Received: 3 October 2020 | Revised: 26 June 2021 | Accepted: 28 June 2021

DOI: 10.1111/epi.17002

Epilepsia®

FULL-LENGTH ORIGINAL RESEARCH

Patient specific prediction of temporal lobe epilepsy surgical outcomes

Marco Benjumeda[1] | **Yee-leng Tan**[2,3,4] | **Karina A. González Otárula**[3] | **Dharshan Chandramohan**[2] | **Edward F. Chang**[5] | **Jeffery A. Hall**[3] | **Concha Bielza**[1] | **Pedro Larrañaga**[1] | **Eliane Kobayashi**[3] | **Robert C. Knowlton**[2]

[1]Computational Intelligence Group, Department of Artificial Intelligence, Universidad Politécnica de Madrid, Madrid, Spain

[2]Department of Neurology, University of California San Francisco Medical Center, San Francisco, CA, USA

[3]Department of Neurology and Neurosurgery, Montreal Neurological Institute and Hospital, McGill University, Montreal, QC, Canada

[4]Department of Neurology, National Neuroscience Institute, Singapore, Singapore

[5]Department of Neurosurgery, University of California San Francisco Medical Center, San Francisco, CA, USA

Correspondence
Marco Benjumeda, Olocip, Calle Chile 10, oficina 210, 28290 Las Rozas de Madrid, Madrid, Spain.
Email: marcobb8@gmail.com

Funding information
Ministerio de Ciencia e Innovación, Grant/Award Number: BES-2014-068637 and PID2019-109247GB-I00; Ministry of Health Singapore; Horizon 2020 Framework Programme, Grant/Award Number: 785907 and 945539; Montreal Neurological Institute; Fundación BBVA

Abstract

Objective: Drug-resistant temporal lobe epilepsy (TLE) is the most common type of epilepsy for which patients undergo surgery. Despite the best clinical judgment and currently available prediction algorithms, surgical outcomes remain variable. We aimed to build and to evaluate the performance of multidimensional Bayesian network classifiers (MBCs), a type of probabilistic graphical model, at predicting probability of seizure freedom after TLE surgery.

Methods: Clinical, neurophysiological, and imaging variables were collected from 231 TLE patients who underwent surgery at the University of California, San Francisco (UCSF) or the Montreal Neurological Institute (MNI) over a 15-year period. Postsurgical Engel outcomes at year 1 (Y1), Y2, and Y5 were analyzed as primary end points. We trained an MBC model on combined data sets from both institutions. Bootstrap bias corrected cross-validation (BBC-CV) was used to evaluate the performance of the models.

Results: The MBC was compared with logistic regression and Cox proportional hazards according to the area under the receiver-operating characteristic curve (AUC). The MBC achieved an AUC of 0.67 at Y1, 0.72 at Y2, and 0.67 at Y5, which indicates modest performance yet superior to what has been reported in the state-of-the-art studies to date.

Significance: The MBC can more precisely encode probabilistic relationships between predictors and class variables (Engel outcomes), achieving promising experimental results compared to other well-known statistical methods. Multisite application of the MBC could further optimize its classification accuracy with prospective data sets. Online access to the MBC is provided, paving the way for its use as an adjunct clinical tool in aiding pre-operative TLE surgical counseling.

KEYWORDS

all epilepsy/seizures, electroencephalography, epilepsy surgery, hippocampal sclerosis, prognosis

Marco Benjumeda, Yee-leng Tan, and Karina A. González Otárula contributed equally to the manuscript.

1 | INTRODUCTION

Temporal lobe epilepsy (TLE) is the most common form of focal epilepsy in adults, with drug resistance developing in approximately one third of patients.[1] Despite being potentially curable with surgery, postoperative seizure-freedom rates vary, ranging from 53% to 84% at 1 year follow-up.[2] Even the so-called "best surgical candidates" might develop seizure recurrence rates reaching 40%–50% 10 years after surgery.[3,4] Meaningful improvement in quality of life is observed primarily only in patients who achieve sustained, that is, long-term, seizure-freedom.[5]

Multiple variables have been shown to influence outcomes in TLE surgery. For example, an antecedent history of febrile seizures, mesial temporal auras, presence of unilateral hippocampal atrophy on imaging, strictly unilateral anterior temporal interictal epileptiform discharges, type I Ebersole ictal electroencephalography (EEG) pattern, and concordant 2-[18F]fluoro-2-deoxy-D-glucose positron emission tomography (FDG-PET) relative hypometabolism have been correlated with good surgical outcomes,[6–9] whereas focal to bilateral tonic-clonic seizures, older age, longer epilepsy duration, frequent pre-operative seizures, bilateral magnetic resonance imaging (MRI) abnormalities, and use of invasive EEG monitoring have been associated with seizure recurrence.[10–12] However, these variables have weak predictive power,[13] and even combining these predictors via multivariate logistic regression modeling achieves only a modest discriminative ability.[14]

An epilepsy surgery nomogram (ESN) and seizure-freedom score have been developed independently to predict success of epilepsy surgery.[15,16] Both models were tested in a group of 20 patients in a recent study (75% of which underwent TLE surgery), and were found to have poor predictive value (area under the curve [AUC] 0.53 for both 2-year and 5-year predictions).[17] However, the same study showed that the best clinical judgment from 24 epilepsy experts was not superior to these models when it came to predicting postoperative outcome either (AUC 0.47 for both 2-year and 5-year predictions). Therefore, a better clinical prediction model designed for patients undergoing TLE surgery remains a very important need in clinical practice.

