doi:10.1093/brain/awy251

BRAIN
A JOURNAL OF NEUROLOGY

Robust EEG-based cross-site and cross-protocol classification of states of consciousness

Denis A. Engemann,[1,2,3,*] Federico Raimondo,[3,4,5,6,*] Jean-Rémi King,[2,7,8] Benjamin Rohaut,[3,9] Gilles Louppe,[7] Frédéric Faugeras,[3] Jitka Annen,[10] Helena Cassol,[10] Olivia Gosseries,[10] Diego Fernandez-Slezak,[4,5] Steven Laureys,[10] Lionel Naccache,[3,6] Stanislas Dehaene[2,11] and Jacobo D. Sitt[3,6]

*These authors contributed equally to this work.

Determining the state of consciousness in patients with disorders of consciousness is a challenging practical and theoretical problem. Recent findings suggest that multiple markers of brain activity extracted from the EEG may index the state of consciousness in the human brain. Furthermore, machine learning has been found to optimize their capacity to discriminate different states of consciousness in clinical practice. However, it is unknown how dependable these EEG markers are in the face of signal variability because of different EEG configurations, EEG protocols and subpopulations from different centres encountered in practice. In this study we analysed 327 recordings of patients with disorders of consciousness (148 unresponsive wakefulness syndrome and 179 minimally conscious state) and 66 healthy controls obtained in two independent research centres (Paris Pitié-Salpêtrière and Liège). We first show that a non-parametric classifier based on ensembles of decision trees provides robust out-of-sample performance on unseen data with a predictive area under the curve (AUC) of ~0.77 that was only marginally affected when using alternative EEG configurations (different numbers and positions of sensors, numbers of epochs, average AUC = 0.750 ± 0.014). In a second step, we observed that classifiers based on multiple as well as single EEG features generalize to recordings obtained from different patient cohorts, EEG protocols and different centres. However, the multivariate model always performed best with a predictive AUC of 0.73 for generalization from Paris 1 to Paris 2 datasets, and an AUC of 0.78 from Paris to Liège datasets. Using simulations, we subsequently demonstrate that multivariate pattern classification has a decisive performance advantage over univariate classification as the stability of EEG features decreases, as different EEG configurations are used for feature-extraction or as noise is added. Moreover, we show that the generalization performance from Paris to Liège remains stable even if up to 20% of the diagnostic labels are randomly flipped. Finally, consistent with recent literature, analysis of the learned decision rules of our classifier suggested that markers related to dynamic fluctuations in theta and alpha frequency bands carried independent information and were most influential. Our findings demonstrate that EEG markers of consciousness can be reliably, economically and automatically identified with machine learning in various clinical and acquisition contexts.

1 Parietal project-team, INRIA Saclay – Île de France, France
2 Cognitive Neuroimaging Unit, CEA DSV/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif sur Yvette, France
3 Inserm U 1127, CNRS UMR 7225, Institut du Cerveau et de la Moelle épinière, ICM, F-75013, Paris, France
4 Laboratorio de Inteligencia Artificial Aplicada, Departamento de Computación FCEyN, UBA, Argentina
5 CONICET – Universidad de Buenos Aires, Instituto de Investigación en Ciencias de la Computación, Godoy Cruz 2290, C1425FQB, Ciudad Autónoma de Buenos Aires, Argentina
6 Sorbonne Universités, UPMC Université Paris 06, Faculté de Médecine Pitié-Salpêtrière, Paris, France
7 New York University, 6 Washington Place, New York, NY, USA
8 Frankfurt Institute for Advanced Studies, Frankfurt, Germany

Received February 1, 2018. Revised August 3, 2018. Accepted August 20, 2018.

9 Department of Neurology, Columbia University, New York, NY, USA
10 Coma Science Group, GIGA Consciousness, University and University Hospital of Liège, Liège, Belgium
11 Collège de France, Paris, France

Correspondence to: Denis A. Engemann
1 Rue Honoré d'Estienne d'Orves
91120 Saclay, France
E-mail: denis-alexander.engemann@inria.fr

Correspondence may also be addressed to: Federico Raimondo
ICM - Hôpital Pitié Salpêtrière, 47 bd de l'hôpital, 75013 Paris, France
E-mail: fraimondo@dc.uba.ar

Jacobo Sitt
ICM - Hôpital Pitié Salpêtrière, 47 bd de l'hôpital, 75013 Paris, France
E-mail: jacobo.sitt@inserm.fr

Keywords: electroencephalography; disorders of consciousness; biomarker; machine learning; diagnosis

Abbreviations: AUC = area under the curve; DOC = disorders of consciousness; MCS = minimally conscious state; MPVA = multivariate pattern analysis; UWS = unresponsive wakefulness syndrome; wSMI = weighted symbolic mutual information

Introduction

Patients suffering from disorders of consciousness (DOC) demonstrate that it is possible to be awake in the absence of behavioural evidence of consciousness (Laureys *et al.*, 2010). Despite best efforts for consistency, current diagnostic procedures rely on human interaction and are, therefore, error-prone (Rohaut and Claassen, 2018). The degree of misdiagnosis in patients with DOC may exceed 40% when relying on the clinician's judgement without standardized behavioural assessments (Schnakers *et al.*, 2009). Even when using diagnostic instruments such as the Coma Recovery Scale-Revised (CRS-R) (Giacino *et al.*, 2004), misdiagnosis can remain high if patients are not assessed repeatedly within a short time window (Wannez *et al.*, 2017). Furthermore, in some cases evidence of conscious processing in these patients can only be obtained using functional neuroimaging where patients sometimes demonstrate wilful modulations of their brain activity (Owen *et al.*, 2006; Monti *et al.*, 2010). These patients have been labelled as 'covert awareness' or 'cognitive motor dissociation (CMD)' patients (Gosseries *et al.*, 2014; Schiff, 2015; Curley *et al.*, 2018).

Among the DOC one distinguishes the comatose state, the unresponsive wakefulness syndrome (UWS, historically vegetative state), and the minimally conscious state (MCS) (Giacino *et al.*, 2002; Laureys *et al.*, 2010). The presence of eye-opening helps to distinguish UWS patients from comatose ones (Jennett and Plum, 1972). Additionally, MCS but not UWS patients show signs of awareness (i.e. visual pursuit in MCS– and command following in MCS+) (Bruno *et al.*, 2011) while neither achieving functional communication nor object use. It is nevertheless believed that these patients can have a partial and fluctuating awareness of themselves and their surroundings and are more likely to recover (Luauté *et al.*, 2010; Faugeras *et al.*, 2018), which emphasizes the importance of reliable diagnostic tools.

In the past two decades, non-invasive brain imaging has supplemented behavioural assessments for detection of consciousness. Sleep studies and neurological assessments have early on revealed preferentially altered EEG amplitudes in the delta (2–4 Hz), theta (4–8 Hz) and alpha (8–12 Hz) frequency ranges (Emmons and Simon, 1956; Rosenberg *et al.*, 1977). PET revealed globally decreased glucose uptake in patients with DOC as compared to healthy controls (Stender *et al.*, 2014). Several functional MRI studies have documented disruption of functional connectivity along diverse subcortical and neocortical pathways in patients with DOC (Demertzi *et al.*, 2014). Ever since, advances in cognitive science have allowed one to infer consciousness from increasingly fine-grained patterns of brain activity. Accordingly, recurrent interactions between higher-order neocortical networks, as well as the morphology and complexity of brain dynamics in response to stimulation have been related to the states of consciousness (Tononi and Edelman, 1998; Dehaene and Naccache, 2001; Casali *et al.*, 2013; Iotzov *et al.*, 2017), which has led to various types of putative markers of consciousness.

Following recent trends in neuroimaging, the increasing number of neural markers of consciousness is likely to be best approached with multivariate pattern analysis (MVPA) (Naci *et al.*, 2012; King *et al.*, 2013*b*; Claassen *et al.*, 2016). Indeed, machine learning algorithms can be trained to best predict the medical status of individual patients from unknown combinations of physiological markers (for example, Chang *et al.*, 2005). Typically, a classifier is trained to optimally discriminate clinical labels based on brain data. Generalization performance is then assessed by comparing the predictions of the classifier to the actual diagnosis when presented with unseen data. In the absence of independent datasets, cross-validation is performed to estimate

the out-of-sample performance by subdividing the data into training and testing sets and averaging over testing set scores. It is, however, noteworthy that cross-validation tends to be too optimistic when sample sizes are small (Varoquaux *et al.*, 2016; Varoquaux, 2018; Woo *et al.*, 2017), rendering face-value interpretation of scores futile for a significant proportion of neuroimaging studies. Examples of MVPA for the study of patients with DOC include the analysis of patterns of resting state functional MRI functional connectivity (Demertzi *et al.*, 2015), spectral responses to command following (Goldfine *et al.*, 2011; Cruse *et al.*, 2012) and cerebral metabolism to distinguish locked-in patients from UWS (Phillips *et al.*, 2011).

