

RESEARCH ARTICLE

Predicting seizure outcome after epilepsy surgery: Do we need more complex models, larger samples, or better data?

Maria H. Eriksson^{1,2,3,4}  | Mathilde Ripart¹  | Rory J. Piper^{1,5}  |
 Friederike Moeller⁶  | Krishna B. Das^{3,6}  | Christin Eltze⁶  | Gerald Cooray^{6,7}  |
 John Booth⁸  | Kirstie J. Whitaker⁴  | Aswin Chari^{1,5}  | Patricia Martin Sanfilippo^{1,2}  |
 Ana Perez Caballero⁹  | Lara Menzies¹⁰  | Amy McTague^{1,3}  | Martin M. Tisdall^{1,5}  |
 J. Helen Cross^{1,3,5,11}  | Torsten Baldeweg^{1,2}  | Sophie Adler¹  | Konrad Wagstyl¹² 

¹Developmental Neurosciences Research & Teaching Department, UCL Great Ormond Street Institute of Child Health, London, UK

²Department of Neuropsychology, Great Ormond Street Hospital, London, UK

³Department of Neurology, Great Ormond Street Hospital, London, UK

⁴The Alan Turing Institute, London, UK

⁵Department of Neurosurgery, Great Ormond Street Hospital, London, UK

⁶Department of Neurophysiology, Great Ormond Street Hospital, London, UK

⁷Clinical Neuroscience, Karolinska Institute, Solna, Sweden

⁸Digital Research Environment, Great Ormond Street Hospital, London, UK

⁹North Thames Genomic Laboratory Hub, Great Ormond Street Hospital, London, UK

¹⁰Department of Clinical Genetics, Great Ormond Street Hospital, London, UK

¹¹Young Epilepsy, Lingfield, UK

¹²Imaging Neuroscience, UCL Queen Square Institute of Neurology, London, UK

Correspondence

Maria H. Eriksson, Developmental Neurosciences Research & Teaching Department, UCL Great Ormond Street Institute of Child Health, 30 Guilford Street, London WC1N 1EH, UK.
 Email: m.eriksson.16@ucl.ac.uk

Funding information

National Institute for Health and Care Research Biomedical Research Centre Great Ormond Street Hospital; Rosetrees Trust, Grant/Award Number: A2665; Wellcome Trust, Grant/Award Number: 215901/Z/19/Z

Abstract

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

Sophie Adler and Konrad Wagstyl are joint last authors.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Epilepsia* published by Wiley Periodicals LLC on behalf of International League Against Epilepsy.

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% $CI_{MLP} = 67\%–74\%$, $AUC_{MLP} = .70$; 95% $CI_{XGBoost\ own} = 68\%–75\%$, $AUC_{XGBoost\ own} = .70$). 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.

KEYWORDS

epilepsy surgery, machine learning, pediatric, prediction

1 | INTRODUCTION

Despite careful evaluation, up to one third of patients with drug-resistant epilepsy are not rendered seizure-free through surgery.¹ This underscores the need to identify which patients are likely to benefit from surgery before the intervention has been carried out.

Surgical candidate selection is typically decided by a multidisciplinary team. This form of expert clinical judgment relies on experience and available evidence, and achieves a moderate degree of accuracy when predicting surgical success.² To aid clinical judgment, some studies have reported average estimates of seizure freedom for specific types of epilepsy (e.g., temporal lobe epilepsy).¹ Other studies have focused on identifying multiple predictors of postoperative seizure outcome, without taking into account how these predictors may interact.¹

In an effort to synthesize patient characteristics and provide objective predictions of seizure freedom, researchers have developed statistical models and calculated risk scores that can generate individualized predictions of outcome.^{3–5} These have included the Epilepsy Surgery Nomogram,³ the modified Seizure Freedom Score,⁴ and the

Key Points

- We trained three models – a logistic regression, a multilayer perceptron, and an XGBoost model – to predict seizure outcome and found that they performed equally well (AUC = .70-.72).
- We applied a previously published machine learning model to our center's patients and found that it underperformed (AUC = .62 on our cohort vs AUC = .73-.74 on the original cohorts).
- Expanding our cohort beyond its current size, up to sample sizes of $N = 2000$, would not provide substantial gains in model performance.
- We were able to improve model performance through data-driven feature selection.
- Future improvements in our ability to predict outcome will require improved feature selection, collaboration between epilepsy surgery services, data standardization, and model sharing.

Epilepsy Surgery Grading Scale.⁵ These tools do not, however, perform better than clinical judgment.^{2,6} Researchers are, therefore, increasingly turning to machine learning in an attempt to improve prediction accuracy.

Machine learning is being leveraged within the realm of clinical research at an exponential pace. The epilepsy surgery pathway generates a plethora of diverse data. As such, it would seem to create an ideal opportunity for the application of machine learning technology. Several machine learning models have indeed been developed to date to predict seizure outcome.^{7–29} The majority of these models have, however, been trained on relatively small sample sizes ($N < 100$)^{7,8,10,11,13–16,18–20,22,24–29} and, therefore, have a high risk of “overfitting” (a model overfits when it models the training data set too closely, performing well on this data set but consequently underperforming on new, “unseen” data sets).^{30,31} Model training sets have also been composed almost exclusively of temporal lobe surgery patients,^{8–11,13–18,21–27,29} often relied on postoperative factors,^{7,10,12,23,25–27} and frequently utilized postprocessing neuroimaging analyses that cannot be replicated readily by others.^{11,13–18,20,21,25,27–29} As such, many existing models may be difficult to incorporate into routine preoperative evaluation. Perhaps more importantly still, none of these models have been externally validated on a substantial cohort.^{17,20} It is, therefore, unknown how well they would perform if used by another surgery center, and whether their adoption as a replacement for traditional statistical modeling approaches is justified.

To advance this field, we asked whether (1) more complex models, (2) larger sample sizes, or (3) better selection

of clinical predictors would improve our ability to predict postoperative seizure outcome (Figure 1). To address the first question, we trained three different models—a traditional logistic regression and two machine learning models—to predict seizure outcome on our data set. We also tested the performance of an external, pre-trained machine learning model¹² on our data set and compared its performance to that of our models. To address the influence of sample size, we investigated how varying sample size—both within and extrapolating beyond our current cohort—impacted model performance. To address the influence of number and type of clinical predictors, we investigated how the inclusion of different predictors affected model performance.

2 | MATERIALS AND METHODS

2.1 | Patient cohort

We retrospectively reviewed medical records for all children who underwent epilepsy surgery at Great Ormond Street Hospital (GOSH; London, UK) from January 1, 2000, through December 31, 2018. We included patients who underwent surgical resection or disconnection. We excluded palliative procedures (corpus callosotomy and multiple subpial transections), as well as neuromodulation (deep brain stimulation and vagus nerve stimulation) and thermocoagulation procedures. If patients had undergone multiple surgeries over the course of the study period, we included only their first surgery.

What do we need to improve the prediction of seizure outcome after epilepsy surgery?

