Received: 18 October 2021 | Revised: 9 February 2022 | Accepted: 7 March 2022

DOI: 10.1111/epi.17217

Epilepsia®

RESEARCH ARTICLE

Machine learning approaches for imaging-based prognostication of the outcome of surgery for mesial temporal lobe epilepsy

Benjamin Sinclair[1,2] | Varduhi Cahill[3,4,5,6] | Jarrel Seah[7] | Andy Kitchen[1] | Lucy E. Vivash[1,2] | Zhibin Chen[1,3] | Charles B. Malpas[1,2,3,6,8] | Marie F. O'Shea[8,9] | Patricia M. Desmond[10] | Rodney J. Hicks[11] | Andrew P. Morokoff[12] | James A. King[12] | Gavin C. Fabinyi[13] | Andrew H. Kaye[14] | Patrick Kwan[1,2,3,6] | Samuel F. Berkovic[9,15] | Meng Law[1,7] | Terence J. O'Brien[1,2,3,6]

[1] Department of Neuroscience, Central Clinical School, Monash University, Melbourne, Victoria, Australia

[2] Department Neurology, Alfred Health, Melbourne, Victoria, Australia

[3] Department of Medicine, University of Melbourne, Melbourne, Victoria, Australia

[4] Academic Neurology Unit, Royal Hallamshire Hospital, University of Sheffield, Sheffield, UK

[5] Division of Neuroscience and Experimental Psychology, School of Biological Sciences, University of Manchester, Manchester, UK

[6] Department of Neurology, Melbourne Brain Centre, Royal Melbourne Hospital, Melbourne, Victoria, Australia

[7] Department of Radiology, Alfred Health, Melbourne, Victoria, Australia

[8] Melbourne School of Psychological Sciences, University of Melbourne, Melbourne, Victoria, Australia

[9] Comprehensive Epilepsy Program, Austin Health, Melbourne, Victoria, Australia

[10] Department of Radiology, Royal Melbourne Hospital, University of Melbourne, Melbourne, Victoria, Australia

[11] Peter MacCallum Cancer Centre and Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia

[12] Department of Surgery, Royal Melbourne Hospital, University of Melbourne, Melbourne, Victoria, Australia

[13] Department of Surgery, Austin Hospital, University of Melbourne, Melbourne, Victoria, Australia

[14] Department of Neurosurgery, Hadassah Hebrew University Hospital, Jerusalem, Israel

[15] Epilepsy Research Centre, Austin Hospital, University of Melbourne, Melbourne, Victoria, Australia

Correspondence
Benjamin Sinclair, Department of Neuroscience, Central Clinical School, Monash University, 99 Commercial Road, Melbourne, Victoria 3004, Australia.
Email: ben.sinclair@monash.edu

Funding information
National Health and Medical Research Council, Grant/Award Number: #APP1176426 and #APP1091593.

Abstract

Objectives: Around 30% of patients undergoing surgical resection for drug-resistant mesial temporal lobe epilepsy (MTLE) do not obtain seizure freedom. Success of anterior temporal lobe resection (ATLR) critically depends on the careful selection of surgical candidates, aiming at optimizing seizure freedom while minimizing postoperative morbidity. Structural MRI and FDG-PET neuroimaging are routinely used in presurgical assessment and guide the decision to proceed to surgery. In this study, we evaluate the potential of machine learning techniques applied to standard presurgical MRI and PET imaging features to provide enhanced prognostic value relative to current practice.

Methods: Eighty two patients with drug resistant MTLE were scanned with FDG-PET pre-surgery and T1-weighted MRI pre- and postsurgery. From these images the following features of interest were derived: volume of temporal lobe (TL) hypometabolism, % of extratemporal hypometabolism, presence of contralateral TL hypometabolism, presence of hippocampal sclerosis, laterality of seizure onset volume of tissue resected and % of temporal lobe hypometabolism resected. These measures were used as predictor variables in logistic regression, support vector machines, random forests and artificial neural networks.

Results: In the study cohort, 24 of 82 (28.3%) who underwent an ATLR for drug-resistant MTLE did not achieve Engel Class I (i.e., free of disabling seizures) outcome at a minimum of 2 years of postoperative follow-up. We found that machine learning approaches were able to predict up to 73% of the 24 ATLR surgical patients who did not achieve a Class I outcome, at the expense of incorrect prediction for up to 31% of patients who did achieve a Class I outcome. Overall accuracies ranged from 70% to 80%, with an area under the receiver operating characteristic curve (AUC) of .75–.81. We additionally found that information regarding overall extent of both total and significantly hypometabolic tissue resected was crucial to predictive performance, with AUC dropping to .59–.62 using presurgical information alone. Incorporating the laterality of seizure onset and the choice of machine learning algorithm did not significantly change predictive performance.

Significance: Collectively, these results indicate that "acceptable" to "good" patient-specific prognostication for drug-resistant MTLE surgery is feasible with machine learning approaches utilizing commonly collected imaging modalities, but that information on the surgical resection region is critical for optimal prognostication.

KEYWORDS

epilepsy, FDG-PET, machine learning, surgery

1 | INTRODUCTION

Surgical resection is the treatment of choice for drug-resistant mesial temporal lobe epilepsy (MTLE) and provides the patient with the best chance of seizure control.[1] However, a significant proportion (approximately 30%) of apparently good surgical candidates do not obtain sustained freedom from disabling seizures postoperatively.[2,3] Furthermore, in terms of risk–benefit analysis, a meta-analysis of surgical complications in epilepsy found that medical complications occur in 7.0% of temporal lobe epilepsy (TLE) surgeries, and neurological complications occur in 15.6%.[4] Complications from epilepsy surgery include infection, hemorrhage, depression, memory loss, language deficit, and visual impairment, and thus surgery is best avoided if it is unlikely to provide relief from seizures. Thorough presurgical evaluation is a prerequisite for estimating the rates of patients having sustained freedom from disabling seizures postoperatively and associated surgical morbidity. Brain imaging, in particular using structural magnetic resonance imaging (MRI) and positron emission tomography (PET), is critical to patient selection and surgical planning,[5] primarily as a means of