In this context, EEG is particularly interesting as this neurophysiological technique conveys rich temporal information on cognitive operations and can be economically operated in a wide range of situations, potentially enabling bedside or home assessment. The challenge of processing large amounts of EEG data at scale can nowadays be addressed using automated EEG processing methods (Engemann *et al.*, 2015; Jas *et al.*, 2017). However, preferences for cognitive theories and EEG methodologies are heterogeneous across laboratories, which significantly obstructs the development of large-scale data resources well suited for high-fidelity machine learning. The emerging EEG markers, so far, fall into four conceptual families. Evoked markers are based on time-locked event-related analysis of cognitive experiments. The other families contain markers defined independently from protocols, including, connectivity markers exploiting brain–network interactions, information theory markers capitalizing on information properties of time series and spectral markers quantifying neuronal oscillations or stochastic band-limited dynamics. Yet, the situation is further complicated by the fact that DOC reflect several cognitive and neurological components rather than a single dimension, motivating the consideration of marker profiles (Bayne *et al.*, 2016; Sergent *et al.*, 2017). In a recent study, using a support vector machine (SVM) classifier, Sitt *et al.* (2014) analysed dozens of EEG markers from more than 150 high-density EEG recordings during an auditory novelty task. Interestingly, combinations of markers synergistically outperformed single markers. Similarly, using graph-theoretical summaries of alpha-band connectivity, Chennu *et al.* (2017) presented an alternative SVM approach cross-validated on 104 patients with severe brain injury (among those 89 with DOC).

Nevertheless, a generalized large-scale attempt for cross-laboratory predictions of state of consciousness in brain-injured patients is missing, and several practical questions remain unanswered: what is the optimal duration for individual EEG recordings? Which task should the patient undergo, if any? How many sensors should be used, and where should they be located? Can a single machine learning algorithm perform on data from different clinical centres? Do models based on current EEG markers achieve prospective generalization on independent data (Woo *et al.*, 2017)? Are single markers sufficiently powerful and when does multivariate classification provide the clearest advantage?

To address these questions, we rigorously probed the robustness and validity of EEG markers of consciousness. Using the robust *Extra-Trees* algorithm (Geurts *et al.*, 2006) we developed a classifier to differentiate UWS from MCS patients (which we termed 'DOC-Forest'). This classifier was trained and tested using 28 potential EEG markers of consciousness from 249 patients recorded at the Paris Pitié-Salpêtrière and 78 patients from the University Hospital of Liège. We first show that different EEG configurations (sensor number, sensor position and numbers of epochs) and experimental protocols (auditory stimulation or resting state) induce significant changes in the distribution and performance of the EEG markers. Yet, we found that the DOC-Forest is relatively immune to such variations by exploiting the information conveyed by reliable EEG markers. We subsequently demonstrate out-of-sample generalization to two independent datasets: a new cohort of 107 task-EEG recordings (not previously analysed) from Paris and 78 resting state EEG recordings from the University Hospital of Liège. Moreover, we show that our DOC-Forest's generalization performance is decisively superior to univariate markers. Finally, by investigating the influence of individual markers on the decisions of DOC-Forest, we found that alpha-band power, theta-band connectivity and time series complexity carry complementary information about states of consciousness.

Materials and methods

Ethics statement

This research project was approved by the ethical committee of the Pitié-Salpêtrière hospital under the code 'Recherche en soins courants' (routine care research). All investigations were carried out in accordance with the Declaration of Helsinki on ethical principles for medical research involving human subjects. For the dataset from the Coma Science Group, the family of the patient gave their informed consent for participation in the study, and the Ethics Committee of the University hospital of Liège approved the study.

Participants

In total, 327 EEG recordings from 268 distinct patients from our expert centres were included in the current study (Table 1). Patients were assessed at variable delays (sub-acute or chronic stage following the brain injury) in order to clarify the actual state of consciousness. Clinical assessments were performed at least three times in the Paris dataset and five times in the Liège dataset, in all cases on different days by trained clinicians (see 'Acknowledgements' section) and included systematically the CRS-R. CRS-R scores range from 0 to 23 and reflect the presence or absence of response on a set of hierarchically ordered items testing auditory, visual, motor,

Table 1 Patient characteristics in the three datasets

	Auditory local global task		Resting state
	Paris 1	Paris 2	Liège
$n_{(EEG)}$	142	107	78
$n_{(patients)}$	98	92	78
$n_{(UWS)}$	75	52	21
$n_{(MCS)}$	67	55	57
Gender ratio, male/female	2.06	1.93	1.26
Age, years, mean (SD)	46.5 (17.8)	45.4 (17.7)	38.0 (14.3)
Delay, days, mean (SD)	126.0 (372.9)	299.6 (823.6)	1040.6 (1227.6)
Delay, days, min–max	6–2611	8–6570	11–5380
Anoxia, %	29.6	30.4	21.7
Stroke, %	29.6	15.2	3.84
Traumatic brain injury, %	23.5	28.2	48.1
Other, %	18.4	29.4	21.8

oromotor, communication, and arousal functions (Giacino *et al.*, 2004). According to the best assessment, each patient was diagnosed with UWS or MCS. The data acquisition protocol included, in all centres, multiple clinical assessments and at least one EEG recording. For some patients, several EEG recordings were available, which we later accounted for by statistical modelling. The number of recordings varied considerably across datasets; however, the ratio of MCS to UWS patients was roughly balanced. Across all datasets more male than female patients were observed. Age distributions were similar; however, the delay from accident was visibly higher for the resting state dataset. Likewise, the distribution of aetiologies was different for the resting state dataset while proportions were consistent with the literature.

Experimental paradigm

In the Paris 1 and 2 datasets, task-related EEG signals were obtained using the 'Local-Global' protocol (Bekinschtein *et al.*, 2009) designed to study unconscious and conscious auditory processing. In the Liège dataset, EEG recordings were task-free (see the online Supplementary material for details)

Selection and computation of putative EEG markers of consciousness

We extracted 28 putative EEG biomarkers detailed in Sitt *et al.* (2014). The markers can be grouped into four conceptual families, i.e. information theory, connectivity, spectral, and evoked response markers (Table 2). Among several connectivity metrics described in Sitt *et al.* (2014), we only considered the weighted symbolic mutual information (wSMI) metric in theta frequency band as previous research had suggested that the long-range connectivity patterns theoretically related to consciousness are most robustly and accurately assessed by this metric (King *et al.*, 2013*a*). Note that for the analysis of resting state EEG we did not make use of the evoked response markers as those are only defined for the task used in the Paris datasets. For a detailed description and discussion of the markers, see Sitt *et al.* (2014).

The markers commonly used in clinical neuroscience are often defined at a general level and can be observed over multiple electrodes, time points or frequency bands. To delineate low-level features, we computed four summary statistics from each marker (Fig. 1). To summarize epochs, we either computed the 80% trimmed mean, or the standard deviation (SD). The sensor dimension was then summarized using a mean or the standard deviation, yielding four combinations in total (Fig. 1A). Throughout the manuscript we refer to these marker subtypes as 'mean,mean', 'std,mean', 'mean,std' and 'std,std' and in figures, for brevity, 'm,m', 's,m', 'm,s', 's,s'. For a full list and abbreviations, see Table 2.

Computation was carried out using a designated Python software library implementing the biomarker extraction functionality from Sitt *et al.* (2014). The extracted markers closely matched the original values and group results for the reference datasets were qualitatively reproduced (Engemann *et al.*, 2015).

