Supplemental Online Content

Hakeem H, Feng W, Chen Z, et al. Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. *JAMA Neurol*. Published online August 29, 2022. doi:10.1001/jamaneurol.2022.2514

This supplemental material has been provided by the authors to give readers additional information about their work.

eAppendix: Description of the transformer model

We utilized a deep learning approach to predict patient responses to their initial ASM treatment. To predict the probability of a successful ASM treatment, we passed the patient and regimen information into a transformer model. All input variables were normalized to ensure fair weighting for each variable before being used for training and validation. The ASM variables were binarized to 0 or 1 as opposed to dosage values. The transformer model outputs a probability of success for a patient's response to individual ASM treatments. The model can be adapted for time series structure such as a longitudinal study of several treatments for an individual patient.

Flow of information through the transformer network: Each patient's individual clinical characteristics and the first ASM regimen taken were used as a separate input to the transformer. As we were only predicting the response to the first regimen to determine seizure freedom, there was only one data point per patient. Drug use was binarized as 0 or 1. Age was trichotomized in to tertiles. The input was passed through the encoder, where the transformer captured the relationship between the different input variables for each patient. The information from the encoder was then passed through into the decoder of the transformer, which modelled the response of the patient information to the prescribed ASM. The patient and regimen data were passed through a multi-head attention layer described in **eEquation 1**. The multi-attention layer allowed the transformer to focus on important aspects of the data when predicting the patient's response to the ASM treatment. The output from the transformer was passed into a sigmoid function where it output values between 0 and 1. **eFigure 1** provides an overview of a single encoder-decoder pair of the transformer network.

Model training, testing, and validation: The model was trained on each first prescribed monotherapy regimen on the training cohort. In experiment 1, we randomly selected 80% of the data to train the model and evaluated the model's performance on the remaining 20% of the data. In experiment 2, the model was developed using only the Glasgow cohort as the development set and externally validated on each of the remaining cohorts. We performed four five-fold cross-validations and obtained a total of 20 experimental results to calculate mean and standard deviation. The formula for calculation of CI is $\bar{X} \pm Z\frac{S}{\sqrt{n}}$, where ' $\bar{X}$' is mean, 'Z' is standard deviation from the mean, 'S' is standard deviation and 'n' is the number of observations. Validation was performed by predicting a probability of achieving seizure freedom for each patient's response to the utilized ASM in their first trial. The threshold probability was tested by determining the optimal weighted average between specificity and sensitivity of the trained model. Patients in the validation set with predicted probabilities above the threshold probability of their prescribed ASM treatments were predicted to have a successful treatment outcome. This was tested against the actual outcome of the validation patient's ASM treatment. In addition, we introduced a data augmentation technique SMOTE[1] to improve the robustness and generalization performance of the model, and also to alleviate the category imbalance problem.

Generalization of associations learnt by the model: Generalizability of the model was analyzed through a separate validation cohort. T-distributed stochastic neighbor embedding (t-SNE) analysis[2] was performed to identify any distinct cluster of patients between the cohorts, and to evaluate how well the model generalized the associations learnt during training. **eFigure 4** is a visualization of the t-SNE analysis by projecting patients within the development and validation cohorts captured by the last encoder layer (treated as the feature layer) of the transformer network. The analysis supports the overall generalizability of the model to patients in the validation cohorts although there are distinct clusters not captured, likely due to the differences in the distribution of baseline characteristics between the cohorts.

References:

