NeuroImage 145 (2017) 288–303

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

# "Look at my classifier's result": Disentangling unresponsive from (minimally) conscious patients

Quentin Noirhomme [a,b,c,*], Ralph Brecheisen [a,b], Damien Lesenfants [d], Georgios Antonopoulos [e], Steven Laureys [e]

[a] Brain Innovation BV, Maastricht, Netherlands
[b] Department of Cognitive Neuroscience, Faculty Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands
[c] Cyclotron Research Centre, University of Liege, Liege, Belgium
[d] School of Engineering and Institute for Brain Science, Brown University, Providence, Rhode Island, USA
[e] Coma Science Group, University Hospital of Liege, Liege, Belgium

ARTICLE INFO

*Article history:*
Accepted 4 December 2015
Available online 12 December 2015

*Keywords:*
Machine learning
Classifier
Disorders of consciousness
Coma
Vegetative state
Unresponsive wakefulness syndrome
Minimally conscious state
Locked-in syndrome
Diagnosis
Prognosis

ABSTRACT

Given the fact that clinical bedside examinations can have a high rate of misdiagnosis, machine learning techniques based on neuroimaging and electrophysiological measurements are increasingly being considered for comatose patients and patients with unresponsive wakefulness syndrome, a minimally conscious state or locked-in syndrome. Machine learning techniques have the potential to move from group-level statistical results to personalized predictions in a clinical setting. They have been applied for the purpose of (1) detecting changes in brain activation during functional tasks, equivalent to a behavioral command-following test and (2) estimating signs of consciousness by analyzing measurement data obtained from multiple subjects in resting state. In this review, we provide a comprehensive overview of the literature on both approaches and discuss the translation of present findings to clinical practice. We found that most studies struggle with the difficulty of establishing a reliable behavioral assessment and fluctuations in the patient's levels of arousal. Both these factors affect the training and validation of machine learning methods to a considerable degree. In studies involving more than 50 patients, small to moderate evidence was found for the presence of signs of consciousness or good outcome, where one study even showed strong evidence for good outcome.



## Introduction

"Look up. Look down. Squeeze my hand". These simple commands behaviorally assess the state of consciousness of a patient following a coma. To date, the diagnostic assessment of patients with disorders of consciousness (DOC) is mainly based on the observation of motor and oromotor behavior at the bedside (Giacino et al., 2014). The evaluation of non-reflex behavior, however, is not straightforward because the patient's level of vigilance may fluctuate over time. Also, he or she may suffer from cognitive deficits (e.g., aphasia or apraxia) and/or sensory impairments (e.g., blindness, deafness, paralysis). Reduced, or easily exhausted, motor activity and pain are other factors that may complicate the evaluation. In all these cases, a lack of responsiveness does not necessarily correspond to absence of awareness (Sanders et al., 2012). The identification of unambiguous signs of consciousness in patients with DOC is clinically challenging and of critical importance for establishing a diagnosis, guiding therapeutic decisions and predicting outcome. Therefore, recognizing the subtle difference between unresponsive wakefulness syndrome (UWS) patients (where patients "awaken" from a coma, meaning they open their eyes, but only show reflex behavior, formerly known as vegetative state or apallic syndrome; Laureys et al., 2010) and minimally conscious state (MCS) patients (who show non-reflex movement, e.g., visual fixation or pursuit, localization to pain or following simple commands like "look up" and "squeeze my hand"; Bruno et al., 2011; Giacino et al., 2002) requires repeated evaluations by skilled examiners. Furthermore, it is relatively easy to confuse UWS and locked-in syndrome patients (LIS; Plum and Posner, 1971) who are fully conscious but completely paralyzed except for small movements of the eyes or eyelids. Not surprisingly, up to 40% of patients with UWS are misdiagnosed (Schnakers et al., 2009a). Key elements in the diagnosis are the acquisition of voluntary responses, such as command following, and functional communication which indicates an emergence from UWS (Schnakers et al., 2009a) and MCS, respectively.

Neuroimaging and electrophysiological approaches have been proposed to complement the bedside examination. They offer motor-independent information to improve clinical differentiation and prognostic predictions. Nevertheless, while significant differences have been reported at the group level, most of these results do not allow

* Corresponding author at: Brain Innovation BV Oxfordlaan 55 6229 EV Maastricht The Netherlands.
*E-mail address:* noirhomme@brainvoyager.com (Q. Noirhomme).

http://dx.doi.org/10.1016/j.neuroimage.2015.12.006
1053-8119/

distinguishing patients at the single-subject level. Some studies, however, extend standard statistical analysis at the single-subject level with expert visual inspection (Stender et al., 2014) or prior hypotheses (for example, Owen et al., 2006; Schnakers et al., 2008; Monti et al., 2010). For example, Owen and colleagues (Owen et al., 2006) instructed a patient to alternate 30-second periods of mental imagery of playing tennis with 30-second periods of rest following a block-design protocol. A single trial consisted of 5 rest vs. imagery cycles. Then, a general linear model contrasting periods of active imagery with periods of rest was computed. Contrasts were constrained by prior hypotheses on activated brain locations; in this case, the supplementary motor area. Significant activation in the predefined brain locations is then used as indication that the patient is correctly performing the task. Calculating and thresholding a single variable or group of variables has also been proposed. For example, spectral entropy summarizes EEG signals as a single value which can distinguish patients in acute state with good accuracy (Gosseries et al., 2011).

Machine learning techniques have the potential to make more effective use of neuroimaging and electrophysiological data and allow diagnosis and prognosis at the single-subject level. Instead of considering features/activations univariately, they combine information in a multivariate way which allows them to highlight differences that might otherwise remain undetected. They are also not biased by prior hypotheses on location or time because they do not focus on the detection of a specific activation pattern but rather on a data-driven estimation of the most discriminative pattern within a trial or class. This can be an advantage given that prior hypotheses may no longer hold in pathological situations. It has been shown that data obtained from patients often exhibits higher inter-trial as well as inter-individual variability than data obtained from controls (Goldfine et al., 2011; King et al., 2013a; Lulé et al., 2013). Machine learning techniques provide a way to quantify differences in neural responses at the level of the single patient. Also, their statistical validation is limited to a single test which is independent of the number of features. This has the added advantage of also limiting the multiple comparison problem.

Until now, machine learning techniques have been applied to individual diagnosis using two main approaches: (1) detection of command-following and (2) prediction of diagnosis and outcome using structural or functional data. The first approach uses data from only a single subject measured over time and has its origin in brain-computer interface (BCI) research. The goal is to assess whether a subject is capable of following commands by measuring his/her brain activity during a functional task. For example, in the tennis paradigm mentioned above, the level of activation in each gray matter voxel could be averaged over a short time period. The average activations of all gray matter voxels could then be used as input features to train a classifier that detects the transition between rest and active imagery states. If the classification accuracy exceeds a given threshold, the subject can be considered to be able to correctly modulate his/her brain activity according to the given commands. This would be equivalent to behavioral command-following, which is a sign of MCS.

The second approach uses data obtained from multiple subjects and tries to derive a prediction model that can be used on individual subjects. For example, resting-state fMRI data might be acquired from a group of patients and healthy controls after which connectivity matrices of certain resting-state networks are calculated for all subjects. Since each group of subjects is likely to have a specific pattern of fMRI connectivity a classifier can be trained which uses the connectivity features to distinguish between the groups, for example, controls versus unresponsive patients. If classification accuracy is high enough the resulting model can then be used to classify (diagnose) new patients. Instead of resting-state networks, features can also be derived from EEG, structural MRI, diffusion tensor imaging (DTI) or positron emission tomography (PET).

In this paper, we present a survey of the literature on the use of electrophysiology and neuroimaging for diagnosing patients with disorders of consciousness (DOC). We compare machine learning techniques with studies based on univariate analysis and simple thresholding. We first provide a brief introduction of the diagnosis of DOC and list key points to take into account when machine learning techniques are used to improve the diagnosis in a clinically useful way. We will then give an overview of previous work done in the two areas mentioned earlier: (1) detection of command-following and (2) prediction of diagnosis and outcome based on multi-subject data. We will highlight the main limitations common to many studies and offer a number of suggestions for further investigation. Finally, we discuss several challenges which the field needs to overcome in order to translate machine learning techniques into clinical practice.

## Machine learning for diagnosis of disorders of consciousness

### *Current practice in diagnosing DOC*

Disorders of consciousness are currently mostly based on consensus diagnosis or using the Coma Recovery Scale-Revised (CRS-R; Giacino et al., 2004). Consensus diagnosis is based on behavioral observations of caregivers and is the most common type of diagnostic procedure in non-specialized centers. These centers would likely benefit the most from an automated diagnostic procedure given the fact that they do not usually employ DOC specialists. The rate of misdiagnosis of UWS patients by clinical consensus methods is up to 40% (Andrews et al., 1996; Childs and Mercer, 1996; Schnakers et al., 2009a). However, this error rate can be reduced by using standardized scoring systems such as the CRS-R, which is currently the most validated and sensitive method for behavioral discrimination of patients with DOC. Diagnosis remains challenging, however, because patients typically show considerable fluctuations in the level of consciousness or arousal over time. The examiner may obtain clear evidence of volitional behavior during one examination but fail to do so in another examination conducted hours or even minutes later (Giacino et al., 2014). For this reason, repeated CRS-R assessments performed by trained and experienced caregivers are essential to establish a reliable final diagnosis (Giacino et al., 2004). Repeating the assessment at least 5 times within a short period (e.g., 2 weeks) has been shown to be most accurate for establishing a diagnosis in chronic DOC patients (Wannez et al., 2016).

Despite the fact that the repeated CRS-R assessment is becoming a standard for diagnosing DOC there is still a chance that patients are incorrectly diagnosed as behaviorally unresponsive. Neuroimaging studies have shown that up to 20% of behaviorally unresponsive patients still show signs of awareness based on their brain activations (see Table 1, specificity). This is one reason why some diagnoses depend on the outcome of imaging or electrophysiological experiments. For example, functional locked-in syndrome patients show extreme behavioral motor dysfunction but still have preserved higher cognitive functions as measured by functional imaging (Bruno et al., 2011). Also, the results of repeated CRS-R assessments are reported in varying ways. Some report the patient's consciousness state only on the day of data recording while others only report the highest consciousness state measured across the multiple assessments. Ideally, results from the CRS-R assessment at the time of data recording and results of any repeated assessments should be reported together to give the clearest picture of a patient's consciousness state over time.

### *Challenges and limitations of current practice*

Diagnosing disorders of consciousness is a challenging problem for a number of reasons which we will discuss shortly. These challenges will also affect any machine learning methods applied to the data.

