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

Authors' names, affiliations, email addresses, and ORCID IDs

Maria H Eriksson [1,2,3,4]	m.eriksson.16@ucl.ac.uk	0000-0001-6869-5804
Mathilde Ripart [1]	m.ripart@ucl.ac.uk	0000-0002-1761-5859
Rory J Piper [1,5]	rory.piper.20@ucl.ac.uk	0000-0002-6422-5853
Friederike Moeller [6]	friederike.moeller@gosh.nhs.uk	-
Krishna B Das [3,6]	krishna.das@gosh.nhs.uk	-
Christin Eltze [6]	christin.eltze@gosh.nhs.uk	-
Gerald Cooray [6,7]	gerald.cooray@gosh.nhs.uk	0000-0001-8321-7085
John Booth [8]	john.booth@gosh.nhs.uk	0000-0003-4357-1324
Kirstie J Whitaker [4]	kwhitaker@turing.ac.uk	0000-0001-8498-4059
Aswin Chari [1,5]	aswin.chari.18@ucl.ac.uk	0000-0003-0053-147X
Patricia Martin Sanfilippo [1,2]	patricia.martin-sanfilippo@gosh.nhs.uk	-
Ana Perez Caballero [9]	ana.perezcaballero@gosh.nhs.uk	-
Lara Menzies [10]	lara.menzies@gosh.nhs.uk	-
Amy McTague [1,3]	a.mctague@ucl.ac.uk	0000-0002-0334-2909
Martin M Tisdall [1,5]	martin.tisdall@gosh.nhs.uk	0000-0001-8880-8386
J Helen Cross [1,3,5,11]	h.cross@ucl.ac.uk	0000-0001-7345-4829
Torsten Baldeweg [1,2]	t.baldeweg@ucl.ac.uk	0000-0002-5724-1679
Sophie Adler [1†]	sophie.adler.13@ucl.ac.uk	0000-0002-3978-7424
Konrad Wagstyl [12†]	k.wagstyl@ucl.ac.uk	0000-0003-3439-5808

[†]These authors are joint last authors.

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

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

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

[4] The Alan Turing Institute, London, UK

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

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

[7] Clinical Neuroscience, Karolinska Institutet, Solna, Sweden

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

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

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

[11] Young Epilepsy, Lingfield, UK

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

Corresponding author's contact information

Maria Helena Eriksson

Developmental Neurosciences Research & Teaching Department,

UCL Great Ormond Street Institute of Child Health,

30 Guilford Street, London, UK, WC1N 1EH

Telephone: (+44) 0770 4233 537

E-mail: m.eriksson.16@ucl.ac.uk

Disclosures

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.

Funding statement

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. All other authors report no competing interests.

Author contribution statement

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.

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.

Data availability statement

The study's data dictionary, statistical analysis plan, and analytic code will be made available on GitHub (https://github.com/MariaEriksson/Predicting-seizure-outcome-paper). The full data are not publicly available due to their sensitive nature. Deidentified data will be made available from the corresponding author upon reasonable request.

Ethical publication statement

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

Word count

Abstract: 308 words

Manuscript: 3,995 words

Number of figures and tables

Figures: 3

Tables: 2

Key words

Epilepsy surgery, Pediatric, Prediction, Machine learning

Glossary

ASM = antiseizure medication; AUC = Area under the (ROC) curve; CI = confidence interval; CNV = copy number variation; EEG = Electroencephalography; fMRI = Functional magnetic resonance imaging; IQR = interquartile range; LR = Logistic regression; MEG = Magnetoencephalography; MLP = Multilayer perceptron; MRI = Magnetic resonance imaging; OR = odds ratio; PET = Positron emission tomography; ROC = Receiver operating characteristic; SNV = single nucleotide variation; SPECT = Single-photon emission computed tomography

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 post-operative seizure outcome. We also conducted the first external validation of a machine learning model trained to predict post-operative seizure outcome.

Methods: We performed a retrospective cohort study of 797 children who had undergone resective or disconnective epilepsy surgery at a single 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 one-year post-operative seizure outcome on our dataset. 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=2,000. Finally, we examined the impact of predictor selection on model performance.

Results: Our logistic regression achieved an accuracy of 72% (95% CI=68-75%, AUC=0.72), while our multilayer perceptron and XGBoost both achieved accuracies of 71% (95% CI_{MLP}=67-74%, AUC_{MLP}=0.70; 95% $CI_{XGBoost\ own}$=68-75%, $AUC_{XGBoost\ own}$=0.70). There was no significant difference in performance between our three models (all P>0.4) and they all performed better than the external XGBoost, which achieved an accuracy of 63% (95% CI=59-67%, AUC=0.62; P_{LR}=0.005, P_{MLP}=0.01, $P_{XGBoost\ own}$=0.01) on our data. All models showed improved performance with increasing sample size, with limited improvements above N=400. 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 post-operative seizure freedom. We instead propose that improved feature selection alongside collaboration, data standardization, and model sharing is required to advance the field.