Key Points

- Machine learning approaches utilizing FDG-PET hypometabolism distribution, structural MRI, and surgical resection region information can successfully predict surgical outcome “better than chance” following an ATLR for drug-resistant MTLE
- Classification performance was "acceptable" to "good" (AUC = .75–.81)
- Information on surgical resection region is critical for optimal classification performance

assisting in localizing the epileptogenic zone and prognosticating the surgical outcome. Complete resection of the epileptogenic zone is strongly related to postsurgical outcome with respect to seizures.[6,7] Although presence of hippocampal sclerosis on MRI has long been used as an indicator of good prognosis for surgery,[8,9] it has been shown that patients with no identifiable potentially epileptogenic lesion on MRI, but focal temporal hypometabolism detected on fluorodeoxyglucose (FDG) PET, can also enjoy comparable rates of excellent outcomes with respect to seizure control following surgery.[10–14]

Imaging-derived features currently used in clinical practice are generally used in a qualitative, rather than a quantitative, manner. Although multiple streams of imaging evidence are considered in tandem when making surgical decisions, these multiple imaging sources are not routinely combined in any quantitative multivariate way to objectively enhance their localizing or predictive value. Furthermore, studies investigating the predictive value of imaging studies for postoperative seizure control typically measure cohort-wide associations, and do not give personalized outcome predictions. A fully automated prognostication tool that could predict the probability of seizure control based on imaging data for a given patient would be an important step toward delivering personalized patient care. This problem lends itself to the application of a machine learning approach. "Machine learning" is an umbrella term for a set of computing techniques aimed at learning patterns from data. Classification tasks, such as outcomes following epilepsy surgery (i.e., sustained freedom from disabling seizure free vs. ongoing disabling seizures postoperatively), are well suited to machine learning approaches, where a priori unknown features in data may classify outcomes with greater accuracy than traditional statistical associations.

In recent work, our group investigated a number of associations between the distribution of hypometabolism on a preoperative FDG-PET and postsurgical seizure control in a cohort of 82 TLE patients from two comprehensive epilepsy programs in Melbourne, Australia.[15] We found imaging features, such as the presence of contralateral FDG-PET hypometabolism and lower volumes of FDG-PET temporal lobe (TL) hypometabolism resected in surgery, were significantly associated with poorer surgical outcome, and interestingly that these associations were dependent on the laterality of the epileptogenic focus.

In this study, we utilized this same cohort to explore the extent to which those same imaging features can predict sustained freedom from disabling seizures (i.e., an Engel Class I outcome at 2 years postoperatively[16]) for individual patients, using a range of the most widely used and powerful methods in machine learning: logistic regression (LR), support vector machine (SVM), random forest (RF), and artificial neural network (ANN). We addressed three further pertinent research questions. First, do models that are able to detect hidden interactions between variables improve classification performance compared to fully specified models? Second, to what extent does the previously reported interaction between TLE laterality, imaging variables, and seizure outcome impact on machine learning classification performance? Third, can preoperative information alone yield satisfactory predictions, or is information about the surgical resection necessary?

2 | MATERIALS AND METHODS

2.1 | Subjects

The sample for this study was previously described by Cahill et al.[15] Subjects were patients in the Comprehensive Epilepsy Programs of the Royal Melbourne and Austin Hospitals, Melbourne, Australia, who underwent an anterior TL resection (ATLR) for treatment of drug-resistant MTLE between 2001 and 2014. Inclusion criteria were age > 16 years, hippocampal sclerosis or no identifiable lesion on an epilepsy protocol MRI (i.e., patients with focal cortical dysplasia, cavernomas, etc. were excluded), concordant results of presurgical investigations including seizure semiology and interictal/ictal electroencephalography, presence of ipsilateral hypometabolism on FDG-PET, and at least 2 years of follow-up following surgery. The study was approved by the Melbourne Health and Austin Health Human Research Ethics Committees.

2.2 | Imaging

As part of their presurgical evaluation, patients had an FDG-PET and a T1-weighted MRI scan (magnetization-prepared rapid acquisition gradient echo) acquired. FDG-PET scans were acquired on an Allegro (Phillips Medical Systems) at Austin Hospital with a voxel size of 2 × 2 × 2 mm or a Discovery 690 (GE Medical Systems) at Peter MacCallum Cancer Centre with a voxel size of 1.82 × 1.82 × 3.27 mm as described previously.[17] The median timing of the FDG-PET scans was 5 months preceding surgery (interquartile range = 3–10.25, range = 1–23 months). At Austin Hospital, MRI examinations were carried out on a Genesis Signa 1.5 T (GE Medical Systems) until 2005, and on a Magnetom Avanto 1.5 T (Siemens Medical Solutions) thereafter. Voxel sizes were .41 × .41 × 1.50 mm until 2006, .65 × .65 × 1.50 mm until 2011, and .77 × .77 × .80 mm thereafter. At Royal Melbourne Hospital, MRI examinations were carried out on a Genesis Signa 1.5 T (GE Medical Systems) until

2005 and on a Magnetom Trio Tim 3 T (Siemens Medical Solutions) thereafter. Voxel sizes varied depending on clinical requirements and scanner upgrades; all were higher resolution than 1 × 1 × 1 mm^3.

Following the surgery, patients underwent a repeat MRI to assess the extent of the surgical resection. Eighty-two patients were identified who had a full set of presurgical FDG-PET and MRI and postsurgical MRI images.

2.3 | Image processing

The image processing was as for Cahill et al.[15] Briefly, in SPM12,[18] pre- and postoperative MRIs were nonlinearly registered and masked using brain tissue segmentation, and the difference between masks was calculated to give the region of resection. FDG-PETs were normalized to the Montreal Neurological Institute template and compared to 20 healthy controls. Regions of hypometabolism were extracted using a two-sample t-test, thresholded at $p < .001$. Statistical maps were inverted back to subject space and then transformed to MRI space by coregistration of the FDG-PET to preoperative MRI. Hypometabolism maps were then compared to resection regions to calculate percentage of TL hypometabolism resected (Figure 1). The presence of hypometabolism in the contralateral TL was visually assessed on the statistical maps by two fellowship-trained neurologists/epileptologists. All continuous variables were standardized to zero mean and unit variance.