#### *Lack of gold standard*

Difficulty in establishing a reliable behavioral diagnosis of DOC, as mentioned briefly before, is one of the main reasons for a lack of

**Table 1**
Command-following studies.

| References | controls | LIS/conscious | MCS+ | Sensitivity (CI) | MCS- | UWS | Specificity (CI) | LR (CI) | Paradigm | Method |
|---|---|---|---|---|---|---|---|---|---|---|
| Mental Imagery - fMRI | 47/47 | 2/3 | 9/30 | 0.33 (.19–.52) | 1/20 | 7/31 | 0.84 (.71–.93) | 2.1 (1.0–4.7) | Motor imagery and spatial navigation | Statistical activation in pre-selected roi |
| MentalImagery – fMRI Stender et al. (2014) | / | 4/4 | 19/42? | 0.50 (.35–.65) | ? | 3/28 | 0.89 (.71–.97) | 4.7 (1.5–14.1) | Motor imagery and spatial navigation | Statistical activation in pre-selected roi |
| Cruse et al. (2011) | 9/12 | / | 2/15 | 0.13 (.02–.42) | 3/8 | 3/16 | 0.75 (.53–.89) | 0.5 (0.2–1.3) | right-hand imagery and toe imagery – block design | SVM classification + binomial test |
| Goldfine et al. (2013) | 3/5 | / | / | / | / | 0/16 | 1.00 (.76–1) | / | right-hand imagery and toe imagery – block design | SVM classification + permutation test |
| Henriques et al. (2014); Gabriel et al. (2015) | 6/20 | / | / | / | / | / | / | / | right-hand imagery and toe imagery – block design | SVM classification + permutation test |
| Coyle et al. (2015) | / | / | 4/4? | / | ? | / | / | / | right-hand imagery and toe imagery – block design | LDA classification + Wilcoxon signed rank test |
| Goldfine et al. (2011) | 5/5 | 1/2 | 1/3 | 0.40 (.07–.83) | / | / | / | / | Motor imagery and spatial navigation - single trial | Frequency analysis with statistical test |
| Cruse et al. (2012a); Gibson et al. (2014) | 5/6 | / | 1/1 | 1.00 | 0/1 | 1/4 | .80 (.30–.99) | / | Motor imagery - single trial | Frequency analysis with statistical test |
| Höller et al. (2013) | 20/22 | / | 0/2 | 0.00 | 0/3 | 0/9 | 1.00 (.70–1) | / | Motor imagery - single trial | SVM classification + proportional chance criteria |
| Schnakers et al. (2008 2009b) | 12/12 | 1/1 | 6/8 | 0.78 (.40–.96) | 3/6 | 0/8 | 0.79 (.49–.94) | 3.6 (1.3–10.5) | P3 - Attention to subject own name | Averaged ERP analysis |
| Local–global | 19/19 | 7/13 | 5/22 | 0.34 (.20–.52) | 2/10 | 2/24 | 0.88 (.72–.96) | 2.9 (1.0–8.2) | P3b – attention to global deviation | Averaged ERP analysis |
| King et al. (2013a) | 27/28 | 12/23 | 20/65? | 0.36 (.27–.47) | ? | 10/70 | 0.86 (.75–.93) | 2.5 (1.3–4.8) | P3b – attention to global deviation | Support vector classifier |
| Lulé et al. (2013) | 14/16 | 1/2 | 0/4 | 0.17 (.01–.64) | 1/9 | 0/3 | 0.92 (.60–1) | 2.0 (0.1–26.7) | 4-choice P3 | LDA + Chi-square test |
| Pokorny et al. (2013) | 8/10 | / | 0/12? | 0.00 (0–.30) | ? | / | / | / | tone stream segregation | stepwise LDA + binomial test |
| Chennu et al. (2013) | 8/8 | / | 2/7 | 0.29 (.05–.70) | 1/5 | 1/9 | 0.86 (.56–.97) | 2.0 (0.4–11.2) | P3b yes-no-distractors | ERP analysis |
| Naci and Owen (2013) | / | / | 1/1 | / | 1/1 | 2/2 | / | / | fMRI attention to sound | Statistical activation |
| Pan et al. (2014) | 4/4 | 1/1 | / | / | 1/3 | 1/4 | 0.71 (.30–.95) | / | Visual P3 and SSVEP | Amplitude and frequency features, and SVM |

Summary of command-following studies reviewed. The x/y ratio corresponds to the number of subjects tested positive (x) on the total number of subjects tested (y). Sensitivity scores were calculated on the behaviorally responsive patients (MCS+, LIS and conscious patients) while specificity scores were calculated on the behaviorally unresponsive (MCS- and UWS) patients. Detection of command-following in behaviorally unresponsive patients may aggregate false-positives with non-behaviorally MCS or undetected LIS patients. We calculated 95% confidence intervals according to the efficient score method (Newcombe, 1998; http://vassarstats.net/clin1.html) when at least 5 patients were reported. Likelihood ratio were calculated from the patients' results. Infinite likelihood ratio and likelihood ratio for studies of less than 15 subjects were not reported. We combined data on fMRI mental imagery from the following publications: Owen et al., 2006; Boly et al., 2007; Monti et al., 2010; Bardin et al., 2011; Fernández-Espejo and Owen, 2013; Chennu et al., 2013; Fernandez-Espejo et al., 2014; Gibson et al., 2014; Gabriel et al., 2015. Results from Stender et al., 2014 are presented separately as twenty-three of the patients have already been reported in Monti et al., 2010. We combined data on Local–Global from the following publications: Bekinschtein et al., 2009; Faugeras et al., 2011; Faugeras et al., 2012. The interrogation mark '?' means that we could not disentangle patients with MCS+ from MCS-. The patients are then all reported in the MCS+ column. LIS: locked-in syndrome patients; CS: patients having recovered consciousness after an acute brain injury; MCS+: patients with minimally conscious state demonstrating command-following behaviorally; MCS-: patients with minimally conscious state not demonstrating command-following behaviorally; UWS: unresponsive wakefulness syndrome; SVM: support vector machine classifier; LDA: linear discriminant analysis classifier; ERP: event related potential; P3: P3 potential; SSVEP: steady-state visually evoked potential. Motor Imagery – fMRI: data from.

"gold-standard", diagnostic labels. As we will see later this is also a major stumbling block when attempting to apply machine learning techniques on DOC patients because these techniques rely on known, ground-truth examples. One way to deal with this problem is to focus on prediction of patient outcome (Galanaud et al., 2013; Luyt et al., 2012) which is also important for caregivers and families because it allows optimal use of limited health care resources (Bodart and Laureys, 2014). At present, there is considerable uncertainty, however, regarding long-term outcome in terms of cognitive, behavioral and functional impairments (Galanaud et al., 2013). Also, very few objective indicators of outcome have been defined. The most common indicators are level of residual consciousness, age, etiology, time after onset and a combination of neuroimaging and electrophysiological tests (Gosseries et al., 2014; Steppacher et al., 2014).

*Patient limitations*

As stated earlier, patients with DOC often show fluctuations in arousal over time. For this reason repeated CRS-R assessments are recommended (Giacino et al., 2004). Unfortunately, most neuroimaging and electrophysiological tests are done only once which increases the risk of the patient being assessed in a suboptimal vigilance state. There is a reason for this because DOC patients are easily exhausted and have a limited memory capacity and span of attention. This means that the duration of the assessment and the complexity and cognitive workload of the task performed may adversely affect the results. Data recording sessions should, therefore, try to find a balance between maximizing the amount of information obtained with the duration of the session and the patient's fatigue and vigilance state. Also, evaluation of communication skills should be done with simple questions and answers that are known a priori. Patients with severe brain damage may be confused, disoriented and have great difficulty in giving accurate answers to trivial yes/no questions (Nakase-Richardson et al., 2009). Finally, measurements may be affected by involuntary body movements, especially in fMRI experiments, or eye movements and muscle artifacts in case of EEG experiments. Such effects cannot always be corrected and will therefore add noise to the recorded data.

*Heterogeneity of patient population*

Disorders of consciousness are often caused by different brain lesions across individuals but still show very similar (un)consciousness states. This increases the difficulty to find common diagnostic patterns. The acute (≤ one month post-injury) or chronic (> one month post-injury) state of a patient may also influence the development of the brain lesion or its impact on consciousness. Within diagnostic categories a distinction can be made based on etiology. For example, traumatic brain injury and anoxia (extreme oxygen-deprivation) are the most common causes of a coma but have different prognoses, anoxia patients often having the worst prognosis (The Multi-Society Task Force on PVS, 1994). Behaviorally a distinction can be made between MCS patients showing only non-reflex behavior (MCS-) and those capable of command-following (MCS+; Bruno et al., 2011). Altogether, the heterogeneity of the patient population results in a high variability in the measured or extracted features as well as in the diagnostic labels. This will, of course, also affect any machine learning methods applied to the data as we will discuss in the next sections.

*Machine learning overview*

Using machine learning techniques for classifying disorders of consciousness involves learning the relation between a set of *input features* and a discrete set of target labels or *classes*, e.g., healthy, minimally conscious and unresponsive wakefulness. The classifier typically learns *by example*, which is also called supervised learning. The examples can be a set of subjects whose diagnosis is known beforehand or multiple trials of imagery tasks performed by a single subject. The more examples we have to learn from, the more reliable will be the classifier's output. For each example a number of features is measured or extracted from the data. Given that in imaging or electrophysiological experiments features are commonly numeric, each example can be represented as a data point in a $N$-dimensional space where $N$ is the number of input features. Learning by example basically involves finding a plane (or manifold) in $N$-dimensional feature space that separates data points corresponding to one group of examples from those of another group. Once this plane has been found, we can classify a new subject by determining on which side of the plane it falls. Note that there is no explicit detection of consciousness, only a search for an optimal separating plane. The resulting classifier depends on the examples provided for learning and the assignment that has been defined.

*Performance estimation*

To evaluate whether the classifier performs well the set of examples is commonly split into training and test sets. Searching for the optimal separation of examples is done on the training set while performance is evaluated on an independent test set. Classification performance can be expressed in many different ways but the most common method is to use classification *accuracy*, which is the percentage of examples correctly classified in the test set. To find a reliable separation between the different groups of examples, you need as many training examples as possible. On the other hand, to reliably estimate a classifier's performance you also need many test examples. Separating the data into a test and a training set is the best way to estimate the accuracy of a classifier and overcome the problem of overfitting or 'double dipping' (Kriegeskorte et al., 2009). By the latter, we mean selecting parameters or features based on information from the whole sample before training and testing a classifier on the same data. In that case, training and test sets are no longer independent and, therefore, accuracy may be overestimated. In practice, the total number of examples is limited and a procedure called cross-validation (CV) is commonly used. It involves splitting up the data set into $K$ parts, using the $K$-th part as test set and the remaining ($K$-$1$) parts as training set. After testing, another part is selected as a test set while the remainder is used for training. This procedure is repeated $K$ times until all parts have been used as a test set once. This results in $K$ performance scores which can be averaged to obtain the final score (Lemm et al., 2011). There are several flavors of CV, for example, leave-one-out (LOO) CV where each iteration uses a only single example for testing and the rest for training. In practice, however, this may be too computationally demanding because you need as many iterations as you have examples. Another option is to split the data set into 10 parts, use 1 part for testing and 9 parts for training. This is commonly called 10-fold CV. You only need 10 iterations although it is recommended to repeat the whole procedure a number of times to obtain more reliable performance estimates (Efron and Tibshirani, 1997; Etzel et al., 2009; Lemm et al., 2011). Note that in case the experimental design introduces dependencies between trials, for example, in a block design, special care is needed to ensure that the test set remains independent (Lemm et al., 2011).