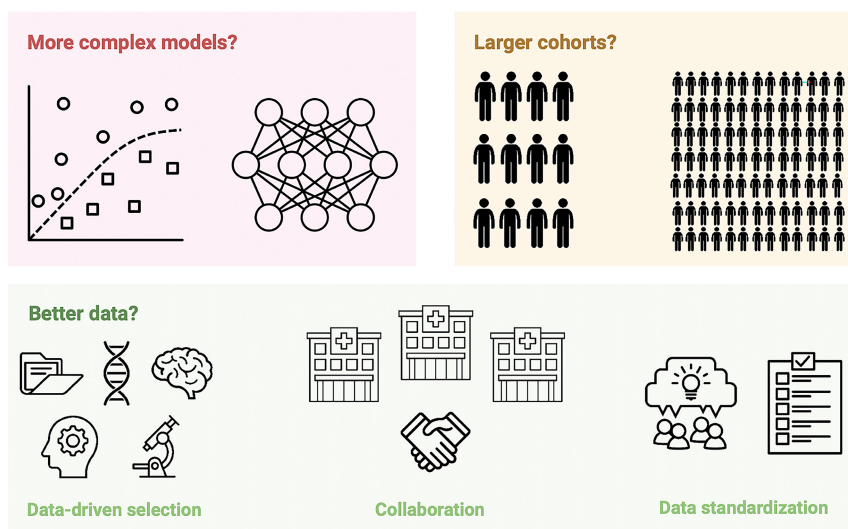


FIGURE 1 Study overview. We investigated the impact of model type, sample size, and feature selection on our ability to accurately predict postoperative seizure outcome.

2.2 | Data set description

We retrieved medical records and extracted the following information: patient demographics, epilepsy characteristics, preoperative magnetic resonance imaging (MRI) findings, preoperative interictal and ictal electroencephalography (EEG) characteristics, preoperative antiseizure medication (ASM; including both total number of ASM trialed from time of epilepsy onset to time of preoperative evaluation, as well as number of ASM at time of preoperative evaluation), surgery details, genetic results, and histopathology diagnosis. A complete list of variables extracted and information about how we categorized these data can be found in Appendix S1.

We classified patients as either seizure-free (including no auras) or not seizure-free at 1-year postoperative follow-up. We also recorded if patients were receiving, weaning, or off ASMs at this time point.

2.3 | Statistical analysis

We calculated the descriptive statistics for the cohort and presented these using mean with standard deviation, median with interquartile range, and count with proportion, as appropriate.

We checked if continuous data were normally distributed using Shapiro–Wilk tests.³² None of the continuous variables were normally distributed. We, therefore, investigated associations between demographic, clinical, and surgical variables using the Mann–Whitney *U*, Kruskal–Wallis *H*, chi-square test of independence, and Spearman's rank correlation coefficient, as appropriate. All tests were two-tailed, and we set the threshold for significance a priori at $p < .05$. We corrected for multiple comparisons using the Holm method.³³

We performed univariable logistic regression analyses to investigate which clinical variables predicted seizure outcome at 1-year postoperative follow-up. In the case of categorical variables, the group known to have the highest seizure freedom rate (according to past literature) was used as the reference category. All other groups were then compared to this reference category to determine if they were significantly less (or more) likely to achieve seizure freedom through surgery. For example, “unilateral MRI abnormalities” was selected as the reference category for the categorical variable “MRI bilaterality,” and we investigated whether those with “bilateral MRI abnormalities” were significantly less (or more) likely to be seizure-free after surgery. We again corrected for multiple comparisons using the Holm method.³³

2.3.1 | Effect of model type on model performance

We performed a multivariable logistic regression (LR) with independent variables that (1) could be obtained preoperatively and (2) were found to be predictive of seizure outcome. We developed a second version of this model, in which MRI diagnosis was replaced with histopathology diagnosis, to determine if this affected model performance.

We used the same predictors to train two machine learning models: a multilayer perceptron (MLP) and an XGBoost model. We chose an MLP due to its high predictive performance, allowing for nonlinear interactions between input variables. We trained the MLP with two hidden layers, with 5 and 10 hidden neurons respectively, balancing the need for sufficient complexity to learn feature interactions across multiple features, while limiting the capacity of the network to overfit to the training data. We chose an XGBoost model to ensure that we could compare the performance of this to the performance of the XGBoost model published by Yossofzai et al.¹²

After training our own three models, we applied the XGBoost model by Yossofzai et al.¹² to the same patient cohort. We evaluated the performance of all models using stratified 10-fold cross-validation. We used a stratified approach to address the outcome imbalance observed in our cohort. We calculated the null accuracy (the accuracy the model would achieve if it always predicted the more commonly occurring outcome in our cohort, i.e., seizure-free), the tested model accuracy, and the area under the receiver-operating characteristic (ROC) curve (AUC) for each model. We reported both the mean AUC obtained across all 10 folds as well as the AUC obtained from each individual fold. We compared the accuracies of the respective models using McNemar's test.

2.3.2 | Effect of sample size on model performance

We investigated how sample size affected model performance by using a previously described learning curve analysis approach.³⁴ First, we trained our models on 38 different sample sizes, starting at $N=20$ and finishing at $N=700$ patients. At each sample size, we evaluated model performance, specifically model accuracy. This allowed us to create a learning curve, plotting model performance against sample size. We then chose an inverse power law function to model the learning curve. We used this function to predict model performance on expanded sample sizes of up to $N=2000$.

2.3.3 | Effect of clinical predictors on model performance

We explored how the number of included predictors, as well as their nature, affected model performance. We used the coefficients from our univariable logistic regression analyses to determine how informative different predictors were. We then added significant predictors one-by-one into our models, from the most informative to the least informative. At each point, we plotted model AUC and confidence intervals (CIs; obtained across the 10 folds).

We performed all statistical analyses and visualizations in Python version 3.7.2 and R version 3.6.3. Our MLP and XGBoost models were implemented using the scikit-learn library.³⁵ The study's analytic code is available on GitHub (<https://github.com/MariaEriksson/Predicting-seizure-outcome-paper>).

3 | RESULTS

3.1 | Patient cohort

A total of 797 children were identified as having undergone first-time surgical resection or disconnection. Demographic information and clinical characteristics for these patients are displayed in [Table S1](#). Data relating to semiology (past seizures and seizures at time of preoperative evaluation) as well as interictal and ictal EEG characteristics are displayed in [Table S2](#). Genetic diagnoses are listed in [Tables S3](#) and [S4](#).

Seizure outcome at 1-year follow-up was available for 709 patients, of which 67% were seizure-free. Of these, 51% were receiving ASM, 34% were weaning ASM, and 15% were not receiving ASM.

3.2 | Relationships between variables

Relationships between demographic, clinical, and surgical variables are displayed in [Figure 2](#). Full statistics are reported in [Table S5](#).