This work used a supervised machine learning approach to improve surgical outcome predictions based on clinical, neurophysiological, and imaging features collected retrospectively from a TLE surgical data set from the University of California, San Francisco (UCSF) and Montreal Neurological Institute (MNI) over a 15-year period. For that, we trained and validated a multidimensional Bayesian network classifier (MBC) model on combined data sets from both institutions. MBCs offer an explicit and interpretable representation of uncertain knowledge based on the sound concept of probabilistic conditional independence. These classifiers can represent multivariate relationships among the class variables and the features. As probabilistic models, they provide a confidence measure on the predicted labels. Finally, we compared our approach with well-known statistical methods using bootstrap bias corrected cross-validation (BBC-CV) to evaluate the performance of the models.

Key Points

- Currently, there are no reliable clinical tools to predict seizure freedom outcomes after temporal lobe epilepsy (TLE) surgery at the individual patient level
- We built and validated a probabilistic model that can predict short and long-term TLE surgical outcomes
- The multidimensional Bayesian network classifier model was based on clinical, neurophysiological, and imaging variables collected from 231 patients who had undergone TLE surgery
- The model showed promising results compared to other well-known statistical methods
- We developed an online calculator providing individualized surgical outcome predictions that can be used at the bedside

2 | METHODS

2.1 | Study design, participants, and procedures

This study was approved by the research ethics committee from both institutions.

We retrospectively studied a cohort of 231 consecutive TLE patients who underwent surgical treatment at UCSF ($n = 167$) and at the MNI ($n = 64$) between years 2000 and 2015. All patients fulfilled the International League Against Epilepsy (ILAE) definition of drug-resistant epilepsy.[18] Adults were included if (1) unilateral temporal lobe seizure onset was demonstrated during scalp and/or intracranial EEG monitoring; (2) pre-surgical 1.5 or 3.0 Tesla MRI revealed no lesions, or showed hippocampal atrophy but no other developmental or space-occupying lesions in the neocortex; and (3) at least 1 year of postsurgical follow-up was available. Surgical candidacy was discussed at a multidisciplinary team meeting, during which seizure history, clinical findings, video-EEG, and neuroimaging scans were reviewed, and intracranial EEG implantation and/or surgical strategies planned.

Patient demographics and clinical, neurophysiological, and imaging variables were collected (Table S1). Prior to data collection, both institutions (UCSF and MNI) agreed on the definitions of all variables to optimize inter-rater reliability. Because additional pre-surgical imaging modalities were available only for a small subset of patients (such as single-photon emission computed tomography), or not routinely performed as part of pre-surgical evaluation in TLE (ie, magnetoencephalography), these variables were not included. Engel outcomes at years 1, 2, and 5 (Y1, Y2, Y5, when available) following surgery were analyzed as primary end points.[19]

2.2 | Statistical analysis

We used a machine learning approach[20] to predict surgical outcomes at Y1, Y2, and Y5. We distinguished between patients with an Engel I score (ie, free of disabling seizures) vs patients with an Engel score of II-IV (ie, persistence of disabling seizures). "Disabling seizures" included focal aware seizures, which either interfered with function or were noticeable by an observer, focal-onset impaired awareness seizures (FIAS), or focal to bilateral tonic-clonic seizures (FBTCS).[21]

The presence of both categorical and continuous variables in the data sets greatly increases the complexity of the prediction of TLE surgery outcomes. Thus we discretized the continuous variables into categorical variables with a reduced number of intervals by means of fixed frequency discretization.[22] Given a sufficient interval frequency (m), this method discretizes the values in ascending order into intervals of approximately m instances. The main difference between fixed frequency discretization and the well-known equal frequency discretization[23,24] is that the former adapts the number of intervals to the number of observed values, helping to control the discretization variance.

One of the main challenges of this analysis is the high proportion of missing values in the data. To address this problem, we trained a Bayesian network[25] (BN), which explicitly represents a joint probability distribution over a set of random variables. BNs are generative models, and they allow sampling from the posterior distribution of the missing values given the observations. Thus they are well suited for missing data imputation.[26–28] We used the tractable structural expectation-maximization algorithm[29,30] to learn the BN from the incomplete data. This method provides reliable results under the missing at random assumption, which is met when the missingness mechanism is fully accounted for by the variables where there is complete information. The resulting model was used to perform multiple imputations of the data.

We used MBCs, an extension of Bayesian network classifiers,[31] to build the predictive model. MBCs[32,33] adapt BNs to the problem of multidimensional classification, obtaining the most probable joint configuration of the class variables conditioned to an instance of the feature variables. Outcomes at different time scales being evidently related, we trained an MBC that considered these relationships, in which the three class variables were TLE surgery outcome at Y1, Y2, and Y5, and the feature variables constituted the predictors.

To train the model, we used the hill-climbing Bayesian network learning method.[34] To ensure that the output was always an MBC, the arcs from the feature variables to the class variables were included in a blacklist. In addition, we did not allow arcs between features that were not connected to any class variable.

When the parameters of the MBC were estimated by maximum likelihood, the resulting model predicted probabilities that fluctuated excessively when the value of certain individual variables changed. We set uniform Dirichlet priors to the parameters (Bayesian estimation) to improve the stability of the predicted probabilities. The training process is outlined in Figure 1A.

To predict surgery outcomes for new cases, each classifier estimated the probability of Engel I at all the time scales, and the predictions were pooled computing the mean probability for each time period. In addition, when the value of any feature was missing, the BN was used to impute these values according to the information available.

2.3 | Validation

We used BBC-CV[35] to estimate the predictive performance of the models and to select the hyperparameters of the training methods. Intuitively, when several hyperparameter configurations are compared using cross-validation, the performance of the best configuration is an optimistically biased estimate of the performance of the final model. BBC-CV corrects this bias by bootstrapping the process of selecting the best performing hyperparameter configuration. Figure 1B summarizes this procedure.

The training process requires selecting several hyperparameters. To score the structures of the BNs we considered the Bayesian information criterion[36] (BIC) and the Akaike information criterion[37] (AIC). Both scoring functions consist of the log-likelihood of the structure penalized by the number of parameters of the models. However, BIC also penalizes each structure with the size of the data, and usually leads to sparser structures than the AIC. The BNs were used to impute the data 20 times.

