Physiological Assessment of Delirium Severity using Quantitative EEG: The E-CAM-S

Running head: Delirium and Outcomes from EEG

Supplemental Digital Content

1. Material and Methods

This section describes the methods, particularly technical aspects, in more detail.

1.1. Study Setting and Participants

We conducted a single-center, retrospective, observational cohort study of adult inpatients who underwent EEG monitoring as a part of routine clinical care at Massachusetts General Hospital (MGH) between August 2015 to December 2019. Patients were considered from all wards, including medical, surgical, and neurological floors, as well as intensive care units (ICUs). Patients were excluded if they had a recorded history of dementia, other intellectual disability, deafness, aphasia, or were non-English speaking (if non-comatose). The study was conducted under a protocol approved by the Institutional Review Board using a waiver of consent.

1.2 Clinical Data

Patients were assessed to determine delirium presence and severity at the bedside by study staff. Staff were trained to perform assessments through a combination of didactics, literature review, in person case reviews, and ongoing discussions. Assessments were performed either during or within 1 hour of beginning or discontinuation of EEG recording. Delirium presence was assessed using the CAM short form[S1]. Delirium severity was assessed using the CAM-S scoring method[2]. The CAM-S Short Form scores the severity of four features: (1) Acute onset & fluctuating course (0 or 1 point); (2) Inattention (0, 1, 2 points); (3) Disorganized thinking (0, 1, 2 points); and (4) Altered level of consciousness (0, 1, 2 points). Delirium severity is scored as the sum of the severity of all four features (total between 0 and 7 points). Responses to individual questions were considered "normal" only if there was an unequivocally correct response. In cases where a patient did not answer a question, the question was repeated. If questions remained unanswered (including due to a decreased level of arousal), non-answers were scored as incorrect. Patients were also evaluated with the Richmond Agitation Sedation Scale (RASS; range -5 to +4, normal score 0[S2]) to assess level of consciousness. Under the above framework, comatose patients (RASS score of -4 or -5) were assigned a CAM-S short form score of 7, given that coma and delirium are increasingly considered part of a spectrum of manifestations of underlying encephalopathy pathophysiology[S3,S4]. The analyses below were performed, however, on both the entire cohort (non-delirious, delirious, and comatose patients) and after excluding patients with coma (retaining non-delirious and delirious patients). Clinical outcomes, including length of stay and in-hospital mortality, and Charlson Comorbidity Index (CCI)[S5] were extracted and calculated from the medical record.

1.3 EEG Recording and Pre-Processing

EEGs were recorded with Ag/AgCl scalp electrodes using the standard international 10-20 system for electrode placement; however, we calculated the E-CAM-S using only the four frontal channels, as forehead electrodes are more amenable to application with less technical experience. We selected the following bipolar frontal channels for analysis: Fp1-Fp2, Fp1-F7, Fp2-F8, F7-F8. All EEGs were resampled to 200 Hz and normalized to have zero mean, notch-filtered at 60 Hz, and bandpass filtered from 0.5 Hz to 30 Hz to reduce line noise and myogenic artifacts. For patients referred for routine EEGs, the EEG recordings had durations between 20 and 60 minutes. For patients undergoing long-term EEG monitoring, we selected the one-hour segment in which the clinical CAM-S score was obtained (30 minutes before and after the clinical assessment). The selected EEG signals were segmented into non-overlapping 6-second epochs and automatically checked for artifacts by identifying segments with absolute value > 500 µV or standard deviation < 1 µV in any channel. Epochs not flagged as containing artifacts were subsequently used for feature extraction.

1.4 Feature Extraction

From each 6-second epoch, we extracted 298 features from time and frequency domains, as summarized in **Supplemental Table 1**[S6,S7]. Each feature was calculated for each of the four bipolar channels and then averaged across all channels. This resulted in 298 features for each 6-second epoch. We then calculated four summary statistics for each feature across all 6-second epochs: average, standard deviation, minimum and maximum value, resulting in a total of 298 x 4 = 1192 feature values per EEG recording. We performed pre-processing and feature extraction using MATLAB (version 2019a, Mathworks, Natick, MA, USA).