3.3 | Univariable logistic regression analyses

Univariable logistic regression analyses identified the following features as predictive of 1-year postoperative seizure freedom: handedness, educational status, genetic findings, age of epilepsy onset, history of infantile spasms, spasms at time of preoperative evaluation, number of

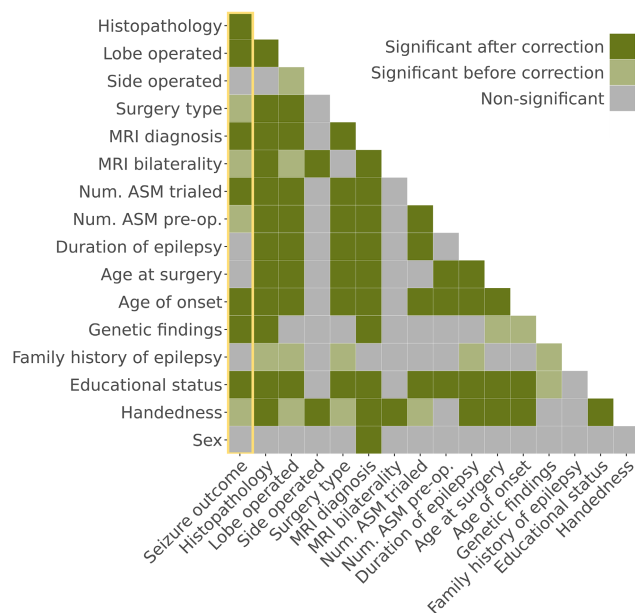


FIGURE 2 Relationships between demographic, clinical, and surgical variables. Relationships are shown both before and after correction for multiple comparison using the Holm method. We have highlighted relationships with seizure outcome using a yellow box. ASM, antiseizure medication; Num. ASM pre-op, number of antiseizure medications at time of preoperative evaluation; Num. ASM trialed, total number of different antiseizure medications trialed from epilepsy onset to preoperative evaluation.

seizure types at time of preoperative evaluation, total number of ASMs trialed (from time of epilepsy onset to time of preoperative evaluation), MRI bilaterality (unilateral vs bilateral MRI abnormalities), MRI diagnosis, type of surgery performed, lobe operated on, and histopathology diagnosis ([Table S6](#)).

3.4 | Effect of model type on model performance

3.4.1 | Logistic regression models

Our multivariable LR achieved an accuracy of 72% (95% CI=68%–75%) and an AUC of .72 (range across the 10 folds: .64–.82). When we assessed whether substituting MRI diagnosis with histopathology diagnosis would improve model performance, we found that this alternative LR achieved a similar accuracy of 73% (95% CI=69%–79%; AUC=.72; range across the 10 folds: .60–.77). There was no significant difference in performance between the LR that included MRI diagnosis and the LR that included histopathology diagnosis (McNemar's test, chi-square=.1, $p=.8$). This was likely due to the high degree of overlap between MRI and histopathology diagnoses ([Figure S1](#)).

3.4.2 | Multilayer perceptron and XGBoost models

Our MLP achieved an accuracy of 71% (95% CI=67%–74%) and an AUC of .70 (range across the 10 folds: .63–.82). Our XGBoost also achieved an accuracy of 71% (95% CI=68%–75%) and an AUC of .70 (range across the 10 folds: .62–.83).

3.4.3 | External XGBoost model

When we applied the XGBoost model developed by Yossofzai et al.¹² to our data, it achieved an accuracy of 63% (95% CI=59%–67%) and an AUC of .62.

3.4.4 | Comparison of model performances

The AUCs of the respective models are compared in Figure 3A. There was no significant difference in performance between our LR and MLP (McNemar's test, chi-square=.8, $p=.4$), our LR and XGBoost (McNemar's test, chi-square=.1, $p=.8$), or our MLP and XGBoost (McNemar's test, chi-square=.1, $p=.8$).

All three models performed better than the external XGBoost model (McNemar's test_{LR}, chi-square=8.0, $p=.005$; McNemar's test_{MLP}, chi-square=6.4, $p=.01$; McNemar's test_{XGB_own}, chi-square=6.8, $p=.01$). Our LR, MLP, and XGBoost models also performed significantly better than model null accuracy (McNemar's

test_{LR}, chi-square=8.7, $p=.003$; McNemar's test_{MLP}, chi-square=5.3, $p=.02$; McNemar's test_{XGB_own}, chi-square=7.6, $p=.006$), whereas the external XGBoost model did not (McNemar's test_{XGB_external}, chi-square=.6, $p=.4$).

3.5 | Effect of sample size on model performance

Increasing our sample size within the limits of our cohort improved the performances of all our models (Figure 3B). However, visual inspection of model performance at increasing sample sizes showed that model performance started to plateau at around $N=400$, after which point increases in sample size followed the law of diminishing returns. In the case of our LR, an increase from $N=20$ to $N=120$ led to a .08 increase in AUC (AUC=.593 vs AUC=.674). However, corresponding increases of 100 patients, from $N=200$ to $N=300$ patients and from $N=300$ to $N=400$ patients, led to .01 and <.01 increases in AUC, respectively (AUC=.689 vs .699 and AUC=.699 vs AUC=.705). Expanding our cohort beyond its current size, up to $N=2000$, did not substantially improve the performances of any of our models (Figure 3B).

3.6 | Effect of data inclusion on model performance

We found that adding more predictor features improved the performances of all models (Figure 4A and Figures S2

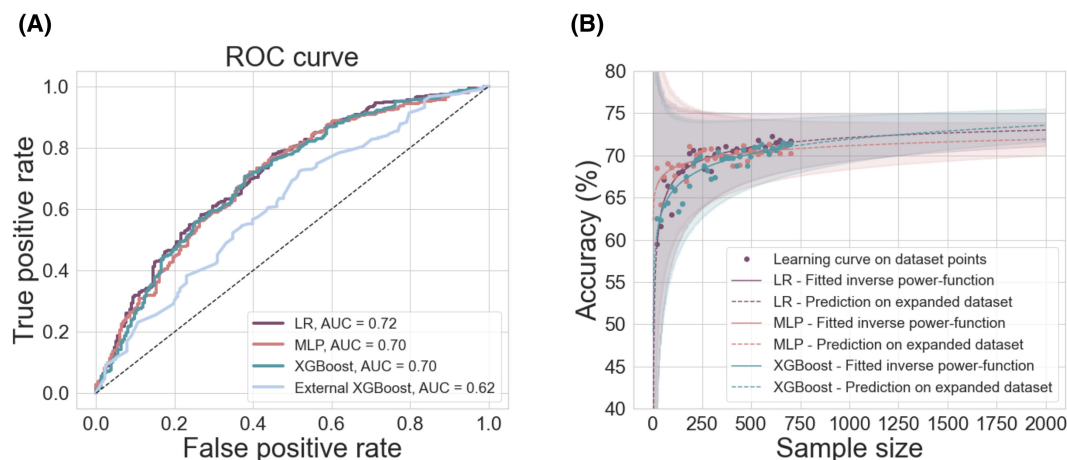


FIGURE 3 Impact of model type and sample size on model performance. (A) Receiver- operating characteristic (ROC) curves showing model performances. There was no significant difference in performance between our LR (purple), MLP (pink), and XGBoost (teal) models. All of our models performed significantly better than the XGBoost model recently developed by Yossofzai et al.¹² (light blue). (B) The effect of sample size on model performance (accuracy). There was an improvement in model performance with increasing sample size for our LR, MLP, and XGBoost models, but only up until a certain point. After this, the models showed only marginal gains in performance. Extrapolating performance for sample sizes up to $N=2000$ did not predict substantial improvement in model performance for any of our models. AUC, area under the (ROC) curve; LR, logistic regression; MLP, multilayer perceptron; ROC, receiver-operating characteristic.