We considered the same scoring functions for learning the MBC structures. The strength of the Dirichlet priors on the MBC parameters was tuned using BBC-CV.

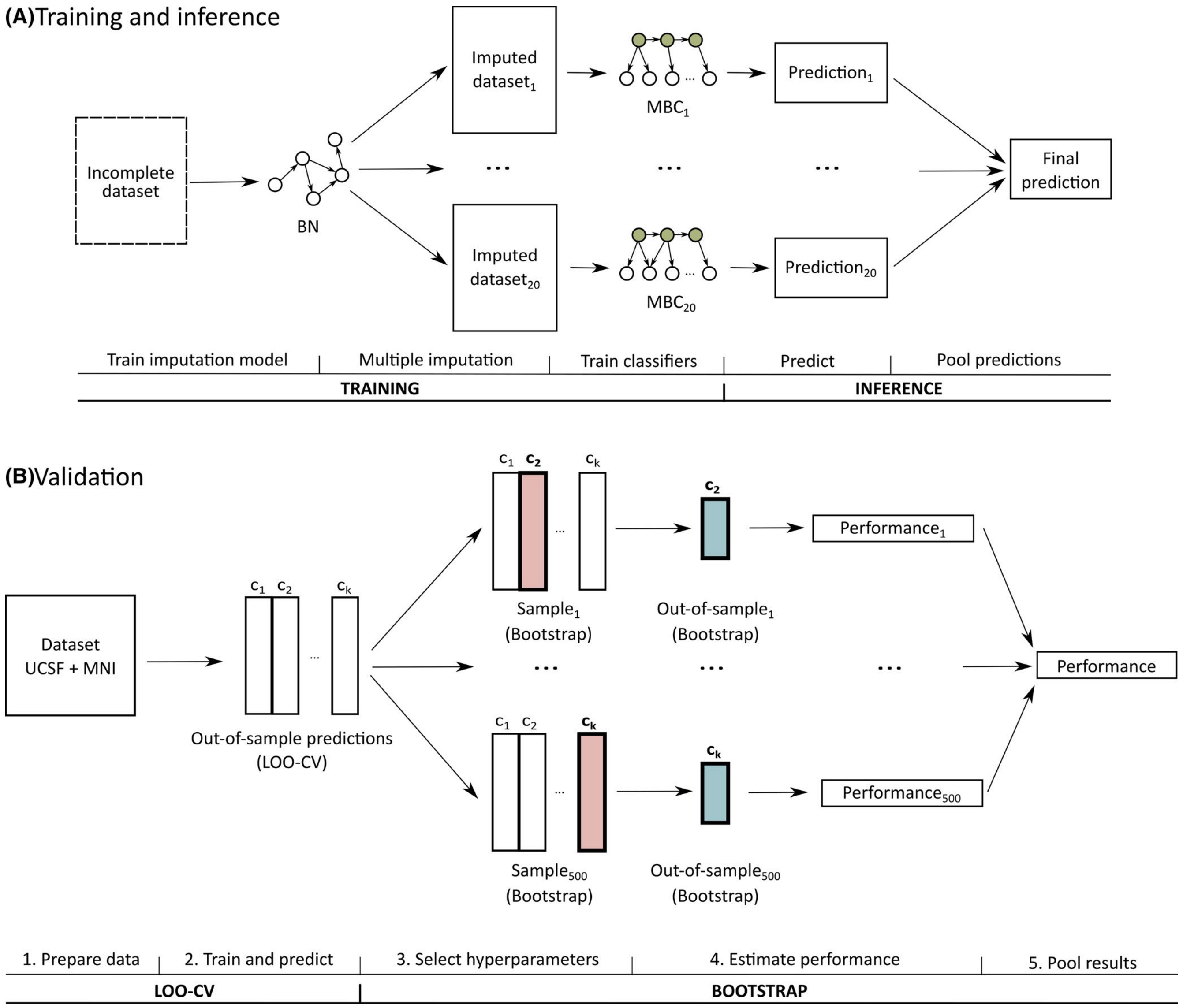

FIGURE 1 Summary of the methods used for data analysis. (A) Outlines the training and inference processes. (B) Summarizes the validation procedure (BBC-CV), which operates as follows: In step 1, the UCSF and MNI data sets are combined and the continuous features are discretized. In step 2, we perform LOO-CV for each hyperparameter configuration (c_1,c_2,…,c_k), and the out-of-sample predictions for each configuration are stored. In this step, the training and inference processes outlined in (A) are used to compute the out-of-sample predictions. Next, the out-of-sample predictions of all configurations are bootstrapped, leading to a matrix of predictions. Subsequently, the configuration with the minimum loss on the bootstrapped data is selected (step 3) and the evaluation metrics are computed on the samples that were not selected by the bootstrap procedure (step 4). The bootstrap process is repeated 500 times, and the average and 95% confidence intervals of the evaluation metrics are returned (step 5). Abbreviations: BBC-CV, bootstrap-bias-corrected cross-validation; BN, Bayesian network; LOO-CV, Leave-one-out cross-validation; MBC, Multidimensional Bayesian network classifier; MNI, Montreal Neurological Institute; UCSF, University of California San Francisco

We compared the MBC with logistic regression and Cox proportional hazards,[38] models previously used to address the prediction of epilepsy surgery outcomes.[14,15] In both cases, the MBC was replaced with the corresponding classifier in the training procedure described in the previous section, and the rest of the process remained unchanged. For logistic regression, we fitted an independent binary classifier to each time period. To avoid overfitting, we penalized the models using Lasso and Ridge regularization. We refer to these methods as Logreg-L1 and Logreg-L2, respectively. In both cases, the regularization strength was tuned with BBC-CV. To train the Cox models, we formulated the problem in terms of survival analysis, and response variables measured the time until the patient developed an Engel score of II-V. We used Ridge regularization to penalize the complexity of the Cox models. The regularization strength was tuned with BBC-CV. Lasso regularization was not available in the software package used to train the Cox models. To evaluate the model performance, we computed the AUC, the Brier score (Brier), the classification accuracy (ACC), the sensitivity or true positive rate (TPR), and the specificity or true negative rate (TNR) of the models. The Brier score measures the accuracy of the probabilities returned by the models, which accounts for their calibration. It has values between 0 and 1, and the predicted probabilities are more accurate as the Brier score takes lower values.