1.5 Model training and cross validation

Data was split into training and testing data at the patient level, with 90% of EEGs (n = 336) used for training, and 10% (n = 37) for testing. This splitting of data and model training was repeated 10 times, allowing model performance to be evaluated once on each EEG (10-fold cross validation). To avoid overfitting, strict separation was maintained between training and testing data, such that all reported model performance statistics reflect performance on held-out testing data. Model training and evaluation was performed in Python (version 3.8.0).

We created the E-CAM-S by training a machine learning model that uses EEG features as inputs and attempts to produce scores that are correlated with the clinical CAM-S score (0 - 7). The model used is a Learning-to-Rank (LTR) ordinal regression model[S8]. Our LTR model is based on a pairwise approach: a binary classifier, using logistic regression with LASSO penalty, was trained to predict, for each pair of patients A and B, whether A is more severely delirious than B. The value of the decision function of the binary classifier for a particular patient results in the E-CAM-S score, a continuous number indicating delirium severity, bounded between 0 and 1.

Feature selection during model training was accomplished using a two-step procedure:

1) We used a simple approach that selects features with large k Spearman’s correlation. We experimented with different values of k (see below).

2) In order to select among the features that remained, we utilized internal cross validation (ICV) to fit the LTR model, using a least absolute shrinkage and selection operator (LASSO) penalty. This penalty is a non-negative number, whose value is determined by the process of ICV.

The two free parameter values (number of univariate selected features k and the LASSO penalty parameter value) are known as hyperparameters, and the process of choosing these parameters using ICV is known as hyperparameter optimization. These hyperparameter values were tuned using internal 10-fold internal cross validation to optimize the Root Mean Squared Error (RMSE) on the internal validation set. The lowest RMSE was achieved when using k = 65 and a LASSO penalty of 100, therefore these hyperparameters were used for final analyses.

Analyses were performed using the entire cohort and a non-comatose subset. The distribution of clinically assessed delirium severity scores as present in both the entire cohort and non-comatose subset can be seen in **Supplemental Figure 1**.

Technical Background on External and Internal Cross Validation.

We trained and evaluated the model using nested cross validation (NCV), a method which employs external and internal cross validation (ECV, ICV). ECV and ICV share similar mechanics, but each serves a unique purpose: ECV is used to avoid inflation of model performance estimates, whereas ICV prevents model overfitting. **Supplemental Figure 3** shows an overview diagram of nested cross validation and further explanation and elaboration on the exact used method is given below.

The ECV approach here is k-fold cross validation, in which we have chosen to use k = 10 folds. The choice of k = 10 is widely used in developing and evaluating machine learning models, as it generally achieves a favorable bias-variance tradeoff[S9,S10] In this method, we split the data into 10 folds, with nine folds (i.e. 90% of the data) reserved for training and one fold (i.e. 10% of the data, 37 subjects in this case) for testing in each round. We then conduct 10 rounds of cross validation, with a different combination of folds in each round, such that each individual fold is used as the testing data in one of the 10 rounds. This approach in turn generates 10 models, due to different subsets of folds being used in each round of training. These models will generally be similar, but not identical, because a different 10% of the data is held out for testing for each round of ECV.

ECV avoids overestimating model performance by strictly separating training and testing data, such that only model predictions generated from *testing* data are used to evaluate model performance. In contrast, if we did not employ ECV, and we instead fitted and tested the model with the same data, our evaluation of performance would be unable to identify overfitting. Rather, we could find an artificially high correlation coefficient (R) between the CAM – S and VE – CAM – S scores, i.e. an upwardly biased estimate, and the model would potentially fail when applied to new data despite a high coefficient. However, if model overfitting has occurred, ECV will reflect this as poor average model performance on the held out data across the multiple rounds of external cross validation, and we will be aware of this overfitting.

Through ECV, we are able to utilize data from all subjects in the evaluation of model performance. Indeed, figure 2 includes all subjects rather than just 37, because the correlation coefficient of 0.68 was calculated based on all subjects. This correlation coefficient is not biased (i.e. over fit) because this estimate was calculated entirely based on the test folds, which contain subjects that were not used in training within a given round of cross validation. In this way, ECV prevents inflated estimates of the correlation coefficient.

