Inferring preoperative cognitive function from intraoperative electroencephalography in elderly patients using machine learning

*Juan C. Pedemonte MD†, Haoqi Sun PhD†, Isaac G. Freedman MD, Isabella Turco BS, Kwame Wiredu PhD, Antonello Penna PhD, Jose Ignacio Egaña PhD, Rodrigo Gutierrez PhD, Mauricio Ibacache PhD, Luis I. Cortínez MD, M. Brandon Westover PhD, Oluwaseun Akeju MD, Gonzalo Boncompte PhD**

***Abstract*—To develop and evaluate machine learning (ML) models that infer preoperative cognitive function from intraoperative electroencephalography (EEG). This was a retrospective ML study that used a training dataset derived from the MINDDS study (306 patients, USA), and an external testing dataset from the Electroencephalographic Biomarker to Predict Acute Post-Operatory Cognitive Dysfunction study (92 patients, Chile). Both contained patients older than 60 years undergoing either cardiac (training dataset) or non-cardiac (testing dataset) surgery under general anesthesia. Preoperative cognitive function was assessed using the Montreal Cognitive Assessment (MoCA) in both cohorts. Four types of ML models were used: logistic regression with L2 penalty, random forest, gradient boosting tree, and extreme gradient boosting. Models were evaluated in terms of weighted root mean square error (WRMSE) and monotonic correlations towards actual MoCA scores (Spearman's rho). A logistic regression model with L2 regularization performed best in the training dataset (WRMSE 2.82 [2.60 – 3.03 95%CI], Spearman's rho 0.18 [0.06 – 0.29], p 0.0015). This performance mostly generalized to the test dataset (WRMSE 2.72 [2.51 – 2.94], Spearman's rho 0.14 [-0.05 – 0.31], p 0.18). This study shows that ML models trained on intraoperative EEG can effectively infer preoperative cognitive function in older patients, with generalizability across distinct populations and relatively low error (<3 MoCA points). However, the correlations were weak, indicating limited ability to capture consistent monotonic relationships. Incorporating this approach into perioperative care could enable early detection and mitigation of neurocognitive disorders, improving surgical outcomes through tailored interventions. Further refinement and validation are required before clinical implementation.**

***Index Terms*—Cognitive impairment, Delirium, Dementia, Perioperative cognition, Electroencephalography, Alpha power, Machine learning**

I. Introduction

Neurocognitive disorders like delirium and cognitive dysfunction can be precipitated by surgery [1]. In turn, episodes of postoperative delirium can trigger the onset of dementia or worsen an already existing one, especially in neurocognitively vulnerable patients [1,2], leading to a general functional decline. The presence of preoperative cognitive impairment is a key factor in determining the likelihood of postoperative neurocognitive disorders [3]. Information about preoperative cognitive function can also aid healthcare providers in deciding on anesthesia and postoperative care strategies that minimize the risk of delirium and other negative postoperative conditions, thus generally improving patients' outcomes [4]. This has been recently emphasized by an expert consensus article that states that baseline cognitive function should be evaluated preoperatively [5]. Moreover, there is ample consensus among geriatric, surgical, and anesthesiology societies' best-practice guidelines in recommending preoperative cognitive screening [6–9], especially in older surgical patients [10].

Despite this, a limited proportion of hospitals include cognitive screening in older adults as a mandatory component of their preoperative evaluation [11]. A web-based questionnaire administered to American Society of Anesthesiologists members reported that screening for preoperative cognitive impairment was completed "rarely" or "never" in 82.6% of the cases [12]. Conducting this type of screening involves a series of monetary, time, and organizational costs, such as the time to perform the assessment and appropriate training of healthcare personnel. These are most likely the main reasons why preoperative

The present work was partially funded by grants from ANID/FONDEF ID19I10345 and IT21I0041. Dr. Westover's laboratory is supported by grants from the NIH (R01AG073410, R01HL161253, R01NS126282, R01AG073598, R01NS131347, R01NS130119, and AWS.

JC Pedemonte, I Cortinez, M Ibacache and G Boncompte are with the División de Anestesiología, Escuela de Medicina, Pontificia Universidad Católica de Chile. JC Pedemonte and M Ibacache are also with the Programa de Farmacología y Toxicología, Facultad de Medicina, Pontificia Universidad Católica de Chile. H Sun, and B Westover are with the Department of Neurology, Beth Israel Deaconess Medical Center, USA. H Sun, I Freedman, I Turco, K Wiredu, B Westover and O Akeju are with the Henry and Allison McCance Center for Brain Health, Boston, MA, USA. I Freedma, I Turco, K Wiredu and O Akeju are with the Department of Anesthesia, Critical Care and Pain Medicine, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. A Penna, JI Egaña and R Gutierrez are with the Departamento de Anestesiología y Medicina Perioperatoria, Hospital Clínico, Universidad de Chile, Santiago, Chile. A Penna and R Gutierrez are also with the Centro de Investigación Clínica Avanzada (CICA), Facultad de Medicina, Hospital Clínico, Universidad de Chile Santiago, Chile.

† JC Pedemonte and H Sun contributed equally to this article.

*Corresponding author is Gonzalo Boncompte, PhD. gnboncompte@gmail.com, +56 2 23543270. Address: Marcoleta 377, División de Anestesiología, Escuela de Medicina, Pontificia Universidad Católica de Chile. Santiago, RM, Chile. 8330024 - Twitter/X handle: @GNBoncompte

cognitive evaluation in older patients appears exceptional rather than the norm in several contexts. Thus, facilitating the collection of information on elderly patients' preoperative cognitive function in an easy, inexpensive, and accessible manner could significantly improve the care of this vulnerable population.

Intraoperative frontal electroencephalography (EEG) is frequently used in older patients as a monitoring tool to guide anesthetic administration. However, EEG could provide information on more than just the anesthetics' effective dosage. Previous studies have seen an association between intraoperative EEG band power in the alpha frequency range (7-13 Hz) and preoperative cognitive impairment [13–18]. Other EEG features, such as the aperiodic exponent, have been linked to age-related cognitive decline and can predict it over a span of 10 years [19]. There are numerous other EEG features that could also contribute to inferring preoperative cognitive status. This may enable the use of EEG as an objective, cost-efficient, and eventually automated tool to inform preoperative cognitive function levels without the costs associated with cognitive screening. Here, we aimed to create a machine learning (ML) model to infer preoperative cognitive impairment across multiple surgical settings in older adults undergoing cardiac and non-cardiac surgery. We used two intraoperative EEG datasets from older patients undergoing general anesthesia for elective surgery who had preoperative cognitive evaluations. To emphasize the generalizability of our results, we trained the model on the first dataset (training dataset) and tested its performance on the second dataset (test dataset).