*Alternative performance measures*

To assess the performance of a classifier different measures can be used, each offering a different perspective. We already mentioned classification accuracy which expresses the percentage of correctly classified examples in the test set. Even though it remains the most widely used performance measure it has several drawbacks which make it less suitable for the evaluation of classifiers in a clinical setting. First, it does not distinguish between false positives (FP) and false negatives (FN). It simply lumps these together and represents the total number of false findings (FP + FN). In a clinical setting, however, very different costs may be associated with these two types of error. Second, it does not take class imbalance into account, for example, when there are many more healthy controls than patients. A wide variety of alternative performance measures, many based on the confusion matrix, have been proposed such as balanced accuracy, sensitivity, specificity and area

under the receiver operating characteristic curve (ROC). We could also calculate positive and negative predictive values (Altman and Bland, 1994). The positive predictive value (PPV) is the proportion of patients correctly classified as showing sign of consciousness. Similarly, the negative predictive value (NPV) is the proportion of patients correctly classified as unconscious. A drawback of these measures is that they depend on the proportion of patients from each category in the study sample and therefore on the center where the data was acquired (Deeks and Altman, 2004). All of the measures described here provide more information than simple accuracy scores but to use a classifier as a diagnostic test we also need to know the probability that it will give the correct diagnosis or outcome prediction. Likelihood ratio is such an alternative statistic. It summarizes classifier performance but does not depend on the above-mentioned class proportions. It has several other interesting properties that make it more useful for clinical applications than most other measures (Sackett et al., 2000; Deeks and Altman, 2004). In short, the likelihood ratio summarizes how many times more likely (minimally) conscious patients are classified as conscious rather than unconscious patients. Formally, it is the ratio between the probability of a positive patient being classified as positive and the probability of a negative patient being classified as positive, that is

$$LR+ = sensitivity/(1–specificity)$$

or, alternatively for detecting unconciousness

$$LR- = (1–sensitivity)/specificity$$

A likelihood ratio ($LR+$) greater than *1* indicates that a positive result of the classifier increases the probability of the patient being (minimally) conscious. The further away the likelihood ratio is from *1* the stronger the evidence. Likelihood ratios above *10* are considered to provide sufficient evidence in most clinical scenarios (Furukawa et al., 2008; Deeks and Altman, 2004). Another useful characteristic of the likelihood ratio is the fact that it represents a probability which can be combined with other probabilistic information using Bayes theorem. Note that, since sensitivity and specificity are combined to give a single likelihood ratio, there is no possibility to assign different costs to either false positives or false negatives.

*Diagnostic thresholds*

In most machine learning scenarios the performance of the classifier is only used to establish whether the resulting classification model is good enough or not. Once you establish this, the performance score is no longer relevant. In command-following paradigms, however, the performance score is actively used to establish whether or not a patient shows signs of consciousness. The performance score should therefore be converted to a binary value using some threshold. Scores above the threshold are considered to indicate the presence of signs of conscious. If the performance score falls below threshold, the classifier is not able to detect signs of command-following and therefore the patient is likely to be unresponsive. The question remains how to select the threshold.

Ideally, a threshold is determined based on average performance scores calculated across a large patient population with known diagnosis. This has not been done so far. In practice, sample sizes are small in which case it is more common to use a statistical significance test on the performance score obtained for an individual subject. An advantage of such a test is that it automatically takes into account the number of trials used to assess command-following, which can vary widely between patients. If the measured data contains no information the classifier should perform no better than chance, e.g., 50% classification accuracy. However, this assumes we have an infinite number of trials for calculating the performance. In practice, we have only a limited number of trials. In that case, it is entirely possible to obtain accuracies $> 50\%$ even though the data contains no information. A statistical significance test can help to ensure that the observed performance actually matches reality.

Instead of statistical tests it is also possible to fix the diagnostic threshold based on domain knowledge. For example, an accuracy of 70% is considered the minimum needed to effectively communicate with a Brain-Computer Interface (Kübler and Birbaumer, 2008).

*Statistical Evaluation of Classifier Performance*

After estimating a classifier's performance we should verify whether it is significantly better than chance. For this reason, a statistical evaluation is needed. Assuming that the target label is binary (e.g., healthy vs. MCS) the output of a classifier is similar to a coin toss and can be modeled as a Bernouilli trial with a probability $p_0$ of success. The probability of achieving $K$ successes out of $N$ indepenent trials is given by the Binomial distribution. This means that a binomial test is the most appropriate method for assessing the significance of any performance estimate on the independent test set (Martin and Hirschberg, 1996; Berrar et al., 2006; Chow et al., 2007; Mueller-Putz et al., 2008; Pereira et al., 2009; Billinger et al., 2013; Noirhomme et al., 2014b). For a high number of trials/subjects (when the distribution approaches a Normal distribution), the Chi-square test matches the binomial test (Howell, 2012) but for small numbers of trials ($< 20$ for a 2-class problem), the Chi-square test is not reliable (Pereira et al., 2009). For this reason, the Chi-square test should not be used for statistical evaluation of classifier performance.

A key assumption of the binomial test is the independence among test trials. If an independent test set is used, this is not a problem. However, cross-validation affects the distribution of performance scores. The variability of the classification models across the different folds and the fact that data points used for testing one model are included in the training set of another model reduces the independence between test trials and introduces a bias in the binomial test (Noirhomme et al., 2014b; Combrisson and Jerbi, 2015). An alternative for the binomial test is a permutation test (Nichols and Holmes, 2002; Maris and Oostenveld, 2007) which handles the idiosyncracies of the CV procedure better and results in an unbiased significance test (Noirhomme et al., 2014b). For this reason, the binomial test should only be used in case of a truly independent test set. If cross-validation was used, the permutation test is the preferred method of inference. Future studies should investigate the power of bootstrap testing for the assessment of classifier performance on small data sets.

*Confidence Intervals*

Together with accuracy, sensitivity, specificity or likelihood ratio it is always good practice to also report measures of effect size, such as confidence intervals (Klöppel et al., 2009; Cumming, 2014; Wolfers et al., 2015). Performance estimates based on samples of the population are subject to random variations. The resulting uncertainty is strongly related to the sample size and can be estimated using confidence intervals (CI; Gardner and Altman, 2000). CIs can be calculated using the asymptotic Normal approximation but more exact methods have been proposed that are based on an approximation of the Binomial distribution. In the following, 95% CIs are calculated according to the *efficient score* method (Newcombe, 1998) and have already been applied to classification results (Klöppel et al., 2009) using an automatic calculator (http://vassarstats.net/clin1.html).

*Challenges for machine learning in DOC*

Machine learning techniques have the potential to improve the diagnosis of disorders of consciousness especially in clinics that do not employee DOC specialists. However, there still remain considerable challenges due to the lack of a gold standard and the heterogeneity of the patient population. Here we will discuss briefly in what way these problems affect machine learning algorithms.

Supervised machine learning methods rely on known examples, that is, patients whose diagnosis is known beforehand. It is assumed the diagnostic labels are correct. If they are not, the learning algorithm is unlikely to produce a reliable classifier. Mislabeled patients will generally impair training of the classifier and subsequent interpretation of the classifier results. For this reason, it is best to exclude from the training set patients for whom the diagnostic tests do not agree. A mismatch between the classifier output and the clinical label may point to an error of the classifier, but it may also indicate the presence of supplementary information not accessible at the bedside (Sitt et al., 2014). We should, therefore, always be careful when interpreting the results of automated classification procedures. As mentioned before, an alternative can be to focus on patient outcome instead. However, it is not straightforward to define reliable indicators of outcome, let alone being able to predict them.

Note that because we lack a gold standard for the absence of consciousness, sensitivity in detecting minimally conscious states and specificity in detecting unresponsive wakefulness syndrome may never be estimated with great reliability (Cruse et al., 2014; Stender et al., 2014). Even so, we can evaluate classifier performance based on the classifier's agreement with the reported diagnosis. Inclusion of non-behavioral minimally conscious patients may have a negative impact on the classifier's performance estimate. Even if both the behavioral test and the classifier are correct, a patient in a non-behavioral minimally conscious state will count as a diagnostic error thereby making the classifier's performance look worse than it really is (Stender et al., 2014).

Another factor that may affect automated classification of DOC is the large heterogeneity in the patient population. Different brain lesions in different brain areas may result in a very similar disorders of consciousness. This introduces additional variability in the measured data and extracted features. In order to compensate for our lack of knowledge about the exact brain mechanisms involved we tend to use as many features as possible (e.g., all gray matter voxels in the brain) in the hope that we capture the relevant patterns. However, this tends to adversely affect classifier performance because most features will contain little or no information about the patient's consciousness state. Some approaches to deal with this are to focus on more specific subgroups of patients, increase patient sample sizes and improving methods for feature selection.

## Detection of command-following

In the previous sections we discussed how machine learning techniques can be used for the diagnosis of disorders of consciousness. We also looked at several challenges that must be overcome if machine learning is to be successfully applied in clinical practice. In this section we will specifically focus on detection of command-following.

Command-following paradigms involve the recording of a single subject's brain activity while performing some functional task. Using either a standard statistical analysis or machine learning we assess whether the subject is indeed modulating his/her brain activity according to the commands given. In the following we will review the literature on command-following paradigms and describe the various studies from the following perspectives:

- Measurement paradigm – Describes specific method of measuring brain activity, for example, fMRI or EEG.
- Diagnostic protocol – Defines how each subject's consciousness state was assessed.
- Dealing with patient limitations – Describes how to deal with physical and clinical constraints of patients as opposed to healthy subjects.
- Data analysis – Describes different ways in which the data was analyzed, for example, using standard statistical analysis, machine learning or a combination of both.

Sample sizes in the studies vary widely. The fMRI-based paradigms and Local-Global paradigms (see Section 3.1) used more than 100 subjects. All other published paradigms relied on much smaller samples (range 3 – 158, median 20). Also, all study samples consisted of heterogeneous patients both in terms of etiology and time of assessment after injury. In the following, we separately discuss behaviorally responsive patients (LIS and MCS+) and behaviorally unresponsive patients (MCS- and UWS). Based on the publications included in this review and other available information (Chatelle et al., 2014), sensitivity scores were calculated on the responsive patients while specificity scores were calculated on the unresponsive (UWS) patients. As discussed earlier, detection of command-following in behaviorally unresponsive patients may aggregate false-positives with non-behaviorally MCS or undetected LIS patients. We calculated 95% confidence intervals according to the efficient score method (Newcombe, 1998). All publications in this review are summarized in Table 1.

### *Measurement paradigms*

Measurement paradigms can be based on either functional MRI or electrophysiology. Paradigms use tasks based on mental imagery, or attention to auditory or visual stimuli. We will discuss these paradigms in order.

#### *FMRI mental imagery*

In their seminal paper Owen and colleagues proposed to use an imagery paradigm for a patient with clinically diagnosed UWS (Owen et al., 2006). The paradigm consisted of a motor task ("imagine you are playing tennis") and a spatial navigation task ("imagine yourself walking through your house"). Tasks were presented using a block-design and the data were analyzed using a standard general linear model. Despite the fact that this patient was diagnosed as unconscious, she showed task-related brain activation similar to that observed in a control cohort of healthy subjects. In the weeks following the experiment, the patient transitioned from UWS to MCS. The experiment was subsequently repeated and reproduced using either the same paradigm (Boly et al., 2007; Monti et al., 2010; Fernández-Espejo and Owen, 2013; Chennu et al., 2013; Fernandez-Espejo et al., 2014; Gibson et al., 2014; Stender et al., 2014; Gabriel et al., 2015) or other imagery paradigms (Bardin et al., 2011). This paradigm was also successfully used on different cohorts of healthy subjects even if, for one of the two tasks, a few subjects failed to show a statistically significant activation in the selected brain areas (Fernandez-Espejo et al., 2014; Gabriel et al., 2015). Altogether these studies involved a relatively large samples of more than 100 subjects and obtained a specificity of 84 to 89% (Table 1). However, this score should be taken with caution given that all studies interpreted a positive result in behaviorally unresponsive patients as proof of command-following. These patients may therefore be considered as non-behavioral MCS patients instead of false positives. Sensitivity was notably lower for behaviorally responsive patients with 33% to 45% (see Table 1).

