in determining “success” than pretreatment seizure count or EEG/imaging abnormalities, each nonmodifiable. One potential explanation could be that ASM choices were already optimized, thus little room for improvement. This seems overly optimistic. For example, that precursor claims-based algorithm² suggested the specific ASM regimen mattered a great deal, and very few were prescribed the predicted “optimal” regimen. That said, as above, claims cannot reliably measure seizures or test results, so perhaps these cohort-based data provide the “real” answer, that we should focus our counseling on patients with the highest baseline risk, rather than feeling too much pressure about choosing the one and only “best” ASM. Other data³ likewise reflect that much of adherence is about differences between patients rather than differences between drugs.

Additional research take-home messages feel like variations on a theme.⁴


The definition of “success” matters. The authors chose 1-year seizure-freedom given a study showing 100% seizure reduction increased quality-of-life slightly more than a 75% to 99% reduction.⁵ Still, even incomplete seizure reduction may predict improved quality of life,⁶⁻⁸ relapse could just as well reflect true inefficacy as it could reflect titrating too slowly which cannot be disentangled here, and in a composite outcome it is difficult to

know exactly what we are measuring particularly when all components are not equal and could have different predictors if examined individually. Thus, one wonders about what might have happened if using less extreme or more granular outcomes.

Next, all that glitters is not gold. Once again, enormously complex machine learning models fail to meaningfully outperform logistic regression. “Black box” models may discover interactions or higher-order effects that would have been undetectable using traditional generalized linear models. Yet, unfettered by assumptions, detecting noise is the rule rather than the exception (note the expected drop in testing vs training performance), and anyways detecting such interactions typically requires huge sample sizes. Furthermore, all models, no matter how fancy, are beholden to the same data limitations. For example, seizure diaries may undercount seizures.⁹ Though, admittedly this is a difficult problem to overcome in absence of perfect seizure detection devices. Regarding interpretability—AUC, despite having a precise mathematical definition, is not intuitive. It is also no secret that clinicians often struggle with concepts such as sensitivity and specificity,¹⁰ which do not calculate a patient’s success probability. Whereas, calibration plots may more intuitively assess model performance, and medicine is optimally performed with a pretest probability in mind, then applying a likelihood ratio, to obtain a post-test probability.


How can we boost performance next time? The top model in the study by Hakeem et al had an AUC actually identical to that precursor claims-based model (0.72), despite probably more accurate diagnostic coding, plus adding in seizure counts and EEG/imaging results. Perhaps the precursor model’s large sample size balanced out addition of cohort-level data in the study by Hakeem et al. One answer would be to say—maybe we still aren’t measuring all the right variables. Some examples: more granular measurement of anxiety and how well a patient has tolerated previous medications, sleep habits, ASM adherence and its predictors, hepatorenal function, drug–drug interactions, drug–disease interactions (e.g., cognitive impairment), additional ASM coincidences (e.g., migraine, neuropathic pain), the specific type of epileptogenic brain lesion, and last but not least a patient’s attitudes and beliefs about medications and which side effects would be of greatest concern. This list is by no means exhaustive. In pursuit of our next top model, it may be worth going back to the drawing board to consider the full complement of biopsychosocial factors driving medication-taking behaviors in addition to biological response.

Predictive modeling is hard work. These investigators completed a heroic effort pooling a large volume of cohort data across centers, analyzing routinely available data using sophisticated models to address an important question. Ultimately, though, we still have much work to do discovering the right drug, for the right patient, at the right time, and translating such knowledge from publications to real-world usable decision support tools.

Samuel W. Terman, MD, MS 
Department of Neurology,
University of Michigan



ORCID iD

Samuel W. Terman  <https://orcid.org/0000-0001-6179-9467>

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. Hakeem H, Feng W, Chen Z, et al. Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. *JAMA Neurol.* 2022;79(10):986-996. doi:10.1001/jamaneurol.2022.2514
2. Devinsky O, Dilley C, Ozery-flato M, et al. Changing the approach to treatment choice in epilepsy using big data. *Epilepsy Behav.* 2016;56:32-37. doi:10.1016/j.yebeh.2015.12.039
3. Terman SW, Kerr WT, Aubert CE, Hill CE, Marcum ZA, Burke JF. Adherence to antiseizure vs other medications among us Medicare beneficiaries with and without epilepsy. *Neurology.* 2021;98(4):e427-e436.
4. Terman SW. Rise of the machines? Predicting brivaracetam response using machine learning. *Epilepsy Curr.* 2021;22(2):111-113.
5. Birbeck GL, Hays RD, Cui X, Vickrey BG. Seizure reduction and quality of life improvements in people with epilepsy. *Epilepsia.* 2002;43(5):535-538.
6. Geitona M, Stamuli E, Giannakodimos S, et al. Lacosamide as a first-line treatment option in focal epilepsy: a cost-utility analysis for the Greek healthcare system. *J Med Econ.* 2019;22(4):359-364. doi:10.1080/13696998.2019.1571499
7. Phumart P, Limwattananon C, Kitwitee P, Unnwongse K, Tiamkao S. EQ-5D-based utilities and healthcare utilization in Thai adults with chronic epilepsy. *Epilepsy Behav.* 2018;83:140-146. doi:10.1016/j.yebeh.2018.03.039.
8. Messori A, Trippoli S, Becagli P, Cincotta M, Labbate M, Zaccara G. Adjunctive lamotrigine therapy in patients with refractory seizures: a lifetime cost-utility analysis. *Eur J Clin Pharmacol.* 1998;53(6):421-427.
9. Hoppe C, Poepel A, Elger CE. Accuracy of patient seizure counts. *Arch Neurol.* 2007;64(11):1595-1599.
10. Whiting PF, Davenport C, Jameson C, et al. How well do health professionals interpret diagnostic information? A systematic review. *BMJ Open.* 2015;5(7):1-8.