Statistical analysis

Classification of disorders of consciousness from EEG markers

Diagnosis was classified based on EEG markers using a univariate and a multivariate machine learning strategy. To enable comparisons across studies, we also computed model-free performance on single markers as in Sitt *et al.* (2014). Performance was assessed using the area under the curve (AUC). For details see Supplementary material 'Area under the curve metric' section. For multivariate and univariate pattern analysis, we chose the *Extra-Trees* algorithm (Geurts *et al.*, 2006) whose non-parametric design facilitates robust classification. To complement insights from univariate classification, we extracted the so-called variable importance metric from the *Extra-Trees* following best practice recommendations for enhanced interpretability (Louppe *et al.*, 2013; Louppe, 2014). Accordingly, our variable importance scores reflect mutual information between a variable and the diagnosis while conditioning out the other variables. For background information on parameters and model tuning, see Supplementary material 'Multivariate pattern classification' section. To use a common currency when comparing

Table 2 Potential EEG biomarkers of consciousness

Abbreviation	Marker	Conceptual family	Protocol
PEΘ	Permutation entropy	Information theory	Task, rest
K	Kolmogorov complexity	Information theory	
wSMI Θ	Weighted symbolic mutual information	Connectivity	
α	Alpha PSD	Spectral	
|α|	Normalized alpha PSD	Spectral	
β	Beta PSD	Spectral	
|β|	Normalized beta PSD	Spectral	
δ	Delta PSD	Spectral	
|δ|	Normalized delta PSD	Spectral	
γ	Gamma PSD	Spectral	
|γ|	Normalized gamma PSD	Spectral	
θ	Theta PSD	Spectral	
|θ|	Normalized theta PSD	Spectral	
MSF	Median power frequency	Spectral	
SE90	Spectral entropy 90	Spectral	
SE95	Spectral entropy 95	Spectral	
SE	Spectral entropy	Spectral	
CNV	Contingent negative variation	Evoked	Task
P1	Short-latency auditory potential to the first sound	Evoked	
P3a	Mid-latency auditory potential to the first sound	Evoked	
P3b	Mid-latency auditory potential to the first sound	Evoked	
GD–GS	Full contrast	Evoked	
LD–LS	Full contrast	Evoked	
LSGD–LDGS	Full contrast	Evoked	
LSGS–LDGD	Full contrast	Evoked	
ΔMMN	Contrasted MNN (local deviant versus local standard)	Evoked	
ΔP3a	Contrasted P3a (local deviant versus local standard)	Evoked	
ΔP3b	Contrasted P3b (global deviant versus global standard)	Evoked	

GD = global deviant; GS = global standard; LD = local deviant; LS = local standard; MMN = mismatch negativity; PSD = power spectral density.

univariate with multivariate marker performance, we turned single markers into fully functional classification models by using the identical recipe as for the DOC-Forest, effectively only changing the features passed to the classifier. This allowed us to predict the probability of DOC diagnosis from single markers using the same framework as for multivariate analysis.

Statistical inference

We extended our visualizations into hypothesis tests by using the percentile bootstrap (Efron and Tibshirani, 1993) (Supplementary material). To assess out-of-sample generalization we used two complementary approaches: a conservative validation on independent data (new cohorts, different protocols and laboratories) and cross-validation (Supplementary material).

Software

All data were processed using the Python programming language. To simplify preprocessing and feature extraction for machine learning, we developed a designated software library (available at https://github.com/nice-tools/nice) built on top of the open source software libraries MNE (Gramfort *et al.*, 2014) and scikit-learn (Pedregosa *et al.*, 2011). The DOC-Forest recipe is publicly available (https://github.com/nice-tools/nice) to encourage community efforts in building predictive models of DOC patients' state of consciousness.

Data availability

The clinical data used in this paper can be made available upon reasonable request, but because of the sensitive nature of the clinical information concerning the patients the ethics protocol does not allow open data sharing.

Results

Robust detection of state-of-consciousness from EEG features

Multivariate classification of UWS versus MCS is robust across EEG configurations

The DOC-Forest classifier exhibited an average performance of AUC = 0.75 (SD = 0.014) and performed better and more robustly than most other markers did individually (Fig. 2A, B, Supplementary Figs 1 and 2). Moreover, its discrimination performance increased with the number of sensors ($rho_{Spearman}$ = 0.803, 95% CI: 0.646–0.891; $P < 0.001$) and epochs ($rho_{Spearman}$ = 0.40, 95% CI: 0.07–0.668; $P < 0.05$) (Fig. 2B) but was already strong with 16 sensors and 5% of epochs. Importantly, using the full EEG configuration, the performance closely resembled previous

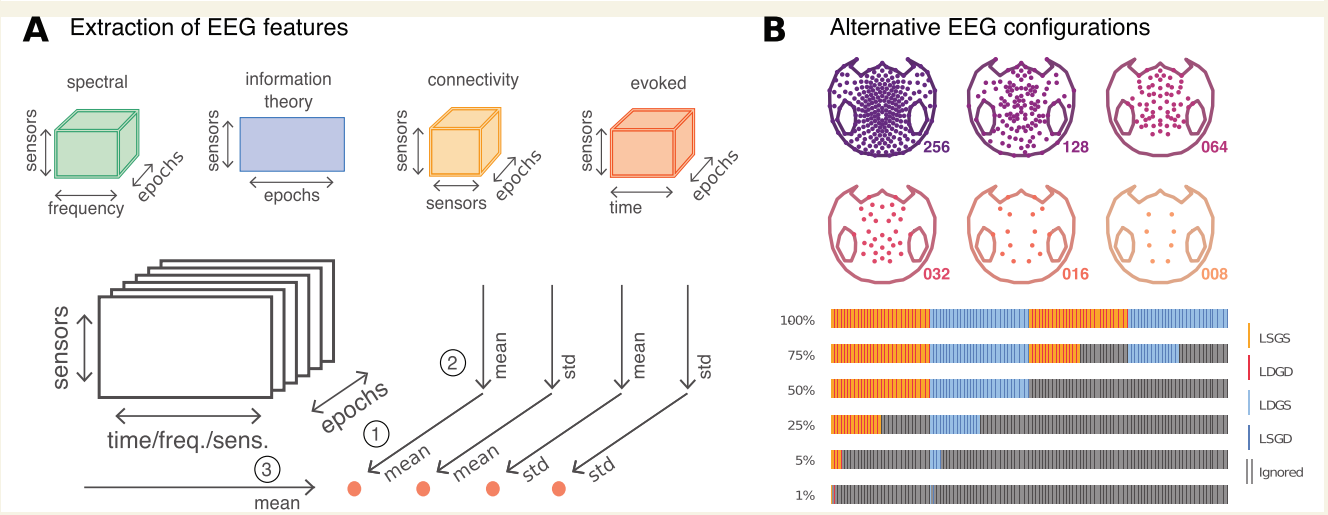

Figure 1 Extraction of EEG features. (**A**) The EEG markers fell into four conceptual families, i.e. spectral, information theory, connectivity and evoked responses. When computing the markers from the preprocessed EEG, we obtained several observations for channels, epochs, time points and frequency bins, depending on the family. Following Sitt *et al.* (2014), we extracted four features from each marker (indicated by the red dots) by summarizing the observations systematically: we computed either the mean or the standard deviation first across epochs (1) and then across sensors (2). If a third dimension was present (3), we summarized it using the mean. We, hence, referred to the ensuring four features as 'mean,mean', 'mean,std', 'std,mean' and 'std,std'. (**B**) We repeated this process using six alternative sensor configurations (256,128, 64, 32, 16, 8) and six alternative percentages of consecutive epochs (1, 5, 25, 50, 75, 100) with about seven epochs at 1% and about 700 epochs at 100%. Sensors were selected such that they approximated realistic EEG caps respecting the international 10-20 system. Selection of epochs respected the relative proportions of conditions used in the task. This allowed us to compute markers based on experimental contrasts at any point. This yielded 36 alternative EEG configurations. D = deviant; freq. = frequency; G = global; L = local; S = standard; sens. = sensor; std = standard deviation.

results reported by Sitt *et al.* (2014) and beat any other marker (Supplementary Fig. 2). These results suggest that the DOC-Forest preferentially tracks information conveyed by a few robust markers over a variety of EEG configurations.

Using the full configuration, we subsequently assessed the consistency of classification success for different aetiological groups and different levels of chronicity (Supplementary material 'Consistency of classification results in diagnostic subgroups' section). Comparable results were obtained for the chronic (delay $>$ 30 days) and acute (delay $\leqslant$ 30 days) groups. The classification performance was significant for all the aetiology groups (i.e. anoxia, stroke and traumatic brain injury). Yet, in the case of traumatic brain injury patients the performance was slightly lower, suggesting that the heterogeneity of this group makes it more difficult to classify. For additional fine-grained comparisons between single markers and the DOC-Forest, see Supplementary material 'Detailed comparison between individual markers and DOC-Forest' section.

Classification is preferentially driven by distinct theta- and alpha-band dynamics

While it is not convenient to reason separately about each of the 2000 decision trees grown inside our DOC-Forest, we can still analyse the relative contributions of EEG markers to classification performance by considering the variable importance. This multivariate metric approximates the mutual information between a marker and the diagnosis while controlling for the contribution of other markers. The variable importance can deviate systematically from the univariate AUC whenever information is shared between markers or the model has identified non-linear interaction effects. Inspecting all DOC-Forest classifiers for the 36 configurations, we observed that markers contributing most strongly on average belonged to different conceptual families (Fig. 2C). Specifically, permutation entropy and long-range connectivity in the theta band and the alpha frequency band power were top ranked in terms of univariate discrimination and variable importance. In contrast, evoked markers, on average, often assumed values below 0.89%, which is less than would be expected if all markers were equally influential. We observed a positive but non-linear relationship between average AUC and average variable importance ($rho_{Spearman}$ = 0.817, 95% CI: 0.727–0.880; P $<$0.001). It can be seen that highly performing markers were disproportionally more important than expected for a linear association (Fig. 2C).