#### *EEG motor imagery*

Previous work has shown that only 30 to 60% of patients can be successfully assessed using functional MRI paradigms (Stender et al., 2014; Chatelle et al., 2015). This is mainly caused by image artifacts resulting from ferrous metallic implants and involuntary movements or because the patient required sedation in order to control such movements. In either case, a reliable analysis of the measured data is no longer possible. EEG-based motor imagery paradigms do not suffer from these particular problems and have therefore been proposed as an alternative (Goldfine et al., 2011; Cruse et al., 2011, 2012b; Höller et al., 2013; Gibson et al., 2014; Coyle et al., 2015; see Table 1). EEG is compact, inexpensive and available in most clinical environments. Also, it can be easily deployed at the bedside and is not affected by metallic implants or patient motion. Except for block-design paradigms (Cruse et al., 2011), EEG motor paradigms have been tested on only a limited number of subjects. Even so, they have the ability to detect command-following

in healthy subjects (30 to 100%) with a sensitivity similar to fMRI paradigms.

*Auditory and visual attention*

EEG has also been used to assess the response to auditory stimuli such the oddball paradigm. For example, subjects may be instructed to count the number of times they hear a specific target sound, e.g., the subject's own name, among a number of auditory distractors generating an event related potential P3 (Schnakers et al., 2008, 2009b; Chennu et al., 2013; Lulé et al., 2013). Other paradigms assess the subject's ability to pay attention to global violations of temporal regularities, such as the Local-Global paradigm (Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; King et al., 2013a). This particular paradigm involves sequences of 5 auditory stimuli, for example, a sequence of 5 identical tones ("xxxxx"), called locally standard, or a sequence of 4 identical and 1 deviant tone ("xxxxY"), called locally deviant. Here, the term "local" refers to a single sequence. Alternatively, the term "global" refers to irregularities between sequences. For example, if 80% of sequences contain the pattern "xxxxY", these are the ones considered globally standard while the remaining 20%, with a pattern "xxxxx", will be the globally deviant sequences since these are the minority. Note that locally deviant sequences typically lead to a mismatch negativity (MMN; Näätänen et al., 2007). Global deviant sequences generate a late P3b response, which is related to conscious processing.

In auditory paradigms command-following is assessed (1) by detecting an increase in amplitude of the P3 component as the subject moves from an unattentive to an attentive state (Schnakers et al., 2008, 2009b; Pokorny et al., 2013) or (2) by detecting the subcomponent P3b which is linked to attention (Chennu et al., 2013; Bekinschtein et al., 2009; Faugeras et al., 2011, 2012). The auditory paradigms were all successfully tested in healthy subjects. Only the Local-Global paradigms was extensively tested on patients and resulted in 34% sensitivity and 88% specificity, similar to the fMRI mental imagery paradigm. One study used visual P3 in combination with steady-state, visually evoked potentials in a small group of healthy subjects and patients (Pan et al., 2014). Another study used fMRI to measure attention to a given sound repeated multiple times but alternated with distractors (Naci and Owen, 2013).

*Paradigms focusing on communication*

Four studies attempted communication with the patients using a fMRI paradigm (Monti et al., 2010; Bardin et al., 2011; Naci and Owen, 2013; Fernández-Espejo and Owen, 2013) and one using EEG (Lulé et al., 2013). Two of the four fMRI studies each tested a single UWS patient using yes/no questions with "yes" being associated with motor imagery and "no" with spatial navigation imagery" (Monti et al., 2010; Fernández-Espejo and Owen, 2013). One patient answered 5 out of 6 questions correctly (the last question did not get a response) while the other patient answered 12 different questions correctly across multiple sessions. Note that during some sessions no significant brain activity was recorded. In the two other fMRI studies binary communication was tested by instructing patients to perform a mental task if the answer to a question is "yes" (e.g., "is your father's name John?") and do nothing if the answer is "no" (Bardin et al., 2011; Naci and Owen, 2013). One study included two patients who had shown behavioral communication skills on previous occasions and one patient who did not. None of the three showed signs of communication during the fMRI experiment. Another patient in this study, who was previously unable to communicate, showed communication ability based on a GLM analysis but gave incorrect answers (Bardin et al., 2011). The other study included two patients who were unable to behaviorally communicate on previous occasions but succeeded to correctly answer questions during the fMRI session (Naci et al., 2013). In all fMRI studies responses were quantitatively assessed based on recorded brain activity in predefined regions of interest. One other study used a combination of EEG and machine learning to assess communication skills in 2 LIS patients, 13 MCS patients and 3 UWS patients using 10 yes/no questions within an oddball 4-choice paradigm. None of them, except one LIS patient, could communicate during the recording session (Lulé et al., 2013).

*Diagnostic protocol*

As explained earlier a subject's state of consciousness is commonly assessed using the CRS-R protocol. The majority of studies we reviewed used the repeated CRS-R method (Goldfine et al., 2011; Cruse et al., 2011, 2012a; Lulé et al., 2013; Chennu et al., 2013; Fernández-Espejo and Owen, 2013; Pan et al., 2014). One study reported the use of one CRS-R assessment just after data recording and one confirmatory assessment after 24 hours (Schnakers et al., 2008). Patients in the Local-Global paradigms (Bekinschtein et al., 2009; Faugeras et al., 2012; King et al., 2013a) as well as the fMRI motor imagery study (Bardin et al., 2011) were assessed with a single CRS-R just before data recording. More assessments may have been done but were not reported. One other fMRI motor imagery study relied on a clinical consensus diagnosis (Monti et al., 2010). In this particular study, 2 out of 4 patients initially diagnosed with UWS using consensus diagnosis later showed signs of consciousness when assessed with repeated CRS-R. Repeated CRS-R and clinical consensus diagnosis were directly compared in another study (Stender et al., 2014). The authors noted that consensus diagnosis was imprecise and failed to correctly identify 33% of patients who were diagnosed a minimally conscious using repeated CRS-R. Not a single study reported the use of repeated CRS-R assessments together with a single CRS-R at the time of data recording.

*Dealing with patient limitations*

As described earlier patients with a disorder of conscious have several physical and cognitive limitations that may affect data recording and the diagnostic assessment. In this following paragraphs we discuss how these issues are dealt with in the studies under review.

*Fluctuations in arousal levels*

Fluctuations in arousal levels are generally managed by taking longer breaks in between test blocks and using verbal stimulation to maintain the level of arousal (King et al., 2013a). In general, the goal is to keep the recording session as short as possible. This results in a small number of data sets (6 to 16 blocks of trials, or 14 to 103 trials). Also, some studies have proposed to present trials block-wise (Cruse et al., 2011, 2012a; Coyle et al., 2015) in which the subject receives instruction on the task to perform and must then repeat that task several times within a block. Using this approach, data from individual trials are more closely matched within a block than across multiple blocks. Furthermore, temporal dependence between blocks may have a confounding effect (Goldfine et al., 2013; Henriques et al., 2014) especially if the number of blocks is small. In general, the block structure of the data must be taken into account when statistically evaluating a classifier's performance (Lemm et al., 2011; Goldfine et al., 2013). Arousal is sometimes assessed by monitoring eye opening. This is commonly done in EEG recordings (Goldfine et al., 2011; Höller et al., 2013; Chennu et al., 2013) but not in fMRI sessions so patients may be unconscious (either permanently or transiently) during these scans (Monti et al., 2010).

*Fatigue and memory*

To deal with fatigue paradigms may be shortened as compared to healthy subjects (Bardin et al., 2011; Lulé et al., 2013; Pokorny et al., 2013). The number of blocks/trials being recorded also depends on the level of arousal of the patient (Cruse et al., 2011, 2012a; Lulé et al., 2013). The duration of a single block of EEG recording varied from 1 minute (Lulé et al., 2013) to 5 minutes except for one study that used a block of approx. 15 minutes (Höller et al., 2013). FMRI recording sessions varied in length from 4 minutes (Bardin et al., 2011) for a single-block recording to 10 minutes when two blocks are recorded together

(Owen et al., 2006; Monti et al., 2010; Stender et al., 2014). EEG recording sessions can reach up to 45 minutes (Schnakers et al., 2009b; Bekinschtein et al., 2009). Trials are generally shorter in duration than block recordings and varied from 4 seconds (Cruse et al., 2012a) to 5 minutes (fMRI mental imagery paradigm). Several authors recommend using short trials with the full instruction repeated before each trial because the patient may be unable to remember the task instruction for more than 30 seconds (Cruse et al., 2012a; Pan et al., 2014).

*Visual and auditory deficits*

Another study explicitly tested for visual acuity using visual evoked potentials, the reasoning being that visual deficits may prevent the patient from successfully completing the task (Pan et al., 2014). Testing for deficits in the auditory system was reported in only one study (Naci and Owen, 2013).

*Noise and movement artifacts*

All studies reported suboptimal quality in the EEG recordings due to ocular and respiratory movement artifacts. Also, in most cases respiratory and nutritional life-support systems were present at the time of recoding thereby increasing ambient noise levels. Because of such artifacts, some studies decided to exclude patients from the final results (Bardin et al., 2011; King et al., 2013a; Höller et al., 2013; Chennu et al., 2013). Other studies either removed or corrected trials affected by artifacts before they ran the statistical analysis (Schnakers et al., 2008; Goldfine et al., 2011; Höller et al., 2013; Bekinschtein et al., 2009; Faugeras et al., 2012; Chennu et al., 2013; Lulé et al., 2013). Three studies reported removing corrupted trials before training and testing the classifier (Cruse et al., 2011, 2012a; King et al., 2013a).

*Multiple recordings*

Despite the costs involved, two studies attempted multiple recording sessions to better capture the patient's consciousness state. In one study, the consistency of differences detected across sessions is used as an outcome measure (Goldfine et al., 2011). In the other study, 2 sessions were recorded for almost all subjects but analyzed separately (Pokorny et al., 2013). Other studies report multiple recordings on some of the included patients, for example, when the diagnosis changed over time.

*Data analysis*

The first studies investigating detection of command-following relied for their data analysis either on the general linear model (Owen et al., 2006; Boly et al., 2007; Monti et al., 2010; Bardin et al., 2011; Fernández-Espejo and Owen, 2013; Chennu et al., 2013; Fernandez-Espejo et al., 2014; Gibson et al., 2014; Stender et al., 2014; Gabriel et al., 2015) or event-related potentials analysis (Schnakers et al., 2008, 2009b; Bekinschtein et al., 2009; Faugeras et al., 2011, 2012; Chennu et al., 2013), both of which remain popular methods today. Inspired by research in brain-computer interfaces (BCI) several researchers have also investigated the potential of detecting command-following using machine learning techniques. Such approaches are less affected by a priori information and are able to highlight brain patterns that are only apparent when combinations of activations/features are considered. Also, they do not depend on expert availability and, furthermore, their output can be directly converted to a communication aid (Lulé et al., 2013).

In the following paragraphs we will describe how the different studies included in this review handle data analysis. We will discuss this from various perspectives. A first question that comes to mind is "how does standard statistical analysis compare to machine learning approaches?". We will describe studies that have looked at exactly this question. Also, we describe different types of features used, the classification procedure used and how a diagnostic threshold is determined for detection of command-following.