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 judgement relies on experience and available evidence, and achieves a moderate degree of accuracy when predicting surgical success.[2] To aid clinical judgement, 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 post-operative 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 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 (**Table 1**). The majority of these models have, however, been trained on relatively small sample sizes ($N < 100$)[7–18] and therefore have a high risk of 'overfitting' (a model overfits when it models the training dataset too closely, performing well on this dataset but consequently underperforming on new, 'unseen' datasets).[19,20] Model training sets have also been comprised almost exclusively of temporal lobe surgery patients[7,8,10–13,15,17,18,21,22], often

relied on post-surgical factors[11,12,14,23], and frequently utilized post-processing neuroimaging analyses that cannot be readily replicated by others.[7,9–11,13,16–18,21] As such, many existing models may be difficult to incorporate into routine pre-surgical evaluation. Perhaps more importantly still, none of these models have been externally validated. 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 modelling 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 post-operative seizure outcome (**Fig. 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 dataset. We also tested the performance of an external, pre-trained machine learning model[23] on our dataset, 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.

Materials and methods

Patient cohort

We retrospectively reviewed medical records for all children who underwent epilepsy surgery at Great Ormond Street Hospital (GOSH; London, UK) from 2000 through 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 vagal nerve stimulation) and thermocoagulation procedures. If patients had undergone multiple surgeries over the course of the study period, we included only their first surgery.

Dataset description

We retrieved medical records and extracted the following information: patient demographics, epilepsy characteristics, pre-operative MRI findings, pre-operative interictal and ictal EEG characteristics, pre-operative antiseizure medication (ASM; including both total number of ASMs trialed from time of epilepsy onset to time of pre-surgical evaluation, as well as number of ASMs at time of pre-operative 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 **Supplementary Material** (p. 2-6).

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

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.[24] None of the continuous variables were normally distributed. We therefore investigated associations between demographic, clinical and surgical variables using 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 $P \leq 0.05$. We corrected for multiple comparisons using the Holm method.[25]

We performed univariable logistic regression analyses to investigate which clinical variables predicted seizure outcome at one-year post-operative 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.[25]

Effect of model type on model performance

We performed a multivariable logistic regression (LR) with independent variables that 1) could be obtained pre-surgically 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 accuracy.

We used the same predictors to train two machine learning models: a multilayer perceptron (MLP) model and an XGBoost model. We chose an MLP due to its high predictive performance, allowing for non-linear interactions between input variables. We trained the MLP with two hidden layers and ten 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.*[23]

After training our own three models, we applied the XGBoost model by Yossofzai *et al.*[23] 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 ROC (Receiver Operating Characteristic) 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.

Effect of sample size on model performance

We investigated how sample size affected model performance by using a previously described learning curve analysis approach.[26] First, we trained our models on ten different sample sizes, starting at *N*=70 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*=2,000.

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 (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.[27]

Results

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 **Supplementary Table 1**. Data relating to semiology (past seizures and seizures at time of pre-surgical evaluation) as well as interictal and ictal EEG characteristics are displayed in **Supplementary Table 2**. Genetic diagnoses are listed in **Supplementary Tables 3** and **4**.

Seizure outcome at one-year follow-up was available for 709 patients, of which 67% were seizure-free. Of these, 51% were on ASM, 34% were weaning ASM, and 15% were on no ASM.

Relationships between variables

Relationships between demographic, clinical and surgical variables are displayed in **Fig. 2**. Full statistics are reported in **Supplementary Table 5**.

Univariable logistic regression analyses

Univariable logistic regression analyses identified the following features as predictive of one-year post-operative seizure freedom: handedness, educational status, genetic findings, age of epilepsy onset, history of infantile spasms, spasms at time of pre-operative evaluation, number of seizure types at time of pre-operative evaluation, total number of ASMs trialed (from time of epilepsy onset to time of pre-operative evaluation), MRI bilaterality (unilateral versus bilateral abnormalities), MRI diagnosis, type of surgery performed, lobe operated on, and histopathology diagnosis (**Supplementary Table 6**).

Effect of model type on model performance

Logistic regression models

Our multivariable LR achieved an accuracy of 72% (95% CI=68-75%) and an AUC of 0.72 (range across the 10 folds: 0.64-0.82). 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=0.72; range across the 10 folds: 0.60-0.77). 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=0.1, *P*=0.8). This was likely due to the high degree of overlap between MRI and histopathology diagnosis (**Supplementary Fig. 1**).

Multilayer perceptron and XGBoost models

Our MLP achieved an accuracy of 71% (95% CI=67-74%) and an AUC of 0.70 (range across the 10 folds: 0.63-0.82). Our XGBoost also achieved an accuracy of 71% (95% CI=68-75%) and an AUC of 0.70 (range across the 10 folds: 0.62-0.83).

External XGBoost model

When we applied the XGBoost model developed by Yossofzai *et al.*[23] to our data, it achieved an accuracy of 63% (95% CI=59-67%) and an AUC of 0.62.