Exploiting invariant EEG features of consciousness for generalization

Generalization to independent data, protocols and configurations

Here we considered two independent cohorts: 107 task-EEG recordings from the Paris Pitié-Salpêtrière Hospital

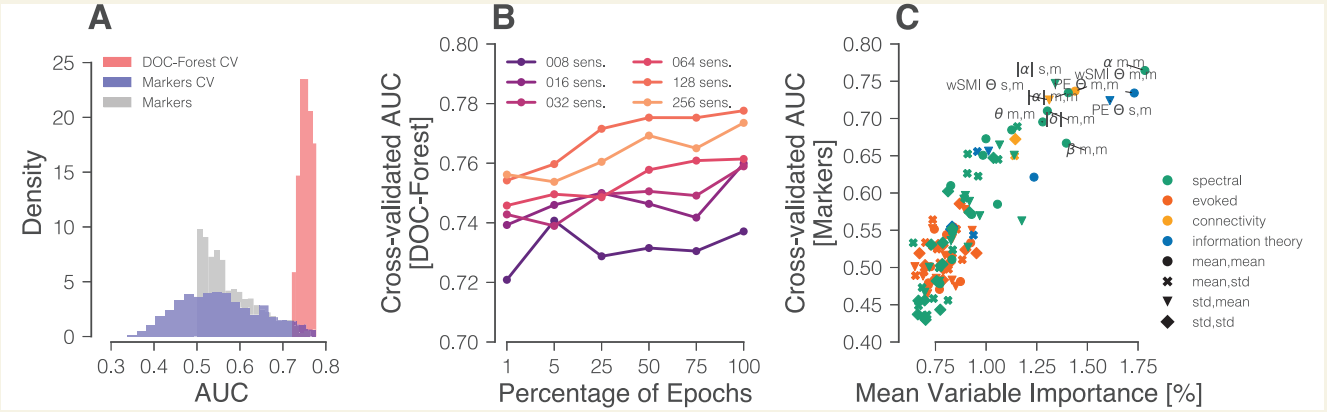

Figure 2 Performance of EEG markers of consciousness across different EEG configurations. (**A**) Performance distribution over markers (grey: model-free in-sample performance; blue: cross-validation with univariate forests) and the multivariate DOC-Forest pattern classifier (red) across 36 EEG configurations on the Paris 1 dataset. (**B**) DOC-Forest tended to improve as more epochs and sensors were used. Although optimal performance was achieved with 128 electrodes, reasonable performance could still be obtained with only 16 electrodes and a minimum of epochs. (**C**) Cross-validated univariate performance as a function of multivariate variable importance in the DOC-Forest, both averaged across EEG-configurations. Marker subtype and conceptual family are indicated by shape and colour, respectively. A positive but non-linear relationship emerged. The best univariate markers were disproportionally more important to the DOC-Forest as a linear relationship would predict. It is noteworthy that markers from the spectral, connectivity and information theory families had the highest univariate performance and were assigned the highest importance by the classifier while the evoked markers systematically less important. See also Table 2; m,m = mean,mean; m,s = mean,std; PE = permutation entropy; sens. = sensor; s,m = std,mean; s,s = std,std.

(Paris 2) and 78 resting state EEG recordings by an independent research group (Coma Science Group, Liège, Belgium; see Table 2 for an overview). When training the DOC-Forest on the Paris 1 dataset, and testing the algorithm on the Paris 2 dataset, each time using the full EEG configuration, we observed significant classification performance with an AUC around 0.73 [standard deviation (SD) = 0.05, 95% CI: 0.63–0.82] (Fig. 3A). Likewise, when trained on all available data from Paris (Paris 1 and Paris 2) but ignoring the evoked markers (Table 1 and Fig. 1A), the DOC-Forest scored an AUC of 0.78 (SD = 0.06, 95% CI: 0.66–0.89) on the Liège resting state data (Fig. 3B).

We subsequently assessed generalization of our classifier trained on the Paris dataset to distinguish UWS versus MCS to a dataset of 66 conscious controls. The DOC-Forest classified 94% of the controls (Paris local-global paradigm: 34 of 36, Liège resting state: 28 of 30) as MCS. This result suggests that the patterns used by the classifier to distinguish UWS versus MCS patients extrapolate to normal controls.

Furthermore, we detected two cognitive-motor dissociation patients in the Liège dataset. These patients were originally labelled as UWS from their behaviour but showed evidence of conscious processing using an active functional MRI paradigm (see the Supplementary material for a brief description of the two cases). Both cases were classified as MCS by DOC-Forest.

Generalization using univariate markers

Less consistent results were obtained when using univariate forests based on the markers from the connectivity, information theory and spectral families, which showed the highest cross-validation performance on the training set. For Paris 1 (Fig. 3A) these were wSMI (mean,mean), theta permutation entropy (mean,mean) and normalized alpha power (std,mean) with scores of 0.75, 0.74 and 0.77, respectively. For the combined Paris 1 and 2 dataset these were: theta wSMI (std,mean), theta permutation entropy (mean,mean) and alpha band power (mean,mean) with cross-validated scores of 0.69, 0.69 and 0.73, respectively. All univariate models showed lower generalization performance (0.04 to 0.14 AUC points) compared to the DOC-Forest and only the alpha band classifiers performed meaningfully better than a dummy classifier (Fig. 3, middle). Comparing the variable importance to each marker's out-of-sample performance, again, revealed positive non-linear associations (Fig. 3A and B, right, $rho_{Spearman\ Paris\ 1\rightarrow 2}$ = 0.477, 95% CI: 0.312–0.620; $P < 0.001$; $rho_{Spearman\ Paris\rightarrow Liège}$ = 0.521, 95% CI: 0.309–0.684; $P < 0.001$). The display reveals that several univariate models showed reasonable generalization performance with AUC values beyond 0.70. Highly performing markers were disproportionally more important for the DOC-Forest than would have been expected assuming a linear relationship.

Strikingly, generalization was even successful when different EEG configurations were combined, e.g. training with 100% of the epochs and 32 sensors and testing with 50% of the epochs and eight sensors, although this induced decodable differences between training and testing sets (Supplementary Fig. 3). On average, the DOC-Forest performed significantly higher than any of the three corresponding univariate forests (Table 3). Inspection of the