2.4 | Machine learning

All algorithms were coded and executed in sci-kit learn, a Python machine learning toolbox. Predictor variables for the machine learning algorithms were those considered in Cahill et al.[15] Presurgical variables were volume of TL hypometabolism, percentage of extratemporal hypometabolism, presence of contralateral TL hypometabolism, presence of hippocampal sclerosis, and laterality of seizure onset. Surgical variables were volume of tissue resected and percentage of TL hypometabolism resected. Engel classification of seizure outcomes[16] was used as the outcome variable (i.e., Engel Class I vs. Class II–IV). Distribution of predictor variables between outcome classes is shown in Table 1. Logistic regression was trained with each input variable included as a linear term and no interaction terms applied. The SVM was trained with radial basis function used as the kernel. The ANN was constructed in TensorFlow using the keras application

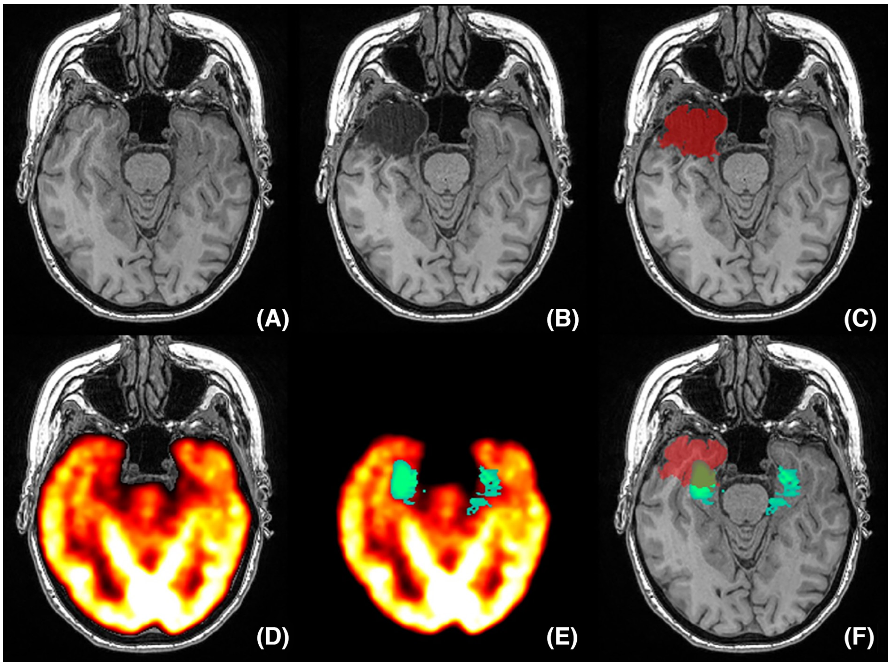

FIGURE 1 Images for a patient who underwent a right anterior temporal lobe resection for drug-resistant mesial temporal lobe epilepsy, who had contralateral mesial hypometabolism (in addition to the ipsilateral hypometabolism) on a preoperative fluorodeoxyglucose positron emission tomography (FDG-PET), and who did not achieve seizure freedom at 2-year follow up. (A) Preoperative magnetic resonance imaging (MRI). (B) postoperative MRI. (C) Subtraction of segmented preoperative and postoperative MRIs (red), used to calculate volume of tissue resected. (D) FDG-PET coregistered to MRI. (E) Hypometabolism (green–blue) measured by comparison to 20 healthy controls. (F) Overlay of resection region with hypometabolism (shaded green–blue), used to calculate the percentage of temporal lobe hypometabolism resected

TABLE 1 Overview of predictive variables for seizure-free patients (Engel I) and non-seizure-free patients (Engel II–IV)

Variable	Engel Class I	Engel Class II–IV	*p*
Presurgical			
Volume of TL hypometabolism, mm^3, median (IQR)	5.82 (2.62–12.55) × 10^3	6.06 (2.77–9.34) × 10^3	.710
Extratemporal hypometabolism, %, median (IQR)	54.2 (35.4–70.1)	61.8 (40.0–77.5)	.180
Presence of contralateral TL hypometabolism, *n* (%)	10/58 (82.8)	11/24 (54.2)	.012[a]
Presence of hippocampal sclerosis, *n* (%)	53/58 (91.4)	17/24 (70.8)	.034[a]
Laterality of seizure onset, left, *n* (%)	28/58 (48.3)	11/24 (45.8)	1.000[a]
Surgical			
Volume of tissue resected, mm^3, median (IQR)	20.66 (15.81–24.14) × 10^3	14.77 (11.17–21.34) × 10^3	.034
% of TL hypometabolism resected, median (IQR)	50.4 (34.5–67.2)	32.9 (20.9–61.3)	.070

Note: Statistical significance of group differences is presented.

Abbreviations: IQR, interquartile range; TL, temporal lobe.

[a] Fisher exact test, otherwise Mann–Whitney *U* test.

programming interface. It had a shallow architecture, with the number of input nodes equal to the (variable) number of input variables, two hidden layers with seven nodes in each layer, and a single output node. Activations functions were ReLU for hidden layers and sigmoid for the output layer, with Adam used as the optimizer, and batch normalization applied with 100 epochs. Additional hyperparameters are described in Section 2.7.

2.5 | Data augmentation

Because our data are imbalanced, with 70.7% achieving seizure freedom, and only 29.3% were not seizure-free, SMOTE resampling (synthetic minority oversampling technique[19]) was performed on the training set. This interpolates within the minority class to generate additional data points for the minority class, ensuring an approximately equal number of training data points for each outcome.

2.6 | Classification performance

To estimate the out-of-sample performance, an estimate of how the algorithm would perform on unseen data, a stratified 10-fold cross-validation with 10 random repeats was performed on each of the classification algorithms. Briefly, the data are split into approximately 10 equally sized samples, with stratification such that an equal proportion of seizure-free and non-seizure-free patients are in each sample. In each fold, nine samples are combined to form the training set to fit model parameters, and the remaining sample is used as a test set to measure prediction performance in an unseen sample. The classification performance was evaluated using area under the receiver operating characteristic curve (AUC), accuracy (total proportion correctly predicted), sensitivity (proportion of Engel Class I correctly predicted), specificity (proportion of Engel Class II–IV correctly predicted), positive predictive value (PPV; probability positive prognosis is correct), and negative predictive value (NPV; probability negative prognosis is correct).

