



# Back to the Basics in Predictive Modeling—Predicting Surgical Success

## Predicting Seizure Outcome After Epilepsy Surgery: Do We Need More Complex Models, Larger Samples, or Better Data?

Eriksson MH, Ripart M, Piper RJ, Piper RJ, Moeller F, Das KB, Eltze C, Cooray G, Booth J, Whitaker KJ, Chari A, Sanfilippo PM, Caballero AP, Menzies L, McTague A, Tisdall MM, Cross JH, Baldeweg T, Adler S, Wagstyl K. *Epilepsia*. 2023;64(8):2014-2026. doi:10.1111/epi.17637

**Objective:** The accurate prediction of seizure freedom after epilepsy surgery remains challenging. We investigated if (1) training more complex models, (2) recruiting larger sample sizes, or (3) using data-driven selection of clinical predictors would improve our ability to predict postoperative seizure outcome using clinical features. We also conducted the first substantial external validation of a machine learning model trained to predict postoperative seizure outcome. **Methods:** We performed a retrospective cohort study of 797 children who had undergone resective or disconnective epilepsy surgery at a tertiary center. We extracted patient information from medical records and trained three models—a logistic regression, a multilayer perceptron, and an XGBoost model—to predict 1-year postoperative seizure outcome on our data set. We evaluated the performance of a recently published XGBoost model on the same patients. We further investigated the impact of sample size on model performance, using learning curve analysis to estimate performance at samples up to  $N = 2000$ . Finally, we examined the impact of predictor selection on model performance. **Results:** Our logistic regression achieved an accuracy of 72% (95% confidence interval [CI] = 68%-75%, area under the curve [AUC] = .72), whereas our multilayer perceptron and XGBoost both achieved accuracies of 71% (95% CI<sub>MLP</sub> = 67%-74%, AUC<sub>MLP</sub> = .70; 95% CI<sub>XGBoost own</sub> = 68%-75%, AUC<sub>XGBoost own</sub> = .70). There was no significant difference in performance between our three models (all  $p > .4$ ) and they all performed better than the external XGBoost, which achieved an accuracy of 63% (95% CI = 59%-67%, AUC = .62;  $p_{LR} = .005$ ,  $p_{MLP} = .01$ ,  $p_{XGBoost own} = .01$ ) on our data. All models showed improved performance with increasing sample size, but limited improvements beyond our current sample. The best model performance was achieved with data-driven feature selection. **Significance:** We show that neither the deployment of complex machine learning models nor the assembly of thousands of patients alone is likely to generate significant improvements in our ability to predict postoperative seizure freedom. We instead propose that improved feature selection alongside collaboration, data standardization, and model sharing is required to advance the field.

## Commentary

Hindsight is 20/20. But foresight is notoriously difficult. The game is consider characteristics of the patient in front of you, and then estimate outcome probabilities under different treatment scenarios to make an informed choice.

Take choosing candidates for epilepsy surgery. While resection substantially increases the chance of seizure freedom in appropriate drug-refractory patients, one-third of patients relapse postoperatively.<sup>1</sup> Clinical judgment is far from perfect at predicting which patients are destined for postoperative seizure freedom. For example, in one study, 20 epilepsy experts were asked to guess the probability of an Engel 1 outcome in 20 patients. Unfortunately, experts performed equivalent to chance (area under the curve [AUC]: 0.49).<sup>2</sup> Also

unfortunately, existing data-driven risk calculators performed no better. Thus, the search remains how to improve postsurgical outcome prediction. If we had perfect foresight, whether by human intuition or predictive models, we could offer surgery only to those who will benefit.

Eriksson et al<sup>3</sup> took a step back and asked which of several common issues most limit a model's accuracy. They retrospectively extracted medical records from 797 children who received either surgical resection or a disconnection procedure between 2000 and 2018 at the Great Ormond Street Hospital.

First, do complex machine learning models outperform logistic regression? Despite the hype, the answer was . . . no, as we should suspect by now.<sup>4,5</sup> The allure of breaking free of parametric assumptions is strong. Yet, even with the ability





to search higher order interactions unfettered, their 2 machine learning models (both: accuracy 72%, AUC 0.70) performed no better than logistic regression.

Second, how many patients is “enough?” They found a plateau in accuracy beyond around  $N = 400$ . This is smaller than what typically comes to mind, as we often think that vast samples are required to adequately train models.

Third, what was the impact of choosing different predictors? Discrimination was worst when including only MRI diagnosis (0.59 = poor), a bit better when including all available predictors (0.69 = modest), and trivially better (0.72) when also omitting variables that were nonsignificant in univariable models. Note that they sought to include ubiquitous variables, thus did not consider single-photon emission computerized tomography (SPECT), magnetoencephalography (MEG), or positron emission tomography (PET) results (except they did consider genetic testing, when obtained).

Fourth, how well would an example external algorithm perform in their sample? Yossofzai et al<sup>6</sup> developed a machine-learning model using 801 children who underwent surgery (in-sample: AUC 0.73). Out-of-sample performance tested in Eriksson’s sample yielded an AUC of 0.62, a noticeable drop.

These results raise several questions.