II. METHODS

This study is reported following the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guidelines and the Minimum Information about Clinical Artificial Intelligence Modelling (MI-CLAIM) guidelines.

A. *Training and Test Datasets*

The training dataset came from the MINDDS [20,21] study (NCT02856594), which was a single-center, parallel-arm, randomized, placebo-controlled superiority trial that evaluated the effectiveness of intraoperative dexmedetomidine to reduce postoperative delirium in patients older than 60 years undergoing cardiac surgery with cardiopulmonary bypass. This study was conducted in Boston, USA. Anesthesia maintenance was performed with Isoflurane as the main hypnotic agent and fentanyl and/or hydromorphone as analgesic agents. Further details on this study are described elsewhere [20,21].

The test dataset was collected from the Electroencephalographic Biomarker to Predict Acute Post-Operatory Cognitive Dysfunction study (NCT04214496). This study was a two-center observational trial aimed to determine the predictive capacity of an intraoperative electroencephalographic biomarker for postoperative delirium in patients older than 60 years programmed to undergo non-cardiac surgery. This study was conducted in Santiago, Chile. Anesthesia maintenance was performed with Sevoflurane. For analgesia, fentanyl and/or remifentanil were employed. In both datasets, a modern volatile agent (Isoflurane or Sevoflurane, both primarily GABAergic) was used together with an opioid agent, following classical clinical practice. The two datasets had distinct and diverse populations. The inclusion/exclusion criteria for both studies are detailed in Appendix D.

B. Outcomes: Cognitive Status and Postoperative Delirium Severity

Our primary outcome for model training was preoperative cognitive status, evaluated with the abbreviated Montreal Cognitive Assessment (aMoCA). Assessment was performed at baseline (preoperatively) and postoperatively at 30, 90, and 180 days in the MINDDS trial. The aMoCA ranges from 0 (worst) to 22 (best) points and can be administered remotely. It is a validated screening tool for assessing cognitive impairment [22,23] with a cutoff score of <18 points for mild cognitive impairment (MCI) [24].

In the Electroencephalographic Biomarker to Predict Acute Post-Operatory Cognitive Dysfunction study, the authorized formal Spanish translation of the Montreal Cognitive Assessment (MoCA) version 7.2 [25] was performed at baseline (preoperatively) and on the 5th postoperative day or at discharge, whichever came first. The 7.2 Spanish version of MoCA ranges from 0 (worst) to 30 (best) and was administered in person by a trained examiner.

To harmonize the different cognitive scales used, a conversion from MoCA in Spanish to abbreviated MoCA in English was carried out following a previously validated and published strategy by Katz et al., designed specifically to convert scores between these two scales [24]. However, this conversion does not include all possible values observed in MoCA in Spanish. Thus, for non-existent values, we used a round-up method (i.e., MoCA=30 is aMoCA=22, MoCA=28 is aMoCA=21, therefore, MoCA=29 is aMoCA=21.5, rounding it up to aMoCA=22).

Our secondary outcome was Postoperative Delirium Severity, evaluated using the long Confusion Assessment Method Severity (CAM-S) score in both the training and test datasets [26]. Delirium severity can be derived from components of the CAM. The CAM-S ranges from 0 (no delirium features) to 19, with higher scores indicating worsening delirium severity. This variable was obtained for both datasets to study its association with the inferred aMoCA. We did not include mortality in our outcomes because of insufficient incidence for proper analysis (Table I).

C. *Electroencephalography (EEG)*

In both studies, all patients received intraoperative EEG monitoring using a SedLine monitor (Masimo Inc., Irvine, CA). Sedtrace electrode arrays were placed on the forehead at approximately Fp1, Fp2, F7, and F8. The ground electrode was approximately at the Fpz position, and the reference electrode was approximately 1 centimeter above it. The training dataset had a sampling frequency of 250 Hz, and the external validation dataset of 178 Hz. We performed a down-sampling procedure to equalize both sampling frequencies to 170 Hz. Further details on EEG processing and feature extraction can be found below (Statistical analysis) and in Appendix A.

D. Covariates

Clinical and demographic data were used as covariates to account for possible confounding factors. The covariates were age (years), sex (male or female), body mass index (kg m-2), and educational level. In the training dataset, education was classified according to the following levels: 1) less than 8th grade, 2) 8th grade but less than High school graduate, 3) High school graduate, 4) some College/Associate's degree, 5) Bachelor's degree, 6) Masters' degree, 7) Doctoral degree. One participant has an unknown/missing educational level, which was imputed by the cohort median. In contrast, in the test dataset, educational level was collected as a continuous variable (years of education) starting at 1st grade (5 years old). Thus, to harmonize education, we categorized the test dataset educational level continuous variable into the following categories matching the above: 1) <9 years of education, 2) 9 to 11 years of education, 3) 12 to 13 years of education, 4) 14 to 16 years of education, 5) 17 to 18 years of education, 6) 18 to 20 years of education, 7) >21 years of education. There is no missing educational level in the test dataset.

E. Machine Learning

Machine learning (ML) was used to infer preoperative aMoCA. To utilize the ordinal nature of the aMoCA outcome, we used ordinal regression in combination with ML, using an incremental approach. The labels (aMoCA) were encoded in an incremental binary manner, where aMoCA $\leq$ 15 was encoded as [0,0,0,0,0,0,0], 16 as [1,0,0,0,0,0,0], 17 as [1,1,0,0,0,0,0], 18 as [1,1,1,0,0,0,0], 19 as [1,1,1,1,0,0,0], 20 as [1,1,1,1,1,0,0], 21 as [1,1,1,1,1,1,0], and 22 as [1,1,1,1,1,1,1]. This converts the ordinal regression into multiple binary classification problems, where each binary classification problem corresponds to an element in the encoded vector. Then, we fitted one classifier for each binary classification problem, totaling 7 classifiers. Due to the different class ratios for each binary classification problem, we calibrated each classifier using Platt's method. We then found the optimal probability cutoff using Youden's Index. At the inference stage, we took the sum of 7 model inferences, where each was binarized using its optimal probability cutoff. The sum ranges from 0 to 7, then mapped back to aMoCA $\leq$ 15, 16, …, 22. This approach is better than plain binary classification since the incremental binary labels encode ordered information; it is also better than continuous regression since it avoids encoding the numeric value. Therefore, this approach is appropriate for modeling scores in a clinical setting [43]. The justification by results is provided in the Supplementary Results. We used four machine learning models for the underlying binary classification component: logistic regression with L2 penalty (LR), random forest (RF), gradient boosting tree (GBT), and extreme gradient boosting (XGB). The models were trained using nested 10-fold cross-validation, as detailed in Appendix B.