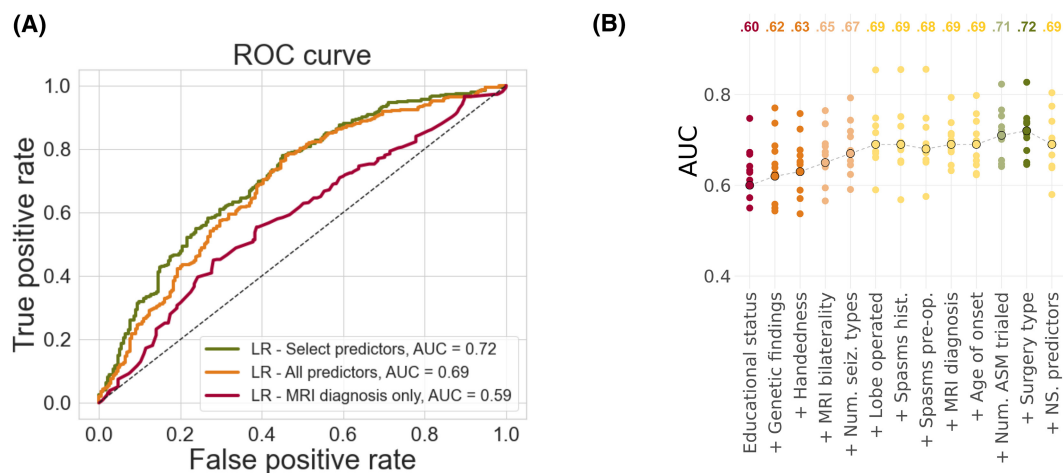


FIGURE 4 Impact of feature selection on model performance. (A) Receiver-operating characteristic (ROC) curves showing model performance for our LR models containing (1) only MRI diagnosis (red), (2) all predictors (orange), and (3) predictors identified through data-driven feature selection (green). Data-driven selection involved including only predictors that were significantly predictive of 1-year postoperative seizure outcome as identified in univariable logistic regression analyses. Corresponding ROC curves showing model performances for our MLP and XGBoost models are displayed in Figures S2 and S3. (B) Effect of data-driven feature selection on model performance (AUC). Variables found to be significantly predictive of seizure outcome from univariable logistic regression analyses were added to the LR, from most information to least informative according to their coefficients. Model performance was best when all significantly predictive features were included in the model. Adding the remaining predictors collected for the study, that is, those that were not significantly predictive of seizure outcome, worsened model performance (far right). Points circled in black represent mean AUC obtained across all 10 folds. Noncircled points represent the AUCs obtained from each of the individual 10 folds. ASM, antiseizure medication; AUC, area under the (ROC) curve; LR, logistic regression; NS. predictors, non-significant predictors; Num. ASM trialed, total number of different antiseizure medication trialed from epilepsy onset to preoperative evaluation; Num. seiz. types, number of seizure types at time of preoperative evaluation; ROC, receiver-operating characteristic; Spasms hist., history of spasms; Spasms pre-op, spasms at time of preoperative evaluation.

and S3). However, the greatest accuracy was achieved when data-driven feature selection was used to filter which clinical predictors should be included in the models (i.e., when the models included only the variables that were found to be significantly predictive of seizure outcome in our univariable logistic regression analyses; Figure 4B). When we added variables that were not significantly predictive of seizure outcome in our univariable logistic regression analyses, model performance worsened (Figure 4B).

4 | DISCUSSION

Up to one third of patients do not achieve seizure freedom through epilepsy surgery despite careful evaluation.¹ There has been a longstanding history of trying to identify these patients preoperatively, both through traditional statistical modeling approaches and more complex machine learning techniques.^{3–5,7–29} These attempts have, however, had limited success. In this study, we explored if we could improve our ability to predict seizure outcome by training more complex models, recruiting larger training sample sizes, or incorporating more or different types of clinical predictors.

To investigate the effect of model type on our ability to predict seizure outcome, we trained three different models, a logistic regression (or LR) and two machine learning models—a multilayer perceptron (or MLP) and an XGBoost—on the same cohort. We showed that our LR performed as well as our MLP and XGBoost models. We also applied a recently published XGBoost model by Yosssofzai et al.¹² to our cohort and found that this model performed worse than our models (AUC = .62 vs AUC = .70–.72). It also performed worse on our cohort compared to the cohorts it was trained and tested on (AUC = .62 vs AUC = .73–.74).

To address the value of larger patient sample sizes, we investigated model performance on a range of sample sizes, up to $N=2000$. We found that the performances of all models improved until around $N=400$, after which point they began to plateau.

To address the influence of clinical predictors, we varied both the number of predictors included in the models as well as the nature of these predictors. We demonstrated that using data-driven feature selection (i.e., including only variables that were predictive of seizure outcome in univariable logistic regression analyses) resulted in the best model performance, while including all collected

predictors led to a deterioration in model performance. Of interest, neither EEG nor semiology characteristics were predictive of seizure outcome in our univariable logistic regression analyses and were, therefore, not included in our models.

4.1 | The illusory superiority of more complex models

There is a growing tendency to favor machine learning technology over traditional statistical modeling approaches when training models to predict postoperative seizure outcome. This is presumably due to an assumed superiority of highly sophisticated or complex models. As a result, a plethora of machine learning techniques have been deployed.^{7–29} It is, however, also increasingly recognized that the potential gains in predictive accuracy that have been attributed to more complex algorithms may have been inflated,^{31,36} and that minor improvements observed “in the laboratory” may not translate into the real world.³¹

Previous studies that have used both machine learning techniques and traditional statistical modeling approaches to predict postoperative seizure outcome have found that logistic regression models perform as well as, or even better than, machine learning ones.^{8,9,25} To our knowledge, only one study by Yossofzai et al.¹² has found that a machine learning model outperforms a logistic regression; however, this was a .01–.02 difference in AUC (.72 vs .73 in the train data set; .72 vs .74 in the test data set). This small improvement is unlikely to deliver an advantage in clinical practice. At the same time, using machine learning models introduces complexity, which in turn complicates their interpretation, implementation, and validation, and increases the risk of overfitting.