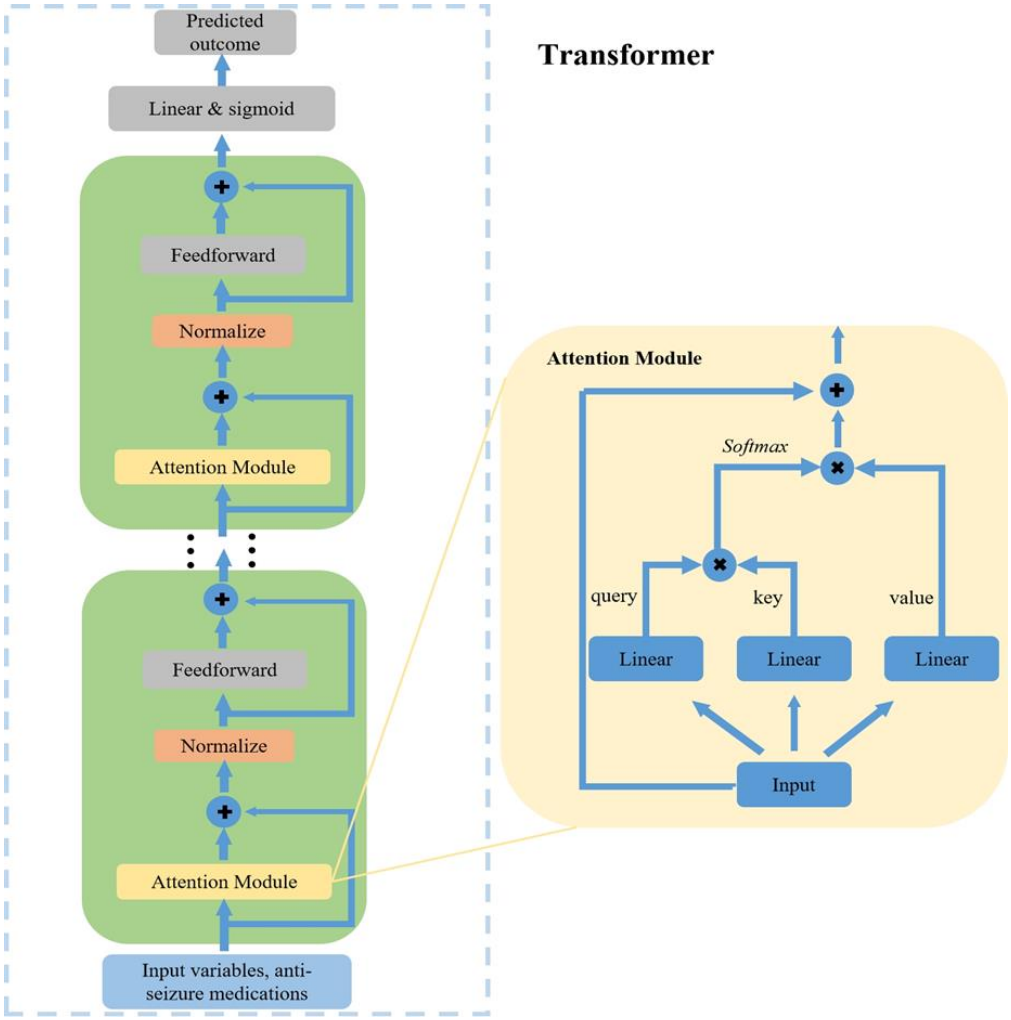
1. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002, 16: 321-357.
2. Van Der Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9(11):2579-2605.

eEquation 1. Output of a multi-head attention layer.

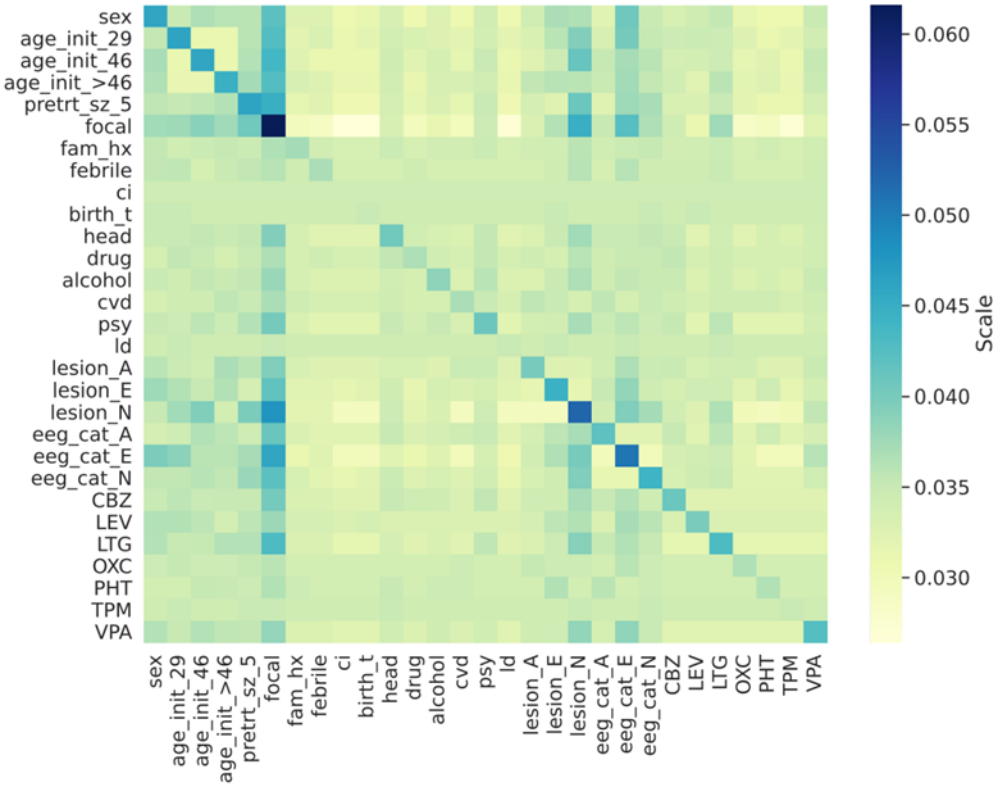
$$T_{out} = softmax\left(\frac{(T_{in}K)(T_{in}Q)^T}{\sqrt{d_k}}\right)(T_{in}V))$$