2.7 | Hyperparameter optimization

LR, SVM, and ANN contain a regularization parameter specifying the penalization of model complexity in the cost function, and RF contains a number of model structure parameters that perform a similar function (number of trees in the forest, maximum depth, number of features per split). These parameters are arbitrary but affect model performance. To choose the best hyperparameters, we optimized on the training set of each fold of the parent cross-validation using a Bayesian search cross-validation. This searches (hyper)parameter space using Bayesian optimization. For each hyperparameter set in the search, a fivefold cross-validation was used to measure classification performance. The following hyperparameter search spaces were specified: LR, penalty term (C) = .1–10; SVM, penalty term (C) = 1–10; RF, n_trees = 3–100, maximum (max) depth = 3–30, max features = sqrt(*n* features) or n features; ANN, L2 regularisation (l2) = .0001–.01. The optimal set of hyperparameters and their associated model weights (parameters) were used to measure classification performance on the held-out test sample of the parent 10 × 10-fold cross-validation.

2.8 | Model comparisons

We hypothesize that (1) machine learning algorithms able to detect not explicitly hypothesized interactions between variables (SVM, RF, ANN) will improve predictive performance compared to logistic regression, (2) information on laterality of seizure onset improves predictive performance, and (3) including surgical information (volume of tissue resected, percentage of TL hypometabolism resected) meaningfully outperforms models using presurgical information only.

To test each of our hypotheses, we statistically compared the AUC, accuracy, sensitivity, specificity, PPV, and NPV of the classifier against a control classifier with the feature of interest omitted (Hypotheses 2 and 3), or using a different algorithm (Hypothesis 1; overview of model comparisons is provided in Table 2). The significance of differences between classifiers was assessed using a one-tailed paired *t*-test on performance measures derived from repeated cross-validation with 10 repeats of 10 folds, as recommended by Bouckaert and Frank,[20] due to its high replicability. This method violates the independence assumption of the paired *t*-test and has a high type 1 error rate. To address this issue, we applied the Nadeau and Bengio[21] correction on degrees of freedom. Finally, within each hypothesis, statistical significance was corrected for multiple comparisons over models and performance measures using the false discovery rate (FDR),[22] and an FDR < .05 was considered significant.

3 | RESULTS

3.1 | Patient characteristics and seizure outcomes

The patient characteristics and seizure outcomes of the 82 MTLE patients who underwent an ATLR in this cohort were reported in Cahill et al.[15] In brief, 43 patients underwent a right ATLR and 39 patients a left ATLR with comparable gender composition, age at epilepsy onset, epilepsy duration, age at surgery, preoperative seizure frequency, and duration of postoperative follow-up. The median postoperative follow-up period was 4 years (range = 2–10 years) in patients who underwent a right ATLR and 5 years (range = 2–14 years) in patients who had a left ATLR. An Engel Class I outcome was achieved in 58 of 82 (70.7%) patients, which did not differ in patients who had a right versus left ATLR ($p = .51$).

3.2 | Classification performance

Table 3 shows the classification performance of each machine learning algorithm. Accuracy ranged between 71% and 80%, with AUCs ranging between .75 and .81. Specificity, the proportion of non-seizure-free patients

TABLE 2 Description of models compared for hypothesis testing

Hypothesis	Model	Model variables	Control model	Control model variables
1.1	SVM	Presurgical Surgical Laterality	LR	Presurgical Surgical Laterality
1.2	RF	Presurgical Surgical Laterality	LR	Presurgical Surgical Laterality
1.3	ANN	Presurgical Surgical Laterality	LR	Presurgical Surgical Laterality
2	LR SVM RF ANN	Presurgical Surgical Laterality	LR SVM RF ANN	Presurgical Surgical
3	LR SVM RF ANN	Presurgical Surgical Laterality	LR SVM RF ANN	Presurgical Laterality

Note: Presurgical variables are percentage of extratemporal hypometabolism, presence of contralateral hypometabolism, and presence of hippocampal sclerosis. Surgical variables are volume of tissue resected and percentage of temporal lobe hypometabolism resected.

Abbreviations: ANN, artificial neural network; LR, logistic regression; RF, random forest; SVM, support vector machine.

TABLE 3 Classification performance measures for each model with all variables included, and statistical comparison of hypothesis-free models (SVM, RF, ANN) against hypothesis-driven logistic regression

						SVM > LR		RF > LR		ANN > LR	
Variables	**Performance**	**LR**	**SVM**	**RF**	**ANN**	***t***	***p***	***t***	***p***	***t***	***p***
Presurgical + surgical + laterality	AUC	.75	.78	.81	.76	.64	.262	.95	.172	.32	.374
	Accuracy	.71	.75	.80	.70	.84	.203	1.60	.057	−.36	.639
	Sensitivity	.75	.78	.86	.69	.38	.352	1.77	.040	−1.53	.936
	Specificity	.61	.71	.65	.73	.86	.195	.20	.420	1.32	.094
	PPV	.83	.87	.86	.87	.76	.223	.59	.279	.97	.168
	NPV	.54	.60	.72	.53	.54	.294	1.44	.076	−.33	.629

Abbreviations: ANN, artificial neural network; AUC, area under receiver operating characteristic curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine.

correctly predicted, ranged from .61 to .73, suggesting that the algorithm could deter the decision to undergo surgery for up to 73% of patients for whom surgery will not result in seizure freedom. However, the NPV was lower, at .53–.72. The best performing (highest AUC and accuracy) algorithm, RF, correctly predicted (on average) 49.9 of 58 (86%) patients with an Engel Class I outcome with respect to seizures at 2 years, and 15.6 of 24 (65%) patients with an Engel Class II–IV outcome.

3.3 | Model comparisons

SVM and RF had higher AUCs, accuracies, sensitivities, and specificities than LR (Table 3), but these were not statistically significant after correction for multiple comparisons. Likewise, ANN had no significant improvements compared to LR. Omission of laterality (Table 4, FIgure 2) did not significantly reduce any performance measure. Omission of surgical information substantially reduced all performance measures. AUC and specificity for SVM, accuracy for RF, and AUC for ANN were statistically significantly lower (after FDR correction) with surgical information omitted (Table 5).