1. How can we build more accurate models? First, the low-tech suggestion. Despite the strong causal plausibility of sleep, substances, and possible anti-seizure medication (ASM) adherence/withdrawal influencing seizure relapse risk, I have never seen such factors included in model development. Challenges include variables are often documented incompletely or inconsistently in the chart, self-report is imperfectly reliable, and each characteristic is complex and time-varying, in contrast to easy to extract, static, indisputable values like age or sex. Yet, given seizure risk prediction models have nearly all demonstrated at best modest performance, maybe the time has come to consider including such variables in prospective efforts to boost performance. Sadly, the most easily/accurately obtained factors are often the least predictive. None of age, sex, handedness, family history, or history of febrile seizures were even significant in unadjusted analyses, which should prompt our search for more predictive variables. Also, variables that are amenable to clean categorization may not be granular enough. Groupings like “vascular,” “genetic,” or “dysplasia” may be too broad to be useful without considering their full anatomical specifics or extent of resection in relation to the lesion or epileptogenic network, and “number of ASMs” does not capture which if any of those have exerted any benefit. Hence, modest performance is unsurprising. The investigators provided a nice touch including whether EEG and semiology variables were “localizing-concordant,” “lateralizing concordant,” or “non-concordant.” However, even those variables were not significant. This leads us to our high-tech solution. As emphasized previously,<sup>7</sup> information relevant to the epileptogenic network,

connectivity, or volumetrics of lesion versus resection are not easily captured as simple as “yes or no” for entering into a bedside prediction calculator. I believe that supplementing “yes or no” or routine continuous variables with more complex electrographic or imaging analytics plus the above critical behavioral and ASM questions could help us predict the future better than we currently can.

2. How can we develop more useful models? The outcome here was 1-year seizure freedom. Though, this does not distinguish timing (day 1 = day 364), magnitude (99% reduction = 0% reduction), or severity (generalized convulsions = auras). This is also a 1-armed model (“outcome chance if treated”), which I would usually worry about as being incomplete without a comparison to “no treatment.” Fortunately, presuming that the chance of seizure freedom without surgery is close to 0% for most subjects, this criticism is weak in this case. Still, issues related to the outcome abound: What chance of seizure freedom is sufficient to justify brain surgery? What AUC is “good enough” to justify basing surgical decisions upon risk calculator output? Is 1-year seizure freedom the right outcome, or should we be considering alternative model types (eg, time to event or count) or outcomes (eg, Engel IA versus Engel 1) that might correct existing blind spots? I don’t have easy answers, but we need to be asking the right questions.
3. How can we build more generalizable models? Models nearly always perform worse during external validation. This is a serious criticism, given a model is only helpful if it applies to the patient in front of you. I point out an underutilized strategy—techniques exist to recalibrate an externally derived model to a local population.<sup>8,9</sup> In other words, take the best externally developed model, then tweak/update its coefficients to improve performance in your population without having to completely abandon ship.

This paper’s message is that the problem isn’t lack of vast sample sizes or fancy black box models. Rather, to the degree that accurate predictive models is a useful goal (which is an entirely separate discussion), the key is to include a set of sufficiently predictive features. I think the way forward involves combining the above low- and high-tech perspectives, not just confined to what is most readily documented from the chart, thinking as carefully as possible about what outcomes would be most meaningful, and adapting externally derived models to a given local population.

Forecasting the future is no easy task. But with collaboration and additional, intentional, rigorous standardized retrospective or prospective data, this paper provides a roadmap for the path forward.

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

**ORCID iD**

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

**Declaration of Conflicting Interests**

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

**References**

1. Widjaja E, Jain P, Demoe L, Guttmann A, Tomlinson G, Sander B. Seizure outcome of pediatric epilepsy surgery. *Neurology*. 2020;94(7):311-321. doi:10.1212/WNL.0000000000008966
2. Gracia CG, Chagin K, Kattan MW, et al. Predicting seizure freedom after epilepsy surgery, a challenge in clinical practice. *Epilepsy Behav*. 2019;95:124-130. doi:10.1016/j.yebeh.2019.03.047
3. Eriksson MH, Ripart M, Piper RJ, et al. Predicting seizure outcome after epilepsy surgery: do we need more complex models, larger samples, or better data? *Epilepsia*. 2023;6(8):2014-2026. doi:10.1111/epi.17637
4. Terman SW. Rise of the machines? Predicting brivaracetam response using machine learning. *Epilepsy Curr*. 2021;22(2):111-113. doi:10.1177/15357597211049052
5. Terman SW. Deep thoughts—predicting initial treatment response in newly diagnosed epilepsy. *Epilepsia Curr*. 2023;23(2):90-92. doi:10.1177/15357597221139365
6. Yossofzai O, Fallah A, Maniquis C, et al. Development and validation of machine learning models for prediction of seizure outcome after pediatric epilepsy surgery. *Epilepsia*. 2022;63(8):1956-1969. doi:10.1111/epi.17320
7. Jehi L. Machine learning for precision epilepsy surgery. *Epilepsy Curr*. 2023;23(2):78-83. doi:10.1177/15357597221150055
8. Janssen KJM, Moons KGM, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol*. 2008;61(1):76-86. doi:10.1016/j.jclinepi.2007.04.018
9. Sussman JB, Wiitala WL, Zawistowski M, Hofer TP, Bentley D, Hayward RA. The Veterans Affairs Cardiac Risk Score: recalibrating the atherosclerotic cardiovascular disease score for applied use. *Med Care*. 2017;55(9):864-870. doi:10.1097/mlr.0000000000000781