ICV is one of several approaches to avoiding overfitting during model training

The mechanics of ICV are similar to ECV; however, it is important to keep in mind that ICV operates entirely on the training data; the held-out testing data is not used. In ICV, the training data are split into a series of folds, and each fold takes a turn serving as the (internal) testing set while the remaining training data is used to fit the model. In the standard approach, which we follow, ICV is repeated multiple times using different values of the regularization parameter. The results of ICV are used to construct a curve of average model performance on the internal test data as a function of the regularization parameter value. This curve is used to select the value of the regularization parameter that achieves the best average performance during ICV. This parameter value is then used to fit a final model on the entire set of training data. This fit model is then tested on the external held out testing data. The entire ICV procedure is repeated across the 10 folds of external cross validation (ECV).

In summary, ECV and ICV have similar methods, but serve distinct purposes. ECV aims to obtain accurate (unbiased) estimates of model performance, regardless of how the model was fit. By contrast, ICV is one of several techniques that aim to avoid overfitting during model development. ICV helps to avoid model overfitting by separating the training data into folds, internally simulating the model testing process, in hopes of finding parameter values that will lead to good model performance during testing on new data.

Other methods used in this paper to avoid model overfitting during model training

In the present paper, we use three approaches to avoid model overfitting:

1) We filtered out features (in training data) with minimal correlation to the outcome of interest.

2) We applied a penalty term when fitting our ordinal logistic regression model.

3) We employed *internal* cross validation (ICV) to select the optimal value of the regularization parameter in the logistic regression model, as described above.

1.6 Association of E-CAM-S with Mortality and Hospital Length of Stay

To evaluate the clinical significance of E-CAM-S scores, we assessed their association with in-hospital mortality using multivariable logistic regression, including age, sex, and CCI as additional covariates. Age, CAM-S, and CCI were z-normalized prior to model fitting. To compensate for data imbalance (see **Supplemental Figure 2** for histograms of the data set distribution), we assigned a weight to each patient inversely proportional to the number of patients with that mortality status. Association with in-hospital mortality was calculated as the average area under the receiver operating curve (AUROC) using 10-fold cross validation under three conditions: without any delirium information, with E-CAM-S scores included, and with clinically assessed CAM-S scores included.

To determine associations with hospital length of stay, we used log-transformed length of stay as the dependent variable. We then performed multivariable linear regression with three models: without any delirium information, with E-CAM-S scores included, and with clinically assessed CAM-S scores included. Results are reported as Spearman correlations of each multivariable prediction model.

1.7 Statistical Reporting

Medians, interquartile ranges, and proportions were calculated for descriptive analysis given that most of the data was not normally distributed. Groups were compared with Mann-Whitney rank-sum tests and proportions with chi-squared tests. To estimate the 95% CI of the performance metrics and the coefficients of the prediction models, we used 1000 rounds of bootstrapping. In each iteration of bootstrapping, 10-fold cross-validation was performed. The significance level for all tests was set at alpha = 0.05.

To evaluate the correlation between E-CAM-S and clinically assessed CAM-S, we used Spearman correlation coefficients. To evaluate the ability of E-CAM-S to discriminate between patients with vs. without delirium, we used AUROC. We also compared the E-CAM-S with a previously published method[S11] for assessing delirium based on the EEG, using Spearman correlations with the CAM-S and AUROC for predicting delirium presence as evaluation metrics.

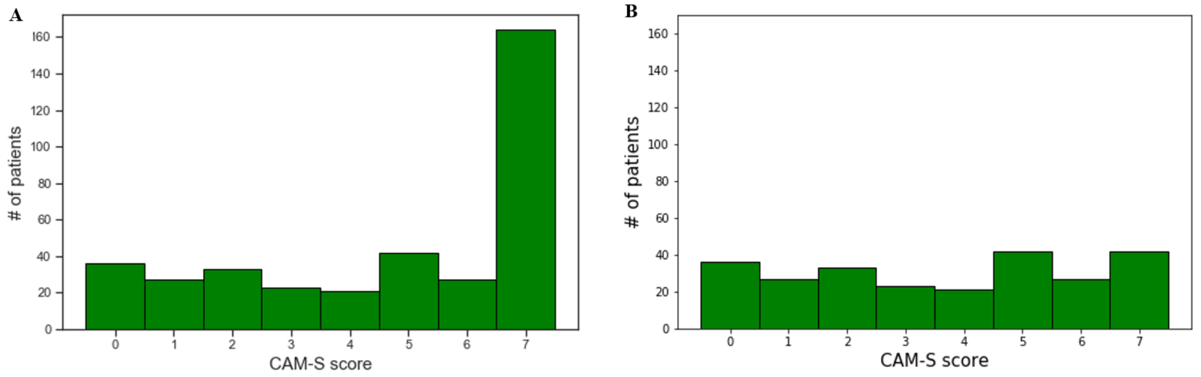
2. Overview of EEG Features

Supplemental Table 1: Extracted EEG features used for the prediction models.