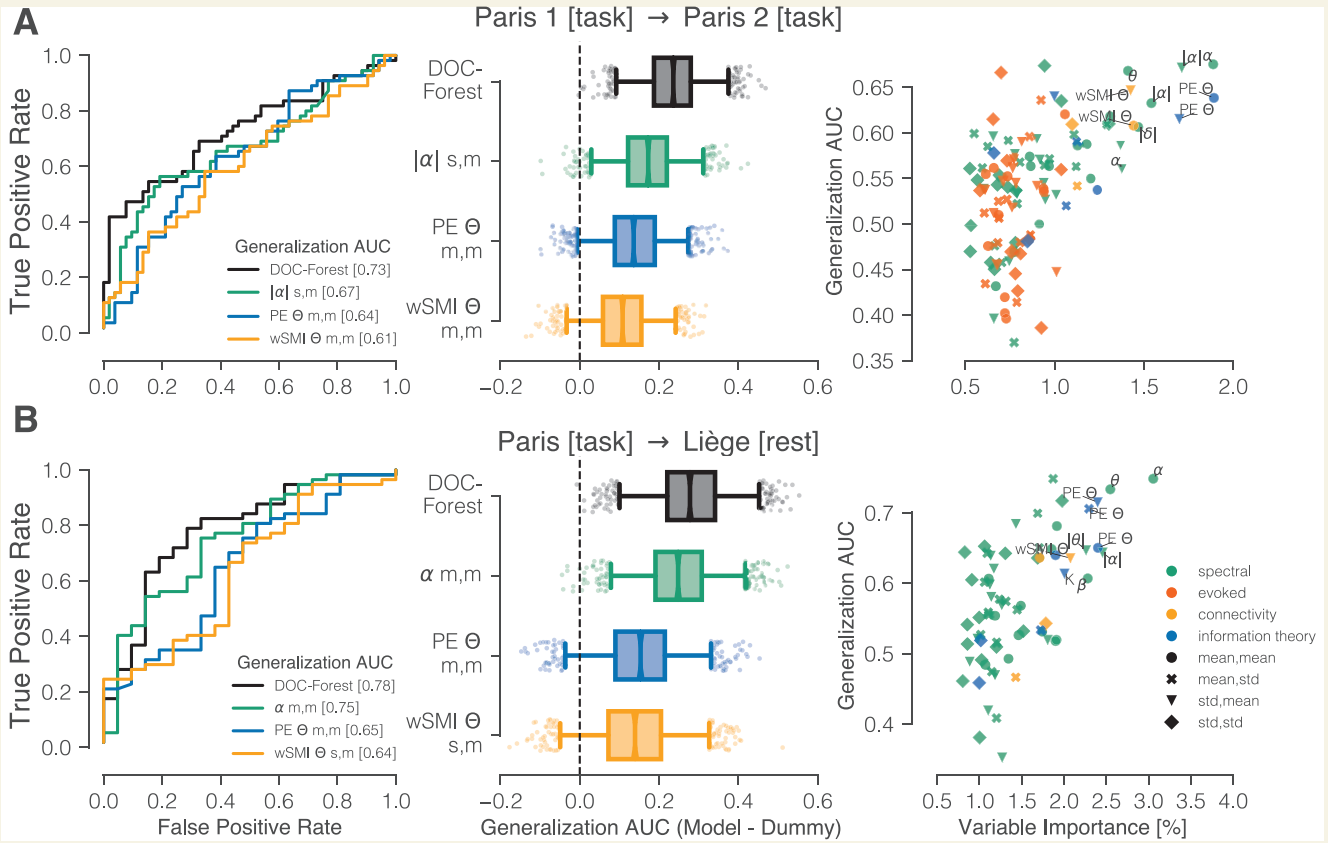

Figure 3 Generalization between datasets and protocols. (**A**) Generalization from the Paris 1 cohort to 107 new EEG recordings from Paris (task-EEG in both cases). *Left*: The ROC curves for the multivariate DOC-Forest and three univariate forests based on the feature that performed best (cross-validation) on the training set corresponding to the connectivity, information and spectral families. *Middle*: Bootstrap distributions of improvements over a dummy classifier based on paired differences, ordered by performance. Positive values indicate performance better than the dummy model. Boxplot whiskers show the 95% CI. *Right*: The generalization performance of each marker against training-set importance. The 10 most important features are labelled for convenience. (**B**) Generalization from 249 task-EEG (Paris 1 + Paris 2) to 78 resting state EEG recordings (Liège) depicting an equivalent analysis as in **A** but not including the evoked response features. The results suggest meaningful prospective generalization for the DOC-Forest while the univariate models were overall less successful. See also Table 2. m,m = mean,mean; m,s = mean,std; PE = permutation entropy; sens. = sensor; s,m = std,mean; s,s = std,std.

cross-configuration generalization patterns revealed that the performance changes were far from random, favouring specific but distinct combinations of sensors and epochs for both generalization tasks (Supplementary Fig. 4).

Robustness to noise

As the DOC-Forest seemed resilient to mismatching EEG configurations, we conducted a computational stress-test by adding noise to the markers in the testing set until classification broke down (Fig. 5A). Unsurprisingly, across generalization tasks, the univariate classifiers collapsed earlier at signal-to-noise ratios (SNRs) between 1/10 and 1/100, whereas the DOC-Forest endured longer, eventually failing at SNR values of 1/1000. Another concern potentially limiting generalization performance is the quality of the diagnostic information. We empirically assessed in a second computational stress-test the stability of generalization from Paris to Liège in the face of increasingly inaccurate diagnostic training labels (Fig. 5B). By design, this simulation forced the DOC-Forest to collapse and eventually yield systematically wrong predictions. However, the classifier still delivered reasonable predictions even if up to 30% of the diagnostic labels were flipped. Moreover, the literature would predict between 6% and 17% of misdiagnoses (Wannez *et al.*, 2017) for the three to five CRS-R repetitions used in this study and, here, fall into the range of resilient generalization. These results demonstrate that the DOC-Forest is not only relatively robust to noise in the data but also to noise in the diagnostic labels.

Discussion

We evaluated the robustness to different EEG configurations and recording conditions of univariate and multivariate pattern based on 28 putative EEG biomarkers of consciousness using the *Extra-Trees* algorithm. To the best of our knowledge, our study represents the most extensive

Table 3 Average generalization performance over different EEG configurations

Generalization	Contrast	Difference	95% CI
Paris 1 → 2	DOC-Forest - wSMI Θ (m,m)	$D = 0.124^{***}$	0.122–0.125
Paris 1 → 2	DOC-Forest - PE Θ (m,m)	$D = 0.097^{***}$	0.096–0.098
Paris 1 → 2	DOC-Forest - |α|(s,m)	$D = 0.035^{***}$	0.033–0.037
Paris → Liège	DOC-Forest - wSMI Θ (s,m)	$D = 0.140^{***}$	0.139–0.142
Paris → Liège	DOC-Forest - PE Θ (m,m)	$D = 0.118^{***}$	0.115–0.120
Paris → Liège	DOC-Forest - α (m,m)	$D = 0.035^{***}$	0.034–0.037

$^{***}p < 0.001$.
See also Table 2. m,m = mean,mean; m,s = mean,std; PE = permutation entropy; s,m = std,mean.

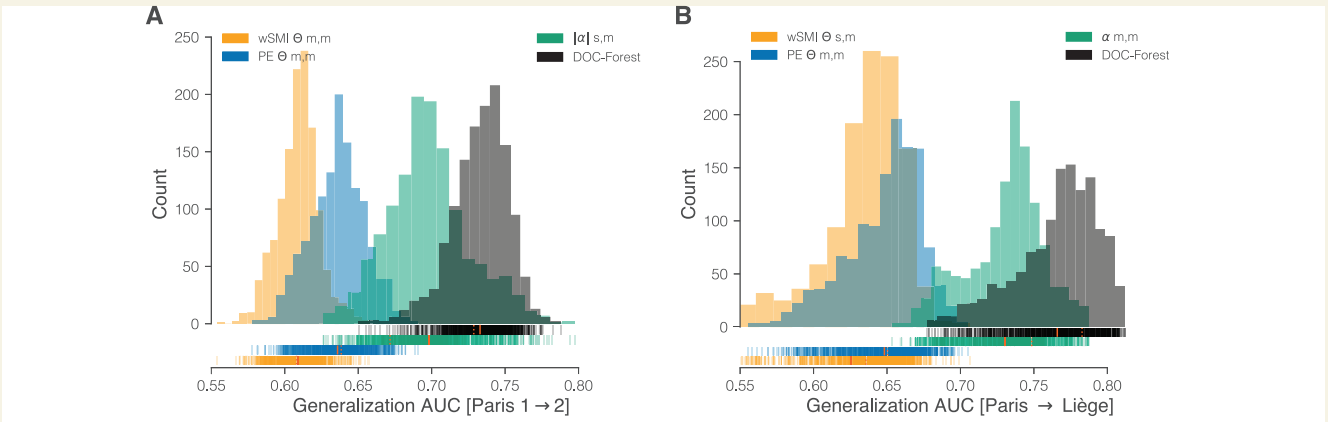

Figure 4 Generalization between datasets and protocols when EEG configurations differ. (**A**) Generalization from Paris 1 to Paris 2 when 1296 different combinations of EEG configurations were used for training and testing (six sensors × six epoch configurations for each set). The same univariate forest models as in Fig. 3 were considered next to the multivariate DOC-Forest. The distribution of AUC scores is indicated by the histograms, single observations are indicated by the rug plot. The orange solid lines indicate the mean of the distribution, the orange dotted line the performance when the reference configuration of 100% epochs and 256 sensors is used on both training and testing. (**B**) The same analysis for the generalization from the joint Paris 1 and 2 dataset to the Liège dataset. It can be seen that, on average, the DOC-Forest outperforms any of the univariate models. See also Table 2. m,m = mean,mean; m,s = mean,std; PE = permutation entropy; sens. = sensor; s,m = std,mean; s,s = std,std.

validation of a machine learning approach to diagnose UWS versus MCS patients for two reasons. Our findings are based on the currently largest EEG dataset of patients suffering from DOC, comprising 327 recordings. Second, in the context of DOC, the present study is the first to demonstrate prospective generalization of multivariate pattern classification between different centres, EEG configurations, and protocols. We demonstrated that robust generalization can be achieved despite non-trivial changes in the spatio-temporal configuration of the EEG and that this generalization can be resistant to certain degree of uncertainty in the training labels (up to 20%). We showed that by relying on a robust classification algorithm, meaningful generalization could be achieved even if the performance of individual markers varied systematically between datasets. While certain EEG markers, i.e. alpha band power and its fluctuations turned out to be useful as stand alone classifiers we found that the advantage of multivariate over univariate classification was most striking when systematic differences between the training and testing sets were present. Moreover, we found the DOC-Forest to preferentially base its predictions on diverse aspects of alpha and theta frequency band dynamics. Importantly, our results show that EEG-markers of consciousness can be accessed equivalently from task and resting state EEG.