*Standard statistical analysis vs. machine learning*

A comparison between standard statistical analysis (GLM or event-related potential analysis) and machine learning has been performed by several studies. In a first study by Goldfine et al. (Goldfine et al., 2011) a statistical test was found to be more sensitive in detecting spectral changes than linear discriminant analysis (LDA). This means that a combination of frequency band changes did not improve detection compared to individual changes. However, it should be noted that classification accuracy was calculated for each channel separately and subjected to a multiple comparison correction. A combination of channels was not attempted. In a study investigating auditory P3, multivariate analysis detected significant within-subject differences while event-related potential analysis did not (Lulé et al., 2013). However, further analysis showed that event-related potential analysis did find a difference through an electrode that was not included in the original analysis. Another auditory P3 study showed opposite results where multiple significant results at different time points and locations were observed using event-related potential analysis while multivariate analysis failed to show such results. The discrepancy between these studies may be partly explained by the fact that spatial and temporal constraints had been relaxed in order to deal with the multiple comparison problem. In future studies, the use of cluster-based permutation tests may overcome such problems (Maris and Oostenveld, 2007). In a single-patient study using a motor imagery paradigm significant results were obtained in both the univariate and multivariate approaches (Cruse et al., 2012b). The most elaborate comparison between these approaches was likely done using the global auditory paradigm in an event-related potential analysis (Bekinschtein et al., 2009; Faugeras et al., 2012) and a multivariate analysis (King et al., 2013a) on the same population. Similar sensitivity scores were observed between univariate and multivariate analysis, even though within-patient comparison was not mentioned. Based on these studies we cannot draw any conclusions as to the superiority of one approach over the other. They provide different and complementary perspectives on the data and we therefore recommend using a combination of univariate statistical analysis and multivariate machine learning.

*Features*

Different types of input features can be used. Both motor imagery (Cruse et al., 2011, 2012a; Höller et al., 2013; Coyle et al., 2015) and steady-state visual evoked potential (Pan et al., 2014) paradigms use features derived from frequency power. P3-based paradigms rely on signal amplitudes (Lulé et al., 2013; King et al., 2013a; Pokorny et al., 2013). Features may also be derived from connectivity and complexity measures (Höller et al., 2013).

*Classification procedure*

In all studies, except one, performance scores such as classification accuracy were calculated offline using a cross-validation (CV) procedure. Only one study attempted online analysis of the data (Lulé et al., 2013) but this had a detrimental effect on the results. Despite the encouraging results on healthy subjects described in the BCI literature (Guger et al., 2003, 2009, 2012) a considerable reduction in performance is experienced when the methods are applied to patients. The observed sensitivity scores remain relatively low, ranging from 0 to 88% (Table 1). As explained earlier, this is likely caused by fluctuating levels of arousal, increased cognitive workload and the use of single recording session.

*Selection of diagnostic threshold*

In command-following paradigms a classifier's performance score is directly used to assess whether the patient is able to following commands or not. Selecting a performance threshold for the classifier is commonly done using either a binomial test, Chi-square test, permutation test or other tests based on a normal approximation. As explained earlier, the permutation test is recommended when the classifier has

been trained and tested using a cross-validation procedure. One study used a Wilcoxon signed-rank test to compare the top performance score across trials with the score obtained at baseline (Coyle et al., 2015). However, given the fact that parameter selection was not done independently the study results are likely to be over-optimistic. To establish a diagnostic threshold an alternative to the statistical significance test has not been proposed so far. Some researchers suggest relaxing the test (Cruse et al., 2013; Peterson et al., 2015) while others advocate adding multiple comparison correction where each patient tested increases the chance of a false positive (Goldfine et al., 2013).

## Predicting diagnosis and outcome

In the previous section we discussed paradigms for detecting command-following in DOC patients. A typical characteristic of such paradigms is that they involve only single subjects. Data is recorded over time and the classifier attempts to learn how brain states change according to the commands given. Predicting diagnosis and outcome relies on a different approach. Here, the examples needed to train the classifier come from multiple subjects. Each subject has a known target label, which can be a diagnosis (e.g., minimally conscious vs. unresponsive wakefulness) or an outcome measure (patient will recover or not). The measurement paradigm provides a set of input features for each subject and the classifier tries to learn the relation between the features and the target label. After training and testing, the classifier should be able to predict the target label of a previously unseen case, that is, a subject for which we only have input features but no label.

The multi-subject approach has several advantages over a single-subject command- following paradigm. As explained previously, command-following requires that the patient actively participates in the experiment and performs certain tasks, such as mental imagery or responding to auditory stimuli. However, if the patient suffers from hearing, linguistic, attentional or working memory deficits this will negatively affect the results of the assessment and increase the risk of false negative findings (Giacino et al., 2014). Ideally, we would like the patient to do nothing at all but still obtain useful information for determining a diagnosis or outcome. Resting-state methods offer this possibility. They do not require a sophisticated setup and do not need active participation of the subject (Demertzi et al., 2015). Note that in this review the term “resting-state” refers to any measurement method that does not require active participation of the subject.

In the following paragraphs we will provide an overview of resting-state methods for the purpose of predicting diagnosis and outcome in DOC patients. Similar to the previous section, we will describe the various studies from the following perspectives:

- Measurement paradigm
- Diagnostic protocol
- Dealing with patient limitations
- Data analysis

We calculated sensitivity scores for (1) predicting diagnosis based on the highest level of consciousness in diagnostic studies and (2) predicting favorable outcome in prognostic studies. For those studies providing enough information we also calculated 95% confidence intervals of sensitivity, specificity and likelihood ratio based on the efficient score method (Newcombe, 1998). Some of these studies calculated their own confidence intervals whose bounds were equal or within 1% of the bounds reported in Table 2.

The number of patients included in the studies under review varied widely (range 19 – 169; median 48) with 3 studies including > 100 patients (Galanaud et al., 2013; Sitt et al., 2014; Stender et al., 2014). The 95% CIs for these studies will be more reliable. Five studies specifically aimed at acute patients (<72 hours post-onset, Tzovara et al., 2013; ≤1 month post-onset, Gosseries et al., 2011; ≤45 days post-injury,

**Table 2**
Prediction of diagnosis.

| References | #controls | #patients | | | | Controls - patients | | | MCS - UWS | | | Independent test set | | | Likelihood ratio (CI) | Modality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LIS/CS | MCS | UWS | coma | Sensitivity (CI) | Specificity (CI) | Accuracy | Sensitivity (CI) | Specificity (CI) | Accuracy | Sensitivity (CI) | Specificity (CI) | Accuracy | | |
| Stender et al. (2014) | / | 4 | 76 | 36 | / | / | / | / | .93 (.85–.98) | .67 (.49–.81) | .85 | / | / | / | 2.8 (1.8–4.5) | PET |
| Phillips et al. (2011) | 37 | 0/8 | / | 13/0 | / | 1 (.88–1) | 1 (.72–1) | 1 | / | / | / | 1 (.60–1) | / | 1 | Inf | PET |
| Fernández-Espejo et al. (2011)* | 12 | / | 12 | 7 | / | / | / | / | .92 (.66–1) | 1 (.56–1) | .95 | / | / | / | Inf | DTI |
| Gosseries et al. (2011) | | | 9 | 14 | 6 | / | / | / | .89 (.51–.99) | .90 (.67–.98) | .90 | / | / | / | 8.9 (2.3–33.8) | EEG – spectral entropy acute |
| Sitt et al. (2014) | 14 | 24 | 68 | 75 | / | / | / | / | .76 (.64–.86) | .67 (.55–.77) | .71 | / | / | / | 2.3 (1.6–3.2) | EEG- combined features |
| Höller et al. (2014) | 23 | / | 20 | 24 | / | / | / | .96 | / | / | .88 | / | / | / | 7.2 (2.5–21.0) | EEG – partial coherence |
| Demertzi et al. (2014) | 27 | / | 24 | 24 | 5 | .81 (.61–.93) | .87 (.74–.94) | .85 | / | / | / | / | / | / | 6.2 (3.0–12.6) | Rs-fMRI networks |
| Demertzi et al. (2015) | 21 | / | 26/16 | 19/6 | 6/0 | / | / | / | .96 (.78–1) | .95 (.72–1) | .96 | .94 (.68–1) | .83 (.36–.99) | .91 | 5.6 (0.9–33.8) | Rs-fMRI networks |
| Casali et al. (2013) | 32 | 8 | 6 | 6 | / | / | / | / | / | / | / | 1 | 1 | 1 | / | TMS-EEG PCI |

Höller et al., 2014 LR estimated by attributing same accuracy to both classes (MCS 18/20; UWS 21/24).
Summary of diagnosis prediction studies reviewed. When two numbers are reported in a cell regarding the number of subjects included, the first number corresponds to number of subjects in the training set and the second number corresponds to the number of subjects in the independent test set. We calculated 95% confidence intervals according to the efficient score method (Newcombe, 1998; http://vassarstats.net/clin1.html). LIS: locked-in syndrome patients; CS: patients having recovered consciousness after an acute brain injury; MCS: patients with minimally conscious state UWS: unresponsive wakefulness syndrome; PET: positron emission tomography; DTI: diffusion tensor imaging; EEG: electroencephalography; rs-fMRI: resting-state functional magnetic resonance imaging; TMS: transcranial magnetic stimulation; PCI: perturbational complexity index.
* sensitivity and specificity estimated based on one missclassified MCS patient (personal interpretation of the paper).

Galanaud et al., 2013; Luyt et al., 2012; ≤53 days post-injury, Perlbarg et al., 2009) while others focus on chronic patients or a mix of chronic and acute patients. One study used spectral entropy derived from EEG to classify MCS and UWS in both acute and chronic patients (Gosseries et al., 2011) but succeeded only with acute patients. This further highlights the differences between these patient groups. Another four studies aimed at predicting outcome in a homogeneous sample of acute patients following traumatic brain injury (Perlbarg et al., 2009; Galanaud et al., 2013) and cardiac arrest (Luyt et al., 2012; Tzovara et al., 2013).

*Measurement paradigms*

As mentioned earlier, a resting-state assessment can refer to any measurement method that does not require active participation of the subject. The following methods were considered in the studies under review: Positron Emission Tomography (PET; Phillips et al., 2011; Stender et al., 2014), Diffusion Tensor Imaging (DTI; Perlbarg et al., 2009; Fernández-Espejo et al., 2011; Galanaud et al., 2013; Luyt et al., 2012), EEG (Gosseries et al., 2011; Sitt et al., 2014; Höller et al., 2014) and resting-state fMRI (Demertzi et al., 2014, 2015).

*Diagnostic paradigms*

Most resting-state studies involving patients with disorders of consciousness aim at differentiating patients at the group level. In this review we only discuss the handful of studies that have attempted classification at the single-subject level using machine learning or alternative methods. Some of these studies have focused on distinguishing between healthy controls and patients ((Phillips et al., 2011; Demertzi et al., 2014) while others have attempted to separate MCS from UWS patients (Fernández-Espejo et al., 2011; Gosseries et al., 2011; Sitt et al., 2014; Höller et al., 2014; Stender et al., 2014; Demertzi et al., 2015; Table 2). The latter case is more challenging because the patient groups are more similar to each other. Even so, classifying between patient groups is more useful for clinical settings. Note that none of the studies attempted to distinguish between MCS + and MCS- patients. Sensitivity ranged from 0.76 to 1.0 (median 0.93) and specificity from 0.67 to 1.0 (median 0.89).