Comparison of model performances

The AUCs of the respective models are compared in **Fig. 3A**. There was no significant difference in performance between our LR and MLP (McNemar's test, chi-square=0.8, *P*=0.4), our LR and XGBoost (McNemar's test, chi-square=0.1, *P*=0.8), or our MLP and XGBoost (McNemar's test, chi-square=0.1, *P*=0.8).

All three models performed better than the external XGBoost model (McNemar's $test_{LR}$, chi-square=8.0, *P*=0.005; McNemar's $test_{MLP}$, chi-square=6.4, *P*=0.01; McNemar's $test_{XGB\ own}$, chi-square=6.8, *P*=0.01). Our LR, MLP and XGBoost models also performed significantly better than model null accuracy (McNemar's $test_{LR}$, chi-square=8.7, *P*=0.003; McNemar's $test_{MLP}$, chi-square=5.3, *P*=0.02; McNemar's $test_{XGB\ own}$, chi-square=7.6, *P*=0.006), whereas the external XGBoost model did not (McNemar's $test_{XGB\ external}$, chi-square=0.6, *P*=0.4).

Effect of sample size on model performance

Increasing our sample size within the limits of our cohort improved the performances of all our models (**Fig. 3B**). This was, however, only true up until around *N*=400, at which point performance started to plateau for all models. Expanding our cohort beyond its current size, up to *N*=2,000, did not substantially improve the performances for any of our models (**Fig. 3B**).

Effect of data inclusion on model performance

We found that adding more predictor features improved the performances of all models (**Fig. 3C** and **Supplementary Fig. 2** and **3**). 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; **Fig. 3D**). When we added variables that were not significantly predictive of seizure outcome in our univariable logistic regression analyses, model performance worsened (**Fig. 3D**).

Discussion

Up to one third of patients do not achieve seizure freedom through epilepsy surgery despite careful evaluation.[1] There has been a long-standing history of trying to identify these patients pre-operatively, both through traditional statistical modelling approaches and more complex machine learning techniques (**Table 1**). 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 (LR) and two machine learning models – a multilayer perceptron (MLP) and an XGBoost – on the same cohort. We showed that our LR performed as well as our MLP and XGBoost models. Importantly, we also applied a recently published XGBoost model by Yossofzai *et al.*[23] to our cohort, and found that this model performed worse than our models (AUC=0.62 versus AUC=0.70-0.72). It also performed worse on our cohort compared to the cohorts it was trained and tested on (AUC=0.62 versus AUC=0.73-0.74).

To address the value of larger patient sample sizes, we investigated model performance on a range of sample sizes, up to N=2,000. 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. Interestingly, neither EEG nor semiology characteristics were predictive of seizure outcome in our univariable logistic regression analyses and therefore not included in our models.

The illusory superiority of more complex models

There is a growing tendency to favor machine learning technology over traditional statistical modelling approaches when training models to predict post-operative 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 (**Table 1**). 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[20,28], and that minor improvements observed "in the laboratory" may not translate into the real-world.[20]

Previous studies that have used both machine learning techniques and traditional statistical modelling approaches to predict post-operative seizure outcome have found that logistic regression models perform as well as, or even better than, machine learning ones.[11,15,22] To our knowledge, only one study by Yossofzai *et al.*[23] has found that a machine learning model outperforms a logistic regression; however, this was a 0.1-0.2 difference in AUC (0.72 versus 0.73 in the train dataset; 0.72 versus 0.74 in the test dataset). 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.

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.[29,30] However, researchers have started to show that this is not always the case.[31] 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 – whilst there is an abundance of unlabeled clinical data in our era of Big Data – (human) annotated clinical data remains scarce. Its creation is time-consuming and requires the expertise of several clinical groups. Nevertheless, annotated datasets 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.[32] Still, it is important to note that this learning curve is only an estimate and that actual model performance could exceed these predictions.

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.*[23] 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.*[23] – not only in terms of sample size, but also in terms of patient characteristics and variables found to be predictive of 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.[33] 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. Whilst official classification systems for annotation procedures exist[34–39], individual studies often choose to – or are forced to – categorize their data ad hoc, often 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[40–42], MRI[43], PET[43] and histopathological data[34]. It is thus possible that while our cohort and the cohort of Yossofzai *et al.*[23] look similar on the surface, they may represent patients who have been characterized in a subtly different manner.

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.

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.[44] To mitigate the occurrence of bias, we used a minimal dataset, which included only clinical variables typically obtained for all epilepsy surgery patients. As such, we did not train our model using positron emission tomography (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.[45] We therefore sought to contribute to this emerging area of research and provide initial evidence for its importance.

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 datasets, 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.[46] We therefore chose to drop instances where continuous data points were missing before including them into the model training datasets, and classified missing categorical data points as such, rather than using imputation.