Robust learning of UWS versus MCS diagnosis from EEG markers of consciousness

Our results demonstrate that diagnosis of UWS versus MCS patients can be robustly inferred from multivariate pattern classification using a wide array of EEG configurations (Fig. 2A and B). This was also the case with a minimum of sensors (~16) and epochs (10–50) and even when EEG configurations differed on the training and testing data (Fig. 4, Supplementary Figs 3 and 4), e.g. when training on 10% of the epochs with eight sensors and testing on all epochs with 256 sensors. We observed that many

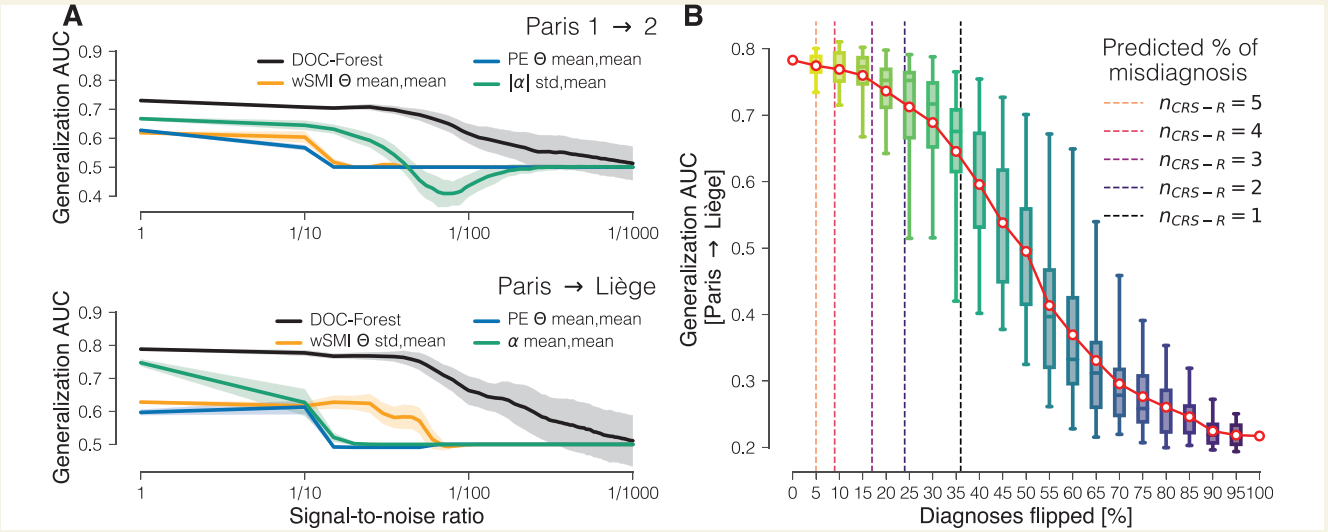

Figure 5 Computational stress tests. (**A**) The generalization performance of the DOC Forest and three univariate models as signal-to-noise ratio is gradually reduced on the testing set. The noise was generated independently from Gaussian distributions with mean and variance parameters from each feature with 50 realizations, scaled by the signal-to-noise ratio parameter and added to the testing set, such that at 1/10 the noise was 10 times stronger than the signal. The standard deviation of performance over realizations is indicated by the shaded areas. It can be readily seen that the DOC-Forest survives longest while at the same time decreasing its performance more slowly than each of the three univariate models. In general, univariate models did not survive a signal to noise ratio of 1/100 or smaller while the DOC-Forest still showed meaningful generalization performance beyond such low SNR values. (**B**) We estimated the impact of misdiagnosis on generalization empirically by flipping the diagnosis labels for an increasing percentage of patients (0 to 100 in steps of five). To avoid bias and estimate variability, we randomly draw patients at each percentage level and repeated the process 50 times. The median generalization performance is depicted by the boxplots (whiskers show the 2.5 and 97.5 percentiles) and the mean performance by the superimposed red circles. The performance at 0% and 100% flipping is shown by the red circles. For convenience, the percentage of misdiagnoses predicted from the number of CRS-R assessments reported by Wannez *et al.* (2017) is superimposed by the coloured dotted lines. It can be seen that the mean generalization performance drops more slowly between 10 and 30% than between 30 and 50% and remains reasonable even if up to 30% of the diagnoses are flipped. PE = permutation entropy.

individual markers were highly variable (Fig. 2A, Supplementary Figs 1 and 2). Nonetheless, our DOC-Forest fluctuated narrowly between AUC scores of 0.72 and 0.77 (Fig. 2D). Inspection of our classifier in terms of the variable importance revealed a striking pattern (Fig. 2C and Supplementary Fig. 1B). Markers that were most influential for its classifications not only were the ones with the greatest individual discrimination performance, but also turned out to be less susceptible to changes in the EEG configuration, noise on the EEG features and noise in the diagnostic labels (Figs 4 and 5). Interestingly, the overall relationship between univariate performance and variable importance was not linear. As univariate marker performance increased, marker importance increased disproportionally, i.e. at the top of the distribution, a change in univariate AUC lead to a bigger change in importance than at the bottom of the distribution. Our findings, therefore, suggest that our DOC-Forest provides robust learning of UWS versus MCS diagnosis by enhancing the impact of robust EEG markers.

In this context, it may be interesting to consider the recently issued warning that predictive variables are not necessarily the ones that differ significantly (Lo *et al.*, 2015; Bzdok *et al.*, 2018). As the AUC can be regarded as a rescaled Mann-Whitney U-test (Supplementary material), significant univariate classification as in Sitt *et al.* (2014) implies significant differences in a marker between the diagnoses. The presence of univariate classification success and its positive correlation with multivariate variable importance suggests that, in the present study, more significant variables were more predictive while less predictive variables were less significant.

Robust classification was driven by distinct alpha and theta frequency band dimensions

Our findings suggested that protocol-general markers were, overall, more reliable. Strikingly, these markers, belonging to different conceptual families, were all related to neuronal dynamics in the theta and alpha range (Figs 3 and 4). The robustness of these markers may be explained by the fact that no excessive averaging is needed for their extraction and their characteristic EEG topographies are simple and easy to capture with few sensors. However, the tight relationship between variable importance and conditional mutual information (Louppe, 2014) suggests that these top performing markers carry independent information. Indeed, recent research has suggested a rather complex picture of functional

and pathophysiological landscapes. The complexity of theta-band signals and their long-range interactions could reflect distinct memory processes underlying consciousness, such as access and maintenance (Axmacher *et al.*, 2010). Similarly, alpha-band power may reflect global arousal and demands for dynamic inhibition required for functional encapsulation of cortical networks (for an overview see Sadaghiani and Kleinschmidt, 2016). Moreover, intact consciousness has been related to the peak frequency of alpha and theta band oscillations originating from distinct cerebral generators (Schiff, 2010; Williams *et al.*, 2013). In fact, the mesocircuit model predicts that the downregulation of the thalamo-cortical circuits following a brain injury should be directly associated to changes in the interactions within these frequency bands observed in this study (Victor *et al.*, 2011; Schiff *et al.*, 2014). Yet, this is further complicated by the fact that these generators can be selectively disrupted for different aetiologies and can show a variety of regional effects during anaesthesia (Purdon *et al.*, 2013). While future experimental research is desirable to disentangle these facets, our findings suggest that the presence of independent physiological sources of information may enhance generalization as it is unlikely that all of their measurements will be corrupted at the same time on new data.

But do our results imply that less important variables were useless? Not necessarily. Many evoked markers enjoy a high degree of neuroscientific validation and intuitively support clinical reasoning. The P3 markers, for example, belong to the most studied indices of consciousness in the EEG literature and are commonly used in brain computer interfaces settings (Lulé *et al.*, 2013). They have been related to processing novelty in bottom-up information, the global neuronal workspace, access consciousness, and context-updating (Donchin and Coles, 1988; Pins, 2003; Sergent *et al.*, 2005; Dehaene *et al.*, 2006; Polich, 2007). Considering such markers for MVPA may, thus, improve interpretability. Additionally, evoked markers indexing auditory novelty have been shown to be rather specific than sensitive (King *et al.*, 2013*b*). Likewise, it could be the case that candidate markers of conscious access, e.g. P3b, may be more relevant to distinguish MCS+ from MCS– patients (Naccache, 2018). Although being de-emphasized by the DOC-Forest, evoked markers may still have contributed positively. Indeed, excluding all evoked markers from the Paris 1 to Paris 2 generalization actually reduced DOC-Forest performance marginally (AUC = 0.71, 95% CI: 0.618–0.807, SD = 0.049). One could, therefore, argue that, evoked markers should be considered for MVPA of DOC whenever available, alongside a few robust markers.