*Prognostic paradigms*

Instead of predicting diagnosis, some studies used the outcome as a target for classification (Perlbarg et al., 2009; Gosseries et al., 2011; Galanaud et al., 2013; Luyt et al., 2012; Tzovara et al., 2012; Stender et al., 2014) (Table 3). Favorable outcome can be defined in different ways such as a Glasgow Coma Scale > 3 (Perlbarg et al., 2009), a Glasgow Coma Scale (Extended) > 4 (Luyt et al., 2012), a Modified Glasgow Coma Scale > 3 + (>= MCS +; Galanaud et al., 2013), recovery of functional communication (i.e., emergence from MCS; Gosseries et al., 2011) and awakening (Tzovara et al., 2013). Other studies reported the outcome of "misclassified" patients, showing that unconscious patients classified as conscious showed better outcome than unconscious patients classified as unconscious (Stender et al., 2014; Sitt et al., 2014; Demertzi et al., 2015). Sensitivity ranged from 0.50 to 1.0 (median 0.91) and specificity from 0.49 to 1.0 (median 0.82).

*Diagnostic protocol*

All studies reported the use of CRS-R assessments to establish a diagnosis. Three studies only included patients with a stable diagnosis meaning that the CRS-R assessment at the time of data recording matched the other assessments (Casali et al., 2013; Höller et al., 2014; Demertzi et al., 2015). One study complemented repeated CRS-R assessment with a PET-based diagnosis (Demertzi et al., 2015). A CRS-R assessment at the time of data recording was performed in three studies (Phillips et al., 2011; Gosseries et al., 2011; Sitt et al., 2014).

*Dealing with patient limitations*

Patient limitations, such as described in previous sections, also need to be handled in resting-state paradigms. Fluctuations in the patient's level of arousal was mentioned in one fMRI study but the authors did not explicitly manage it (Demertzi et al., 2014). One EEG study used multiple, short recording blocks with verbal stimulation before each block (i.e., Local-Global paradigm, Sitt et al., 2014). Another EEG study attempted to reduce the effect of fluctuating levels of arousal by limiting the duration of the recording sessions to 2-3 minutes (Höller et al., 2014). In the FDG-PET studies patient arousal was monitored during the 45 minutes between FDG injection and the PET scan (Phillips et al., 2011; Stender et al., 2014). In case of prolonged closure of the eyes the CRS-R arousal procedure was used. Given that EEG recordings are more sensitive to artifacts, several EEG studies reported rejection of trials and/or patients because of such artifacts (Höller et al., 2014; Sitt et al., 2014). One fMRI study used additional analysis to deal with movement artifacts and rejected patients with too much movements (Demertzi et al., 2015).

*Data analysis*

As mentioned before, we focus on studies that attempt to obtain single-subject classifications of patients with a disorder of consciousness. In the following we will give an overview of the data analysis approaches used in these studies.

*Features*

Among the studies reviewed different types of features have been reported, such as PET activation (Phillips et al., 2011; Stender et al.,

**Table 3**
Prediction of prognosis.

| References | #controls | #patients | | | | Favorable - Unfavorable | | | Independent test set | | | Likelihood ratio (CI) | Modality |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | LIS/CS | MCS | UWS | coma | Sensitivity (CI) | Specificity (CI) | Accuracy | Sensitivity (CI) | Specificity (CI) | Accuracy | | |
| Gosseries et al. (2011) | / | / | 9 | 14 | 6 | .60 | .78 | .72* | / | / | / | 2.9 (1.0–7.8)* | EEG |
| Perlbarg et al. (2009) | / | / | 30 | | | .86 | .86 | .86 | / | / | / | 6.1 | DTI |
| Galanaud et al. (2013) | / | / | 105 (acute TBI) | | | .95 | .64 | .84 | / | / | / | 2.7 (1.8–4.2)* | DTI |
| Luyt et al. (2012) | 70 | | 57 (acute CA) | | | 1 (.60–1) | .94 (.82–.98) | .97 | / | / | / | 16.3 (5.5–48.9) | DTI |
| Tzovara et al. (2013) | / | / | | | 12/18 | .58 (.29–.84) | 1 (.48–1) | .75 | .50 (.22–.78) | 1 (.52–1) | .67 | / | ERP |
| Stender et al. (2014) | / | 4 | 69 | 33 | / | .96 (.86–.99) | .49 (.35–.63) | .74 | / | / | / | 1.9 (1.4–2.5) | PET |

CA: cardiac arrest.
Summary of outcome prediction studies reviewed. Favorable and unfavorable outcome were based on each study definition. We calculated 95% confidence intervals according to the efficient score method (Newcombe, 1998; http://vassarstats.net/clin1.html). LIS: locked-in syndrome patients; CS: patients having recovered consciousness after an acute brain injury; MCS: patients with minimally conscious state UWS: unresponsive wakefulness syndrome; PET: positron emission tomography; DTI: diffusion tensor imaging; EEG: electroencephalography; TBI: traumatic brain injury; CA: cardiac arrest.
* estimation based on number of patients in each category and reported specificity and sensitivity.

2014), DTI measures, e.g., apparent diffusion coefficient, mean diffusivity and fractional anisotropy, (Perlbarg et al., 2009; Fernández-Espejo et al., 2011; Galanaud et al., 2013; Luyt et al., 2012), EEG complexity, connectivity and amplitude measures (Gosseries et al., 2011; Sitt et al., 2014; Höller et al., 2014) and connectivity derived from resting-state fMRI experiments (Demertzi et al., 2014, 2015). None of the studies attempted to combine measurement modalities although two studies mention combining different kinds of features within the EEG domain, e.g., features derived from event-related potential analysis, connectivity, frequency power and complexity (Höller et al., 2014; Sitt et al., 2014). Such a combination of different features was found to be more effective than using features of a single type, e.g., connectivity (Sitt et al., 2014).

*Classification procedures*

All studies relied on binary classification, e.g., separating between healthy controls and patients or between MCS and UWS patients. Multi-class classification was not reported in any study. In all cases, classifier performance is estimated using a cross-validation (CV) procedure. Only a few studies proceeded to also validate the classifier on an independent test set (Phillips et al., 2011; Galanaud et al., 2013; Tzovara et al., 2013; Demertzi et al., 2015) which resulted in similar estimates as those obtained from the CV procedure.

*Alternatives to machine learning*

Four studies did not use machine learning for predicting diagnosis or outcome. One study used spectral entropy derived from EEG as a single-valued summary measure and used ROC analysis to establish the optimal diagnostic threshold for separating MCS from UWS patients (Gosseries et al., 2011). Another study measured the complexity of the EEG response to a TMS pulse. The diagnostic threshold was selected based on previous studies related to sleep and anesthesia (Casali et al., 2013). Classification performance can also be used directly for prediction. For example, one study used a classifier to detect sound processing at two separate occasions, one at 4-5 hours post-onset, the other after 24-72 hours. An improvement in the classifier's performance over time was only observed in patients with good outcome (Tzovara et al.,2013). Finally, one study using PET combined a GLM analysis with visual inspection of the results within hypometabolic and preserved regions of interest (Stender et al., 2014).

## Overview and future perspectives

In the previous sections we gave a comprehensive summary of previous work involving diagnosis and outcome prediction of patients with disorders of consciousness. We separately discussed command-following paradigms, which involve only single subjects, and prediction paradigms, which require training of a classifier on multiple subjects. In the following sections we provide an integrated view of these approaches. We also propose a number of good practices that can help to extract the most from machine learning techniques and facilitate interpretation of the results. Furthermore, we suggest alternative analysis approaches that may allow for easier translation of machine learning techniques to a clinical setting.

*Propositions for good practices*

A correct assessment of machine learning methods requires reliable labels. The only widely accepted test for level of consciousness is behavioral responsiveness (Giacino et al., 2009). However, this measure has several reliability issues (Stender et al., 2014). Ideally, a combination of repeated CRS-R assessments and a CRS-R at the time of data recording should be performed. This allows comparison of the recording results with the most accurate behavioral examination and also helps dealing with fluctuations in the patient's level of arousal. Additional neuroimaging and electrophysiological findings would further clarify the diagnosis. Similarly, repeated data recordings are highly encouraged to further improve sensitivity, robustness and reproducibility of the paradigm and provide insight into the impact of arousal levels on the recording session. After all, the value of any test is determined by its ability to yield the same result after being applied multiple times to a stable patient (Furukawa et al., 2008). On the other hand, if measurement results vary significantly between sessions, we can select the best ones in the same way as the interpretation of repeated CRS-R assessments.

As explained previously, fluctuations in arousal levels remain a major factor to take into consideration. Monitoring fluctuations not only helps improving the quality of single sessions but also assists in the long-term planning of repeated CRS-R assessments and recording sessions. For example, if fluctuation patterns indicate that arousal levels are highest during the morning, the CRS-R assessments and recording sessions are best planned during that time. Training and testing of classifiers is also affected by changes in arousal levels. In structural MRI paradigms selecting the best diagnosis out of multiple CRS-R assessments is sufficient (since brain structure does not change over time). In functional paradigms, however, one should aim to include only patients with a stable diagnosis (Höller et al., 2014; Demertzi et al., 2015). Patients with an unstable diagnosis will introduce unwanted variability into the experiment. It may also be possible to include "diagnostic stability" as a feature in the classification procedure, e.g., assigning stronger weights to stable subjects than to unstable subjects. Some classifiers naturally deal with label uncertainty (Brodley and Friedl, 1999; Guan et al., 2011). Note that in acute patients outcome is often considered to be more important than diagnosis so diagnostic stability may be less of a problem there. This will be discussed in more detail in a next section.

A careful design of the recording paradigm can also help to mitigate the influence of fluctuating arousal levels. Recording blocks should be kept short with breaks in between that allow for stimulation of the patient. Ideally, the CRS-R arousal facilitation protocol is used for this. Within each block, trials should also be kept short, ideally less than 10s, and systematically preceded by clear instructions given the fact that some patients may suffer from memory deficits. Visual and auditory tests can help to determine whether the patient is able to understand the instructions as well as actually perform them (especially if visual stimuli are used). If the patient is able to proceed, simple questions with known answers should be used for the evaluation of communication skills (Nakase-Richardson et al., 2009). Finally, systematically reporting the results of patients behaviorally following commands (MCS+ and higher) would also help to better establish the sensitivity of the paradigm.

*Classification*

Classification procedures, so far, have been mainly limited to single data modalities, a single type of feature set measured at a single point in time. However, one study used EEG features of different kinds and showed that these were not entirely redundant. A combination of these feature types did lead to a better discrimination of the patient's state of consciousness (Sitt et al., 2014). Another two studies showed that the *difference* between two recordings can be a good predictor of outcome in comatose patients (Tzovara et al., 2013) and patients following severe traumatic brain injury (Lutkenhoff et al., 2015).

Classifiers that are able to combine data from multiple modalities may benefit from both structural and functional information at different temporal and spatial scales. This is important for the analysis of disorders of consciousness because they are usually not associated with specific lesions or biological processes. For this reason, it makes perfect sense to combine different types of data, each capturing a different aspect of consciousness, and analyze them jointly. Still, interpretation of such multivariate, cross-modality results with respect to the underlying pathophysiology of disorders of consciousness may not be straightforward because, in general, it is no longer possible to draw conclusions about the contribution of individual features (Haufe et al., 2014). You

can only say something about the *combined effect* of features. Either way, this will depend on the specific classifier used. Some linear classifiers do allow single feature interpretation but most classifiers, especially non-linear ones, do not.

Many classifiers require that the number of features is considerably smaller than the number of subjects ($p << n$). However, in studies of DOC the situation is often reversed, that is, $p >> n$. In this case, the feature space is very sparsely populated with data points (subjects) thereby increasing the risk of overfitting (fewer points makes it easier to fit them perfectly, which is a bad thing as explained in the machine learning overview). Kernel-based classifiers, on the other hand, have a built-in mechanism for controlling the tendency to overfit and, for this reason, are popular in DOC studies. The most common example of such a classifier is the Support Vector Machine (SVM).