An in-house developed Python 2.7 package was used to train the BN and the MBC.[39] Python package pysurvival,[40] version 0.1.2, was used to train the Cox models. We used scikit-learn, version 0.18.1, to train the logistic regression models and to compute all the evaluation metrics.

2.4 | Data availability

Anonymized data will be shared upon request from any qualified investigator with due approval from her/his institutional ethics board.

3 | RESULTS

Table 1 shows the class variable counts in the UCSF data set (167 patients) and in the MNI data set (64 patients). In both cases there is a high rate of missing values at Y2 and Y5, as well as in some features (see Table S1). Both data sets are fairly balanced, although there is a higher proportion of Engel I outcomes in the MNI data set than in the UCSF data set. The intervals obtained after discretizing the continuous variables are shown in Table 2.

Table 3 provides the performance of the models under a BBC-CV scheme. In addition, the receiver-operating characteristic (ROC) curves for each method and time scale are shown in Figure S1.

The MBC achieved the highest AUC alongside Logreg-L2 at Y1, retaining the highest at Y2 and the second highest at Y5. Logreg-L2 always performed better than Logreg-L1 at all time periods according to this metric. Although the Cox model obtained the best result at Y5, it was the model with lowest AUC at Y1 and Y2.

Logreg-L2 obtained the lowest Brier score overall, and only the MBC performed better at Y5 according to this metric. The Cox model obtained the highest (ie, the worst) Brier scores at each time scale, especially at Y1 and Y2.

The MBC obtained the highest accuracy at the three time scales, and it was the best model at detecting negative cases, whereas it was close to the logistic regression models in terms of TPR. The Cox model had the highest TPR and the lowest TNR in all cases.

Figure 2 shows the calibration plots of all the models, which are consistent with the Brier scores reported in Table 3. Logreg-L2 yielded the best calibration out of the compared models. The MBC model overestimated high probabilities and underestimated low probabilities at Y1 and Y2. The Cox model overestimated the probability of Engel I at Y2 and Y5.

The confidence intervals are wide in most cases, and their width increases with the number of missing values in the class variables. Overall, the differences among the results obtained by the compared methods at the different time scales are not large.

BBC-CV selected the next hyperparameters for training the MBC: The scoring function for evaluating the structures of the BN and the MBC was AIC, and the scale of the Dirichlet prior for the parameters of the MBC was set to 0.25. These hyperparameters were used to train the final models from the complete data set.

Table S2 provides the features selected during BBC-CV by the MBCs. The structure of the BN (imputation model) is shown in Figure S2. We provide an interpretation of these results in the discussion.

4 | DISCUSSION

Our machine learning approach generated meaningful results despite minimal supervision and few constraints on

TABLE 1 Class variable frequencies

MNI				UCSF			
Time scale	E I	E II–IV	Missing	Time scale	E I	E II–IV	Missing
Year 1	40	24	0	Year 1	95	72	0
Year 2	36	14	14	Year 2	62	57	48
Year 5	9	7	48	Year 5	26	25	116

Note: Class variable frequencies in the MNI (left) and UCSF (right) data sets. For each time scale (in rows), the table shows the number of instances belonging to an Engel score of I, the number of instances with an Engel score greater than I, and the number of missing instances.

Abbreviations: E I, Engel score I; E II-IV, Engel score greater than I; MNI, Montreal Neurological Institute; UCSF, University of California San Francisco.

TABLE 2 Intervals obtained after applying the fixed frequency discretization algorithm to discretize the continuous variables

Variable	Intervals
Age at seizure onset, years	[0, 3], (3, 9], (9, 15.6], (15.6, 27], >27
No. of FIAS/month	[0, 1.2], (1.2, 3], (3, 7], (7, 10], >10
No. of GTC seizures/year	[0, 1], (1, 12], >12
No. of disabling seizures captured/duration of vEEG in days	[0, 0.33], (0.33, 0.6], (0.6, 0.83], (0.83, 1.33] >1.33
Proportion of FBTC seizures/no. of disabling seizures captured during vEEG	[0, 0.718], >0.718
Total no. of AEDs tried	[0, 4], 5, 6, 7, >8
No. of AEDs at time of surgery	[0, 2], 3, >4
Age at surgery, years	[0, 25.2], [25.2, 33], [33, 40], [40, 47], >47
Duration of epilepsy until surgery, years	[0, 9], [9, 14], [14,23], [23,31], >31

Abbreviations: AED, antiepileptic drug; EEG, electroencephalography; EFD, equal frequency discretization; FBTC, focal to bilateral tonic-clonic; FIAS, focal impaired awareness seizures; vEEG, video-electroencephalography.