F. Model Performance Metrics

The primary model performance metric was weighted root mean square error (WRMSE), where the weight is inversely proportional to the number of samples in each aMoCA level. WRMSE measures the error between the inferred and the actual aMoCA, with higher scores indicating worse model performance. The secondary model performance metric was Spearman's rank correlation coefficient (rho), which measures the monotonic trend between the inferred and the actual aMoCA, with higher values indicating better performance. These metrics were selected due to the ordinal nature of aMOCA, which is not described by traditional metrics such as accuracy, specificity, and sensitivity that are suitable for categorical outcomes. Also, dichotomizing aMoCA requires an arbitrary threshold and thus a loss of information.

G. External Validation

The best-performing model was then applied to the test dataset from the Electroencephalographic Biomarker to Predict Acute Post-Operatory Cognitive Dysfunction study. The same performance metrics and associations with the same outcomes were measured.

H. Statistical Analysis

Continuous data are presented using the mean ± standard deviation. In addition, groups are compared using Student's t-test for independent samples. Categorical data are presented as absolute counts and their corresponding percentages. Differences between groups are studied using the $\chi 2$ test. A p-value < 0.05 criterion was employed to evaluate statistical significance. For the performance metrics, we also did permutation tests to get the chance-level distributions. Specifically, we permutated the actual aMoCA scores to destroy the correspondence with the predicted aMoCA scores, and then obtained the same performance metrics. This process was repeated 1,000 times to obtain the 2.5% and 97.5% percentiles for the lower and upper bounds of the permutated confidence interval. For confidence intervals, we used bootstrapping, i.e., randomly sampling the dataset with replacement to the same sample size and performing the analysis 1,000 times, where the 2.5% and 97.5% percentiles were used at the lower and upper bounds of the 95% confidence interval. All analyses were performed using Python 3.10. EEG processing was conducted using Python custom-made scripts and publicly available open-source libraries. Numpy, Scipy, and Matplotlib were the main libraries employed throughout the analyses. For EEG files management and spectral estimations, we used the MNE library (e.g., mne.time_frequency.psd_array_multitaper). For entropy-related estimations, we used OrdPy (e.g., ordpy.permutation_entropy). Lempel-Ziv Complexity was evaluated using previously published scripts [39]. Aperiodic components were calculated using Specparam library (previously FOOOF) [28].

III. RESULTS

A. Cohort Characteristics

As shown in Figure 1, 306 participants were included in the final training dataset and 92 participants in the test dataset. The average age for the training dataset was 69.2 years old, with 24.8% female participants (Table I). The average age for the test dataset was 73.5 years old, with 50% female participants. When comparing all measured covariates between the two datasets, patients in the testing dataset were on average slightly older, had lower BMI, and lower educational level (Table I).

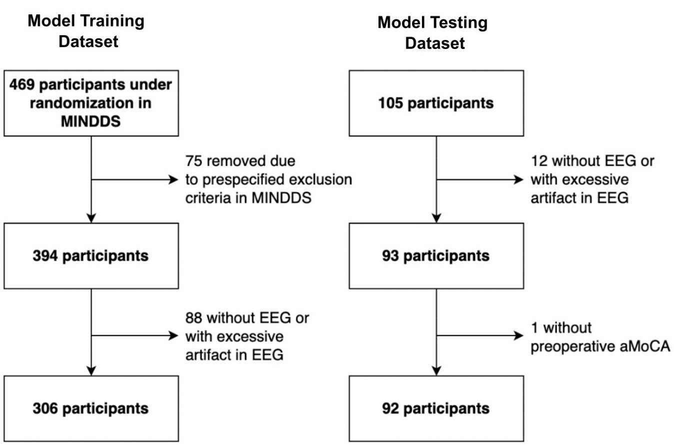

Figure 1. Cohort diagram for both the training dataset and the test dataset.

B. Cross-Validated Performances for Inferring preoperative aMOCA in the Training Dataset

When measured by Spearman's rho, the LR model achieved the best performance (highest correlation = 0.18 (0.06 – 0.29), p = 0.0015, Table II), where the permutation-based chance level is 0 (-0.11 – 0.11). When measured by WRMSE, the RF model achieved the best performance (lowest WRMSE = 2.59 (2.40-2.76)). As shown in Figure 2, the inference using LR showed a wider range (greater confidence interval), while the inference using RF showed a smaller range (regression to the mean). Therefore, we chose LR as the best model of the two for conducting further analyses. The permuted confusion matrices and their comparisons are shown in Supplementary Figure 1.

TABLE I
COHORT CHARACTERISTICS

	Training Dataset (n=306)	**Test Dataset (n=92)**	**P-value**
Age (year): mean±std	69.2±6.3	73.5±6.2	**<0.001**
Female: n (%)	76 (24.8%)	46 (50%)	**<0.001**
BMI (kg/m^2): mean±std	28.3±5.3	26.1±4.4	**<0.001**
Education level: n (%)			**<0.001**
1: less than 8th grade	0	26 (28.3%)	
2: 8th grade but less than high school graduate	2 (0.7%)	16 (17.4%)	
3: high school graduate	44 (14.4%)	17 (18.5%)	
4: Some college / associate	66 (21.6%)	19 (20.7%)	
5: Bachelor degree	90 (29.4%)	6 (6.5%)	
6: Master degree	57 (18.6%)	7 (7.6%)	
7: Doctoral degree	46 (15.0%)	1 (1.1%)	
Unknown/missing	1 (0.3%)	0	
Smoking			**0.036**
Never or Past	297 (97.1%) (154 never, 143 past)	84 (91.3%)	
Current	9 (2.9%)	8 (8.7%)	
Illicit drug	8 (2.6%)	0 (0%)	0.25
Comorbidity: n (%)			
Hypertension	239 (78.1%)	55 (59.8%)	**<0.001**
Diabetes	64 (20.9%)	22 (23.9%)	0.64
Sleep apnea	66 (21.6%)	–	–
Stroke	35 (11.4%)	–	–
Myocardial infarction	34 (11.1%)	6 (6.5%)	0.28
Peripheral artery disease	23 (7.5%)	–	–
Atrial fibrillation	112 (36.6%)	–	–
Previous cardiac intervention	102 (33.3%)	–	–
Chronic lung disease	48 (15.7%)	–	–
Renal failure	3 (1.0%)	3 (3.3%)	0.28
Liver disease	9 (2.9%)	1 (1.1%)	0.54
Mortality: n (%)			
1-month follow-up	4 (1.3%)	0 (0%)	0.61
3-month follow-up	4 (1.3%)	3 (3.3%)	0.42
6-month follow-up	6 (2.0%)	6 (6.5%)	0.058

^ for continuous variables, t-test was used; for categorical variables, χ^2 test was used. The *p*-values are two-sided, where values < 0.05 are bolded. "–" stands for not available.