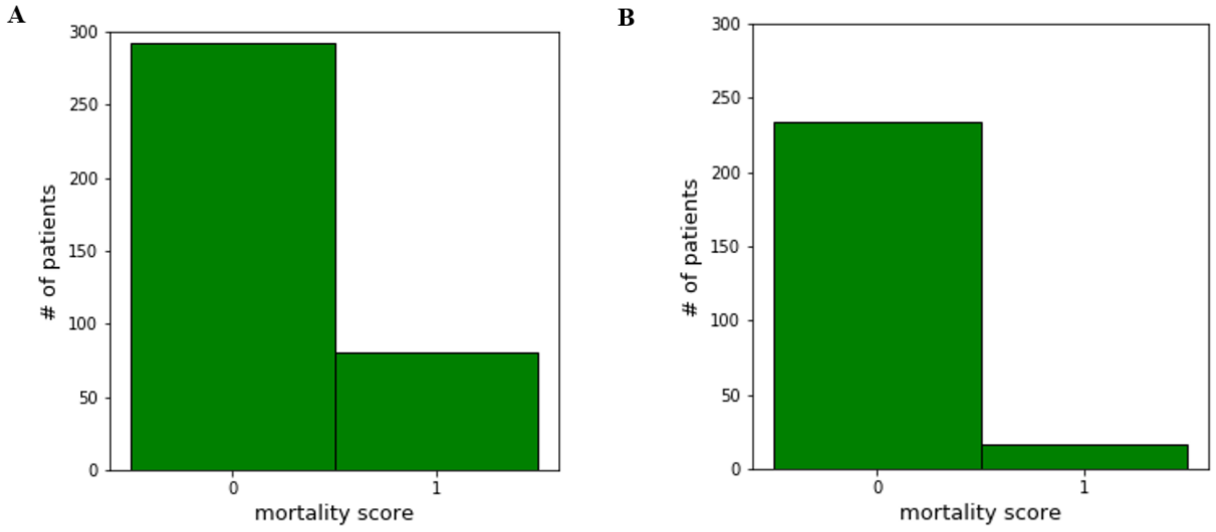
Domain	Feature	Number	Remark
time	Mean, median, 25% percentile, 75% percentile, standard deviation, variance, mean absolute gradient, line-length, Zero Crossing Rate, Hjorth mobility, Hjorth complexity, skewness, kurtosis, Shannon entropy, Higuchi fractal dimension	15	
frequency	Mean spectral (center) frequency, power at center frequency, spectral bandwidth, spectral entropy, spectral edge frequencies: SEF95 and SEF5.	6	Computed over the whole spectrogram (0.5-30 Hz):
	Harmonic indexes computed for specific frequency bands and ratios: mean, median, min, max, std, iqr, 5% percentile, 95% percentile.	8x7=56	Delta (0.5-4), theta (4-8), alpha (8-12), beta (13-20), delta/theta, delta/alpha, theta/alpha.
	Power Spectral Density (PSD) of different frequency bands and band-ratios, calculated for both the PSD value in dB + relative PSD value (PSD value in band x / PSD whole spectrogram). Calculated separately for different frequency bands (26) and the band-ratios (e.g. PSD delta/PSD alpha) (27).	26x2 + 27x2 = 106	Delta: 0.2-3, 0.5-2, 0.5-3, 0.5-4, 1-4, 1-5, 1-6, 2-4, 2-6 Hz Theta: 3-7, 4-6, 4-8, 6-8 Hz Alpha: 7-12, 8-10, 8-11, 8-12, 8-13, 10-12 Hz Beta: 11-16, 11-20, 12-30, 15-25, 21-29 Hz All: 1-20, 1-30 Hz
	Coherence for different frequency bands (26) and band ratios (27), calculated for both the mean and sum of the coherence in the specific bands.	26x2 + 27x2 = 106	Same bands/ratio's as for PSD calculation.