EEG markers of consciousness are shared between protocols and contexts

In the field of clinical neuroscience, cross-validation is commonly used to assess MVPA performance. However, it has been shown that cross-validation can give positively biased performance estimates (Saeb *et al.*, 2017; Varoquaux *et al.*, 2016; Varoquaux, 2018; Woo *et al.*, 2017). Beyond cross-validation, here, we demonstrated significant, positive generalization to independent EEG data from a different EEG protocol recorded by an independent research group (Fig. 4) and did not observe considerable deviations from cross-validation scores. Generalization from the Paris to the Liège dataset even showed marginal improvements over cross-validation. As noted previously, this could not be explained by the absence of evoked markers. Precluding the possibility of random selection bias, this may suggest that either the signal quality or the diagnostic information may have been more favourable on the Liège data. Interestingly, compared to the best markers, i.e. alpha band power and its fluctuations, the advantage of the DOC-Forest was only marginal by a few AUC points. In contrast, the other remaining univariate models (based on theta band permutation entropy and theta wSMI) did not generalize significantly. Thus, our findings demonstrate that single markers can yield reasonable stand-alone classifiers but also expose the difficulty of anticipating which marker will actually succeed. Fortunately, MVPA potentially solves this selection problem with greater success by learning predictive profiles of markers. Indeed, we observed that DOC-Forest was more robust than individual markers when using different combinations of EEG configurations for training and testing. Likewise, we observed that univariate classifiers collapsed earlier and faster than the DOC-Forest as we experimentally corrupted the training data (Fig. 5).

The significant generalization from task to resting state EEG deserves separate consideration. It is conceivable that EEG markers related to the so-called functional axis of consciousness (Sergent *et al.*, 2017), are accessible during task and resting state EEG. Accordingly, changing states of consciousness should impact markers of global house-keeping functions such as alpha band power, global long-range connectivity or signal complexity, irrespective of the context. For instance, for a patient with locked-in syndrome we observed EEG patterns similar to healthy persons during rest (Rohaut *et al.*, 2017) and here we also demonstrate the discrimination of two cognitive motor dissociation patients from UWS patients from their resting state EEG. This can be explained by that fact that we observed significant generalization from task to resting state EEG by several EEG makers, principally for alpha band power (Fig. 3B, right).

Practical implications and suggestions

How long should EEG recordings be to yield a useful feature space for machine learning?

Our results suggest that reasonable results can be achieved with a short duration EEG recording (30 s to 3 min). This

potentially broadens the scope of protocols usable in practice and encourages development of fast, time-resolved, economic screening tasks.

How many EEG sensors should be used?

When high-density nets are available, using the full configurations turns out to be beneficial for model fitting. However, results based on 16 sensors from a 10-20 montage scheme are already encouraging. As a consequence, this supports the idea that data can be successfully pooled over various EEG systems even when the number of electrodes differs.

Which EEG protocol should be used?

Both univariate and multivariate analysis suggested that EEG markers of consciousness are accessible using task and resting state data. This suggests that protocols can be liberally combined in clinical practice and encourages the development of simpler and faster screening routines as compared to a full-blown cognitive experiment encompassing hundreds of trials.

Can classification models generalize to data from other sites?

Our findings demonstrate prospective generalization to new data from younger cohorts and data from other research laboratories. The use of robust methods is particularly recommended to alleviate problem of changing marker distributions between datasets.

When should multivariate analysis be preferred to predict diagnosis?

Multivariate classification is more resilient to changes of marker distributions across datasets, be it because of noise in the signals or in the training labels, differences of populations or differences in EEG configurations and protocols. Beyond optimizing accuracy, multivariate classification models therefore yield more dependable classification performance.

How to extract biological insight from machine learning models

Here we demonstrate how the careful inspection of multivariate variable importance scores supplements the univariate analysis in qualifying interdependencies between EEG markers. While such insight may also be obtained from model coefficients of linear models, the variable importance metric as used in this study is not limited to linear relationships and does not necessitate explicit definition of non-linear effects or interaction effects.

Besides these specific points, we want to emphasize that we did not find one single globally best biomarker and that using machine learning tools to robustly combine theoretically heterogeneous markers is the recommended strategy.

Conclusion

In the current study, we demonstrate that electrophysiological markers of consciousness can be robustly exploited across contexts and protocols by relying on robust machine learning techniques. In this context, the proposed feature-extraction method based on multiple summary statistics was particularly useful as it permits one to abstract away specific sensor layouts, recording protocols and local EEG methodologies. Future work will have to demonstrate if the here-proposed 'robust tool for detecting state-of-consciousness in brain-injured patients' can be extended to a 'robust neurophysiological marker of conscious state'. It will have to be demonstrated that the proposed model can generalize to other loss of consciousness scenarios, such as sleep or anaesthesia. We wish that our findings and our publicly released strategy for classification will contribute to building large datasets that could eventually enable intensely data-driven, cross-centre approaches to treatment of severely brain-injured patients and understanding the neural-underpinnings of conscious processing.

Acknowledgements

We thank Charlene Aubinet, Olivier Bodart, Manon Carrier, Athena Demertzi, Charlotte Martial and Sarah Wannez for their contributions with to clinical evaluation of the patients. We would like to thank the UNICOG and Parietal team at Neurospin for repeated fruitful and stimulating discussion on this research project. We would also like to express our gratitude to the MNE community. The current study would not have been possible without our collaborative software development efforts. We specifically acknowledge helpful discussions and comments on this study by Alexandre Gramfort, Benjamin de Haas, Danilo Bzdok, Johan Stender, Stefania de Vito and Virginie van Wassenhove (alphabetical order). This study is dedicated to the patients and to their close relatives.

Funding

This work was supported by an ERC proof of concept grant issued to S.D., Institut National de la Santé et de la Recherche Médicale (France), the James S. McDonnell Foundation, the Institut du Cerveau et de la Moelle Épinière (France) to L.N., Consejo Nacional de Investigaciones Científicas y Técnicas (Argentina), the FRM Equipe 2015 grant to L.N., STIC-AmSud grants *Complexity as a neural marker: applications to EEG and natural language processing* and *RTBRAIN - Towards Real-time processing of brain signals*, the Belgian Funds for Scientific Research (FRS-FNRS), the European Commission, the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 720270 (Human

Brain Project SGA1) and No. 785907 (Human Brain Project SGA2), the Luminous project (EU-H2020-fetopen-ga686764), the Center-tbi, the European Space Agency, Belspo, 'Fondazione Europea di Ricerca Biomedica', the BIAL Foundation, Wallonia-Brussels Federation Concerted Research Action and the Mind Science Foundation. S.D. gratefully acknowledges additional support from CIFAR. D.E. gratefully acknowledges support by the INRIA starting researcher grant 2016, the Amazon Web Services research grant, and by the ERCYStG-263584 during 2015–16 issued to Virginie van Wassenhove. O.G. is post-doctoral fellow and S.L. is research director at FRS-FNRS.

Supplementary material

Supplementary material is available at *Brain* online.

References

Axmacher N, Henseler MM, Jensen O, Weinreich I, Elger CE, Fell J. Cross-frequency coupling supports multi-item working memory in the human hippocampus. Proc Natl Acad Sci USA 2010; 107: 3228–33.

Bayne T, Hohwy J, Owen AM. Are there levels of consciousness? Trends Cogn Sci 2016; 20: 405–13.

Bekinschtein TA, Dehaene S, Rohaut B, Tadel F, Cohen L, Naccache L. Neural signature of the conscious processing of auditory regularities. Proc Natl Acad Sci USA 2009; 106: 1672–7.

Bruno MA, Vanhaudenhuyse A, Thibaut A, Moonen G, Laureys S. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. J Neurol 2011; 258: 1373–84.

Bzdok D, Engemann D-A, Grisel O, Varoquaux G, Thirion B. Prediction and inference diverge in biomedicine: simulations and real-world data. bioRxiv 2018. doi: 10.1101/327437.

Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, et al. A theoretically based index of consciousness independent of sensory processing and behavior. Sci Transl Med 2013; 5: 198ra105.

Chang HY, Nuyten DSA, Sneddon JB, Hastie T, Tibshirani R, Sørlie T, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA 2005; 102: 3738–43.

Chennu S, Annen J, Wannez S, Thibaut A, Chatelle C, Cassol H, et al. Brain networks predict metabolism, diagnosis and prognosis at the bedside in disorders of consciousness. Brain 2017; 140: 2120–32.

Claassen J, Velazquez A, Meyers E, Witsch J, Falo MC, Park S, et al. Bedside quantitative electroencephalography improves assessment of consciousness in comatose subarachnoid hemorrhage patients. Ann Neurol 2016; 80: 541–53.

Cruse D, Chennu S, Chatelle C, Bekinschtein TA, Fernández-Espejo D, Pickard JD, et al. Bedside detection of awareness in the vegetative state: a cohort study. Lancet 2012; 378: 2088–94.

Curley WH, Forgacs PB, Voss HU, Conte MM, Schiff ND. Characterization of EEG signals revealing covert cognition in the injured brain. Brain 2018; 141: 1404–21.

Dehaene S, Changeux J-P, Naccache L, Sackur J, Sergent C. Conscious, preconscious, and subliminal processing: a testable taxonomy. Trends Cogn Sci 2006; 10: 204–11.

Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 2001; 79: 1–37.

Demertzi A, Antonopoulos G, Heine L, Voss HU, Crone JS, De Los Angeles C, et al. Intrinsic functional connectivity differentiates minimally conscious from unresponsive patients. Brain 2015; 138: 2619–31.