4 | DISCUSSION

In this study, we applied a range of the most widely used machine learning algorithms to the problem of surgical candidate selection in a well-characterized cohort of patients with drug-resistant MTLE who had undergone an ATLR with at least 2 years of postoperative follow-up. Many advances have been made in surgical candidate selection and surgical outcomes since the advent of brain imaging.[23–26] A substantial proportion of patients undergoing an ATLR for drug-resistant MTLE, however, do not achieve postoperative seizure control (29.3% in this cohort). A tool enabling the reliable detection of such patients, utilizing clinically well-established indicators of likely seizure freedom, could optimize the identification of those patients who would (and would not) benefit from invasive surgical intervention.

The accuracies of our machine learning algorithms based on MRI and FDG-PET features ranged between 70% and 80%, a modest improvement from the 71% rate of Engel Class I outcomes in this cohort, and AUCs ranged between .75 and .81, which is roughly considered to be "acceptable" to "good" discrimination.[27] Although these are lower than the levels generally considered to have clinical utility, it should be noted that the starting sample is already highly filtered (patients considered suitable for surgery by the multidisciplinary team members based on the outcome of phase I evaluation, including electroclinical evaluation, structural MRI, FDG-PET, and single photon emission computed tomography, where indicated). As such, it is important to emphasize that the performance measures reported correspond to patients already approved for surgery, and not to all potential surgical candidates. This is a more difficult classification task, because patients with obvious indications of being unsuitable for surgery are not available to classify. To them put into context, the accuracies of 70%–80% mean that if this algorithm were to be used in isolation (which we do not recommend) to select the best treatment option for this subset of patients after the referral to surgery, the “correct treatment” would be administered up to 80% of the time, compared to the current ~70% of the time. Given that the causes of unsuccessful ATLR remain unclear, any marginal increase in correct treatment administration is of value and merits further consideration.

The algorithms were able to detect up to 73% of patients who would not achieve seizure freedom. If used as a clinical tool, this could potentially help stratify patients and assist with decision-making and presurgical counseling to set realistic expectations. However, the trade-off

TABLE 4 Classification performance measures for each model with laterality of seizure onset omitted, and statistical comparison of models including laterality to those without laterality (Table 3)

						Laterality > no laterality							
						LR		**SVM**		**RF**		**ANN**	
Variables	**Performance**	**LR**	**SVM**	**RF**	**ANN**	***t***	***p***	***t***	***p***	***t***	***p***	***t***	***p***
Presurgical + surgical	AUC	.75	.79	.81	.75	−.64	.739	−.29	.615	.01	.498	.03	.486
	Accuracy	.71	.76	.78	.69	−.22	.585	−.45	.675	.23	.410	.11	.455
	Sensitivity	.75	.80	.84	.68	−.10	.539	−.80	.788	.56	.289	.19	.427
	Specificity	.61	.69	.65	.73	−.07	.528	.34	.368	−.35	.636	.00	.500
	PPV	.83	.87	.86	.87	−.20	.58	.19	.426	−.15	.561	.02	.492
	NPV	.55	.62	.67	.52	−.18	.571	−.40	.657	.62	.270	.17	.431

Abbreviations: ANN, artificial neural network; AUC, area under receiver operating characteristic curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine.

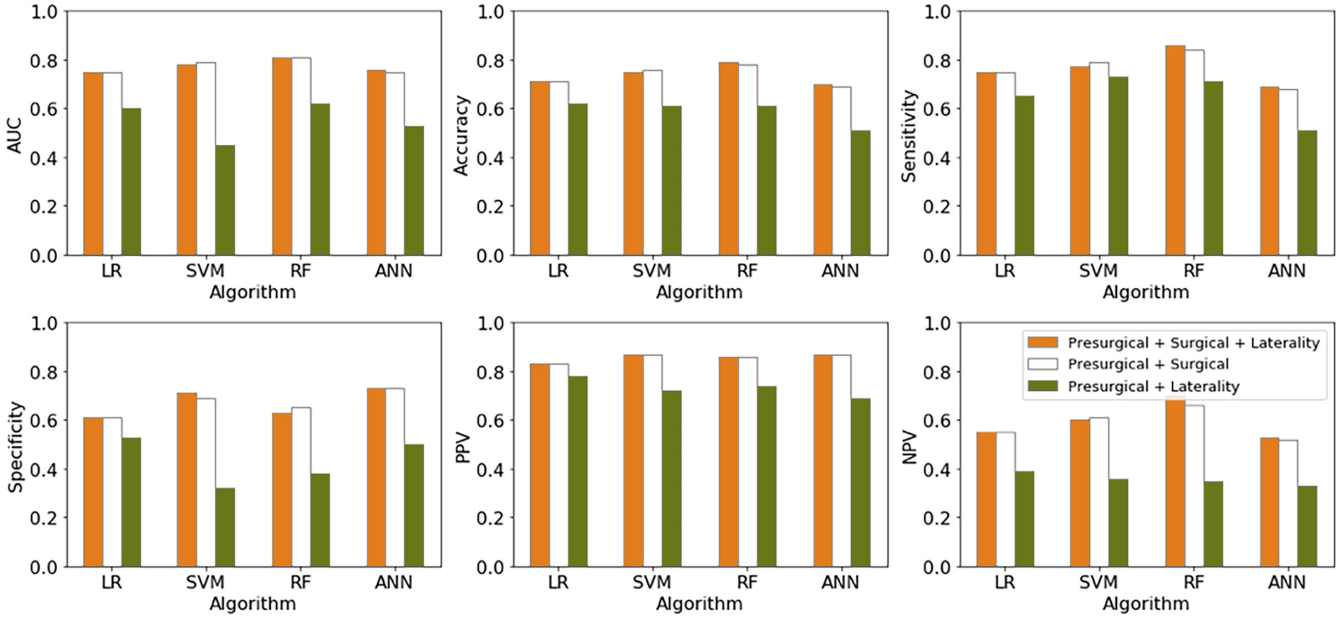

FIGURE 2 Classification performance for each machine learning algorithm: area under receiver operating characteristic curve (AUC), accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). Colors indicate inputs to model (see legend). ANN, artificial neural network; LR, logistic regression; RF, random forest; SVM, support vector machine