TABLE 3 Performance of the models under a BBC-CV scheme

Time scale	Method	AUC	Brier	ACC	TPR	TNR
Y1	MBC	**0.667 (0.564,0.758)**	0.240 (0.196,0.283)	**0.649 (0.576,0.734)**	0.779 (0.673,0.885)	**0.465 (0.308,0.620)**
	Logreg-L1	0.652 (0.551,0.743)	0.234 (0.201,0.272)	0.639 (0.561,0.718)	0.774 (0.649,0.901)	0.449 (0.278,0.600)
	Logreg-L2	0.667 (0.564,0.759)	**0.225 (0.196,0.262)**	0.647 (0.571,0.730)	0.789 (0.660,0.890)	0.446 (0.317,0.591)
	Cox	0.646 (0.545,0.742)	0.264 (0.217,0.316)	0.614 (0.538,0.692)	**0.878 (0.802,0.956)**	0.242 (0.123,0.370)
Y2	MBC	**0.716 (0.607,0.824)**	0.230 (0.173,0.283)	**0.658 (0.568,0.754)**	0.770 (0.65,0.889)	**0.504 (0.333,0.680)**
	Logreg-L1	0.672 (0.566,0.786)	0.234 (0.191,0.282)	0.604 (0.512,0.702)	0.737 (0.611,0.871)	0.420 (0.259,0.579)
	Logreg-L2	0.684 (0.576,0.798)	**0.226 (0.190,0.265)**	0.629 (0.533,0.720)	0.786 (0.650,0.906)	0.413 (0.250,0.600)
	Cox	0.646 (0.532,0.758)	0.264 (0.210,0.328)	0.628 (0.537,0.717)	**0.808 (0.690,0.914)**	0.380 (0.229,0.521)
Y5	MBC	0.673 (0.450,0.850)	**0.225 (0.163,0.287)**	**0.652 (0.483,0.808)**	0.716 (0.500,0.933)	**0.578 (0.273,0.833)**
	Logreg-L1	0.594 (0.362,0.789)	0.238 (0.194,0.284)	0.539 (0.362,0.714)	0.668 (0.392,0.936)	0.395 (0.100,0.667)
	Logreg-L2	0.624 (0.415,0.819)	0.236 (0.197,0.278)	0.577 (0.413,0.745)	0.709 (0.438,0.923)	0.428 (0.167,0.667)
	Cox	**0.696 (0.512,0.898)**	0.243 (0.164,0.321)	0.604 (0.437,0.755)	**0.838 (0.651,1.000)**	0.344 (0.111,0.592)

Note: For each metric we show its expected value and the 95% confidence interval. The best results for each metric and time scale are highlighted in bold.

Abbreviations: ACC, accuracy; AUC, area under the receiver-operating characteristic curve; Brier, Brier score; Cox, Cox proportional hazards; Logreg-L1, logistic regression with Lasso regularization; Logreg-L2, logistic regression with Ridge regularization; MBC, Multidimensional Bayesian network classifier; TNR, true negative rate; TPR, true positive rate; Y1, 1 year after surgery; Y2, 2 years after surgery; Y5, 5 years after surgery.

the model. This study demonstrates the potential for such a model to give us valuable clinical insights and to express complex relationships in patient data beyond what can be offered by expert-designed, highly constrained, hypothesis-driven statistical models.

4.1 | Model performance

The MBC performed the best on average among the models tested in the experiments in terms of discrimination, although the differences were not large. It obtained higher