4.2 | Larger samples mean higher accuracy... but only up until a certain point

There exists a general consensus in the machine learning community that more data, or larger sample sizes, equates to better model performance.^{37,38} However, researchers have started to show that this is not always the case.³⁹ We found that expanding our cohort beyond its current size ($N=797$) nearly three-fold did not provide meaningful gains.

Estimating the point of diminishing returns is invaluable because, although there is an abundance of unlabeled clinical data in our era of Big Data, (human) annotated clinical data remain scarce. Its creation is

time-consuming and requires the expertise of several clinical groups. Nevertheless, annotated data sets are essential in the creation of (supervised) learning algorithms. Generating learning curves can, therefore, inform researchers of the relative costs and benefits of adding additional annotated data to their model.⁴⁰ Still, it is important to note that this learning curve is only an estimate and that actual model performance could exceed these predictions. Oversampling techniques that generate synthetic data could provide a data set that is similar in size to our expanded (predicted) data set; however, these approaches carry a risk of overfitting, as the synthetic data that they generate may closely resemble the original data set in a way that new data may not. The only way to validate this prediction is, therefore, to collect a sample size of several thousands of patients.

4.3 | In pursuit of (geographical) model generalizability

Machine learning in clinical research is placing an increasing emphasis on model generalizability, where the highest level of evidence is achieved from applying models externally—to new centers. When we tested the model by Yossofzai et al.¹² on our data, we found that it did not generalize well. This may at first glance seem surprising, as there is a striking similarity between our cohort and the cohort of Yossofzai et al.¹²—not only in terms of sample size but also in terms of patient characteristics and variables found to be predictive of seizure outcome. However, it also highlights a common issue related to the use of machine learning, namely, the tendency for models to overfit to local data. We, therefore, expect that a similar decrease in model performance would be demonstrated if another center were to use the machine learning models that we trained.

Different epilepsy surgery centers show variation in which diagnostic and therapeutic procedures are available, for which patients they are requested, and with which specifications they are carried out.⁴¹ Local practices also influence how data are annotated. Clinical data are interpreted by experts who assign a wide range of labels, from MRI diagnosis to epilepsy syndromes. Although official classification systems for annotation procedures exist,^{42–47} individual studies often choose to—or are forced to—categorize their data ad hoc, primarily due to the restraints introduced by the retrospective nature of their data. Furthermore, not all experts will agree on the same label, which is evidenced by a lack of agreement regarding interpretation of EEG,^{48–50} MRI,⁵¹ positron emission

tomography (PET),⁵¹ and histopathological data.⁴² It is thus possible that although our cohort and the cohort of Yossofzai et al.¹² look similar on the surface, they may represent patients who have been characterized in a subtly different manner.

4.4 | Limitations of the current study

The primary limitation of our study is that it is a retrospective study, which uses data originally obtained to understand patient disease and support clinical care, rather than to enable data analysis. These data are, therefore, at risk of being biased and incomplete.

4.4.1 | Biased data

Presurgical evaluation is largely standardized in that all patients undergo a full clinical history, structural MRI, and scalp- or video-EEG, but the extent of further investigations will be patient dependent.⁵² To mitigate the occurrence of bias, we used a minimal data set, which included only clinical variables typically obtained for all epilepsy surgery patients. As such, we did not train our model using PET, single-photon emission computed tomography (SPECT), magnetoencephalography (MEG), or functional MRI (fMRI) measures. One exception to this was the inclusion of genetic diagnosis, which we included despite not all patients having undergone genetic testing. The predictive value of genetic information in surgery candidate selection has not been systematically investigated.⁵³ Consequently, we sought to contribute to this emerging area of research and provide initial evidence for its importance.

4.4.2 | Incomplete data

Related to the limitation of biased data is the limitation of incomplete data. Similar to past retrospective studies that have developed models for the prediction of seizure outcome after epilepsy surgery, we had a considerable amount of missing data. There are multiple ways of handling incomplete data sets, including deleting instances or replacing them with estimated values—a method known as imputation. Imputation techniques must, however, be used with caution, as they have limitations and can impact model performance.⁵⁴ We, therefore, chose to drop instances where continuous data points were missing before including them into the model training data sets, and classified missing categorical data points as such, rather than using imputation.

4.5 | Moving forward

Taken together, our findings suggest that (1) traditional statistical approaches such as logistic regression analyses are likely to perform as well as more complex machine learning models (when using routinely collected clinical predictors similar to those described here) and have advantages in interpretability, implementation, and generalizability; (2) collecting a large sample is important because it improves model performance and reduces overfitting, but including more than a thousand patients is unlikely to generate significant returns on data sets similar to ours; (3) model improvement is likely to come from data-driven feature selection and exploring the inclusion of features that have thus far been overlooked or not undergone external validation due to barriers in study replication (discussed below).

Based on these findings, we make recommendations to advance our ability to predict seizure outcome after epilepsy surgery (Table 1). Surgery centers around the world must collaborate to produce high-quality data for research purposes. Although models trained on single-center data sets are likely to produce higher model performances than multi-center data sets, they may not be suitable for use by other surgery centers. Critically, data must be collected and curated in a standardized manner, as highlighted by experts⁵⁵ and similar to recent multi-center endeavors.^{9,56,57} Here it will be important to distinguish between investigating variables that may be predictive of outcome and identifying variables that can (feasibly) be included as predictors in a clinical decision-making tool. For the purpose of developing a clinical decision-making

TABLE 1 Recommendations for future research.

1. Epilepsy surgery services should collaborate to create high-quality data sets for research purposes
2. Data collection, annotation, and categorization should be standardized across surgery centers
3. Variables included as predictors in a clinical decision-making tool should be limited to those that
 - (i) are routinely collected for all epilepsy surgery patients
 - (ii) can be obtained preoperatively, and
 - (iii) are significantly predictive of outcome
4. Data should be harmonized across surgery centers to tackle variability in data acquisition (e.g., variability induced by differences in MRI scanners and protocols)
5. Researchers should openly share their code on platforms (such as GitHub; <https://github.com>) to maximize transparency, support reproducibility, and enable external validation. In cases where code cannot be shared, researchers should share their models in a way that they can be validated by external centers

tool, we suggest including only variables that are routinely collected for all epilepsy surgery patients at most centers, to avoid introducing bias into the model. In other words, researchers should carefully consider the added value of modalities such as MEG, PET, SPECT, and fMRI. It is notable that only variables obtained prior to surgery should be included in the model, as the aim is to create a predictive model. This means excluding variables such as postoperative measurement of resection and histopathology diagnosis. Reassuringly, we have shown that MRI diagnosis provides information similar to histopathology diagnosis. We also echo past recommendations⁵³ in that we suggest avoiding variables that have repeatedly failed to predict outcome, as these have been shown to worsen model performance.