Moving forward

Taken together, our findings suggest that 1) traditional statistical approaches such as logistic regression are likely to perform as well as more complex machine learning models (when using 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 datasets 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 2**). Surgery centers around the world must collaborate to produce high-quality data for *research* purposes. Although models trained on single center data are likely to produce higher model performances than multicenter datasets, 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 experts[47] and similar to recent multicenter endeavours.[22,48,49] 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 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. Importantly, 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 post-operative measurement of resection and histopathology diagnosis. Reassuringly, we have shown that MRI diagnosis provides similar information to histopathology diagnosis. We also echo past recommendations[45] in that we suggest avoiding variables that have repeatedly failed to predict outcome, as these have been shown to worsen model performance.

It is unlikely that clinical information alone will procedure high model performance, as demonstrated both here and by previous studies (**Table 1**). Instead, better data must also entail new data. The inclusion of additional predictors to improve model performance may involve extracting quantitative features from pre-operative MRI or EEG (as several studies detailed in **Table 1** have done), characterizing the epileptogenic network through

computational modelling[50], measuring lesion overlap with eloquent cortex[51], or adopting a network analysis approach.[52]

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 as high as 90-100% using features extracted from MRI and EEG (**Table 1**); however, none of these findings can be reproduced, nor can any of these models be adopted by other centers, as there is insufficient information about how they were generated. Yossofzai *et al.*[23] are to be commended for sharing their model in a way that allowed for it to be externally tested by ourselves and others.

Conclusions

Accurate prediction of seizure outcome after epilepsy surgery remains difficult. We highlight the importance of comparing traditional statistical modelling 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 datasets, the value of carefully choosing predictor variables for modelling, and the benefit of sharing models and code openly.

References

1. 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):311-321. 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 & Behavior*. 2019;95:124-130. doi:10.1016/j.yebeh.2019.03.047

3. Jehi L, Yardi R, Chagin K, et al. Development and validation of nomograms to provide individualised predictions of seizure outcomes after epilepsy surgery: a retrospective analysis. *The Lancet Neurology*. 2015;14(3):283-290. doi:10.1016/S1474-4422(14)70325-4

4. Garcia Gracia C, Yardi R, Kattan MW, et al. Seizure freedom score: A new simple method to predict success of epilepsy surgery. *Epilepsia*. 2015;56(3):359-365. doi:10.1111/epi.12892

5. Dugan P, Carlson C, Jetté N, 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. doi:10.1111/epi.13730

6. Fassin AK, Knake S, Strzelczyk A, et al. Predicting outcome of epilepsy surgery in clinical practice: Prediction models vs. clinical acumen. *Seizure*. 2020;76:79-83. doi:10.1016/j.seizure.2020.01.016

7. Yankam Njiwa J, Gray KR, Costes N, Mauguiere 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: Clinical*. 2015;7:122-131. doi:10.1016/j.nicl.2014.11.013

8. Munsell BC, Wee CY, Keller SS, et al. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. *NeuroImage*. 2015;118:219-230. doi:10.1016/j.neuroimage.2015.06.008

9. Tomlinson SB, Porter BE, Marsh ED. Interictal network synchrony and local heterogeneity predict epilepsy surgery outcome among pediatric patients. *Epilepsia*. 2017;58(3):402-411. doi:10.1111/epi.13657

10. 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: Clinical*. 2013;2:903-911. doi:10.1016/j.nicl.2013.06.010

11. Sinclair B, Cahill V, Seah J, et al. Machine learning approaches for imaging□based prognostication of the outcome of surgery for mesial temporal lobe epilepsy. *Epilepsia*. 2022;63:epi.17217. doi:10.1111/epi.17217

12. Grigsby J, Kramer RE, Schneiders JL, Gates JR, Brewster Smith W. Predicting outcome of anterior temporal lobectomy using simulated neural networks. *Epilepsia*. 1998;39(1):61-66. doi:10.1111/j.1528-1157.1998.tb01275.x

13. Antony AR, Alexopoulos AV, González-Martínez JA, et al. Functional connectivity estimated from intracranial EEG predicts surgical outcome in intractable temporal lobe epilepsy. Zochowski M, ed. *PLoS ONE*. 2013;8(10):e77916. doi:10.1371/journal.pone.0077916

14. Arle JE, Perrine K, Devinsky O, Doyle W. Neural network analysis of preoperative variables and outcome in epilepsy surgery. *J Neurosurg*. 1999;90:12.

15. Armañanzas R, Alonso-Nanclares L, DeFelipe-Oroquieta J, et al. Machine learning approach for the outcome prediction of temporal lobe epilepsy surgery. Gross RE, ed. *PLoS ONE*. 2013;8(4):e62819. doi:10.1371/journal.pone.0062819

16. Ibrahim GM, Sharma P, Hyslopd A, Guillene M, Morgan B, Bhatiaa S. Presurgical thalamocortical connectivity is associated with response to vagus nerve stimulation in children with intractable epilepsy. *Neuroimage Clin*. 2017;16:9.