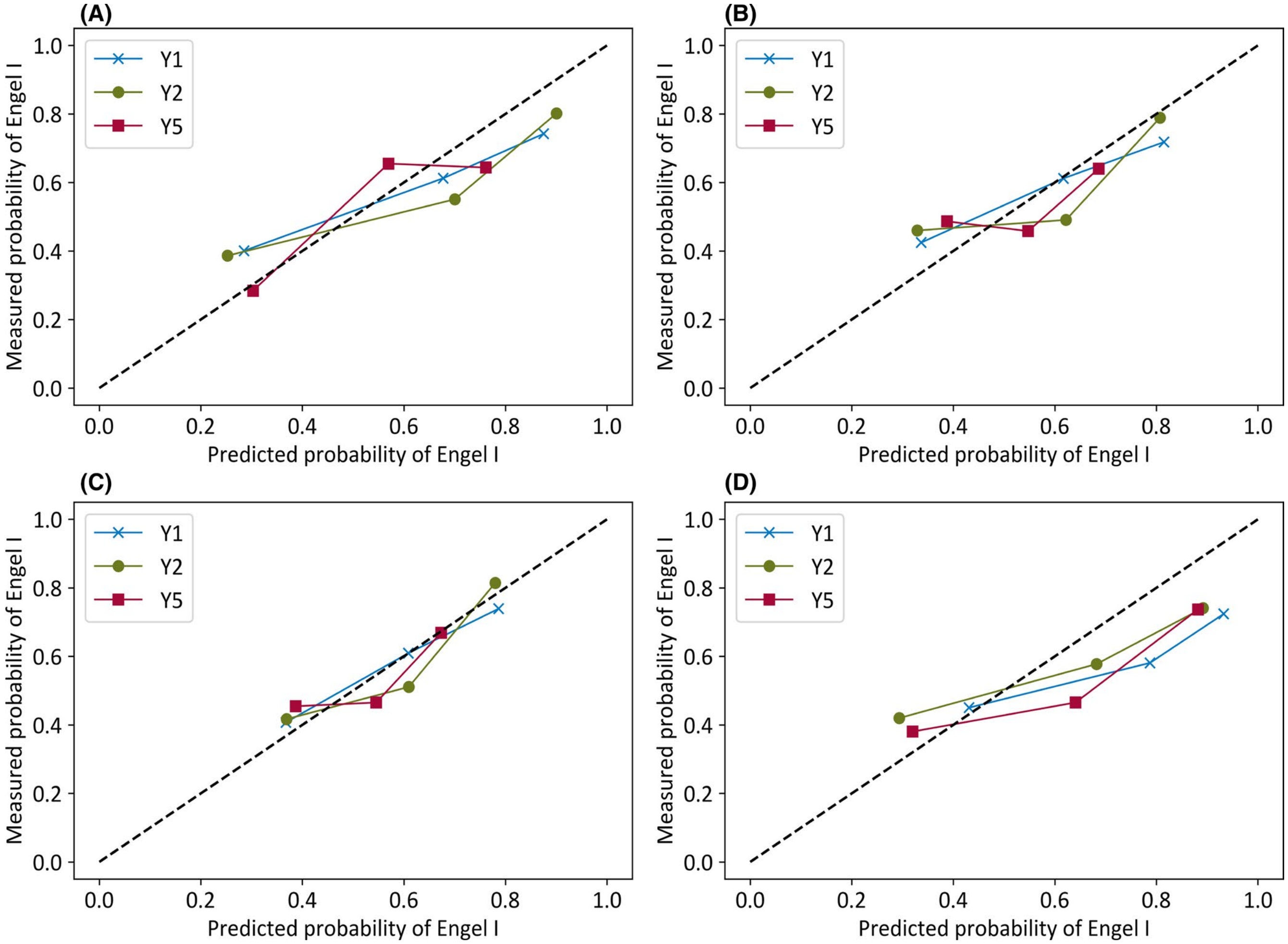

FIGURE 2 Calibration curves of the models in BBC-CV. Calibration of the MBC (A), Logreg-L1 (B), Logreg-L2 (C), and Cox (D). Each plot provides the calibration of a model at Y1 (blue), Y2 (green), and Y5 (red).

Abbreviations: Cox, Cox proportional hazards; Logreg-L1, logistic regression with Lasso regularization; Logreg-L2, logistic regression with Ridge regularization; MBC; Multidimensional Bayesian network classifier; Y1, 1 year after surgery; Y2, 2 years after surgery; Y5, 5 years after surgery

AUC than the logistic regression models in most cases. Although Logreg-L2 obtains the same AUC as the MBC at Y1, its performance worsens at the other time scales, especially at Y5. Given that an independent logistic regression is fitted to each class variable, the model cannot take advantage of the relationships among the surgery outcomes at different time scales. On the contrary, the MBC induces from the data the dependencies among the class variables, and if a feature is connected to a class variable in its structure it also affects the rest. For example, if variable *MRI findings* are connected to Y1, this feature also affects the prediction of Y5.

The MBC achieved better results than the Cox model on average according to all the metrics except TPR. The Cox model assumes a multiplicative relationship between covariates and prediction and cannot represent the individual contribution of each feature to each class variable. The MBC is sufficiently expressive to represent that a feature may influence differently two time scales. For instance, a feature connected to two class variables in the structure of the MBC will probably have a different effect in the prediction of each class.

These results suggest that the model that best describes the underlying statistical relationships between clinical predictor variables and surgical outcomes at the three time scales might be best represented by an MBC-style network. The performance of the MBC (AUC 0.67 at Y1, 0.72 at Y2, and 0.67 at Y5) at all the time scales is modest. However, all models compared in our experiments improve the results reported in the state-of-the-art for predicting epilepsy surgery outcomes.[17] Although the 95% confidence intervals (CIs) are wide, we can draw some conclusions about the short-term performance of the model. The CIs of AUC at Y1 and Y2 are (0.564,0.758) and (0.607,0.824), respectively. Thus it is very likely that the model will discriminate better than chance, but its performance could not be excellent either.

As a deliverable of the current work toward allowing researchers to compare future model performances with the MBC, we have developed an online calculator that is freely available on the web[41] allowing the reader to enter feature variables for individual TLE patients and to obtain an automated individualized prediction of seizure-free probability at different time scales.

4.2 | Clinical interpretation of the models

Our proposal consists of two models, a BN for imputing the missing values and MBCs for predicting surgery outcomes. One of the advantages of these models is their intuitive graphical representation.

The imputation model is explained by the structure of the BN (Figure S2). Each variable in the network is conditionally independent on the rest given its Markov blanket, which is composed of its parents, its children, and the parents of its children in the graph. This means that if the value of a variable is missing, we can estimate it using only the Markov blanket of this variable. For example, if the *Age at surgery* is unknown, the model would only consider *Age at seizure onset*, *Duration of epilepsy until surgery*, *History of heavy alcohol/substance use*, and *Static encephalopathy* to estimate the value of this variable. The BN is used to impute the missing values in the training data and during inference. Reassuringly, most of the arcs included in the structure satisfied clinical intuition.

The same principles apply for MBCs. However, the purpose of these models is to represent the posterior distribution of the class variables given the features rather than to encode the joint distribution of all the variables in the data set. Therefore, an MBC has a restricted topology that is composed of three subgraphs: A class subgraph, which represents the relationships among class variables; a bridge subgraph, which contains the arcs between class variables and features; and a feature subgraph, which contains arcs between features. Note that the bridge subgraph determines the subset of features that are selected to make predictions.

Next, we analyze the MBCs learned under a BBC-CV scheme. The class subgraph of all the MBCs contains an arc from Y1 to Y2, and from Y2 to Y5, which means that the class variables are clearly related. In addition, more than 99% of the models obtained under a BBC-CV scheme contained arcs from Y1, Y2, and Y5 to *Laterality of interictal spike*, *MRI findings*, and *Reoperation case*, respectively (see Table S2). This implies that *Reoperation case* may be a good predictor for long-term seizure freedom, whereas *Laterality of interictal spike* and *MRI findings* provide more information about short-term seizure freedom. The three features that are connected with more class variables on average are *MRI findings*, *Type of surgery*, and *No. of GTC seizures/year*. This is concordant with clinical studies, which have shown seizure-freedom rates a year after TLE surgery to be higher in the presence of a lesion on MRI ("MRI-positive" TLE 75% vs "MRI-negative" TLE 51%).[42] All of these features were selected by the final model.

The interpretation of the predicted probabilities generated by the MBC should be informed by the associated calibration curves (Figure 2). The calibration plots suggest that the MBC may be overconfident, given that it outputs probabilities that are more extreme than they should be. Using a higher complexity penalty on the structure of the MBCs or a stronger prior on the parameters may improve the calibration. However, these changes limit the discriminative power of the MBC according to the experimental results. A larger sample size should mitigate this problem.