In order to better represent diagnostic uncertainty we recommend to investigate probabilistic classifiers. Such classifiers use Bayes Theorem to assign probabilities to each class, e.g., a patient is MCS- with 20% probability and MCS + with 80% probability. Such an output format may be better suited for clinical settings (Phillips et al., 2011; Wolfers et al., 2015). Examples of probabilistic classifiers are Relevance Vector Machine (Tipping, 2001), Gaussian Processes (Rasmussen, 2004) and Logistic Regression (Cox, 1958).

In all studies under review none have attempted to use multi-class classifiers as opposed to binary, two-class classifiers. However, multi-class prediction can be especially important for disorders of consciousness because in most cases there are more than two possible diagnoses. Also, outcome measures are more usefully defined by multiple levels than just "good" versus "bad".

Finally, as stated earlier, most studies report only classification accuracy which provides insufficient insight into the classifier's performance for clinical purposes. We recommend including both the confusion matrix and confidence intervals in the performance report.

### *Clinical validation*

Since no diagnostic test is perfect (Mallett et al., 2012) neither can be any of the classifiers discussed in this review. In disorders of consciousness, however, it is far from trivial how to even define a misdiagnosis. For example, most studies focus on predicting signs of consciousness in patients with unresponsive wakefulness syndrome (UWS). The detection rate (i.e., 1 - specificity) varies from one study to the next with a range of 11-33% in 6 of the most highly populated studies involving command-following (King et al., 2013a; Stender et al., 2014-fMRI; fMRI-mental imagery studies) and diagnosis prediction (Sitt et al., 2014; Stender et al., 2014-PET; Demertzi et al., 2015). In these studies, whenever patients are classified as "positive", i.e., signs of conscious have been detected, this can mean one of two things: either the initial UWS diagnosis was incorrect and, for example, should have been "behaviorally unresponsive" MCS (MCS-), or the classifier made a mistake and the patient is actually a false positive. Because of the general lack of gold standards it is very difficult to establish which of the two scenarios apply.

Another challenge related to the clinical validation of classifiers in the context of DOC is the fact that false negatives and false positives are often assigned different costs. These costs are determined by various factors such as rehabilitation, end-of-life decisions, a fair distribution of medical resources and the emotional cost to the family (Peterson et al., 2015). In most studies presented here, however, the classifier's performance is optimized using classification accuracy which does not take these costs into account. One study optimized the classifier's sensitivity to good outcome at the expense of specificity (Galanaud et al., 2013). Some other studies also argue in favor of lower costs for false positives (Cruse et al., 2014; Peterson et al., 2015). So, even though in research settings we may balance sensitivity and specificity, for clinical settings the goal should be to minimize false negatives, that is, patients who are conscious but are classified as unconscious. This amounts to maximizing sensitivity (TP / (TP + FN)). This may, however, increase the number of false positives. So, if we encounter a patient with UWS who is classified as minimally conscious (positive) we should repeat the recording of the paradigm and do clinical follow-up to ensure that the patient is not a false positive.

In order to evaluate the clinical utility of the different paradigms discussed in this review, we calculated their likelihood ratios. Looking at the 10 most highly populated studies (more than 50 patients with MCS, UWS or coma), the likelihood ratios range from 1.9 to 16.3 (median 2.7). The highest likelihood ratio is obtained in a study including only acute patients following cardiac arrest (Luyt et al., 2012) and provides strong evidence of good outcome. The remaining 9 studies show small to moderate evidence for the detection of good outcome or the presence of signs of consciousness which illustrates their potential usefulness in supporting diagnosis and/or prognosis decisions. It should be noted that, although the likelihood ratio offers a simple way to combine the results of these paradigms with other behavioral and neuroimaging findings, the ratios reported here may be underestimated. As indicated before, there may be cases where the classifier is actually right while the diagnosis based on the behavioral assessment is wrong. In that case, the classifier result is unjustly called a false positive/negative thereby altering its sensitivity and specificity scores.

### *Working towards a gold standard*

In the previous sections we reviewed different ways in which previous studies deal with the lack of a gold standard in the diagnosis of disorders of consciousness. The focus seems to be most on the diagnostic labels, e.g., UWS, MCS-, MCS +, etcetera. In this section we would like to discuss an alternative standard based on outcome after one year. This measure has already been used by a few studies (Luyt et al., 2012; Galanaud et al., 2013) but we believe it deserves much more attention as a potential candidate for a gold standard. First of all, outcome measures help clinicians to make the right treatment decisions. They also provide an indication to family members as to what they can expect in terms of recovery (Randolph et al., 2008; Bodart and Laureys, 2014). After one year patients are mostly stabilized. They had the benefit of an additional year of medical testing and daily caregiving which increases the chance of detecting signs of consciousness if the prognosis is good. Several studies showed that unresponsive (UWS) patients classified as minimally conscious (positive) by a classifier often had a good prognosis. Outcome measures such as these are also important in acute patients where the diagnosis "comatose" is the same for all patients and does not help the clinician very much. For this reason, outcome after one year was used in 3 studies involving only acute patients (Luyt et al., 2012; Galanaud et al., 2013; Tzovara et al.,2013). By increasing the sample sizes of these studies a more fine-grained estimation of outcome may become possible instead of the binary measures "good" versus "bad" which are defined differently across studies.

### *Combining command-following and prediction paradigms*

Command-following and prediction paradigms have different but complementary goals. Because prediction paradigms do not require active participation by the subject, they are able to distinguish between MCS- versus UWS patients. Since command-following paradigms do require active participation they are not able to do this because both diagnoses refer to unresponsive patients. On the other hand, command-following allows distinction between MCS- and MCS + which has, so far, not been attempted using prediction paradigms. Also, emergence from MCS, which is signaled by a recovery in functional communication or the ability to manipulate objects, may be more easily detected by command-following than prediction paradigms. These factors seem to imply that the two paradigms might complement each other and could be integrated into a single paradigm. One study already attempted this by combining features from a command-following paradigms with

those of a prediction paradigm and passing them to a single classifier (Sitt et al., 2014). Ideally, this would be a *multi-class* classifier which is able to distinguish between UWS, MCS-, MCS + and fully conscious state.

From a clinical perspective, command-following paradigms have a number of advantages. They require data from only a single subject and the relationship between patient and researcher is usually much more intensive. If the results do not match the behavioral assessment there is usually more time for additional discussion, testing and reporting of the results. Because of this focus on the individual patient command-following paradigms are more easily integrated into the daily clinical care routine. On the other hand, multi-subject prediction paradigms are more expensive and time-consuming. Analysis is often performed months or even years after the data was recorded so there is little or no opportunity to be actively involved with patients and their daily care routine. The focus is mainly on publication of the results and less on a clinical application of the techniques.

*Data recording and feature extraction*

Despite the fact that all studies reported the presence of artifacts in the data, e.g., due to excessive movement in the scanner, not all of them took steps to mitigate their impact. In principle, artifacts can be dealt with in two ways. Either, the corresponding subjects/trials are excluded from the experiment or specific processing methods are applied in order to reduce or eliminate their effect. Excluding data is probably the safest approach but may significantly reduce the sample size and thereby result in a less reliable performance estimate of the classifier. Correcting for artifacts may allow one to keep most of the data but, ideally, this is done using automated procedures. Additionally, there is the option to choose classifiers that are relatively insensitive to artifacts.

In terms of feature extraction, most studies used signal amplitude, spectral power, signal complexity or connectivity as input features for classification. However, connectivity features are a somewhat special category given the fact that they have been used in both command-following and prediction paradigms and can be derived from structural as well as functional MRI and EEG measurements. Several group studies have shown that connectivity decreases along the spectrum of consciousness being highest in healthy controls and dropping in value as we progress from MCS to UWS and then comatose patients (Schiff et al., 2005; Perlbarg et al., 2009; Vanhaudenhuyse et al., 2010; Fernández-Espejo et al., 2011; Lehembre et al., 2012; Demertzi et al., 2014). Furthermore, evidence exists that connectivity and signal complexity are associated as was shown in two studies on altered consciousness combining transcranial magnetic stimulation (TMS) and EEG (Rosanova et al., 2012; Casali et al., 2013) as well as in the recently published *symbolic weighted mutual information* measure (King et al., 2013b) which, by itself, was the most discriminative among all EEG features proposed by Sitt and colleagues (Sitt et al., 2014). Connectivity and complexity were also used in two other studies (Höller et al., 2013, 2014).

*Statistical power*

In command-following paradigms statistical tests are commonly used to select the accuracy threshold. In this case, it is also recommended to run a power analysis (Cohen, 1988). Power analysis aims to answer the following question: how many trials are needed in order to detect a clinically meaningful difference between brain states? In many studies, this refers to an accuracy score that is significantly higher than chance (50%). In clinical research, a conventional choice of power is 80 to 90% (Chow et al., 2007, p. 16).

The main elements of a power calculation are (1) the significance level, also called alpha, which is the probability of rejecting the null hypothesis when it is actually true, (2) the effect size which indicates how much better the estimated accuracy score is compared to chance level, (3) the number of trials and (4) the statistical test employed. The significance level is usually set to 0.05 or 0.01. Effect sizes can be estimated from the classification accuracy obtained in healthy subjects adjusted downwards for patients. For example, a review of the visual P3 speller in BCI showed that an accuracy score of 90% is attainable in healthy subjects but may drop to 74% in patients (Marchetti and Pfiris, 2014). A similar drop in accuracy was observed between healthy subjects and LIS patients using a covert steady-state visually evoked potential (Lesenfants et al., 2014). If classification accuracy is estimated using an independent test set, the binomial test can be used to assess the significance of the results. For example, in a 2-class problem with alpha = 0.05 and an effect size of 70% (estimated accuracy), 37 trials (24 correct) would be needed to get 80% power and 53 trials (33 correct) for 90% power. Obviously, an increase in sample size also increases statistical power, that is, the probability of detecting consciousness in MCS patients. If instead of an independent test set, cross-validation was used we recommend a permutation test to assess the significance of the results. Unfortunately, there exists no formula for calculating power in permutations tests. We can either estimate the number of trials based on a binomial test or run a simulation. In case the power analysis prescribes an infeasible number of trials it is possible to use a multi-stage design (Chow et al., 2007) where initially a reduced number of trials is recorded to determine if the study holds sufficient promise to warrant a more time-consuming experiment. Using the same parameters as before ($\alpha = .05$, accuracy = 70%, power = 80%), we might perform a first session with 23 trials. If 12 of these are correct we can then run a second session to acquire the remaining trials (Chow et al., 2007). The above-mentioned discussed refers specifically to command-following paradigms. However, a power analysis could also help to determine the minimum number of subjects in a prediction paradigm, even though we are not aware of existing studies doing that.

*Sample size*

Most studies under review used relatively small samples acquired in a single imaging center. For this reason, the reported accuracy scores may be somewhat optimistic. Furthermore, confidence intervals are rarely reported. This makes it difficult to draw more general conclusions about the results except for a few studies that also reported accuracy scores on additional data sets. Clearly, a more elaborate validation using large, carefully acquired samples with detailed diagnostic information obtained using multiple paradigms is necessary. Those studies that did use large patient samples tend also to provide more detailed information about the classifier's performance and validity of its output. Using large patient samples also allows you to capture a broader range of patients instead of focusing only on a specific group of patients in specialized centers. Furthermore, considerable progress can be made in acute DOC patients that often have very similar short-term diagnoses (comatose) but widely differing long-term outcome.