T_{in} and T_{out} are the input and output data. K, Q, V are key, query, and value vectors respectively, that retrieve corresponding values for each of the inputs. $\frac{1}{\sqrt{d_k}}$ represents a scaling factor for the queried value.

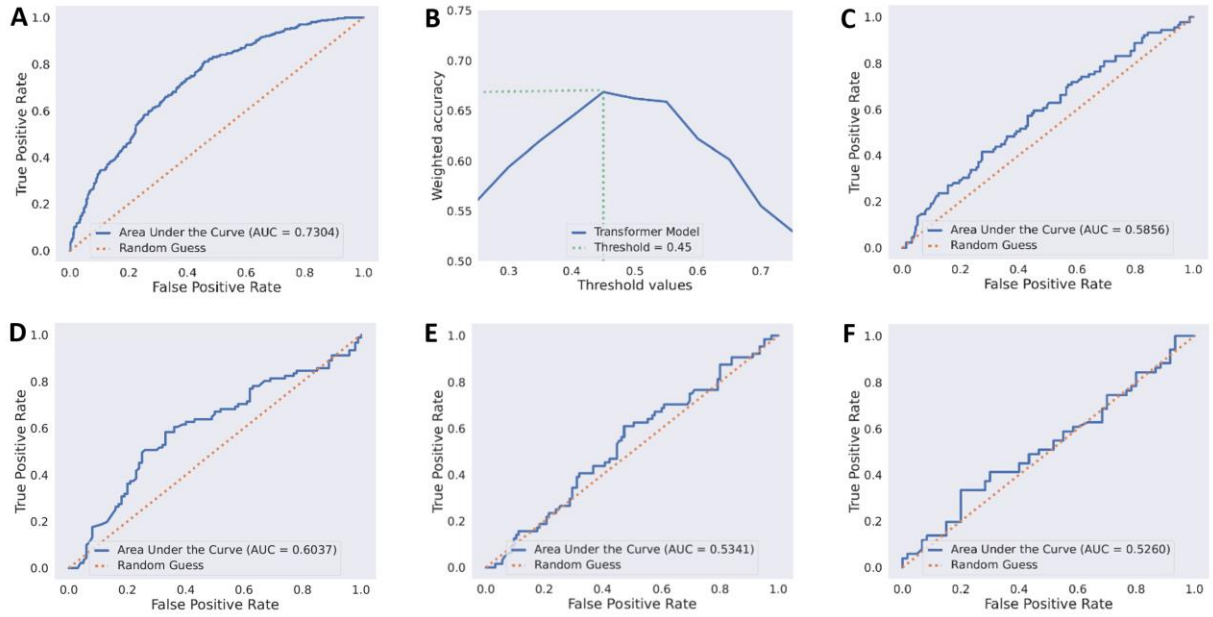
eFigure 1. Overview of a single encoder-decoder pair of the transformer network. The input variables for individual patients are fed into the transformer network, where a attention layer simultaneously focuses on different aspects of the data. Outputs from each layer are typically normalized for each output. The linear layer takes the output from the transformer to perform the final classification through a sigmoid function which compresses the value between 0 and 1.