Training models using only clinical information is unlikely to procedure high model performance. Instead, better data must also entail new data. The inclusion of additional predictors to improve model performance may involve extracting quantitative features from preoperative MRI or EEG (as several studies have done^{13–29}), characterizing the epileptogenic network through computational modeling,⁵⁸ measuring lesion overlap with eloquent cortex,⁵⁹ or adopting a network analysis approach.⁶⁰ Here, it is important to note that machine learning techniques could provide superior performance compared to traditional statistical approaches if quantitative MRI and/or EEG features are used; however, to our knowledge, only one imaging study has to date compared these two approaches and found that they performed similarly well.²⁵

It is important that all model software is made available—either as ready-to-use tools or openly shared code on platforms such as GitHub. Past studies have reported models capable of achieving accuracies of >90% using quantitative features extracted from MRI and EEG^{14,15,19,28}; however, none of these findings can be reproduced, and none of these models can be adopted by other centers, as there is insufficient information about how they were generated. Yossofzai et al.¹² are to be commended for sharing their model in a way that allowed for it to be externally tested by ourselves and others.

5 | CONCLUSIONS

Accurate prediction of seizure outcome after epilepsy surgery remains difficult. We highlight the importance of comparing traditional statistical modeling to complex machine learning techniques, as we show that these two approaches may perform equally well. We also demonstrate the importance of performing external validation of machine learning models, as we show that algorithms may underperform on other centers' data. Based on our

findings, we present recommendations for future research, including the need for epilepsy services to collaborate in the creation of standardized data sets, the value of carefully choosing predictor variables for modeling, and the benefit of sharing models and code openly.

AUTHOR CONTRIBUTIONS

MHE, TB, SA, and KW conceived and designed the study. MHE, JB, FM, KBD, CE, GC, MMT, PMS, RJP, APC, LM, and AM retrieved, anonymized, curated, and verified the data. MHE, MR, TB, SA, and KW analyzed the data, interpreted the results, and produced the figures. MHE wrote the manuscript. All authors edited and approved the final draft of the manuscript.

ACKNOWLEDGMENTS

None.

FUNDING INFORMATION

This research is supported by the National Institute for Health Research Biomedical Research Centre at Great Ormond Street Hospital (NIHR GOSH BRC). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health. The NIHR GOSH BRC had no role in the manuscript or the decision to submit it for publication. AC and RJP are supported by GOSH Children's Charity Surgeon Scientist Fellowships. SA is funded by the Rosetrees Trust (A2665). KW is supported by the Wellcome Trust (215901/Z/19/Z). MHE is supported by a Child Health Research Studentship, funded by NIHR GOSH BRC.

CONFLICT OF INTEREST STATEMENT

JHC has acted as an investigator for studies with GW Pharmaceuticals, Zogenix, Vitaflo, Ovid, Marinius, and Stoke Therapeutics. She has been a speaker and on advisory boards for GW Pharmaceuticals, Zogenix, Biocodex, Stoke Therapeutics, and Nutricia; all remuneration has been paid to her department. She is president of the International League Against Epilepsy (2021–2025), and chair of the medical boards for Dravet UK, Hope 4 Hypothalamic Hamartoma, and Matthew's friends. MT has received grants from Royal Academy of Engineers and LifeArc. He has received honoraria from Medtronic. LM has received personal consultancy fees from Mendelian Ltd, outside the submitted work. AM has received honoraria from Biocodex and Nutricia, and provided consultancy to Biogen, outside the submitted work. All other authors report no disclosures relevant to the manuscript.

STUDY APPROVAL STATEMENT

The study was approved by the National Research Ethics Service and registered with the Joint Research and

Development Office of UCL Great Ormond Street Institute of Child Health and Great Ormond Street Hospital.

PATIENT CONSENT STATEMENT

Informed patient consent for this retrospective assessment of our own clinical data was waived, provided that the data were handled anonymously by the clinical care team.