The outputs of any classification model should be interpreted in the proper clinical context when pre-surgical counseling takes place – even if the algorithm predicts a 40%–50% chance of mid-term seizure freedom after TLE surgery, surgery still confers a far greater chance of seizure freedom compared to additional trials of antiepileptic medications. Less than 5% of patients become seizure-free, with a third medication regimen after drug resistance is observed.[43]

4.3 | Limitations

The inherent limitations of analyzing retrospectively collected data were missing data values and patients lost to follow-up over time. One can argue that these patients who "disappeared" from our clinics were those who most likely were cured by surgery and evaded clinical surveillance. It is unlikely that a patient who continues to experience seizures would not return for reassessment, but that can only be speculated here.

If we assume that the mechanism for "drop-out" of the incomplete variables is accounted for by the observed variables in the data set, then the methods used for dealing with missing data should still suffice and provide a reliable output.

We recognize that although our number of 231 TLE patients is the largest data set to date in machine learning literature for epilepsy surgical outcome prediction,[44] a much larger number is still desirable and needed for model optimization. Of note, only TLE patients with MRI studies showing hippocampal atrophy or normal-appearing hippocampi and temporal neocortex were included in the study, as the number of patients with other MRI findings such as temporal lobe neoplasms or MRI-visible cortical malformations was comparatively much lower and risked being under-represented. Hence the current MBC model is not generalizable to other subtypes of TLEs, until these other underlying etiologies are included in prospective training and validation data sets.

Finally, medical data are inherently noisy, and approximated values may have led to erroneous classification. Noise tolerance is hence a requirement during analysis. These key issues can only be overcome and minimized with large prospectively collected data inputs, before performance levels can become consistent and acceptable.

5 | CONCLUSIONS AND FUTURE DIRECTION

Seizure-freedom outcome prediction for TLE surgery based on MBC modeling obtained results that were comparable to the well-known logistic regression and Cox proportional hazards statistical models. The model yielded 0.67, 0.72, and 0.67 AUC for outcome predictions at years 1, 2, and 5, respectively, indicating promising predictive power compared to previous results for predicting epilepsy surgery outcomes. Before this can become a clinical tool in aiding pre-operative counseling, further testing is needed in prospective large cohorts to ensure its reproducibility. Moreover, follow-up information beyond 5 years would provide a more realistic timeframe for long-term prediction. To further enhance this tool, our long-term goal is to allow the MBC classifier to continuously learn from prospectively entered data, which will optimize its classification accuracy. Future iterations of this model may also utilize priors derived from other studies regarding the effects of individual predictors on overall seizure freedom.

ACKNOWLEDGMENTS

This work has been supported partially by the Spanish Ministry of Science and Innovation through the PID2019-109247GB-I00 project, by the EBRAINS research infrastructure, funded from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), and by Fundación BBVA grants to Scientific Research Teams in Big Data 2019. This project has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under Specific Grant Agreement No. 785907 (HBP SGA2). MB was supported by a predoctoral contract for the formation of doctors from the Spanish Ministry of Science and Innovation (BES-2014-068637). YLT was funded by the Ministry of Health HMDP Scholarship from Singapore. KGO was funded by the Frederick Andermann Epilepsy and EEG fellowship from the MNI, through an unrestricted educational grant awarded by UCB Canada.

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.

AUTHOR CONTRIBUTION

MB, YLT, and KGO had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

ORCID

Marco Benjumeda https://orcid.org/0000-0001-6681-1699
Eliane Kobayashi https://orcid.org/0000-0002-1713-1563
Robert C. Knowlton https://orcid.org/0000-0002-5146-9664

REFERENCES

1. Semah F, Picot M-C, Adam C, Broglin D, Arzimanoglou A, Bazin B, et al. Is the underlying cause of epilepsy a major prognostic factor for recurrence? Neurology. 1998;51(5):1256–62.
2. Spencer S, Huh L. Outcomes of epilepsy surgery in adults and children. Lancet Neurol. 2008;7(6):525–37.
3. Spencer SS. Long-term outcome after epilepsy surgery. Epilepsia. 1996;37(9):807–13.
4. 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(Pt 9):2018–30.
5. Markand ON, Salanova V, Whelihan E, Emsley CL. Health-related quality of life outcome in medically refractory epilepsy treated with anterior temporal lobectomy. Epilepsia. 2000;41(6):749–59.
6. Adry RARC, Meguins LC, Pereira CU, Silva Júnior SC, Araújo Filho GM, Marques LHN. Auras as a prognostic factor in anterior temporal lobe resections for mesial temporal sclerosis. Eur J Neurol. 2018;25(11):1372–7.
7. Tatum WO 4th, Benbadis SR, Hussain A, Al-Saadi S, Kaminski B, Heriaud LS, et al. Ictal EEG remains the prominent predictor of seizure-free outcome after temporal lobectomy in epileptic patients with normal brain MRI. Seizure. 2008;17(7):631–6.
8. Willmann O, Wennberg R, May T, Woermann FG, Pohlmann-Eden B. The contribution of 18F-FDG PET in preoperative epilepsy surgery evaluation for patients with temporal lobe epilepsy: a meta-analysis. Seizure. 2007;16(6):509–20.
9. Sylaja PN, Radhakrishnan K, Kesavadas C, Sarma PS. Seizure outcome after anterior temporal lobectomy and its predictors in patients with apparent temporal lobe epilepsy and normal MRI. Epilepsia. 2004;45(7):803–8.
10. Jeha LE, Najm IM, Bingaman WE, Khandwala F, Widdess-Walsh P, Morris HH, et al. Predictors of outcome after temporal lobectomy for the treatment of intractable epilepsy. Neurology. 2006;66(12):1938–40.
11. Janszky J, Janszky I, Schulz R, Hoppe M, Behne F, Pannek HW, et al. Temporal lobe epilepsy with hippocampal sclerosis: predictors for long-term surgical outcome. Brain. 2005;128(Pt 2):395–404.