	FOOOF parameters: max amplitude, max frequency, max bandwidth, number of peaks, broadband offset and exponent of aperiodic fit (for both 1-15 Hz and 15-30 Hz range).	8	FOOOF parameterizes neural power spectra: https://fooof-tools.github.io/fooof/
# features		= 298	for each 6 seconds epoch
Total number of features per EEG		*298 x 4 = 1192*	*Average, std, min, max across all 6-second epochs.*

3. Data set distribution

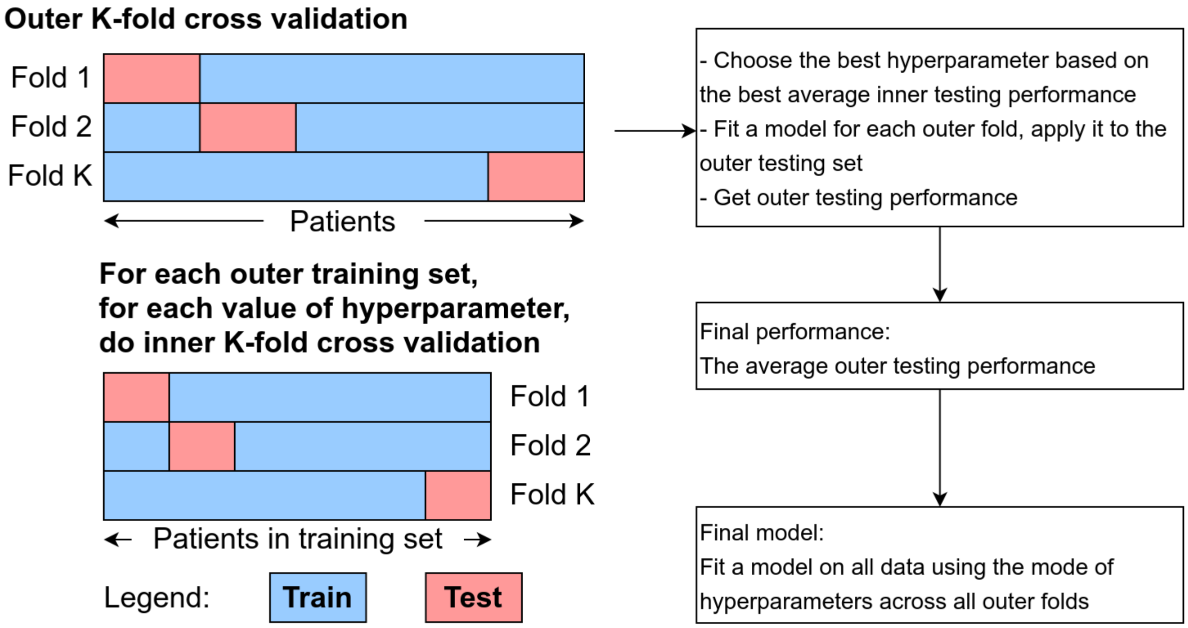


Supplemental Figure 1: Histogram of CAM-S distribution for entire population **(A)** and non-comatose subset **(B).**



Supplemental Figure 2: Histogram of mortality score distribution for entire population **(A)** and non-comatose subset **(B).**