We then compared the inference performance using exclusively alpha band power (8 - 12Hz). As shown in Table III, the WRMSE using alpha band power was higher (worse

This article has been accepted for publication in IEEE Transactions on Biomedical Engineering. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBME.2026.3671187

performance) compared to using multiple EEG features and covariates. However, the difference in Spearman's correlation was smaller. The best performance was obtained when using data from both EEG and clinical covariates. In terms of Spearman's correlation, the best performance was 0.18 (0.06 – 0.29), where the absolute magnitude indicates a weak prediction, and the statistical significance could be driven by sample size. In terms of WRMSE, the best performance was 2.82 (2.60 – 3.03), where the permutation-based chance level is 3.14 (2.93 – 3.34).

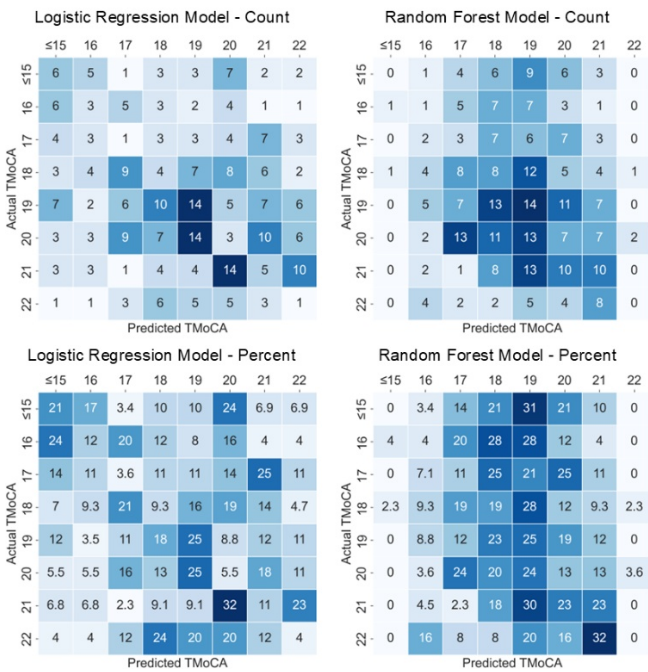

Figure 2. Confusion matrix using Logistic regression (LR, left) and random forest (RF, right). Each row presents the number of participants with different aMoCA inferences at a specific actual aMoCA score. The color in each cell is proportional to the number in the cell.

C. Association with Secondary Outcome: Postoperative CAM-S

We tested the association of the inferred aMoCA with postoperative delirium severity measured using CAM-S. The inferred aMoCA using LR had Spearman's correlation with CAM-S of -0.13 ($p = 0.032$). For reference, the actual aMoCA correlated with CAM-S -0.19 ($p < 0.001$). Since the participants in the MINDDS study had been randomly assigned to receive dexmedetomidine, we further adjusted for the dexmedetomidine use and repeated the same test. The results are similar, as shown in Appendix C. The results indicated that inferred aMoCA using LR significantly correlated with the CAM-S.

As an exploratory analysis, we also directly predicted the ordinal values of CAM-S using the EEG features and covariates. In the training set, the cross-validated performance was WRMSE of 2.29 (2.06 – 2.50) and Spearman's rho of 0.12 (0.01 – 0.23), indicating weak but detectable predictability.

D. External Validation Results

We applied LR to the external test dataset. The LR model had a WRMSE of 2.72 (2.51 – 2.94), similar to its performance on the training dataset (bootstrapped p = 0.87). The permutation-based chance level is 2.86 (2.49 – 3.26). It showed a Spearman's correlation of 0.14 (-0.05 – 0.31; p = 0.18), similar to its performance on the training set (bootstrapped p = 0.32), albeit with a wider confidence interval, likely due to the difference in sample sizes. The permutation-based chance level is 0 (-0.21 – 0.20). The confusion matrix is shown in the Supplementary Figure 2. All performances are shown in the Supplementary Figure 3. For the association with the secondary outcome CAM-S, we did not find a significant association for LR-inferred aMoCA (Spearman's correlation p = 0.43). For directly predicting CAM-S, the WMSE was 3.44 (3.07 – 3.75) and Spearman's rho was 0.00 (-0.22 – 0.20), indicating no predictability. As a sensitivity analysis, we evaluated the model's performance on both the training and test datasets across patient subsets. Specifically in Male and Female patients and in patients younger and older than 70 years of age (Supplementary Table 1).

TABLE II
PERFORMANCES OF PREOPERATIVE AMOCA INFERENCES

Model	WRMSE: measures inference error, on the aMoCA scale (The smaller the better)	Spearman's correlation: measures monotonic trend (The higher the better)
Logistic regression with L2 penalty (LR)	2.82 (2.60 – 3.03)	0.18 (0.06 – 0.29)
Random forest (RF)	2.59 (2.40 – 2.76)	0.14 (0.02 – 0.25)
Gradient boosting tree (GBT)	2.59 (2.42 – 2.77)	0.08 (-0.05 – 0.19)
Extreme gradient boosting (XGB)	2.70 (2.53 – 2.89)	0.02 (-0.10 – 0.12)

TABLE III
PERFORMANCES OF PREOPERATIVE AMOCA INFERENCES USING LR, WITH DIFFERENT INPUTS.

Input	WRMSE: measures inference error, on the aMoCA scale (The smaller the better)	Spearman's correlation: measures monotonic trend (The higher the better)
Alpha band power (classic approach)	3.58 (3.35 – 3.81)	**0.18 (0.06 – 0.28)**
Alpha band power	3.15 (2.87 – 3.40)	0.15 (0.04 – 0.26)

(oscillatory-specific part using the 1/f approach)		
EEG	3.15 (2.92 – 3.38)	0.09 (-0.02 – 0.20)
EEG + Covariates	**2.82 (2.60 – 3.03)**	**0.18 (0.06 – 0.29)**
Best model metrics are highlighted in bold.		