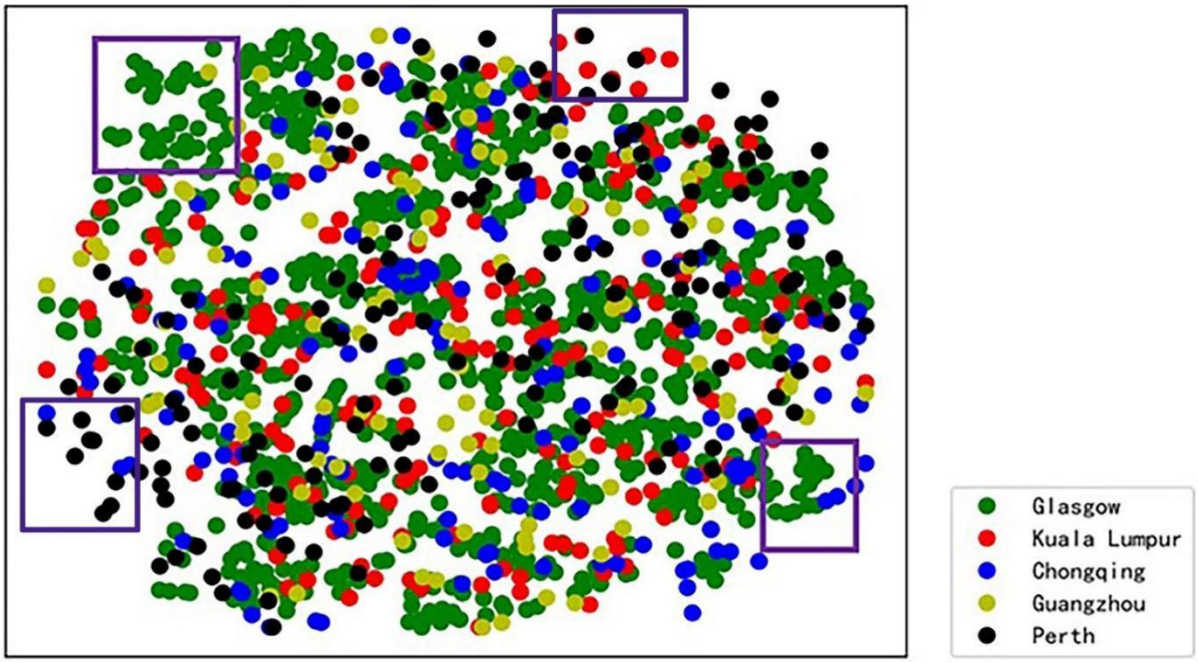
eFigure 2. Attention map of the transformer network. The heat map shows the interdependencies between different input features in the transformer model, with increasing shade of blue indicating stronger association between the corresponding variables. Age_init_29, age at treatment initiation ≤29 years; Age_init_46, age at treatment initiation >29 to ≤46 years; Age_init_>46, age at treatment initiation >46 years; Alcohol, history of alcohol abuse; Birth_t, history of birth trauma; CBZ, carbamazepine; Ci, history of central nervous system infection; CVD, cardiovascular disease; Drug, history of drug abuse; EEG_cat_A, non-epileptiform abnormalities on electroencephalography; EEG_cat_E, epileptiform electroencephalography abnormalities; EEG_cat_N, normal electroencephalography findings; Fam_hx, family history of epilepsy in first-degree relatives; Febrile, history of febrile seizures; Focal, diagnosed with focal epilepsy; Head, history of significant head injury; Id, presence of intellectual disability; Lesion_A, non-epileptiform abnormalities on brain imaging; Lesion_E, epileptiform abnormality on brain imaging; Lesion_N, normal brain imaging; LEV, levetiracetam; LTG, lamotrigine; OXC, oxcarbazepine; PHT, phenytoin; Psy, presence of psychiatric history; Pretrt_sz_5, >5 pre-treatment seizures; TPM, topiramate; VPA, valproate.