17. 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. *Computers in Biology and Medicine*. 2015;64:67-78. doi:10.1016/j.compbiomed.2015.06.008

18. Gleichgerrcht E, Munsell B, Bhatia S, et al. Deep learning applied to whole-brain connectome to determine seizure control after epilepsy surgery. *Epilepsia*. 2018;59(9):1643-1654. doi:10.1111/epi.14528

19. Ying X. An overview of overfitting and its solutions. *J Phys: Conf Ser*. 2019;1168:022022. doi:10.1088/1742-6596/1168/2/022022

20. Hand DJ. Classifier technology and the illusion of progress. *Statist Sci*. 2006;21(1). doi:10.1214/088342306000000060

21. 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):436-446. doi:10.1002/ana.24341

22. Benjumeda M, Tan Y, González Otárula KA, et al. Patient specific prediction of temporal lobe epilepsy surgical outcomes. *Epilepsia*. 2021;62(9):2113-2122. doi:10.1111/epi.17002

23. 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

24. Razali NM, Wah YB. Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. :14.

25. Aickin M, Gensler H. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health*. 1996;86(5):726-728. doi:10.2105/AJPH.86.5.726

26. 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. doi:10.1186/1472-6947-12-8

27. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *MACHINE LEARNING IN PYTHON*.:6.

28. 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. *Journal of Clinical Epidemiology*. 2019;110:12-22. doi:10.1016/j.jclinepi.2019.02.004

29. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. Published online March 18, 2020:m441. doi:10.1136/bmj.m441

30. 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. doi:10.1186/1471-2288-14-137

31. Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *npj Digit Med*. 2022;5(1):48. doi:10.1038/s41746-022-00592-y

32. Richter AN, Khoshgoftaar TM. Sample size determination for biomedical big data with limited labels. *Netw Model Anal Health Inform Bioinforma*. 2020;9(1):12. doi:10.1007/s13721-020-0218-0

33. 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):146-155. doi:10.1111/j.1528-1167.2007.01421.x

34. Blümcke I, Coras R, Busch RM, et al. Toward a better definition of focal cortical dysplasia: An iterative histopathological and genetic agreement trial. *Epilepsia*. 2021;62(6):1416-1428. doi:10.1111/epi.16899

35. Blümcke I, Thom M, Aronica E, 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):1315-1329. doi:10.1111/epi.12220

36. Blümcke I, Thom M, Aronica E, et al. The clinicopathologic spectrum of focal cortical dysplasias: A consensus classification proposed by an ad hoc Task Force of the ILAE Diagnostic Methods Commission1: The ILAE Classification System of FCD. *Epilepsia*. 2011;52(1):158-174. doi:10.1111/j.1528-1167.2010.02777.x

37. Zuberi SM, Wirrell E, Yozawitz E, 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):1349-1397. doi:10.1111/epi.17239

38. Scheffer IE, Berkovic S, Capovilla G, et al. ILAE classification of the epilepsies: Position paper of the ILAE Commission for Classification and Terminology. *Epilepsia*. 2017;58(4):512-521. doi:10.1111/epi.13709

39. Trinka E, Cock H, Hesdorffer D, et al. A definition and classification of status epilepticus - Report of the ILAE Task Force on Classification of Status Epilepticus. *Epilepsia*. 2015;56(10):1515-1523. doi:10.1111/epi.13121

40. Jing J, Herlopian A, Karakis I, et al. Interrater reliability of experts in identifying Interictal epileptiform discharges in electroencephalograms. *JAMA Neurol*. 2020;77(1):49. doi:10.1001/jamaneurol.2019.3531

41. Piccinelli P, Viri M, Zucca C, et al. Inter-rater reliability of the EEG reading in patients with childhood idiopathic epilepsy. *Epilepsy Research*. 2005;66(1-3):195-198. doi:10.1016/j.eplepsyres.2005.07.004

42. Grant AC, Abdel-Baki SG, Weedon J, et al. EEG interpretation reliability and interpreter confidence: A large single-center study. *Epilepsy & Behavior*. 2014;32:102-107. doi:10.1016/j.yebeh.2014.01.011

43. Struck AF, Westover MB. Variability in clinical assessment of neuroimaging in temporal lobe epilepsy. *Seizure*. 2015;30:132-135. doi:10.1016/j.seizure.2015.06.011

44. Cross JH, Reilly C, Gutierrez Delicado E, Smith ML, Malmgren K. Epilepsy surgery for children and adolescents: evidence-based but underused. *The Lancet Child & Adolescent Health*. 2022;6(7):484-494. doi:10.1016/S2352-4642(22)00098-0

45. 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. doi:10.1136/jnnp-2021-327119

46. 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. doi:10.1186/s40537-021-00516-9

47. Litt B. Engineers drive new directions in translational epilepsy research. *Brain*. 2022;145(11):3725-3726. doi:10.1093/brain/awac375