12. Fong JS, Jehi L, Najm I, Prayson RA, Busch R, Bingaman W. Seizure outcome and its predictors after temporal lobe epilepsy surgery in patients with normal MRI. Epilepsia. 2011;52(8):1393–401.
13. Aull-Watschinger S, Pataraia E, Czech T, Baumgartner C. Outcome predictors for surgical treatment of temporal lobe epilepsy with hippocampal sclerosis. Epilepsia. 2008;49(8):1308–16.
14. Uijl SG, Leijten FSS, Arends JBAM, Parra J, van Huffelen AC, Moons KGM. Prognosis after temporal lobe epilepsy surgery: the value of combining predictors. Epilepsia. 2008;49(8):1317–23.
15. Jehi L, Yardi R, Chagin K, Tassi L, Lo Russo G, Worrell G, et al. Development and validation of nomograms to provide individualised predictions of seizure outcomes after epilepsy surgery: a retrospective analysis. Lancet Neurol. 2015;14(3):283–90.
16. Gracia CG, Yardi R, Kattan MW, Nair D, Gupta A, Najm I, et al. Seizure freedom score: a new simple method to predict success of epilepsy surgery. Epilepsia. 2015;56(3):359–65.
17. Gracia CG, Chagin K, Kattan MW, Ji X, Kattan MG, Crotty L, et al. Predicting seizure freedom after epilepsy surgery, a challenge in clinical practice. Epilepsy Behav. 2019;95:124–30.
18. Kwan P, Arzimanoglou A, Berg AT, Brodie MJ, Allen Hauser W, Mathern G, et al. Definition of drug resistant epilepsy: consensus proposal by the ad hoc Task Force of the ILAE Commission on Therapeutic Strategies. Epilepsia. 2009;51(6):1069–77.
19. Van Ness PC, Rasmussen TB, Ojemann LM, Engel J. Outcome with Respect to Epileptic Seizures. New York: Raven Press; 1993.
20. Bielza C, Larrañaga P. Data-Driven Computational Neuroscience: Machine Learning and Statistical Models. Cambridge: Cambridge University Press; 2020.
21. Engel J Jr, McDermott MP, Wiebe S, Langfitt JT, Stern JM, Dewar S, et al. Early surgical therapy for drug-resistant temporal lobe epilepsy: a randomized trial. JAMA. 2012;307(9):922–30.
22. Yang Y, Webb GI. Discretization for naive-Bayes learning: managing discretization bias and variance. Mach Learn. 2009;74(1):39–74.
23. Catlett J. On changing continuous attributes into ordered discrete attributes. In: Proceedings of the European Working Session on Learning. 2009. p. 164–78.
24. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. In: Proceedings of the Twelfth International Conference on Machine Learning. 1995. p. 194–202.
25. Bielza C, Larrañaga P. Bayesian networks in neuroscience: a survey. Front Comput Neurosci. 2014;8:131.
26. Romero V, Salmerón A. Multivariate imputation of qualitative missing data using Bayesian networks. In: López-Díaz M, Gil M, Grzegorzewski P, Hryniewicz O, Lawry J, editors. Soft Methodology and Random Information Systems. Berlin, Heidelberg: Springer; 2004. p. 605–12.
27. Hruschka ER, Hruschka ER, Ebecken NFF. Bayesian networks for imputation in classification problems. J Intell Inf Syst. 2007;29(3):231–52.
28. Niloofar P, Ganjali M. A new multivariate imputation method based on Bayesian networks. J Appl Stat. 2014;41(3):501–18.
29. Benjumeda M, Luengo-Sanchez S, Larrañaga P, Bielza C. Tractable learning of Bayesian networks from partially observed data. Pattern Recognit. 2019;91:190–9.
30. Friedman N. Learning belief networks in the presence of missing values and hidden variables. In: Proceedings of the Fourteenth International Conference on International Conference on Machine Learning. 1997. p. 125–33.
31. Bielza C, Larrañaga P. Discrete Bayesian network classifiers. ACM Comput Surv. 2014;47(1):1–43.
32. Van Der Gaag LC, De Waal PR. Multi-dimensional Bayesian network classifiers. Mach Learn. 2006;3(20):107–14.
33. Bielza C, Li G, Larrañaga P. Multi-dimensional classification with Bayesian networks. Int J Approx Reason. 2011;52(6):705–27.
34. Heckerman D, Geiger D, Chickering DM. Learning Bayesian networks: the combination of knowledge and statistical data. Mach Learn. 1995;20(3):197–243.
35. Tsamardinos I, Greasidou E, Borboudakis G. Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. Mach Learn. 2018;107(12):1895–922.
36. Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6(2):461–4.
37. Akaike H. A new look at the statistical model identification. IEEE Trans Automat Contr. 1974;19(6):716–23.
38. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol. 1972;34(2):187–202.
39. Benjumeda M. (2021, July). *Code of the methods proposed in the PhD thesis "Learning Tractable Bayesian Networks"*. https://github.com/marcobb8/tr_bn.
40. Fotso S. PySurvival: open source package for survival analysis modeling; 2019.
41. Benjumeda M, Tan Y-L, González Otárula KA, Chandramohan D, Chang EF, Hall JA, et al. (2021, July). *Online tool for patient specific prediction of temporal lobe epilepsy surgical outcomes*. https://marcobb8.shinyapps.io/shiny/.
42. Téllez-Zenteno JF, Hernández Ronquillo L, 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.
43. Kwan P, Brodie MJ. Early identification of refractory epilepsy. N Engl J Med. 2000;342(5):314–9.
44. Abbasi B, Goldenholz DM. Machine learning applications in epilepsy. Epilepsia. 2019;60(10):2037–47.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Benjumeda M, Tan Y-L, González Otárula KA, Chandramohan D, Chang EF, Hall JA, et al. Patient specific prediction of temporal lobe epilepsy surgical outcomes. Epilepsia. 2021;62:2113–2122. https://doi.org/10.1111/epi.17002