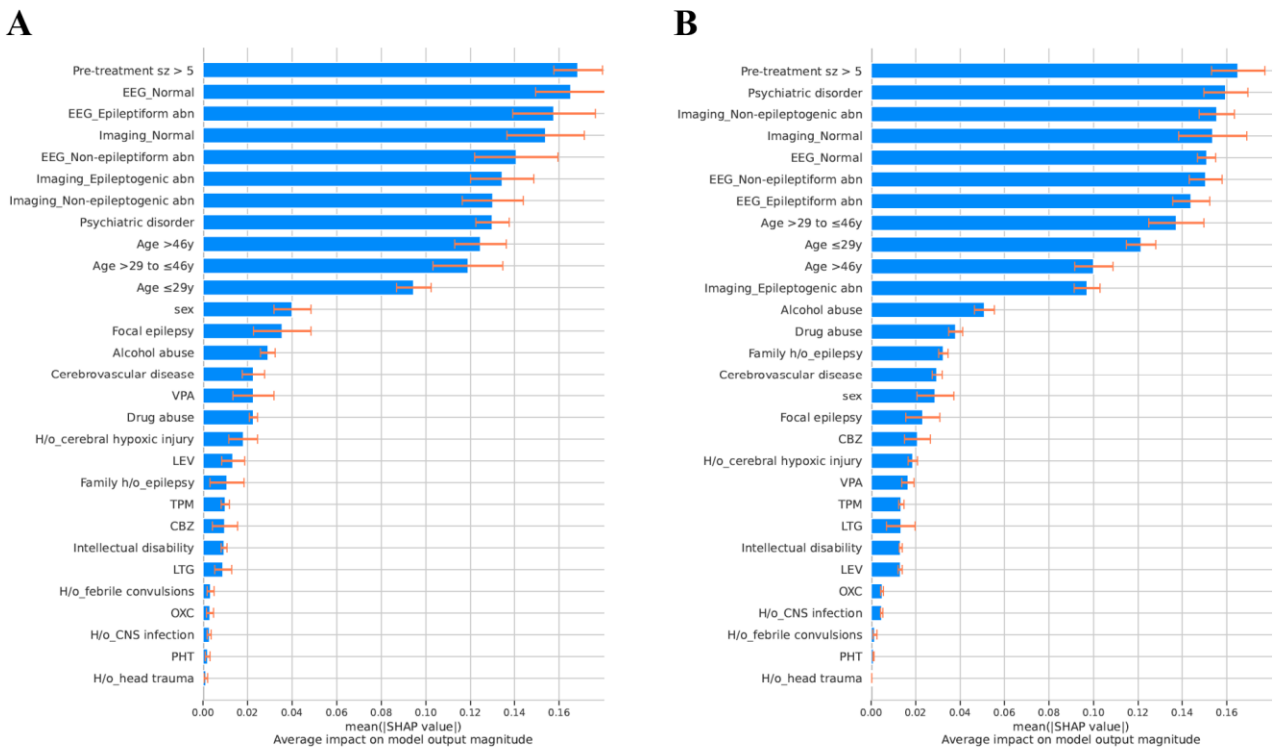
eFigure 3. Receiver operating characteristic curves and weighted accuracy curve for transformer model developed using entire Glasgow cohort. (**A**) Receiver operating characteristic (ROC) curve on the training set. (**B**) Weighted accuracy curve at different threshold values of probability. Highest weighted accuracy was obtained at a threshold of 0.45. Optimal threshold value is indicated by intersection of dashed green lines. (**C**) ROC curve on the Kuala Lumpur cohort. **(D)** ROC curve on the Chongqing cohort. **(E)** ROC curve on the Perth cohort. **(F)** ROC curve on the Guangzhou cohort. AUC, area under receiver operating characteristics curve.



eFigure 4. t-SNE analysis for projection of clusters within the five cohorts captured by the last encoder layer of the transformer network. The figure shows the overall generalizability of the model to patients in the validation cohorts. Some distinct clusters (purple boxes) can be observed in the figure. Those clusters may reflect the differences in the distribution of baseline characteristics between the cohorts.



eFigure 5. SHAP analysis for visualization of the importance of input feature variables for the predicted outcomes of the model. (**A)** Model developed using the pooled cohort (experiment setting 1). **(B)** Model developed using the Glasgow cohort only (experiment setting 2). Blue and error bars represent mean SHAP value and standard deviation, respectively for each variable. Abn; abnormality; EEG, electroencephalography; CBZ, carbamazepine; LEV, levetiracetam; LTG, lamotrigine; OXC, oxcarbazepine; PHT, phenytoin; TPM, topiramate; VPA, valproate; y, years.



eTable 1. Summary of missing data in each cohort. The major sources of missing data are missing EEG and brain imaging in the Glasgow cohort and psychiatric history and learning disability in the Perth cohort. The formal is related to clinical practice and the latter is related to data collection practice. Although the data is not missing at random in these two cohorts, the missing data were evenly distributed among patients with successful and unsuccessful outcomes. Thus the missingness can be considered as conditionally independent of outcome and the complete case analysis has negligible bias and a still valid primary analysis. Little's chi-square test for missing completely at random was performed to assess the missingness in each cohort.