IV. DISCUSSION

In this study, we show that older patients' preoperative cognitive status can be inferred with a machine learning model that employs intraoperative EEG information. ML models using intraoperative EEG predicted preoperative cognitive function with low error, but correlations with actual scores were weak. Logistic Regression with L2 penalty performed best among all types of models tested. In the test dataset of non-cardiac older surgical patients, this model had the lowest error between the inferred and actual values (WRMSE). Importantly, performance increased when covariates (age, sex, body mass index, and educational level) were added. However, differences between models were milder when performance was assessed via the monotonic trend between actual and inferred aMoCA through Spearman's correlation. When the relation of inferred aMoCA with postoperative outcomes was explored, an association with postoperative delirium severity (CAM-S) as well as direct predictability from EEG in the training dataset was found. In summary, our results indicate that intraoperative EEG data could be used to infer cognitive function levels in older patients effectively. This indicates technical feasibility and possible generalizability across cohorts, yet limited predictive performance. Further refinement and validation are needed before clinical application, highlighting potential but not immediate utility in perioperative care. to deliver more personalized postoperative management [4].

The model's aMoCA prediction performances on the test and training datasets were comparable, showing only slightly worse metrics in the test dataset, which illustrates the generalizability of our model. However, the substantial difference between the performance of our best model in both datasets was that it was not able to replicate the association between preoperative cognition and CAM-S. There were important and significant differences between the characteristics of the patients in the training and test datasets (Table I). Foremost, the training dataset consisted of cardiac patients, while the patients in the test dataset were not cardiac patients. Also, the population in the test dataset was older, had a more equilibrated female-male ratio, presented a lower body mass index, and a lower educational level than those in the training dataset. Importantly, the training dataset was obtained from a US-based population, and the test dataset was obtained from a Chile-based one. These differences between groups may explain the difference in performance between datasets. A recent study in a Chilean older population (>66 years) found 19 points as a cutoff score for mild cognitive impairment in this population [29]. In this line, differences in cutoff values for cognitive impairment between populations may also explain these performance differences. However, the use of diverse datasets is necessary to make robust models that can be generalized to the general population.

Cognitive function screening is seldom conducted in everyday practice [11,12]. Patients with low preoperative cognitive performance, especially after cardiac surgery, have increased odds of postoperative bleeding, intensive-care unit length of stay, extended hospital stay, and a greater likelihood of being discharged to a place other than home [30–32]. These patients also have an increased likelihood of developing postoperative delirium. This is of particular relevance because episodes of delirium can be a triggering factor for the onset of various dementias [1,2]. Our results emphasize the usefulness of automatically inferring preoperative cognitive levels from a relatively cheap and non-invasive intraoperative monitor, such as an EEG recording. However, before clinical implementation, further refinement and enhancement of these models must be conducted to improve the quality (effect size) of preoperative cognitive function inference.

Intraoperative EEG has been extensively used to titrate anesthetic drugs, particularly as a correlate of hypnosis and nociception [33]. However, it could also help to infer specific patients' cognitive features [16,18,34]. Intraoperative EEG is a useful tool for anesthesia providers that can lead to more personalized intraoperative and postoperative care. It is a tool that has been shown to safely facilitate the thorough administration of anesthetic drugs. Implementing an automated EEG-based system to classify patients' preoperative cognitive status could lead to better and more individualized perioperative care.

Our results build upon recent evidence [13–18] that indicates that intraoperative alpha power is a particularly valuable tool to infer preoperative cognitive levels (Table III). When we further evaluated this relation, specifically with oscillatory-specific alpha power, the model's performance increased as reflected by lower WRMSE values. However, Spearman's correlation decreased, indicating a milder monotonic association. Classical alpha band power is the combination of oscillatory-specific and 1/f-like aperiodic activity [27]. Oscillations are the result of coordinated periodic modulations of neuronal spiking, while aperiodic activity is thought to be related to the background tone of activity [35–37]. Aperiodic activity has a known relation with aging and cognition [38]. It has also shown promise in several clinical applications [40-41]. Our results suggest that the brain capacity for robust temporal coordination, which is intrinsic to oscillatory-specific alpha activity, is more strongly related to the level of preoperative cognitive function than the combined periodic and aperiodic activity present in classical alpha band power (Table III). This result shows the importance of distinguishing between periodic and aperiodic components when studying brain activity and its relation to clinical features. A detailed table with the feature coefficients is provided in Supplementary Table 2.

One limitation of the current study emerges from the differences between how and which variables were measured in the original studies. There were cases of clinical covariates that were missing in the test dataset compared to the training dataset, which limited the performance of the final model. Also, there was a difference in the instruments used to quantify cognitive function (aMoCA vs MoCA), which required a conversion between scores from the MoCA scale to aMoCA. This process could have led to marginally different scores. For WRMSE, this causes a higher deviation from the ground truth, hence overestimating the error. This was not the case for Spearman's correlation scores because this metric only considers the relative rank. Future studies could include common and more detailed cognitive assessment tools to refine these inferential models. Another limitation that should always be acknowledged in these types of studies is the fact that, although we extracted 26 EEG features, coming from diverse theoretical backgrounds, we could be missing other important EEG features. An intrinsic challenge from the strategy used here is the fact that our model predicts cognitive function intraoperatively, which precludes the possibility of implementing strategies such as preoperative rehabilitation (e.g. [42]), leaving the clinicians space to conduct mitigatory and/or preventive measures only during intraoperative and postoperative periods. From the ML modeling perspective, models aimed at predicting cognitive function from EEG features could be improved by employing more detailed cognitive assessments, EEG data with greater head coverage instead of only frontal electrodes, a greater number of EEG features, or the use of other models.

In conclusion, an intraoperative EEG-based ML model can better aid in inferring preoperative cognitive status in older patients undergoing cardiac surgery than alpha band power. This study highlights the potential of intraoperative EEG as a source of information for cognitive assessment, but current models provide only marginal predictive value. Future work should focus on improving model accuracy and robustness, as well as exploring complementary biomarkers, to enable meaningful integration into perioperative care.

Code Availability

All code and package versions can be found at https://github.com/Hockey86/EEGPreopCognitionML.

Appendix

A. EEG Processing and Feature Extraction

All EEG traces were visually inspected, and artifacts were removed. A 2-minute artifact-free period, collected between the induction of anesthesia and the initiation of cardiopulmonary bypass, was selected for analysis in the training dataset. In the validation dataset, a 1-minute artifact-free EEG segment collected during anesthesia maintenance (1 hour after tracheal intubation) was selected for analysis.