ORCID

Maria H. Eriksson  <https://orcid.org/0000-0001-6869-5804>

Mathilde Ripart  <https://orcid.org/0000-0002-1761-5859>

Rory J. Piper  <https://orcid.org/0000-0002-6422-5853>


Friederike Moeller  <https://orcid.org/0000-0003-2434-2825>

Krishna B. Das  <https://orcid.org/0009-0006-8459-2382>

Christin Eltze  <https://orcid.org/0000-0003-4411-2939>

Gerald Cooray  <https://orcid.org/0000-0001-8321-7085>

John Booth  <https://orcid.org/0000-0003-4357-1324>

Kirstie J. Whitaker  <https://orcid.org/0000-0001-8498-4059>

Aswin Chari  <https://orcid.org/0000-0003-0053-147X>

Amy McTague  <https://orcid.org/0000-0002-0334-2909>

Martin M. Tisdall  <https://orcid.org/0000-0001-8880-8386>

J. Helen Cross  <https://orcid.org/0000-0001-7345-4829>

Torsten Baldeweg  <https://orcid.org/0000-0002-5724-1679>

Sophie Adler  <https://orcid.org/0000-0002-3978-7424>

Konrad Wagstyl  <https://orcid.org/0000-0003-3439-5808>

REFERENCES

- Widjaja E, Jain P, Demoe L, Guttmann A, Tomlinson G, Sander B. Seizure outcome of pediatric epilepsy surgery: systematic review and meta-analyses. *Neurology*. 2020;94(7):311–21. <https://doi.org/10.1212/WNL.0000000000008966>
- Gracia CG, Chagin K, Kattan MW, Ji X, Kattan MG, Crotty L, et al. Predicting seizure freedom after epilepsy surgery, a challenge in clinical practice. *Epilepsy Behav*. 2019;95:124–30. <https://doi.org/10.1016/j.yebeh.2019.03.047>
- Jehi L, Yardi R, Chagin K, Tassi L, Russo GL, Worrell G, et al. Development and validation of nomograms to provide individualised predictions of seizure outcomes after epilepsy surgery: a retrospective analysis. *Lancet Neurol*. 2015;14(3):283–90. [https://doi.org/10.1016/S1474-4422\(14\)70325-4](https://doi.org/10.1016/S1474-4422(14)70325-4)
- Garcia GC, Yardi R, Kattan MW, Nair D, Gupta A, Najm I, et al. Seizure freedom score: a new simple method to predict success of epilepsy surgery. *Epilepsia*. 2015;56(3):359–65. <https://doi.org/10.1111/epi.12892>
- Dugan P, Carlson C, Jetté N, Wiebe S, Bunch M, Kuzniecky R, et al. Derivation and initial validation of a surgical grading scale for the preliminary evaluation of adult patients with drug-resistant focal epilepsy. *Epilepsia*. 2017;58(5):792–800. <https://doi.org/10.1111/epi.13730>
- Fassin AK, Knake S, Strzelczyk A, Josephson CB, Reif PS, Haag A, et al. Predicting outcome of epilepsy surgery in clinical practice: prediction models vs. clinical acumen. *Seizure*. 2020;76:79–83. <https://doi.org/10.1016/j.seizure.2020.01.016>
- Arle JE, Perrine K, Devinsky O, Doyle W. Neural network analysis of preoperative variables and outcome in epilepsy surgery. *J Neurosurg*. 1999;90:998–1004.
- Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, Kastanauskaite A, de Sola RG, Defelipe J, et al. Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. *PLoS One*. 2013;8(4):e62819. <https://doi.org/10.1371/journal.pone.0062819>
- Benjumeda M, Tan Y, González Otárola KA, Chandramohan D, Chang EF, Hall JA, et al. Patient specific prediction of temporal lobe epilepsy surgical outcomes. *Epilepsia*. 2021;62(9):2113–22. <https://doi.org/10.1111/epi.17002>
- Grigsby J, Kramer RE, Schneiders JL, Gates JR, Brewster SW. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia*. 1998;39(1):61–6. <https://doi.org/10.1111/j.1528-1157.1998.tb01275.x>
- Memarian N, Kim S, Dewar S, Engel J, Staba RJ. Multimodal data and machine learning for surgery outcome prediction in complicated cases of mesial temporal lobe epilepsy. *Comput Biol Med*. 2015;64:67–78. <https://doi.org/10.1016/j.compbiomed.2015.06.008>
- Yossofzai O, Fallah A, Maniquis C, Wang S, Ragheb J, Weil AG, et al. Development and validation of machine learning models for prediction of seizure outcome after pediatric epilepsy surgery. *Epilepsia*. 2022;63(8):1956–69. <https://doi.org/10.1111/epi.17320>
- Antony AR, Alexopoulos AV, González-Martínez JA, Mosher JC, Jehi L, Burgess RC, et al. Functional connectivity estimated from intracarotid EEG predicts surgical outcome in intractable temporal lobe epilepsy. *PLoS One*. 2013;8(10):e77916. <https://doi.org/10.1371/journal.pone.0077916>
- Bernhardt BC, Hong SJ, Bernasconi A, Bernasconi N. Magnetic resonance imaging pattern learning in temporal lobe epilepsy: classification and prognostics: MRI profiling in TLE. *Ann Neurol*. 2015;77(3):436–46. <https://doi.org/10.1002/ana.24341>
- Feis DL, Schoene-Bake JC, Elger C, Wagner J, Tittgemeyer M, Weber B. Prediction of post-surgical seizure outcome in left mesial temporal lobe epilepsy. *Neuroimage Clin*. 2013;2:903–11. <https://doi.org/10.1016/j.nicl.2013.06.010>
- Gleichgerrcht E, Munsell B, Bhatia S, Vandergrift WA III, Rorden C, McDonald C, et al. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia*. 2018;59(9):1643–54. <https://doi.org/10.1111/epi.14528>
- Gleichgerrcht E, Keller SS, Drane DL, Munsell BC, Davis KA, Kaestner E, et al. Temporal lobe epilepsy surgical outcomes can be inferred based on structural connectome hubs: a machine learning study. *Ann Neurol*. 2020;88(5):970–83. <https://doi.org/10.1002/ana.25888>
- He X, Doucet GE, Pustina D, Sperling MR, Sharan AD, Tracy JJ. Presurgical thalamic “hubness” predicts surgical outcome in temporal lobe epilepsy. *Neurology*. 2017;88(24):2285–93. <https://doi.org/10.1212/WNL.0000000000004035>
- Hong SJ, Bernhardt BC, Schrader DS, Bernasconi N, Bernasconi A. Whole-brain MRI phenotyping in dysplasia-related frontal lobe epilepsy. *Neurology*. 2016;86(7):643–50. <https://doi.org/10.1212/WNL.0000000000002374>

20. Ibrahim GM, Sharma P, Hyslop A, Guillene M, Morgan B, Bhatia S. Presurgical thalamocortical connectivity is associated with response to vagus nerve stimulation in children with intractable epilepsy. *Neuroimage Clin.* 2017;16:634–42.
21. Johnson GW, Cai LY, Narasimhan S, González HFJ, Wills KE, Morgan VL, et al. Temporal lobe epilepsy lateralisation and surgical outcome prediction using diffusion imaging. *J Neurol Neurosurg Psychiatry.* 2022;93(6):599–608. <https://doi.org/10.1136/jnnp-2021-328185>
22. Kini LG, Thaker AA, Hadar PN, Shinohara RT, Brown MG, Dubroff JG, et al. Quantitative [18F]FDG PET asymmetry features predict long-term seizure recurrence in refractory epilepsy. *Epilepsy Behav.* 2021;116:107714. <https://doi.org/10.1016/j.yebeh.2020.107714>
23. Morita-Sherman M, Li M, Joseph B, Yasuda C, Vegh D, de Campos BM, et al. Incorporation of quantitative MRI in a model to predict temporal lobe epilepsy surgery outcome. *Brain Commun.* 2021;3(3):fcab164. <https://doi.org/10.1093/braincomms/fcab164>
24. Munsell BC, Wee CY, Keller SS, Weber B, Elger C, da Silva LAT, et al. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *Neuroimage.* 2015;118:219–30. <https://doi.org/10.1016/j.neuroimage.2015.06.008>
25. Sinclair B, Cahill V, Seah J, Kitchen A, Vivash LE, Chen Z, et al. Machine learning approaches for imaging-based prognostication of the outcome of surgery for mesial temporal lobe epilepsy. *Epilepsia.* 2022;63:1081–92. <https://doi.org/10.1111/epi.17217>
26. Sinha N, Wang Y, Moreira da Silva N, Miserocchi A, McEvoy AW, de Tisi J, et al. Structural brain network abnormalities and the probability of seizure recurrence after epilepsy surgery. *Neurology.* 2021;96(5):e758–71. <https://doi.org/10.1212/WNL.00000000000011315>
27. Taylor PN, Sinha N, Wang Y, Vos SB, de Tisi J, Miserocchi A, et al. The impact of epilepsy surgery on the structural connectome and its relation to outcome. *Neuroimage Clin.* 2018;18:202–14. <https://doi.org/10.1016/j.nicl.2018.01.028>
28. Tomlinson SB, Porter BE, Marsh ED. Interictal network synchrony and local heterogeneity predict epilepsy surgery outcome among pediatric patients. *Epilepsia.* 2017;58(3):402–11. <https://doi.org/10.1111/epi.13657>
29. Yankam Njiwa J, Gray KR, Costes N, Mauguire F, Ryvlin P, Hammers A. Advanced [18F]FDG and [11C]flumazenil PET analysis for individual outcome prediction after temporal lobe epilepsy surgery for hippocampal sclerosis. *Neuroimage Clin.* 2015;7:122–31. <https://doi.org/10.1016/j.nicl.2014.11.013>
30. Ying X. An overview of overfitting and its solutions. *J Phys Conf Ser.* 2019;1168:022022. <https://doi.org/10.1088/1742-6596/1168/2/022022>
31. Hand DJ. Classifier technology and the illusion of progress. *Stat Sci.* 2006;21(1):1–14. <https://doi.org/10.1214/08834230600000060>
32. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *J Stat Model Anal.* 2011;2(1):21–33.
33. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health.* 1996;86(5):726–8. <https://doi.org/10.2105/AJPH.86.5.726>
34. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. *BMC Med Inform Decis Mak.* 2012;12(1):8. <https://doi.org/10.1186/1472-6947-12-8>
35. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res.* 2011;12:2825–30.
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12–22. <https://doi.org/10.1016/j.jclinepi.2019.02.004>
37. Riley RD, Ensor J, Snell KIE, Harrell FE Jr, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ.* 2020;368:m441. <https://doi.org/10.1136/bmj.m441>
38. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol.* 2014;14(1):137. <https://doi.org/10.1186/1471-2288-14-137>
39. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med.* 2022;5(1):48. <https://doi.org/10.1038/s41746-022-00592-y>
40. Richter AN, Khoshgoftaar TM. Sample size determination for biomedical big data with limited labels. *Netw Model Anal Health Inform Bioinform.* 2020;9(1):12. <https://doi.org/10.1007/s13721-020-0218-0>
41. Harvey AS, Cross JH, Shinnar S, Mathern GW, The Pediatric Epilepsy Surgery Survey Taskforce. Defining the spectrum of international practice in pediatric epilepsy surgery patients. *Epilepsia.* 2008;49(1):146–55. <https://doi.org/10.1111/j.1528-1167.2007.01421.x>
42. Blümcke I, Coras R, Busch RM, Morita-Sherman M, Lal D, Prayson R, et al. Toward a better definition of focal cortical dysplasia: an iterative histopathological and genetic agreement trial. *Epilepsia.* 2021;62(6):1416–28. <https://doi.org/10.1111/epi.16899>
43. Blümcke I, Thom M, Aronica E, Armstrong DD, Bartolomei F, Bernasconi A, et al. International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: a Task Force report from the ILAE Commission on Diagnostic Methods. *Epilepsia.* 2013;54(7):1315–29. <https://doi.org/10.1111/epi.12220>
44. Blümcke I, Thom M, Aronica E, Armstrong DD, Vinters HV, Palmini A, et al. The clinicopathologic spectrum of focal cortical dysplasias: a consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission. *Epilepsia.* 2011;52(1):158–74. <https://doi.org/10.1111/j.1528-1167.2010.02777.x>
45. Zuberi SM, Wirrell E, Yozawitz E, Wilmshurst JM, Specchio N, Riney K, et al. ILAE classification and definition of epilepsy syndromes with onset in neonates and infants: position statement by the ILAE Task Force on Nosology and Definitions. *Epilepsia.* 2022;63(6):1349–97. <https://doi.org/10.1111/epi.17239>
46. Scheffer IE, Berkovic S, Capovilla G, Connolly MB, French J, Guilhoto L, et al. ILAE classification of the epilepsies: position paper of the ILAE Commission for Classification