COHORT	**GLASGOW**				**KUALA LUMPUR**				**CHONGQING**				**PERTH**				**GUANGZHOU**				**TOTAL**			
Eligible individuals, N	**1480**				**242**				**245**				**317**				**120**				**2404**			
Treatment outcome, n (%)	**Success**		**Unsuccess**		**Success**		**Unsuccess**		**Success**		**Unsuccess**		**Success**		**Unsuccess**		**Success**		**Unsuccess**		**Success**		**Unsuccess**	
	423	**(29)**	**1057**	**(71)**	**89**	**(37)**	**153**	**(63)**	**125**	**(51)**	**120**	**(49)**	**118**	**(37)**	**199**	**(63)**	**55**	**(46)**	**65**	**(54)**	**810**	**(34)**	**1594**	**(66)**
Characteristics, n (%)																								
Sex	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Age at treatment initiation	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
History of febrile convulsions	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	3	(2.5)	5	(2.5)	0	(0)	0	(0)	3	(0.4)	5	(0.3)
History of CNS infection in childhood	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	1	(0.8)	1	(0.8)	2	(1.0)	0	(0)	0	(0)	1	(0.1)	3	(0.2)
History of significant head trauma	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	1	(0.8)	3	(1.5)	0	(0)	0	(0)	1	(0.1)	3	(0.2)
History of cerebral hypoxic injury	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	3	(2.5)	4	(2.0)	0	(0)	0	(0)	3	(0.4)	4	(0.3)
History of substance abuse	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
History of alcohol abuse	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	1	(0.8)	4	(2.0)	0	(0)	0	(0)	1	(0.1)	4	(0.3)
History of epilepsy in first degree relatives	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	3	(2.5)	6	(3.0)	0	(0)	0	(0)	3	(0.4)	6	(0.4)
Presence of cerebrovascular disease	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Presence of intellectual disability	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	23	(19)	18	(9.0)	0	(0)	0	(0)	23	(2.8)	18	(1.1)
Presence of psychiatric disorder	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	41	(35)	52	(26)	0	(0)	0	(0)	41	(5.1)	52	(3.3)
Number of pre-treatment seizures	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	1	(0.8)	4	(2.0)	0	(0)	0	(0)	1	(0.1)	4	(0.3)
Type of epilepsy	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)	0	(0)
Electroencephalography findings	96	(23)	258	(24)	0	(0)	0	(0)	10	(8.0)	7	(5.8)	2	(1.7)	8	(4.0)	3	(5.5)	5	(7.7)	111	(14)	278	(17)
Brain imaging findings	32	(8)	74	(7)	0	(0)	0	(0)	30	(24)	17	(14)	3	(2.5)	1	(0.5)	2	(3.6)	1	(1.5)	67	(8.3)	93	(5.8)
Number of individuals with missing data	**116**	**(27)**	**299**	**(28)**	**0**	**(0)**	**0**	**(0)**	**34**	**(27)**	**20**	**(17)**	**54**	**(46)**	**74**	**(37)**	**4**	**(7.3)**	**5**	**(7.7)**	**208**	**(26)**	**398**	**(25)**
Missing data pattern - test for data missing completely at random	**Rejected p<0.001**				**N/A**				**Not rejected p=0.94**				**Rejected p<0.001**				**Not rejected p=0.56**				**Rejected p<0.001**			

eTable 2. Treatment outcome and treatment status at the end of 12 months in each cohort.

Treatment outcome[a] and treatment status	Glasgow (n=1065)	Kuala Lumpur (n=242)	Chongqing (n=191)	Perth (n=189)	Guangzhou (n=111)
Successful	307 (28.8%)	89 (36.8%)	91 (47.6%)	64 (33.9%)	51 (46.0%)
Not successful	758 (71.2%)	153 (63.2%)	100 (52.4%)	125 (66.1%)	60 (54.0%)
Remained on first ASM	522 (49.0%)	100 (41.3%)	90 (47.1%)	95 (50.3%)	17 (15.3%)
First ASM ceased	175 (16.4%)	25 (10.3%)	9 (4.7%)	14 (7.4%)	23 (20.7%)
Lack of efficacy	57 (5.3%)	16 (6.6%)	3 (1.6%)	4 (2.1%)	13 (11.7%)
Intolerable adverse effects	118 (11.1%)	9 (3.7%)	6 (3.1%)	10 (5.3%)	10 (9.0%)
Another ASM added	61 (5.7%)	28 (11.6%)	1 (0.5%)	16 (8.5%)	20 (18.0%)