Demertzi A, Gomez F, Crone JS, Vanhaudenhuyse A, Tshibanda L, Noirhomme Q, et al. Multiple fMRI system-level baseline connectivity is disrupted in patients with consciousness alterations. Cortex 2014; 52: 35–46.

Donchin E, Coles MGH. Is the P300 component a manifestation of context updating. Behav Brain Sci 1988; 11: 357–427.

Efron B, Tibshirani R. An introduction to the bootstrap New York, NY: Chapman & Hall; 1993.

Emmons WH, Simon CW. EEG, consciousness, and sleep. Science 1956; 124: 1066–9.

Engemann D, Raimondo F, King J-R, Jas M, Gramfort A, Dehaene S, et al. Automated measurement and prediction of consciousness in vegetative and minimally conscious patients. In: ICML workshop on statistics, machine learning and neuroscience 2015. Lille, France; 2015.

Faugeras F, Rohaut B, Valente M, Sitt J, Demeret S, Bolgert F, et al. Survival and consciousness recovery are better in the minimally conscious state than in the vegetative state. Brain Inj 2018; 32: 72–7.

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning 2006; 63: 3. Springer/Kluwer Academic Publishers.

Giacino JT, Ashwal S, Childs N, Cranford R, Jennett B, Katz DI, et al. The minimally conscious state definition and diagnostic criteria. Neurology 2002; 58: 349–53.

Giacino JT, Kalmar K, Whyte J. The JFK Coma Recovery Scale-Revised: Measurement characteristics and diagnostic utility11No commercial party having a direct financial interest in the results of the research supporting this article has or will confer a benefit upon the authors or upon. Arch Phys Med Rehabil 2004; 85: 2020–9.

Goldfine AM, Victor JD, Conte MM, Bardin JC, Schiff ND. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. Clin Neurophysiol 2011; 122: 2157–68.

Gosseries O, Zasler ND, Laureys S. Recent advances in disorders of consciousness: focus on the diagnosis. Brain Inj 2014; 28: 1141–50.

Gramfort A, Luessi M, Larson E, Engemann D, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. Neuroimage 2014; 86: 446–60.

Iotzov I, Fidali BC, Petroni A, Conte MM, Schiff ND, Parra LC. Divergent neural responses to narrative speech in disorders of consciousness. Ann Clin Transl Neurol 2017; 4: 784–92.

Jas M, Engemann DA, Bekhti Y, Raimondo F, Gramfort A. Autoreject: automated artifact rejection for MEG and EEG data. Neuroimage 2017; 159: 417–29.

Jennett B, Plum F. Persistent vegetative state after brain damage: a syndrome in search of a name. Lancet 1972; 299: 734–7.

King J-R, Sitt JD, Faugeras F, Rohaut B, Karoui I El, Cohen L, et al. Information sharing in the brain indexes consciousness in noncommunicative patients. Curr Biol 2013a; 23: 1914–19.

King JR, Faugeras F, Gramfort A, Schurger A, El Karoui I, Sitt JD, et al. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. Neuroimage 2013b; 83: 726–38.

Laureys S, Celesia GG, Cohadon F, Lavrijsen J, José L-C, Sannita WG, et al. Unresponsive wakefulness syndrome: a new name for the vegetative state or apallic syndrome. BMC Med 2010; 8: 68.

Lo A, Chernoff H, Zheng T, Lo S-H. Why significant variables aren't automatically good predictors. Pro. Natl Acad Sci USA 2015; 112: 13892–7.

Louppe G, Wehenkel L, Sutera A, Geurts P. Understanding variable importances in forests of randomized trees. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ, editors. Advances in

neural information processing systems 26 (NIPS). Lake Tahoe: Curran Associates, Inc., 2013; p. 431–439.
Louppe G. Understanding random forests: from theory to practice. PhD thesis. University of Liège, Faculty of Applied Sciences, Department of Electrical Engineering & Computer Science, 2014.
Luauté J, Maucort-Boulch D, Tell L, Quelard F, Sarraf T, Iwaz J, et al. Long-term outcomes of chronic minimally conscious and vegetative states. Neurology 2010; 75: 246–52.
Lulé D, Noirhomme Q, Kleih SC, Chatelle C, Halder S, Demertzi A, et al. Probing command following in patients with disorders of consciousness using a brain–computer interface. Clin Neurophysiol 2013; 124: 101–6.
Monti MM, Vanhaudenhuyse A, Coleman MR, Boly M, Pickard JD, Tshibanda L, et al. Willful modulation of brain activity in disorders of consciousness. N Engl J.Med 2010; 362: 579–89.
Naccache L. Minimally conscious state or cortically mediated state? Brain 2018; 141: 949–60.
Naci L, Monti MM, Cruse D, Kübler A, Sorger B, Goebel R, et al. Brain-computer interfaces for communication with nonresponsive patients. Ann Neurol 2012; 72: 312–23.
Owen AM, Coleman MR, Boly M, Davis MH, Laureys S, Pickard JD. Detecting awareness in the vegetative state. Science 2006; 313: 1402.
Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in python. J Mach Learn Res 2011; 12: 2825–30.
Phillips CL, Bruno M-A, Maquet P, Boly M, Noirhomme Q, Schnakers C, et al. "Relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. Neuroimage 2011; 56: 797–808.
Pins D. The neural correlates of conscious vision. Cereb Cortex 2003; 13: 461–74.
Polich J. Updating P300: an integrative theory of P3a and P3b. Clin Neurophysiol 2007; 118: 2128–48.
Purdon PL, Pierce ET, Mukamel EA, Prerau MJ, Walsh JL, Wong KFK, et al. Electroencephalogram signatures of loss and recovery of consciousness from propofol. Proc Natl Acad Sci USA 2013; 110: E1142–51.
Rohaut B, Claassen J. Decision making in perceived devastating brain injury: a call to explore the impact of cognitive biases. Br J Anaesth 2018; 120: 5–9.
Rohaut B, Raimondo F, Galanaud D, Valente M, Sitt JD, Naccache L. Probing consciousness in a sensory-disconnected paralyzed patient. Brain Inj 2017; 31: 1398–403.
Rosenberg GA, Johnson SF, Brenner RP. Recovery of cognition after prolonged vegetative state. Ann Neurol 1977; 2: 167–8.
Sadaghiani S, Kleinschmidt A. Brain networks and/-oscillations: structural and functional foundations of cognitive control. Trends Cogn Sci 2016; 20: 805–17.
Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. Gigascience 2017; 6: 1–9.
Schiff ND. Recovery of consciousness after brain injury: a mesocircuit hypothesis. Trends Neurosci 2010; 33: 1–9.
Schiff ND. Cognitive motor dissociation following severe brain injuries. JAMA Neurol 2015; 72: 1413–15.
Schiff ND, Nauvel T, Victor JD. Large-scale brain dynamics in disorders of consciousness. Curr Opin Neurobiol 2014; 25: 7–14.
Schnakers C, Vanhaudenhuyse A, Giacino J, Ventura M, Boly M, Majerus S, et al. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. BMC Neurol 2009; 9: 35.
Sergent C, Baillet S, Dehaene S. Timing of the brain events underlying access to consciousness during the attentional blink. Nat Neurosci 2005; 8: 1391–400.
Sergent C, Faugeras F, Rohaut B, Perrin F, Valente M, Tallon-Baudry C, et al. Multidimensional cognitive evaluation of patients with disorders of consciousness using EEG: a proof of concept study. Neuroimage Clin 2017; 13: 455–69.
Sitt JD, King J-R, El Karoui I, Rohaut B, Faugeras F, Gramfort A, et al. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. Brain 2014; 137: 2258–70.
Stender J, Gosseries O, Bruno M-A, Charland-Verville V, Vanhaudenhuyse A, Demertzi A, et al. Diagnostic precision of PET imaging and functional MRI in disorders of consciousness: a clinical validation study. Lancet 2014; 384: 514–22.
Tononi G, Edelman GM. Consciousness and complexity. Science 1998; 282: 1846–51.
Varoquaux G. Cross-validation failure: small sample sizes lead to large error bars. Neuroimage 2018; 180: 68–77.
Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. Neuroimage 2016
Victor JD, Drover JD, Conte MM, Schiff ND. Mean-field modeling of thalamocortical dynamics and a model-driven approach to EEG analysis. Proc Natl Acad Sci USA 2011; 108 (Suppl 3): 15631–8.
Wannez S, Heine L, Thonnard M, Gosseries O, Laureys S; Coma Science Group Collaborators. The repetition of behavioral assessments in diagnosis of disorders of consciousness. Ann Neurol 2017; 81: 883–9.
Williams ST, Conte MM, Goldfine AM, Noirhomme Q, Gosseries O, Thonnard M, et al. Common resting brain dynamics indicate a possible mechanism underlying zolpidem response in severe brain injury. Elife 2013; 2: e01157.
Woo C-W, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 2017; 20: 365–77.