is that a substantial proportion (14%–31%) of successful surgeries were predicted by the algorithms to be unsuccessful, and would thus potentially deny these patients a beneficial and life-changing procedure. Although the risks from surgery can be severe, the risks from not doing surgery (continued seizures and associated comorbidities and side effects) are equally severe. Therefore, although this tool could prove helpful in guiding risk assessment and decision-making, the final decision to proceed to epilepsy surgery would ultimately be shared and guided by the patient's choice based on individual circumstances and expectations related to the outcome of epilepsy surgery. It is also important to note that in a clinical setting, to influence the decision not to proceed to surgery, a classification algorithm should ideally have a high NPV. For this analysis on the current dataset, the NPV was the lowest of the performance measures at between .53 and .72, meaning that if a patient were told that the surgery would not yield seizure freedom, there is an up to 72% chance of this being a correct prognosis. Although this is less than ideal, it is still high enough to provide value in the presurgical discussion with the patient. However, clearly a 72% NPV is not high enough to delegate the decision to proceed to surgery entirely to this algorithm, and if it were to be used clinically, it should only be as an additional source of information to take into consideration.

TABLE 5 Classification performance measures for each model with surgical information omitted, and statistical comparison of models including surgical and presurgical information (Table 3) to those with presurgical information alone

						Presurgical + surgical > presurgical only							
						LR		SVM		RF		ANN	
Variables	**Performance**	**LR**	**SVM**	**RF**	**ANN**	***t***	***p***	***t***	***p***	***t***	***p***	***t***	***p***
Presurgical + laterality	AUC	.59	.47	.62	.53	1.88	.031	3.41	<.001[a]	1.84	.034	2.56	.006[a]
	Accuracy	.62	.62	.62	.51	1.59	.057	1.81	.037	2.61	.005[a]	2.27	.013
	Sensitivity	.65	.74	.71	.51	1.3	.098	.44	.330	1.79	.038	1.31	.097
	Specificity	.54	.33	.39	.50	.53	.299	3.08	.001[a]	2.07	.021	1.44	.077
	PPV	.79	.72	.75	.69	.87	.193	3.09	.001	2.55	.006[a]	2.49	.007
	NPV	.39	.37	.36	.33	1.54	.063	1.58	.059	2.58	.006	1.82	.036

Abbreviations: ANN, artificial neural network; AUC, area under receiver operating characteristic curve; LR, logistic regression; NPV, negative predictive value; PPV, positive predictive value; RF, random forest; SVM, support vector machine.

[a]False discovery rate-corrected significant differences.

Only one study has previously used machine learning on PET data to predict surgical outcome in epilepsy.[28] Using RF classifiers on 16 surgical patients, it reported a 62.5% accuracy using FDG-PET, and a 87.5% accuracy using a [11C]flumazenil (FMZ) PET tracer. Our study achieved greater accuracy with FDG-PET, likely due to the larger sample size. The higher accuracies reported for [11C]FMZ PET are intriguing and worthy of further study, but [11C]FMZ PET is not routinely used in presurgical assessment of potential epilepsy surgical candidates, and thus would be harder to translate to clinical application.

Other studies have used machine learning algorithms to predict TLE surgical outcome on a range of other imaging and nonimaging modalities,[29,30] and report a wide range of predictive performances of .63–.98. To our knowledge, the highest performance using imaging data alone was published by Feis et al.[31] These authors used SVM on voxelwise white matter volumes, which gave a classification accuracy of 95%. Munsell et al.[32] and Gleichgerrcht et al.[33] used connectivity maps based on diffusion MRI tractography to predict surgical outcome, achieving classification performances of 70% and 79%–88% using SVM and deep neural networks respectively. This is similar to performance observed in our study, but using only presurgical information. The lower classification performance in the present study compared to the highest performing algorithms in the literature may be due to the modalities used (i.e., PET hypometabolism, resection regions, visual MRI inspection), which contain less prognostic information than more derivative measures (i.e., voxelwise volumetry or whole brain connectomics). Alternatively, the particular features employed (i.e., PET hypometabolism, hippocampal sclerosis, etc.) may have introduced the potential for pretest selection bias. That is, having already been considered by multidisciplinary clinical teams when selecting surgical candidates, the variability in these features was reduced in the dataset upon which we trained our models. These features, therefore, had lower marginal prognostic value on the specific filtered set of patients we considered than features that had not been used to select surgical candidates (e.g., voxelwise volumetry, connectomics). A final consideration is that our models were trained on only seven predictive variables selected via domain-specific knowledge compared to the hundreds to thousands of input features used in those more exploratory analyses. It may be that the more exploratory analyses are picking up high-dimensional hidden features pertinent to surgical outcome, or it may be that they are overfitting. In the future, we aim to conduct exploratory, voxelwise analyses on our PET and MRI datasets.

SVM and RF had higher classification performance than LR. However, these differences were not statistically significant after correction for multiple comparisons, perhaps indicating that a simple model is sufficient for obtaining the classification performance reported here. SVM and RF have an advantage over LR in that they are less hypothesis driven and are able to detect unhypothesized interactions between variables affecting surgical outcome. The higher accuracies of these models may indicate that there are such hidden interactions between our predictor variables. The black-box nature of SVM and RF, however, makes it difficult to discern what these are.

One such expected interaction between the variables was that between laterality with presence of contralateral hypometabolism, and laterality with volume of tissue resected and percentage of TL hypometabolism resected. Our previous study[15] showed that presence of contralateral hypometabolism on FDG PET was significantly associated with a poorer chance of seizure freedom for right MTLE patients but not left MTLE patients. Furthermore, the volume of TL tissue resected and percentage of TL hypometabolism resected were significantly associated with higher chance of seizure freedom following left ATLR but not right ATLR. In the present study, we found that including laterality as a predictor variable did not improve the classification performance (slight decreases were observed). In contrast to regression analysis, addition of predictor variables can reduce performance in machine learning due to overfitting. The apparent discrepancy between the importance of laterality in this study compared to Cahill et al.[15] is most likely due to the high colinearity between laterality and other predictor variables, which would impact our machine learning models, but not the two-sample (Engel Class I vs. Engel Class II–IV postsurgical outcomes) *t*-tests in Cahill et al.[15] For example, in our sample, laterality is correlated with percentage extratemporal hypometabolism ($\rho = .63$) and volume of tissue resected ($\rho = .43$), reflecting the more focal distribution of hypometabolism in left MTLE patients, and the sparing of critical language regions in left ATLR. Based on these two predictor variables alone, our machine learning algorithms are likely to already have a representation of laterality, without this variable being explicitly specified, meaning that the inclusion of laterality in our models does not add substantial marginal predictive information.