For estimating spectral-domain EEG features, we segmented the data into 10-second epochs. From each, we extracted total band power (0.5-40Hz), delta band power (0.5-3Hz), theta band power (3-7Hz), alpha band power (8-12Hz), and beta band power (15-25Hz). The band powers were derived from the power spectral density (PSD), estimated using the multitaper method (7 tapers). We also computed an oscillatory-specific alpha band as reported previously [27] using the specparam algorithm [28], which decomposes PSD into aperiodic and oscillatory components. Then, we extracted the oscillatory-specific alpha band's number of oscillatory peaks, power, central frequency, bandwidth, and peak prevalence. We also extracted the aperiodic exponent (slope) and offset, as well as the deviation from the 1/f component. Features from the complexity domain included permutation entropy, dispersion entropy, Hjorth complexity, number of zero crossings, Higuchi fractal dimension, slope from the detrended fluctuation analysis, and Lempel-Ziv complexity. Finally, synchronization between lateral electrodes was assessed using three complementary measures: coherence of the alpha band, synchronization of the alpha band (weighted phase lag index, wPLI), and mutual information. These measures were calculated for each pair of electrodes from the same brain hemisphere and for pairs from different brain hemispheres (intra- and interhemispheric, respectively). In total, we calculated 26 EEG-derived features.

B. Model Training using Nested 10-Fold Cross-Validation

The nested 10-fold cross-validation consists of an outer and inner loop. The purpose of the outer loop was to estimate the out-of-sample generalization performance. The inner loop aims to select the best hyperparameter of the ML models. The participants were randomly divided into 10 folds in the outer loop with the same distribution of aMoCA scores. Each fold is used as the testing set in rotation, and the other 9 folds are combined as the training set. A model was fitted using the hyperparameter from the inner loop on the training set and then inferred on the testing set. The final performance metrics were based on comparing all true classes to the concatenated test results across the 10 folds. The training participants from an outer loop are further divided into 10 folds in the inner loop. We combined univariate feature selection in the training process. Before fitting the model, we used the Mann-Whitney U test to obtain the p-value of each feature in the training set, and then features with a p-value lower than a threshold (hyperparameter) entered the model fitting. For each hyperparameter choice, the model was trained on the 9 inner folds and tested on the 1 inner fold. The combined testing performance on all inner testing sets was the performance for this hyperparameter choice. We used Bayesian optimization to find the best hyperparameter that maximizes the inner testing performance. For LR, the hyperparameter was the penalty strength selected from the range of 10-3 to 103. For RF, the hyperparameters were the number of trees selected from 10 to 200, maximum depth from 1 to 5, minimum samples at leaf from 1 to 50, and cost complexity pruning from 10-3 to 103. For GBT, the hyperparameters were maximum boosting levels from 10 to 50, maximum tree depth from 2 to 4, L2 penalty from 10-3 to 103, and learning rate from 10-3 to 100. For XGB, the hyperparameters were maximum tree depth from 2 to 4, learning rate from 0 to 1, the minimum sum needed in a child node from 10-2 to 102, and minimum loss reduction for partition leaf node from 10-3 to 103. For all models, the univariate feature selection p-value threshold was selected from 0.1 to 0.5.

C. Association with CAM-S while Adjusting for Dexmedetomidine

Since the participants in the MINDDS study had been randomly assigned to receive dexmedetomidine, we tested the association of the inferred aMoCA with CAM-S while adjusting for dexmedetomidine. The inferred aMoCA using LR had Spearman's correlation with CAM-S of -0.12 ($p = 0.043$). The inferred aMoCA using RF had -0.075 ($p = 0.20$). For reference, the actual aMoCA had -0.19 ($p<0.001$).

D. Inclusion and Exclusion Criteria

Training dataset: Patients were eligible for inclusion if they were 60 years or older, scheduled to undergo a cardiac surgical procedure with planned postoperative admission to the cardiac surgical intensive care unit (CSICU) for 24 h or more, and were scheduled for a same-day surgical admission. Patients were excluded from participation if they were allergic to dexmedetomidine, had renal (requiring dialysis) or liver failure (Child-Pugh score > 5), were on chronic benzodiazepine or antipsychotic therapy, had severe deficit(s) due to structural or anoxic brain damage, were admitted to the ICU for more than 2 days in the month before surgery, previously underwent cardiac surgery within 1 year of surgery, were undergoing a surgical procedure requiring total circulatory arrest, or were SARS-CoV-2 positive or symptomatic (e.g., fever, cough, loss of taste/smell). Participants who were blind, deaf, or unable to communicate in English were excluded due to their inability to complete the cognitive assessments, as were patients experiencing circumstances for which long-term follow-up might be difficult (e.g., homelessness, active psychotic disorder, or substance abuse).

Testing dataset: Patients were eligible if they were 60 years or older, scheduled for high-risk elective surgery, planned for at least 3 days of hospital stay, and undergoing surgery under general anesthesia, with written informed consent required for participation in the trial. Patients with preoperative delirium or dementia, using neuroleptic drugs for the last 6 months, with a history of encephalopathy, psychosis, stroke, or brain trauma with neurologic sequelae were excluded. Cases under use of ketamine or dexmedetomidine during surgery, emergency surgery, patients requiring mechanical ventilation for 72 hours after surgery, and those unable to communicate in Spanish or illiterate were also excluded. Participants could not be included in another clinical trial.

Declaration of interests

Dr Akeju is listed as an inventor on brain monitoring patents assigned to Massachusetts General Hospital and is a consultant with Equity in Reversal Therapeutics. Dr. Westover is a co-founder, serves as a scientific advisor and consultant to, and has a personal equity interest in Beacon Biosignals.

The rest of the authors declare no competing interests.

References

[1] Fong TG, Inouye SK. The inter-relationship between delirium and dementia: the importance of delirium prevention. Nat Rev Neurol 2022;18:579–96. https://doi.org/10.1038/s41582-022-00698-7.

[2] Wang P, Velagapudi R, Kong C, Rodriguiz RM, Wetsel WC, Yang T, et al. Neurovascular and immune mechanisms that regulate postoperative delirium superimposed on dementia. Alzheimer's & Dementia 2020;16:734–49. https://doi.org/10.1002/alz.12064.

[3] Evered L, Scott DA, Silbert B. Cognitive decline associated with anesthesia and surgery in the elderly: does this contribute to dementia prevalence? Current Opinion in Psychiatry 2017;30:220–6. https://doi.org/10.1097/YCO.0000000000000321.