4. Methods: nested cross-validation

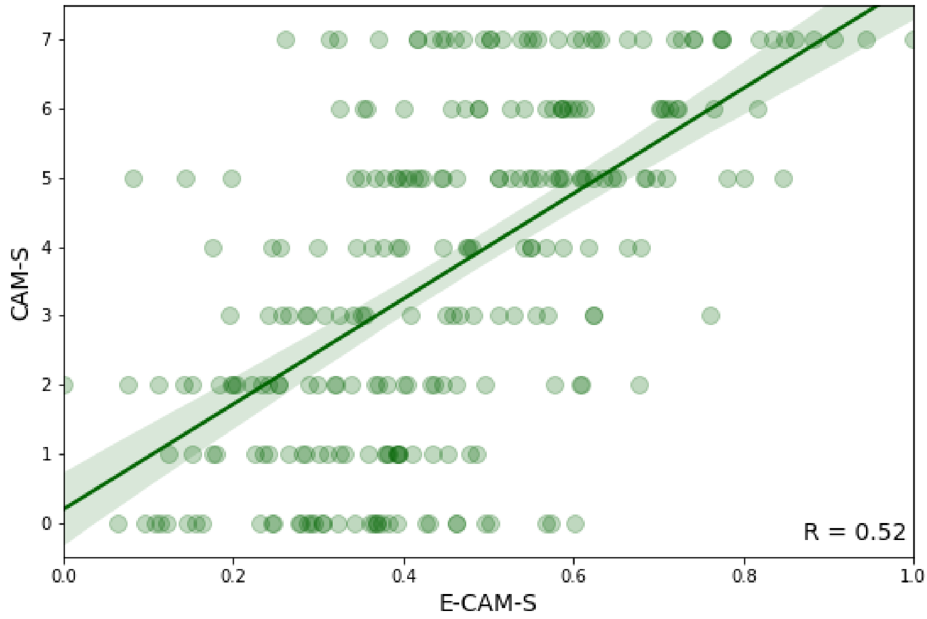

Outer K-fold cross validation
Fold 1
Fold 2
Fold K
Patients
For each outer training set,
for each value of hyperparameter,
do inner K-fold cross validation
Fold 1
Fold 2
Fold K
Patients in training set
Legend:
Train
Test
- Choose the best hyperparameter based on the best average inner testing performance
- Fit a model for each outer fold, apply it to the outer testing set
- Get outer testing performance
Final performance:
The average outer testing performance
Final model:
Fit a model on all data using the mode of hyperparameters across all outer folds

Supplemental Figure 3: Diagram showing nested cross-validation.

5. EEG-based delirium severity prediction for the non-comatose subset

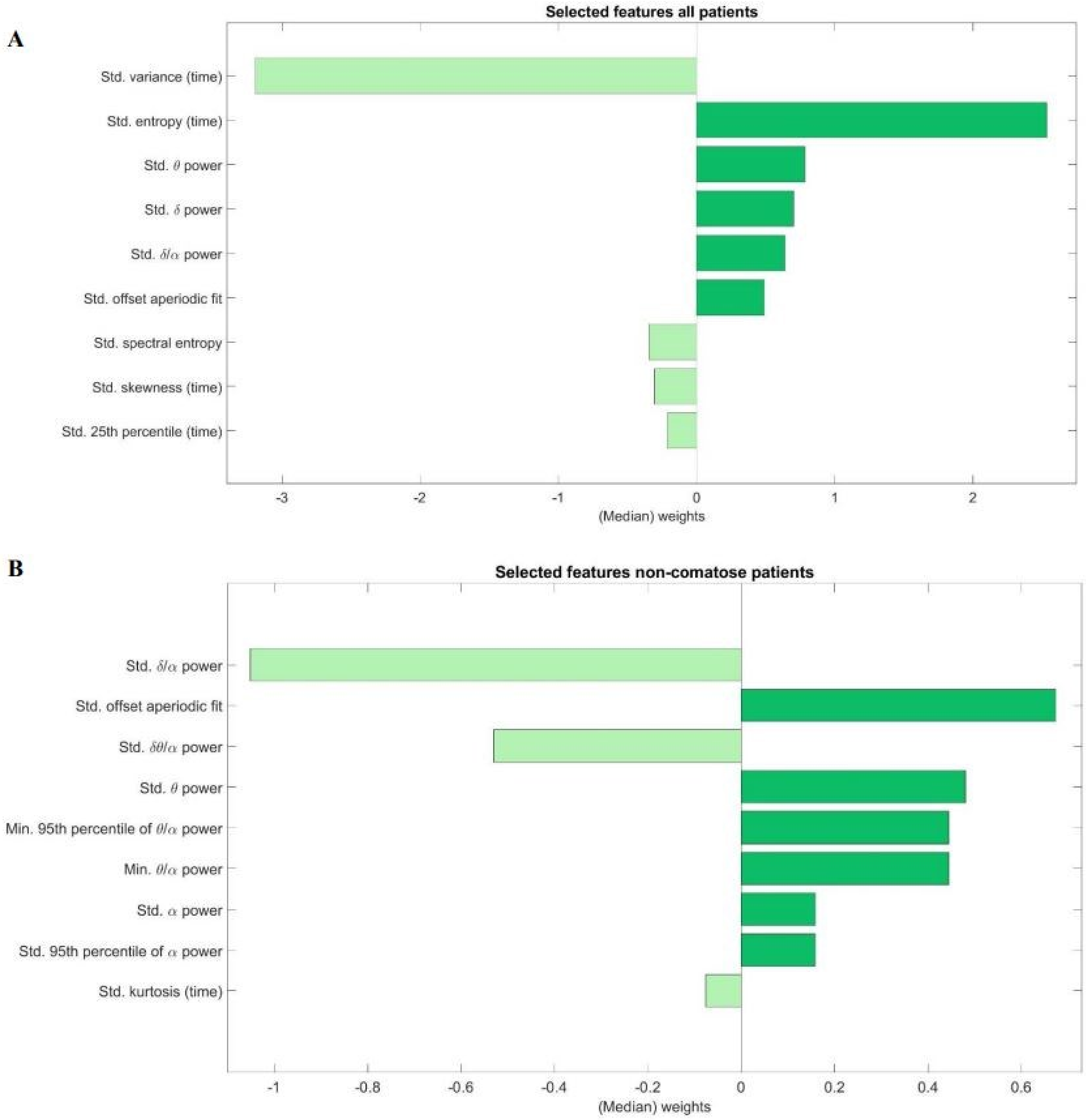
Supplemental Table 2: Performance metrics delirium prediction comparing using entire population and non-comatose subset.