Finally, predictive performance was found to be substantially worse when using presurgical information alone, compared to using surgical information. This indicates that for use as a presurgical tool, some information about the surgical resection region will likely be necessary. This could be achieved by utilizing surgical planning software[34] or even conceivably using machine learning itself to calculate the region of resection most likely to result in positive surgical outcome.[35,36]

The current study has several limitations. Although 82 patients are a relatively large cohort in the context of epilepsy surgery studies, and is one of the larger cohorts used to apply machine learning algorithms to the problem of predicting postsurgical outcome with respect to seizures,[29] the sample size is smaller than typically required for machine learning. A second consideration, which may be a limitation or a strength, is that the predictor variables derived (e.g., localization of hypometabolism within TLs, presence of contralateral hypometabolism, etc.) are based on clinical domain knowledge; they are those considered important when assessing surgical suitability. For our study, this was necessary to reduce dimensionality with low sample sizes, and could improve classifier performance by selecting the most pertinent clinical variables. Conversely, it may reduce classification performance: first, because it removes much information available in the raw data; and second, because these predictors or related features (such as localization of hypometabolism within TLs) are likely to have influenced the decision to perform surgery, leaving a sample with less variance and predictive power in these predictor variables. Similarly, our predictor variables contained little localizing information, limited to laterality of hypometabolism, temporal versus extratemporal location of the hypometabolism, and overlap between PET and MRI features. The precise anatomical locations of resection have been reported to have important associations with seizure outcome, with targeting of the hippocampus, amygdala piriform cortex complex, and entorhinal cortex associated with seizure freedom.[37,38] With larger datasets, we will advance this approach to using whole brain images as input, rather than only derived information as used in this study, allowing more information from the raw data to be extracted and included in the predictive modeling. A final consideration pertains to the statistical methods employed in this study. To test our hypotheses, we compared models using repeated *k*-fold cross-validation with a paired *t*-test. This method is widely used, but violates independence assumptions of the *t*-test, and has an inflated type 1 error rate. We have mitigated this somewhat with the Nadeau and Bengio[21] correction, but the findings should be evaluated with this limitation in mind. Other considerations include those pertaining to the use of machine learning algorithms in general. Validation of the predictive value of the machine learning approaches tested here on external datasets is necessary prior to establishing their utility to be incorporated into clinical practice, internal validation in a single dataset usually underestimates the out-of-sample error, and external validation gives a better insight on generalizability and domain relevance of the model.[39] We have not yet performed such validation and will seek to do so in the future. Furthermore, our machine learning algorithms were trained on seizure freedom alone. It is important to note that although the analyses in this study focused on postoperative seizure control, this is not the only outcome of importance. Minimizing postoperative functional deficits, such as visual, neurocognitive, and neuropsychiatric deficits, along with risk of surgical complications, are all important considerations affecting the decision to operate and the extent of tissue to resect. Future applications of machine learning approaches to predict outcomes from epilepsy surgery could incorporate these postoperative outcomes in addition to seizure control.

5 | CONCLUSIONS

This study showed that machine learning algorithms are able to provide additional prognostic value from clinically available neuroimaging features already incorporated in initial evaluation of epilepsy surgery candidates. Most notably, up to 73% of patients with poor surgical outcome were predicted, potentially providing additional information to incorporate into surgical decision-making and patient counseling. Of our secondary hypotheses, the strongest finding was that performance was substantially reduced without knowledge of the resection region, indicating that prospective incorporation of the planned surgical resection into the machine learning approach is necessary for the optimum prognostication value.

ACKNOWLEDGMENTS

This study was funded in part by a National Health and Medical Research Council (NHMRC) Program Grant (#APP1091593) and NHMRC Investigator Grant (#APP1176426). Open access publishing facilitated by Monash University, as part of the Wiley - Monash University agreement via the Council of Australian University Librarians. [Correction added on 17 May 2022, after first online publication: CAUL funding statement has been added.]

CONFLICT OF INTEREST

None of the authors has any conflict of interest to disclose. 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.

ORCID

Benjamin Sinclair https://orcid.org/0000-0002-0850-3644
Lucy E. Vivash https://orcid.org/0000-0002-1182-0907
Charles B. Malpas https://orcid.org/0000-0003-0534-3718
Patrick Kwan https://orcid.org/0000-0001-7310-276X
Samuel F. Berkovic https://orcid.org/0000-0003-4580-841X
Terence J. O'Brien https://orcid.org/0000-0002-7198-8621