[4] Hauer K, Dutzi I, Werner C, Bauer J, Ullrich P. Delirium Prevention in Early Rehabilitation During Acute Hospitalization and Implementation of Programs Specifically Tailored to Older Patients with Cognitive Impairment: A Scoping Review with Meta-Analysis. JAD 2024;97:3–29. https://doi.org/10.3233/JAD-230644.

[5] Peden CJ, Miller TR, Deiner SG, Eckenhoff RG, Fleisher LA, Fleisher L, et al. Improving perioperative brain health: an expert consensus review of key actions for the perioperative care team. British Journal of anesthesia 2021;126:423–32. https://doi.org/10.1016/j.bja.2020.10.037.

[6] Aldecoa C, Bettelli G, Bilotta F, Sanders RD, Aceto P, Audisio R, et al. Update of the European Society of Anaesthesiology and Intensive Care Medicine evidence-based and consensus-based guideline on postoperative delirium in adult patients. European Journal of Anaesthesiology 2023. https://doi.org/10.1097/EJA.0000000000001876.

[7] Inouye SK, Robinson T, Blaum C, Busby-Whitehead J, Boustani M, Chalian A, et al. Postoperative Delirium in Older Adults: Best Practice Statement from the American Geriatrics Society. Journal of the American College of Surgeons 2015;220:136-148e1. https://doi.org/10.1016/j.jamcollsurg.2014.10.019.

[8] Berger M, Schenning KJ, Brown CH, Deiner SG, Whittington RA, Eckenhoff RG. Best Practices for Postoperative Brain Health: Recommendations From the Fifth International Perioperative Neurotoxicity Working Group. Anesthesia & Analgesia 2018;127:1406–13. https://doi.org/10.1213/ANE.0000000000003841.

[9] Mohanty S, Rosenthal RA, Russell MM, Neuman MD, Ko CY, Esnaola NF. Optimal Perioperative Management of the Geriatric Patient: A Best Practices Guideline from the American College of Surgeons NSQIP and the American Geriatrics Society. Journal of the

American College of Surgeons 2016;222:930–47. https://doi.org/10.1016/j.jamcollsurg.2015.12.026.

[10] Crosby G, Culley DJ, Hyman BT. Preoperative Cognitive Assessment of the Elderly Surgical Patient. Anesthesiology 2011;114:1265–8. https://doi.org/10.1097/ALN.0b013e31821b1bc8.

[11] Mahanna-Gabrielli E, Schenning KJ, Eriksson LI, Browndyke JN, Wright CB, Evered L, et al. State of the clinical science of perioperative brain health: report from the American Society of Anesthesiologists Brain Health Initiative Summit 2018. British Journal of anesthesia 2019;123:464–78. https://doi.org/10.1016/j.bja.2019.07.004.

[12] on behalf of the ASA Committee on Geriatric Anesthesia and the ASA Perioperative Brain Health Initiative, Deiner S, Fleisher LA, Leung JM, Peden C, Miller T, et al. Adherence to recommended practices for perioperative anesthesia care for older adults among US anesthesiologists: results from the ASA Committee on Geriatric Anesthesia-Perioperative Brain Health Initiative ASA member survey. Perioper Med 2020;9:6. https://doi.org/10.1186/s13741-020-0136-9.

[13] Giattino CM, Gardner JE, Sbahi FM, Roberts KC, Cooter M, Moretti E, et al. Intraoperative Frontal Alpha-Band Power Correlates with Preoperative Neurocognitive Function in Older Adults. Front Syst Neurosci 2017;11:24. https://doi.org/10.3389/fnsys.2017.00024.

[14] Gutiérrez RG, Egaña JI, Maldonado FA, Sáez IA, Reyes FI, Soulat H, et al. Association Between Lower Preoperative Cognition With Intraoperative Electroencephalographic Features Consistent With Deep States of Anesthesia in Older Patients: An Observational Cohort Study. Anesthesia & Analgesia 2021;133:205–14. https://doi.org/10.1213/ANE.0000000000005262.

[15] Koch S, Feinkohl I, Chakravarty S, Windmann V, Lichtner G, Pischon T, et al. Cognitive Impairment Is Associated with Absolute Intraoperative Frontal α-Band Power but Not with Baseline α-Band Power: A Pilot Study. Dement Geriatr Cogn Disord 2019;48:83–92. https://doi.org/10.1159/000502950.

[16] Pedemonte JC, Plummer GS, Chamadia S, Locascio JJ, Hahm E, Ethridge B, et al. Electroencephalogram Burst-suppression during Cardiopulmonary Bypass in Elderly Patients Mediates Postoperative Delirium. Anesthesiology 2020;133:280–92. https://doi.org/10.1097/ALN.0000000000003328.

[17] Touchard C, Guimard P, Guessous K, Aubin OS, Levé C, Joachim J, et al. Association of sleep and anesthesia EEG biomarkers with preoperative MOCA score: A pilot study. Acta Anaesthesiol Scand 2023;67:877–84. https://doi.org/10.1111/aas.14251.

[18] Shao YR, Kahali P, Houle TT, Deng H, Colvin C, Dickerson BC, et al. Low Frontal Alpha Power Is Associated With the Propensity for Burst Suppression: An Electroencephalogram Phenotype for a “Vulnerable Brain.” Anesthesia & Analgesia 2020;131:1529–39. https://doi.org/10.1213/ANE.0000000000004781.

[19] Finley AJ, Angus DJ, Knight EL, Van Reekum CM, Lachman ME, Davidson RJ, et al. Resting EEG Periodic and Aperiodic Components Predict Cognitive Decline Over 10 Years. J Neurosci 2024;44:e1332232024. https://doi.org/10.1523/JNEUROSCI.1332-23.2024.

[20] Qu JZ, Mueller A, McKay TB, Westover MB, Shelton KT, Shaefi S, et al. Nighttime dexmedetomidine for delirium prevention in non-mechanically ventilated patients after cardiac surgery (MINDDS): a single-centre, parallel-arm, randomised, placebo-controlled superiority trial. eClinicalMedicine 2023;56:101796. https://doi.org/10.1016/j.eclinm.2022.101796.

[21] Shelton KT, Qu J, Bilotta F, Brown EN, Cudemus G, D’Alessandro DA, et al. Minimizing ICU Neurological Dysfunction with Dexmedetomidine-induced Sleep (MINDDS): protocol for a randomised, double-blind, parallel-arm, placebo-controlled trial. BMJ Open 2018;8:e020316. https://doi.org/10.1136/bmjopen-2017-020316.