48. Lamberink HJ, Otte WM, Blümcke I, et al. Seizure outcome and use of antiepileptic drugs after epilepsy surgery according to histopathological diagnosis: a retrospective multicentre cohort study. *The Lancet Neurology*. 2020;19(9):748-757. doi:10.1016/S1474-4422(20)30220-9

49. Spitzer H, Ripart M, Whitaker K, et al. Interpretable surface-based detection of focal cortical dysplasias: a Multi-centre Epilepsy Lesion Detection study. *Brain*. Published online August 12, 2022:awac224. doi:10.1093/brain/awac224

50. Sinha N, Dauwels J, Kaiser M, et al. Predicting neurosurgical outcomes in focal epilepsy patients using computational modelling. *Brain*. 2017;140(2):319-332. doi:10.1093/brain/aww299

51. Wagstyl K, Whitaker K, Raznahan A, et al. Atlas of lesion locations and postsurgical seizure freedom in focal cortical dysplasia: A MELD study. *Epilepsia*. 2022;63(1):61-74. doi:10.1111/epi.17130

52. Chari A, Seunarine KK, He X, 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. doi:10.1038/s42003-022-03342-8

Figure legends

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

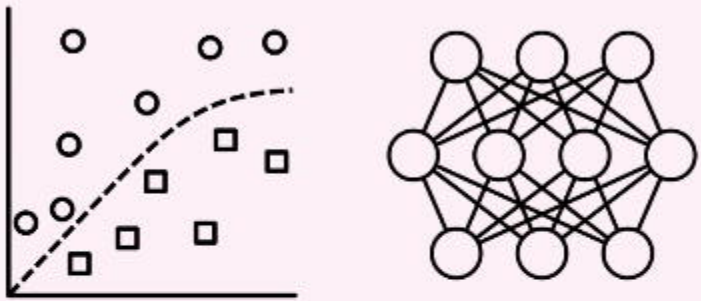
Figure 2 Relationships between demographic, clinical and surgical variables. Relationships are shown both before and after correction for multiple comparison using the Holm method. We have highlighted relationships with seizure outcome using a yellow box.

Figure 3 Impact of model type, sample size, and selection of clinical variables on model performance. (A) Receiver operating characteristic (ROC) curves showing model performances. There was no significant difference in performance between our LR (purple), MLP (turquoise), and XGBoost (blue) models. All of our models performed significantly better than the XGBoost model recently developed by Yossofzai *et al.*[23] (pink). **(B)** Effect of sample size on model performance (accuracy). There was an improvement in model performance with increasing sample size for our LR, MPL and XGBoost models, up until around *N*=400. After this point, the model showed only marginal gains in performance. Extrapolating performance for sample sizes up to *N*=2,000 did not predict substantial improvement in model performance for any of our models. **(C)** 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 one-year seizure outcome as identified in univariable logistic regression analyses. Corresponding ROC curves showing model performances for our MLP and XGBoost models are displayed in **Supplementary Fig. 2** and **3**. **(D)** 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 one-by-one to the LR, from most information to least informative according to their coefficients. Model performance increased in line with the predictors being added. Adding the remaining predictors collected for the study, i.e. those that were not significantly predictive of seizure outcome, worsened model performance (far right). Points circled in black represent mean

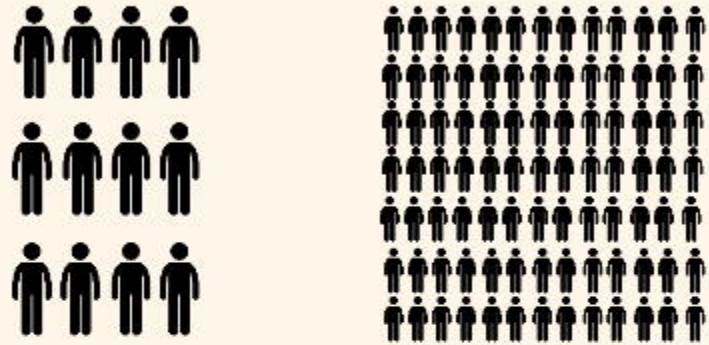
AUC obtained across all 10 folds. Non-circled points represent the AUCs obtained from each of the individual 10 folds.

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

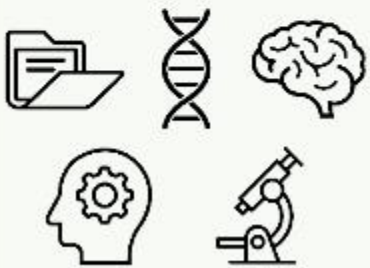
More complex models?



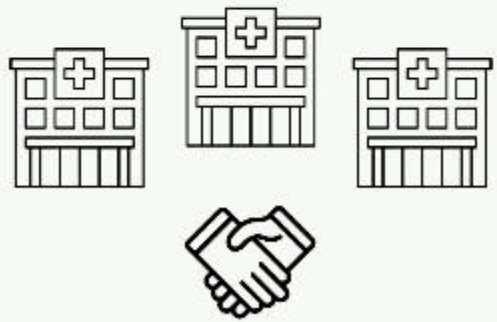
Larger cohorts?