- and Terminology. *Epilepsia*. 2017;58(4):512–21. <https://doi.org/10.1111/epi.13709>
47. Trinka E, Cock H, Hesdorffer D, Rossetti AO, Scheffer IE, Shinnar S, et al. A definition and classification of status epilepticus – report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*. 2015;56(10):1515–23. <https://doi.org/10.1111/epi.13121>
 48. Jing J, Herlopian A, Karakis I, Ng M, Halford JJ, Lam A, et al. Interrater reliability of experts in identifying interictal epileptiform discharges in electroencephalograms. *JAMA Neurol*. 2020;77(1):49. <https://doi.org/10.1001/jamaneurol.2019.3531>
 49. Piccinelli P, Viri M, Zucca C, Borgatti R, Romeo A, Giordano L, et al. Inter-rater reliability of the EEG reading in patients with childhood idiopathic epilepsy. *Epilepsy Res*. 2005;66(1–3):195–8. <https://doi.org/10.1016/j.eplepsyres.2005.07.004>
 50. Grant AC, Abdel-Baki SG, Weedon J, Arnedo V, Chari G, Koziorynska E, et al. EEG interpretation reliability and interpreter confidence: a large single-center study. *Epilepsy Behav*. 2014;32:102–7. <https://doi.org/10.1016/j.yebeh.2014.01.011>
 51. Struck AF, Westover MB. Variability in clinical assessment of neuroimaging in temporal lobe epilepsy. *Seizure*. 2015;30:132–5. <https://doi.org/10.1016/j.seizure.2015.06.011>
 52. Cross JH, Reilly C, Gutierrez Delicado E, Smith ML, Malmgren K. Epilepsy surgery for children and adolescents: evidence-based but underused. *Lancet Child Adolesc Health*. 2022;6(7):484–94. [https://doi.org/10.1016/S2352-4642\(22\)00098-0](https://doi.org/10.1016/S2352-4642(22)00098-0)
 53. Alim-Marvasti A, Vakharia VN, Duncan JS. Multimodal prognostic features of seizure freedom in epilepsy surgery. *J Neurol Neurosurg Psychiatry*. 2022;93(5):499–508. <https://doi.org/10.1136/jnnp-2021-327119>
 54. Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O. A survey on missing data in machine learning. *J Big Data*. 2021;8(1):140. <https://doi.org/10.1186/s40537-021-00516-9>
 55. Litt B. Engineers drive new directions in translational epilepsy research. *Brain*. 2022;145(11):3725–6. <https://doi.org/10.1093/brain/awac375>
 56. Lamberink HJ, Otte WM, Blümcke I, Braun KPJ, Aichholzer M, Amorim I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. *Lancet Neurol*. 2020;19(9):748–57. [https://doi.org/10.1016/S1474-4422\(20\)30220-9](https://doi.org/10.1016/S1474-4422(20)30220-9)
 57. Spitzer H, Ripart M, Whitaker K, D'Arco F, Mankad K, Chen AA, et al. Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection Study. *Brain*. 2022;145(11):3859–71. <https://doi.org/10.1093/brain/awac224>
 58. Sinha N, Dauwels J, Kaiser M, Cash SS, Brandon Westover M, Wang Y, et al. Predicting neurosurgical outcomes in focal epilepsy patients using computational modelling. *Brain*. 2017;140(2):319–32. <https://doi.org/10.1093/brain/aww299>
 59. Wagstyl K, Whitaker K, Raznahan A, Seidlitz J, Vértes PE, Foldes S, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: a MELD study. *Epilepsia*. 2022;63(1):61–74. <https://doi.org/10.1111/epi.17130>
 60. Chari A, Seunarine KK, He X, Tisdall MM, Clark CA, Bassett DS, et al. Drug-resistant focal epilepsy in children is associated with increased modal controllability of the whole brain and epileptogenic regions. *Commun Biol*. 2022;5(1):394. <https://doi.org/10.1038/s42003-022-03342-8>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eriksson MH, Ripart M, Piper RJ, Moeller F, Das KB, Eltze C, et al. Predicting seizure outcome after epilepsy surgery: Do we need more complex models, larger samples, or better data? *Epilepsia*. 2023;64:2014–2026. <https://doi.org/10.1111/epi.17637>