[a]Treatment was deemed successful if the patient was seizure-free while taking the first antiseizure medication during the entire first year of treatment. Treatment was deemed unsuccessful if the patient had recurrent seizures of any type, the first antiseizure medication was ceased and replaced with a different drug (due to persistent seizures or intolerable adverse effects), or other antiseizure medication(s) was added due to persistent seizures. ASM, antiseizure medication

eTable 3. Tuned hyperparameters for the machine learning models.

Algorithm	Parameters
Transformer	'learning rate',1e-2, 'weight decay factor':1e-2, 'number of epochs':50, 'early stopping epochs':50
Multilayered perceptron	'hidden_layer_sizes':(200,200)
Logistic regression	'C': 100, 'tol': 0.001
Support vector machine	'C': 1000, 'kernel': 'linear'
XGBoost	'colsample_bytree': 0.5, 'learning_rate': 0.02, 'min_child_weight': 1, 'n_estimators': 30, 'subsample': 0.7, 'max_depth': 10, 'booster': 'gbtree'
Random forest	'n_estimators': 1000, 'criterion': 'entropy', 'max_depth': 20

eTable 4. Computational time (inference time) for the machine learning models. Inference time of one sample is reported for each model. Python package time (https://pypi.org/project/times/) was used to calculate the running time of each algorithm applied to the test set in experiment 1. All experiments were implemented with Pytorch 0.4.1, python 3.8.1 using one 3090Ti GPU.

Model parameter	**Transformer**	**Multilayered perceptron**	**Logistic regression**	**Support vector machine**	**XGBoost**	**Random forest**
Inference time (seconds)	7.3e-05	7.86e-06	6.04e-06	5.002e-05	5.41e-05	0.0018

eTable 5. Comparison of model performance in subgroups of focal and generalized/unclassified epilepsy.

Model parameter	Transformer[a]	Multilayered perceptron[a]	Logistic regression[a]	Support vector machine[a]	XGBoost[a]	Random forest[a]
Focal epilepsy (n=1392)						
Average AUC[b]	0.64 (0.62-0.66)	0.61 (0.58-0.64)	0.61 (0.59-0.63)	0.61 (0.59-0.63)	0.59 (0.57-0.61)	0.58 (0.56-0.60)
Weighted balanced accuracy	0.62 (0.60-0.64)	0.61 (0.59-0.63)	0.59 (0.58-0.60)	0.61 (0.59-0.63)	0.56 (0.54-0.58)	0.60 (0.58-0.62)
Sensitivity	0.56(0.53-0.59)	0.54 (0.51-0.57)	0.58 (0.56-0.60)	0.48 (0.45-0.51)	0.56 (0.53-0.59)	0.48 (0.45-0.51)
Specificity	0.67 (0.64-0.70)	0.65 (0.61-0.69)	0.60 (0.57-0.63)	0.68 (0.64-0.72)	0.57 (0.54-0.60)	0.66 (0.63-0.69)
Generalized/ unclassified epilepsy (n=406)						
Average AUC[b]	0.58 (0.56-0.60)	0.55 (0.53-0.57)	0.52 (0.50-0.54)	0.51 (0.49-0.53)	0.55 (0.53-0.57)	0.51 (0.49-0.53)
Weighted balanced accuracy	0.60 (0.58-0.62)	0.57 (0.55-0.59)	0.50 (0.48-0.52)	0.50 (0.49-0.51)	0.54 (0.52-0.56)	0.54 (0.52-0.56)
Sensitivity	0.62 (0.61-0.63)	0.57 (0.54-0.60)	0.40 (0.38-0.42)	0.37 (0.34-0.40)	0.45 (0.43-0.47)	0.40 (0.37-0.43)
Specificity	0.66 (0.63-0.69)	0.59 (0.56-0.62)	0.57 (0.55-0.59)	0.59 (0.56-0.62)	0.61 (0.58-0.64)	0.65 (0.62-0.68)

[a]Numbers in parentheses are 95% confidence intervals
[b]Area under receiver operating characteristics curve