Evaluation metrics	All patients (n = 373)	Non-comatose subset (n = 251)
Correlation *(CAM-S, E-CAM)*	0.68 [0.64 – 0.73]	0.52 [0.47 – 0.61]



Supplemental Figure 4: Scatter plot of EEG-based delirium severity prediction (E-CAM-S) vs. CAM-S scores for non-comatose subset. The green line represents a fitted regression line with 95% confidence interval.

6. EEG Features Predictive of Delirium Severity



Supplemental Figure 5: Influence of most important features for E-CAM-S prediction using the entire cohort **(A)** and non-comatose subset **(B)**, with dark green color reflecting positive correlation and light green color negative correlation with clinical CAM-S. For the features that were calculated over the same frequency range, the median of these weights was taken (e.g. different features for delta power, either calculated from 0.5-2 Hz or 2-4 Hz). Abbreviations: std = standard deviation, min = minimum. The first abbreviation reflects the summary statistic (std, mean, min or max) and the second abbreviation reflects the extracted EEG feature. E.g. “std. δ/α power” refers to the standard deviation across all 6-second epochs of the δ/α power (that was calculated for each 6-second epoch).

References

S1. Inouye SK, van Dyck CH, Alessi CA, et al: Clarifying confusion: The confusion assessment method: A new method for detection of delirium. *Annals of Internal Medicine*. 1990;113(12):941-948.

S2. Sessler CN, Gosnell MS, Grap MJ, et al: The Richmond Agitation-Sedation Scale: Validity and reliability in adult intensive care unit patients. *American Journal of Respiratory and Critical Care Medicine*. 2002;166(10):1338-1344.

S3. Oldham MA, Holloway RG: Delirium disorder: Integrating delirium and acute encephalopathy. *Neurology*. 2020;95(4):173-178.

S4. Slooter AJC, Otte WM, Devlin JW, et al: Updated nomenclature of delirium and acute encephalopathy: statement of ten Societies. *Intensive Care Medicine*. 2020;46(5):1020-1022.

S5. Charlson ME, Pompei P, Ales KL, et al: A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*. 1987;40(5):373-383.

S6. Sun H, Jia J, Goparaju B, et al: Large-scale automated sleep staging. *Sleep*. 2017;40(10).

S7. Donoghue T, Haller M, Peterson EJ, et al: Parameterizing neural power spectra into periodic and aperiodic components. *Nat Neurosci*. 2020;23:1655-1665.

S8. Burges C, Shaked T, Renshaw E, et al: Learning to rank using gradient descent. In: *ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning*. ACM Press; 2005:89-96.

S9 Kohavi R: *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*. International Joint Conference on Artificial Intelligence; 1995;14:1137-1143.

S10. James G, Witten D, Hastie T et al: *An Introduction to Statistical Learning: with applications in R*. New York: Springer; 2013. S11. Numan T, van den Boogaard M, Kamper AM, et al: Delirium detection using relative delta power based on 1-minute single-channel EEG: a multicentre study. *British Journal of Anaesthesia*. 2019;122(1):60-68.