Better data?

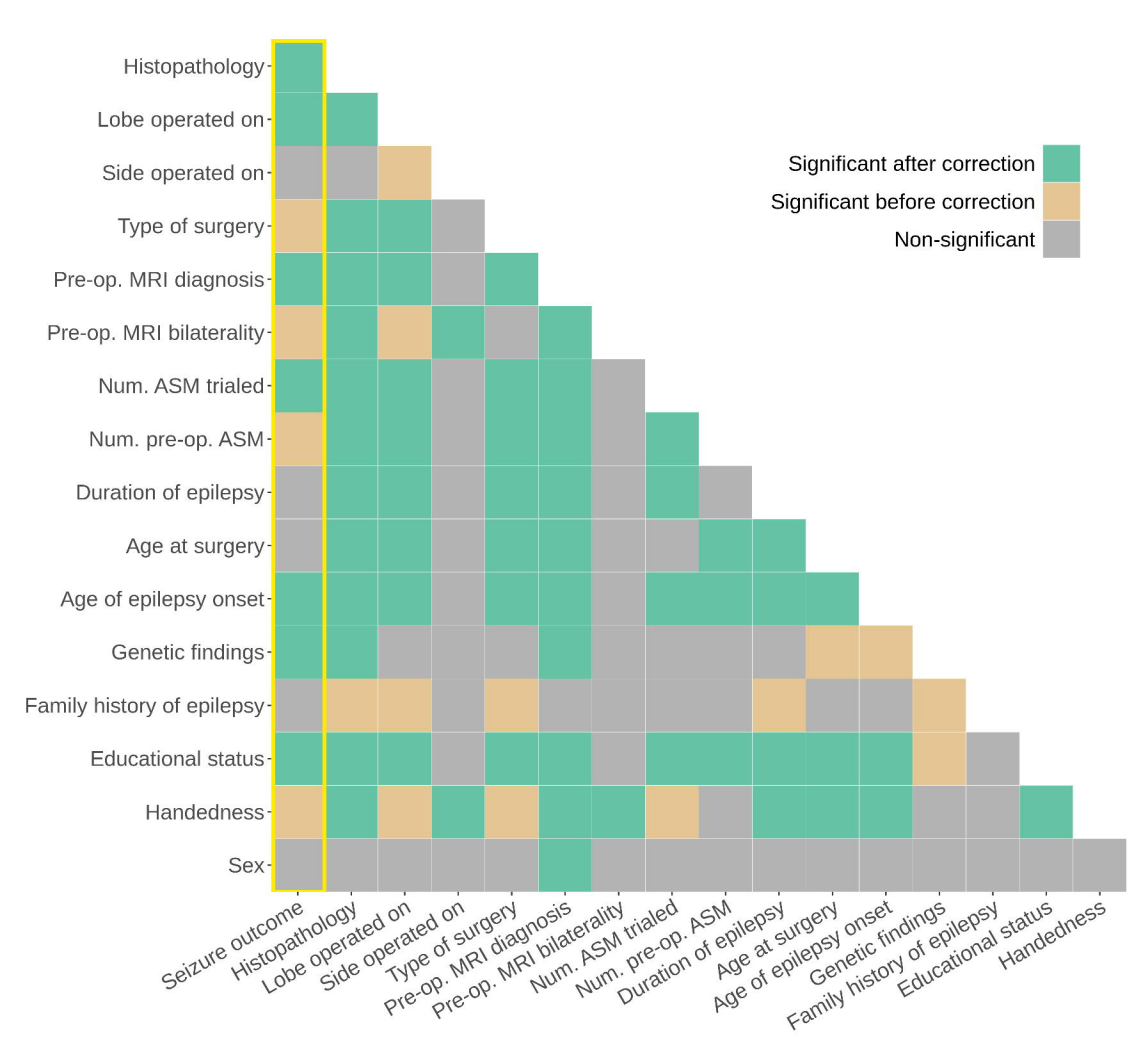


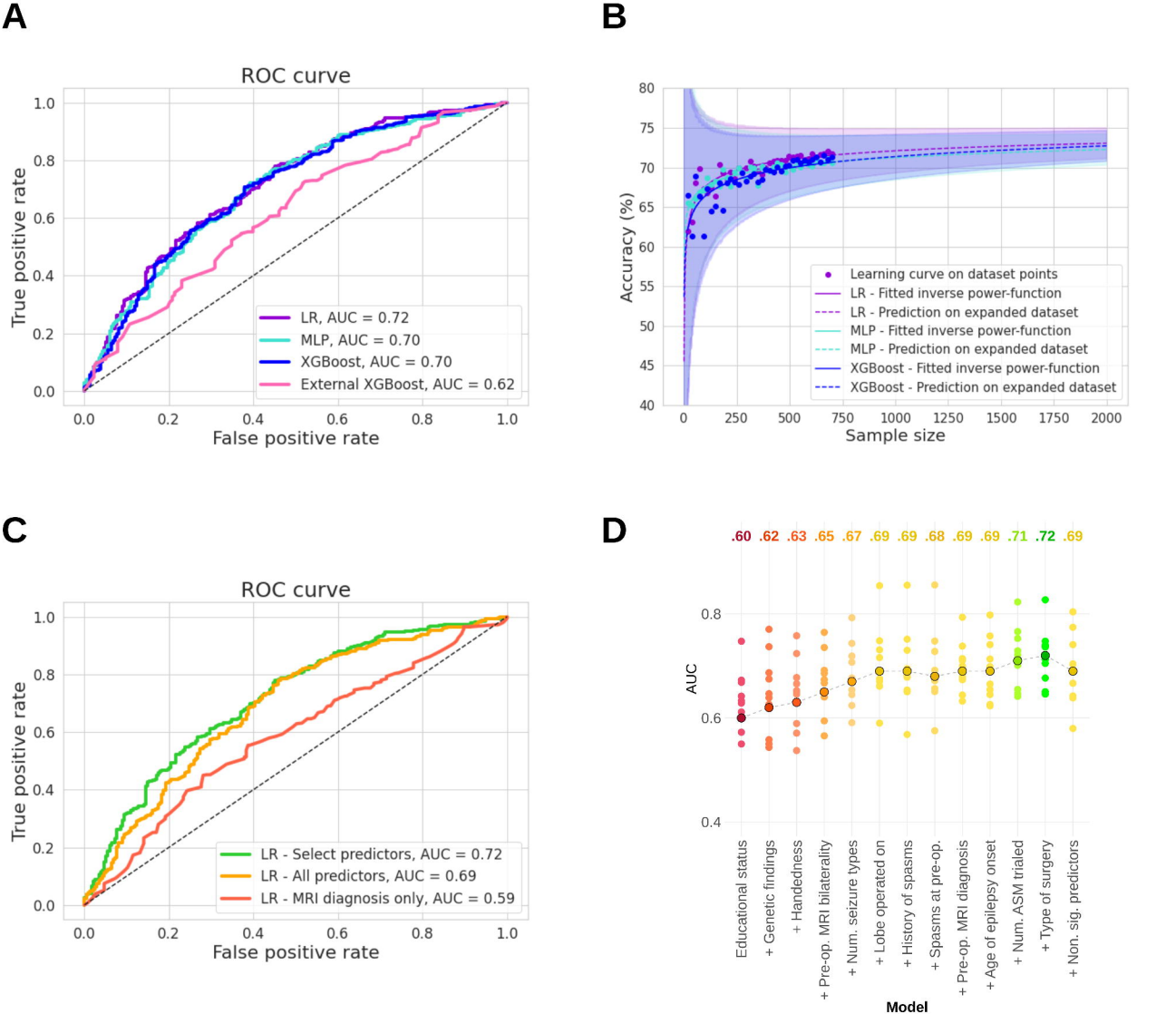
Data-driven selection



Collaboration

Data standardization


Histopathology
Lobe operated on
Side operated on
Type of surgery
Pre-op. MRI diagnosis
Pre-op. MRI bilaterality
Num. ASM trialed
Num. pre-op. ASM
Duration of epilepsy
Age at surgery
Age of epilepsy onset
Genetic findings
Family history of epilepsy
Educational status
Handedness
Sex
Seizure outcome
Histopathology
Lobe operated on
Side operated on
Type of surgery
Pre-op. MRI diagnosis
Pre-op. MRI bilaterality
Num. ASM trialed
Num. pre-op. ASM
Duration of epilepsy
Age at surgery
Age of epilepsy onset
Genetic findings
Family history of epilepsy
Educational status
Handedness
Significant after correction
Significant before correction
Non-significant


A
ROC curve
True positive rate
False positive rate
LR, AUC = 0.72
MLP, AUC = 0.70
XGBoost, AUC = 0.70
External XGBoost, AUC = 0.62
B
Accuracy (%)
Sample size
Learning curve on dataset points
LR - Fitted inverse power-function
LR - Prediction on expanded dataset
MLP - Fitted inverse power-function
MLP - Prediction on expanded dataset
XGBoost - Fitted inverse power-function
XGBoost - Prediction on expanded dataset
C
ROC curve
True positive rate
False positive rate
LR - Select predictors, AUC = 0.72
LR - All predictors, AUC = 0.69
LR - MRI diagnosis only, AUC = 0.59
D
.60 .62 .63 .65 .67 .69 .69 .68 .69 .69 .71 .72 .69
AUC
Educational status
+ Genetic findings
+ Handedness
+ Pre-op. MRI bilaterality
+ Num. seizure types
+ Lobe operated on
+ History of spasms
+ Spasms at pre-op.
+ Pre-op. MRI diagnosis
+ Age of epilepsy onset
+ Num. ASM trialed
+ Type of surgery
+ Non. sig. predictors
Model