REFERENCES

1. Wiebe S, Blume WT, Girvin JP, Eliasziw M, Effectiveness and Efficiency of Surgery for Temporal Lobe Epilepsy Study Group. A randomized, controlled trial of surgery for temporal-lobe epilepsy. N Engl J Med. 2001;345(5):311–8.
2. Téllez-Zenteno JF, Ronquillo LH, Moien-Afshari F, Wiebe S. Surgical outcomes in lesional and non-lesional epilepsy: a systematic review and meta-analysis. Epilepsy Res. 2010;89(2–3):310–8.
3. McIntosh AM, Kalnins RM, Mitchell LA, Fabinyi GCA, Briellmann RS, Berkovic SF. Temporal lobectomy: long-term seizure outcome, late recurrence and risks for seizure recurrence. Brain. 2004;127(9):2018–30.
4. Hader WJ, Tellez-Zenteno J, Metcalfe A, Hernandez-Ronquillo L, Wiebe S, Kwon CS, et al. Complications of epilepsy surgery—a systematic review of focal surgical resections and invasive EEG monitoring. Epilepsia. 2013;54(5):840–7.
5. Duncan JS, Winston GP, Koepp MJ, Ourselin S. Brain imaging in the assessment for epilepsy surgery. Lancet Neurol. 2016;15(4):420–33.
6. Jasper HH, Arfel-Capdeville G, Rasmussen T. Evaluation of EEG and cortical electrographic studies for prognosis of seizures following surgical excision of epileptogenic lesions. Epilepsia. 1961;2:130–7.
7. Jehi L. The epileptogenic zone: concept and definition. Epilepsy Curr. 2018;18(1):12–6.
8. Radhakrishnan K, So EL, Silbert PL, Jack CR, Cascino GD, Sharbrough FW, et al. Predictors of outcome of anterior temporal lobectomy for intractable epilepsy: a multivariate study. Neurology. 1998;51(2):465–71.
9. McIntosh AM, Wilson SJ, Berkovic SF. Seizure outcome after temporal lobectomy: current research practice and findings. Epilepsia. 2001;42(10):1288–307.
10. Carne RP, O'Brien TJ, Kilpatrick CJ, MacGregor LR, Hicks RJ, Murphy MA, et al. MRI-negative PET-positive temporal lobe epilepsy: a distinct surgically remediable syndrome. Brain. 2004;127(10):2276–85.
11. Lopinto-Khoury C, Sperling MR, Skidmore C, Nei M, Evans J, Sharan A, et al. Surgical outcome in PET-positive, MRI-negative patients with temporal lobe epilepsy. Epilepsia. 2012;53(2):342–8.
12. O'Brien TJ, Hicks RJ, Ware R, Binns DS, Murphy M, Cook MJ. The utility of a 3-dimensional, large-field-of-view, sodium iodide crystal-based PET scanner in the presurgical evaluation of partial epilepsy. J Nucl Med. 2001;42(8):1158–65.
13. Vinton AB, Carne R, Hicks RJ, Desmond PM, Kilpatrick C, Kaye AH, et al. The extent of resection of FDG-PET hypometabolism relates to outcome of temporal lobectomy. Brain. 2007;130(2):548–60.
14. O'Brien TJ, Miles K, Ware R, Cook MJ, Binns DS, Hicks RJ. The cost-effective use of 18F-FDG PET in the presurgical evaluation of medically refractory focal epilepsy. J Nucl Med. 2008;49(6):931–7.
15. Cahill V, Sinclair B, Malpas CB, McIntosh AM, Chen Z, Vivash LE, et al. Metabolic patterns and seizure outcomes following anterior temporal lobectomy. Ann Neurol. 2019;85(2):241–50.
16. Engel J Jr, van Ness PC, Rasmussen TB, Ojemann LM. Outcome with respect to epileptic seizures. In: Engel J, editor. Surgical treatment of the epilepsies. New York, NY: Raven Press; 1993. p. 609–21.

17. Sharpe C, Sinclair B, Kwan P, Hicks RJ, O'Brien TJ, Vivash L. Longitudinal changes of focal cortical glucose hypometabolism in adults with chronic drug resistant temporal lobe epilepsy. Brain Imaging Behav. 2021;15(6):2795–803.
18. Penny W, Friston K, Ashburner J, Kiebel S, Nichols T. Statistical parametric mapping: the analysis of functional brain images. New York, NY: Elsevier; 2007.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
20. Bouckaert RR & Frank E. Evaluating the replicability of significance tests for comparing learning algorithms. In: Pacific-Asia conference on knowledge discovery and data mining. Berlin: Springer; 2004.
21. Nadeau C, Bengio Y. Inference for the generalization error. Advances in Nural Information Processing Systems 12. 1999.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B (Methodol). 1995;57(1):289–300.
23. Baud MO, Perneger T, Rácz A, Pensel MC, Elger C, Rydenhag B, et al. European trends in epilepsy surgery. Neurology. 2018;91(2):e96–106.
24. Hemb M, Velasco TR, Parnes MS, Wu JY, Lerner JT, Matsumoto JH, et al. Improved outcomes in pediatric epilepsy surgery: the UCLA experience, 1986-2008. Neurology. 2010;74(22):1768–75.
25. Thom M, Mathern GW, Cross JH, Bertram EH. Mesial temporal lobe epilepsy: how do we improve surgical outcome? Ann Neurol. 2010;68(4):424–34.
26. Vakharia VN, Duncan JS, Witt JA, Elger CE, Staba R, Engel J. Getting the best outcomes from epilepsy surgery. Ann Neurol. 2018;83(4):676–90.
27. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York, NY: John Wiley & Sons; 2000.
28. Yankam Njiwa J, Gray KR, Costes N, Mauguiere F, Ryvlin P, Hammers A. Advanced [(18)F]FDG and [(11)C]flumazenil PET analysis for individual outcome prediction after temporal lobe epilepsy surgery for hippocampal sclerosis. Neuroimage Clin. 2015;7:122–31.
29. Senders JT, Staples PC, Karhade AV, Zaki MM, Gormley WB, Broekman MLD, et al. Machine learning and neurosurgical outcome prediction: a systematic review. World Neurosurg. 2018;109:476–86.e1.
30. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019;60(10):2037–47.
31. 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(1):903–11.
32. 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.
33. Gleichgerrcht E, Munsell B, Bhatia S, Vandergrift WA, 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.
34. Nowell M, Rodionov R, Zombori G, Sparks R, Rizzi M, Ourselin S, et al. A pipeline for 3D multimodality image integration and computer-assisted planning in epilepsy surgery. J Vis Exp. 2016;2016(111):53450.
35. Qu Y, Li X, Yan Z, Zhao L, Zhang L, Liu C, et al. Surgical planning of pelvic tumor using multi-view CNN with relation-context representation learning. Med Image Anal. 2021;69:101954.
36. Bhandari M, Zeffiro T, Reddiboina M. Artificial intelligence and robotic surgery: current perspective and future directions. Curr Opin Urol. 2020;30(1):48–54.
37. Gleichgerrcht E, Drane DL, Keller SS, Davis KA, Gross R, Willie JT, et al. Association between anatomical location of surgically induced lesions and postoperative seizure outcome in temporal lobe epilepsy. Neurology. 2021;98(2):e141–51.
38. Galovic M, Baudracco I, Wright-Goff E, Pillajo G, Nachev P, Wandschneider B, et al. Association of piriform cortex resection with surgical outcomes in patients with temporal lobe epilepsy. JAMA Neurol. 2019;76(6):690.
39. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns. 2020;1(8):100129.

How to cite this article: 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;63:1081–1092. https://doi.org/10.1111/epi.17217