[22] Horton DK, Hynan LS, Lacritz LH, Rossetti HC, Weiner MF, Cullum CM. An Abbreviated Montreal Cognitive Assessment (MoCA) for Dementia Screening. The Clinical Neuropsychologist 2015;29:413–25. https://doi.org/10.1080/13854046.2015.1043349.

[23] Wong A, Nyenhuis D, Black SE, Law LSN, Lo ESK, Kwan PWL, et al. Montreal Cognitive Assessment 5-Minute Protocol Is a Brief, Valid, Reliable, and Feasible Cognitive Screen for Telephone Administration. Stroke 2015;46:1059–64. https://doi.org/10.1161/STROKEAHA.114.007253.

[24] Katz MJ, Wang C, Nester CO, Derby CA, Zimmerman ME, Lipton RB, et al. T-MoCA: A valid phone screen for cognitive impairment in diverse community samples. Alz & Dem Diag Ass & Dis Mo 2021;13. https://doi.org/10.1002/dad2.12144.

[25] Aguilar-Navarro SG, Mimenza-Alvarado AJ, Palacios-García AA, Samudio-Cruz A, Gutiérrez-Gutiérrez LA, Ávila-Funes JA. Validity and reliability of the Spanish Version of the Montreal Cognitive Assessment (MoCA) for the detection of cognitive impairment in Mexico. Revista Colombiana de Psiquiatría (English Ed) 2018;47:237–43. https://doi.org/10.1016/j.rcpeng.2018.10.004.

[26] Inouye SK, Westendorp RG, Saczynski JS. Delirium in elderly people. The Lancet 2014;383:911–22. https://doi.org/10.1016/S0140-6736(13)60688-1.

[27] Boncompte G, Freedman I, Qu J, Turco I, Khawaja ZQ, Cortinez I, et al. Cognitive function mediates the relationship between age and anesthesia-induced oscillatory-specific alpha power. Brain Communications 2023;6:fcae023. https://doi.org/10.1093/braincomms/fcae023.

[28] Donoghue T, Haller M, Peterson EJ, Varma P, Sebastian P, Gao R, et al. Parameterizing neural power spectra into periodic and aperiodic components. Nat Neurosci 2020;23:1655–65. https://doi.org/10.1038/s41593-020-00744-x.

[29] Gaete M, Jorquera S, Bello-Lepe S, Mendoza YM, Véliz M, Alonso-Sanchez MF, et al. Standardised results of the Montreal Cognitive Assessment (MoCA) for neurocognitive screening in a Chilean population. Neurología (English Edition) 2023;38:246–55. https://doi.org/10.1016/j.nrleng.2020.08.021.

[30] Jiang J-L, Zhang L, He L-L, Yu H, Li X-F, Dai S-H, et al. Volatile Versus Total Intravenous Anesthesia on Postoperative Delirium in Adult Patients Undergoing Cardiac Valve Surgery: A Randomized Clinical Trial. Anesthesia & Analgesia 2022. https://doi.org/10.1213/ANE.0000000000006257.

[31] Culley DJ, Flaherty D, Fahey MC, Rudolph JL, Javedan H, Huang C-C, et al. Poor Performance on a Preoperative Cognitive Screening Test Predicts Postoperative Complications in Older Orthopedic Surgical Patients. Anesthesiology 2017;127:765–74. https://doi.org/10.1097/ALN.0000000000001859.

[32] Weiss Y, Zac L, Refaeli E, Ben-Yishai S, Zegerman A, Cohen B, et al. Preoperative Cognitive Impairment and Postoperative Delirium in Elderly Surgical Patients: A Retrospective Large Cohort Study (The CIPOD Study). Annals of Surgery 2023;278:59–64. https://doi.org/10.1097/SLA.0000000000005657.

[33] García PS, Kreuzer M, Hight D, Sleigh JW. Effects of noxious stimulation on the electroencephalogram during general anesthesia: a narrative review and approach to analgesic titration. British Journal of anesthesia 2021;126:445–57. https://doi.org/10.1016/j.bja.2020.10.036.

[34] Fritz BA, Kalarickal PL, Maybrier HR, Muench MR, Dearth D, Chen Y, et al. Intraoperative Electroencephalogram Suppression Predicts Postoperative Delirium. Anesthesia & Analgesia 2016;122:234–42. https://doi.org/10.1213/ANE.0000000000000989.

[35] Allegrini P, Menicucci D, Bedini R, Fronzoni L, Gemignani A, Grigolini P, et al. Spontaneous brain activity as a source of ideal 1 / f noise. Phys Rev E 2009;80:061914. https://doi.org/10.1103/PhysRevE.80.061914.

[36] Ibarra Chaoul A, Siegel M. Cortical correlation structure of aperiodic neuronal population activity. NeuroImage 2021;245:118672. https://doi.org/10.1016/j.neuroimage.2021.118672.

[37] Jacob MS, Roach BJ, Sargent K, Mathalon DH, Ford JM. Aperiodic measures of neural excitability are associated with anticorrelated hemodynamic networks at rest: a combined EEG-fMRI study n.d.

[38] Merkin A, Sghirripa S, Graetz L, Smith AE, Hordacre B, Harris R, et al. Age differences in aperiodic neural activity measured with resting EEG. Neuroscience; 2021. https://doi.org/10.1101/2021.08.31.458328.

[39] Boncompte, G., Medel, V., Cortínez, L. I., & Ossandón, T. Brain activity complexity has a nonlinear relation to the level of propofol sedation. British Journal of Anaesthesia, 127(2), 254-263. https://doi.org/10.1016/j.bja.2021.04.023

[40] Pani SM, Saba L, Fraschini M. Clinical applications of EEG power spectra aperiodic component analysis: A mini-review. Clinical Neurophysiology 2022;143:1–13. https://doi.org/10.1016/j.clinph.2022.08.010.

[41] Donoghue, T. A systematic review of aperiodic neural activity in clinical investigations. MedRxiv, 2024. 2024-10. https://doi.org/10.1101/2024.10.14.24314925

[42] He, Y., Wang, Z., Zhao, Y., Han, X., Guo, K., Sun, N., & Liu, X. (2024). Cognitive prehabilitation for older adults undergoing elective surgery: a systematic review and narrative synthesis. Frontiers in Aging Neuroscience, 16, 1474504. https://doi.org/10.3389/fnagi.2024.1474504

[43] Frank, E., Hall, M. (2001). A Simple Approach to Ordinal Classification. In: De Raedt, L., Flach, P. (eds) Machine Learning: ECML 2001. ECML 2001. Lecture Notes in Computer Science(), vol 2167. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-44795-4_13