*Alternative approaches*

Besides using machine learning for computer-aided diagnosis it has also found its use in detecting the presence of event-related potential components (Blankertzet al., 2011; Tzovara et al., 2013; King et al., 2013a; Sitt et al., 2014). As stated earlier, machine learning techniques are not biased by a priori hypotheses regarding electrode locations or latency of the components. Compared to the traditional techniques in this domain they are also less affected by transient, artifact-contaminated activity recorded at certain electrodes. Furthermore, they provide a way to quantify differences in neural responses at the level of the single patient (Tzovara et al., 2013). Analysis of event-related potential fluctuations across trials may also provide a highly sensitive measure of consciousness state (Sitt et al., 2014) similar to the variability in functional connectivity measured with fMRI (Barttfeld et al., 2015; Demertzi et al., 2015). Automatic analysis of wakefulness (Noirhomme et al., 2014a)

and sleep patterns (Landsness et al., 2011; Malinowska et al., 2013) in DOC patients may also result in use features for prediction (Forgacs et al., 2014). New paradigms using vibro-tactile stimulation (Lugo et al., 2014), simplified visual P3 (Hoffmann et al., 2008) and covert steady-state visual potentials (Lesenfants et al., 2014) have been proposed in the literature but not yet tested on DOC patients. Alternatively, non-brain-based approaches, such as measurement of subclinical electromyography signals (Bekinschtein et al., 2008; Habbal et al., 2014), pupil dilation during mental calculation (Stoll et al., 2013), changes in salivary pH (Wilhelm et al., 2006; Ruf et al., 2013) or changes in respiration patterns (Charland-Verville et al., 2014) can be also be used to identify covert signs of command-following in patients with disorders of consciousness.

## Conclusion

Despite promising results, the use of machine learning techniques for assisting diagnosis and prognosis of disorders of consciousness is still in its infancy. In the future, efforts need to be made to (1) improve diagnostic labeling of the training set, (2) better monitor arousal levels in patients and (3) increase efficiency of the paradigms. Also, it is recommended to investigate the potential of robust, probabilistic, multi-class classification algorithms. This will, overall, improve the clinical utility of machine learning techniques for single-patient diagnosis and prognosis. Then, international research consortia, which are able to bring together both clinical and methodological expertise as well as collect large quantities of well-documented patient data sets, could provide accurate estimations of effect sizes with regard to the diagnostic and prognostic capabilities of the different approaches.

## Acknowledgments

The authors thank the anonymous reviewers for their helpful comments, which improved the quality of this paper. The research leading to these results has received funding from the European Community's Seventh Framework Program under grant agreement n° 602450 (IMAGEMEND), the Belgian National Funds for Scientic Research (FNRS) and James McDonnell Foundation. SL is FNRS research director. This paper reflects only the authors' view and the funding sources are not liable for any use that may be made of the information contained therein.

## References

Altman, D.G., Bland, J.M., 1994. Statistics Notes: Diagnostic tests 2: predictive values. BMJ 309 (6947), 102.

Andrews, K., Murphy, L., Munday, R., Littlewood, C., 1996. Misdiagnosis of the vegetative state: retrospective study in a rehabilitation unit. BMJ 313, 13–16.

Bardin, J.C., Fins, J.J., Katz, D.I., Hersh, J., Heier, L.A., Tabelow, K., Dyke, J.P., Ballon, D.J., Schiff, N.D., Voss, H.U., 2011. Dissociations between behavioural and functional magnetic resonance imaging-based evaluations of cognitive function after brain injury. Brain 134 (3), 769–782.

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., Dehaene, S., 2015. Signature of consciousness in the dynamics of resting-state brain activity. Proceedings of the National Academy of Sciences 112 (3), 887–892.

Bekinschtein, T.A., Coleman, M.R., Niklison, J., Pickard, J.D., Manes, F.F., 2008. Can electromyography objectively detect voluntary movement in disorders of consciousness? J. Neurol. Neurosurg. Psychiatry 79 (7), 826–-828.

Bekinschtein, T.A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., Naccache, L., 2009. Neural signature of the conscious processing of auditory regularities. Proceedings of the National Academy of Sciences 106 (5), 1672–1677.

Berrar, D., Bradbury, I., Dubitzky, W., 2006. Avoiding model selection bias in small-sample genomic datasets. Bioinformatics 22 (10), 1245–1250.

Billinger, M., Daly, I., Kaiser, V., Jin, J., Allison, B.Z., Müller-Putz, G.R., Brunner, C., 2013. Is it significant? Guidelines for reporting BCI performance. In Towards Practical Brain-Computer Interfaces. Springer, Berlin Heidelberg, pp. 333–-354.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2011. Single-trial analysis and classification of ERP components—a tutorial. NeuroImage 56 (2), 814–825.

Bodart, O., Laureys, S., 2014. Predicting outcome from subacute unresponsive wakefulness syndrome or vegetative state. Crit. Care 18 (2), 132. http://dx.doi.org/10.1186/cc13831 (Apr 15).

Boly, M., Coleman, M.R., Davis, M.H., Hampshire, A.A., Bor, D., MMoonen, G., Maquet, P.A., Pickard, J.D., LLaureys, S., Owen, A.A.M., 2007. When thoughts become action: an fMRI paradigm to study volitional brain activity in non-communicative brain injured patients. Neuroimage 36 (3), 979–992.

Brodley, C.E., Friedl, M.A., 1999. Identifying mislabeled training data. Journal of Artificial Intelligence Research 11, 131–167.

Bruno, M.-A., Vanhaudenhuyse, A., Thibaut, A., Moonen, G., Laureys, S., 2011. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. J. Neurol. 258 (7), 1373–1384.

Casali, A.G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K.R., Casarotto, S., et al., 2013. A theoretically based index of consciousness independent of sensory processing and behavior. Sci. Transl. Med. 5 (198) (198ra105-198ra105).

Childs, N.L., Mercer, W.N., 1996. Misdiagnosing the persistent vegetative state. Misdiagnosis certainly occurs. BMJ 313, 944.

Charland-Verville, V., Lesenfants, D., Sela, L., Noirhomme, Q., Ziegler, E., Chatelle, C., Plotkin, A., Sobel, N., Laureys, S., 2014. Detection of response to command using voluntary control of breathing in disorders of consciousness. Front. Hum. Neurosci. 8.

Chatelle, C., Laureys, S., Noirhomme, Q., 2014. Brain-Computer Interfaces and Diagnosis. Brain-Computer-Interfaces in their ethical, social and cultural contexts. Springer, Netherlands, pp. 39–-47.

Chatelle, C., Lesenfants, D., Guller, Y., Laureys, S., Noirhomme, Q., 2015. Brain-Computer Interface for Assessing Consciousness in Severely Brain-Injured Patients. Clinical Neurophysiology in Disorders of Consciousness. Springer, Vienna, pp. 133–-148.

Chennu, S., Finoia, P., Kamau, E., Monti, M.M., Allanson, J., Pickard, J.D., Owen, A.M., Bekinschtein, T.A., 2013. Dissociable endogenous and exogenous attention in disorders of consciousness. NeuroImage 3, 450–461.

Chow, S.-C., 2007. Hansheng Wang, and Jun Shao. CRC Press, Sample size calculations in clinical research.

Cohen, J., 1988. Statistical power for the social sciences. Laurence Erlbaum and Associates, Hillsdale, NJ.

Combrisson, E., Jerbi, K., 2015. Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J. Neurosci. Methods.

Cox, D.R., 1958. The Regression Analysis of Binary Sequences (with discussion). Journal of the Royal Statistical Society 20, 215–242.

Coyle, D., Stow, J., McCreadie, K., McElligott, J., Carroll, Á., 2015. Sensorimotor Modulation Assessment and Brain-Computer Interface Training in Disorders of Consciousness. Arch. Phys. Med. Rehabil. 96 (3), S62–S70.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T.A., Fernández-Espejo, D., Pickard, J.D., Laureys, S., Owen, A.M., 2011. Bedside detection of awareness in the vegetative state: a cohort study. The Lancet 378 (9809), 2088–2094.

Cruse, D., Chennu, S., Chatelle, C., Fernández-Espejo, D., Bekinschtein, T.A., Pickard, J.D., Laureys, S., Owen, A.M., 2012a. Relationship between etiology and covert cognition in the minimally conscious state. Neurology 78 (11), 816–822.

Cruse, D., Chennu, S., Fernández-Espejo, D., Payne, W.L., Young, G.B., Owen, A.M., 2012b. Detecting awareness in the vegetative state: electroencephalographic evidence for attempted movements to command. p. e49933.

Cruse, D., Chennu, S., Chatelle, C., Bekinschtein, T.A., Fernández-Espejo, D., Pickard, J.D., Laureys, S., Owen, A.M., 2013. Reanalysis of "bedside detection of awareness in the vegetative state: a cohort study"–authors' reply. Lancet 381 (9863), 291–292.

Cruse, D., Gantner, I., Soddu, A., Owen, A.M., 2014. Lies, damned lies and diagnoses: estimating the clinical utility of assessments of covert awareness in the vegetative state. Brain Inj. 28 (9), 1197–1201.

Cumming, 2014. The new statistics: why and how. Psychol. Sci. 25 (1), 7–29 (2014).

Deeks, J.J., Altman, D.G., 2004. Diagnostic tests 4: likelihood ratios. BMJ 329 (7458), 168–169.

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

Demertzi, A., Antonopoulos, G., Heine, L., Voss, H.U., Crone, J.S., de Los Angeles, C., Bahri, M.A., et al., 2015. Intrinsic functional connectivity differentiates minimally conscious from unresponsive patients. Brain, awv169.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560.

Etzel, J.A., Gazzola, V., Keysers, C., 2009. An introduction to anatomical ROI-based fMRI classification analysis. Brain Res. 1282, 114–125.

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T.A., Galanaud, D., Puybasset, L., Bolgert, F., et al., 2011. Probing consciousness with event-related potentials in the vegetative state. Neurology 77 (3), 264–268.

Faugeras, F., Rohaut, B., Weiss, N., Bekinschtein, T., Galanaud, D., Puybasset, L.L., Bolgert, F., et al., 2012. Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. Neuropsychologia 50 (3), 403–418.

Fernández-Espejo, D., Bekinschtein, T., Monti, M.M., Pickard, J.D., Junque, C., Coleman, M.R., Owen, A.M., 2011. Diffusion weighted imaging distinguishes the vegetative state from the minimally conscious state. NeuroImage 54 (1), 103–112.

Fernández-Espejo, D., Owen, A.M., 2013. Detecting awareness after severe brain injury. Nat. Rev. Neurosci. 14 (11), 801–809.

Fernandez-Espejo, D., Norton, L., Owen, A.A.M., 2014. The clinical utility of fMRI for identifying covert awareness in the vegetative state: a comparison of sensitivity between 3 T and 1.5 T. PLoS One 9 (4), e95082.

Forgacs, P.B., Conte, M.M., Fridman, E.A., Voss, H.U., Victor, J.D., Schiff, N.D., 2014. Preservation of electroencephalographic organization in patients with impaired consciousness and imaging-based evidence of command-following. Ann. Neurol. 76 (6), 869–879.

Furukawa, T.A., Strauss, S., Bucher, H.C., Guyatt, G., 2008. Diagnostic tests. In: Guyatt, G., Rennie, D. (Eds.), Users' guides to the medical literature, second ed. AMA Press, Chicago, pp. 419–438.
Gabriel, D., Henriques, J., Comte, A., Lyudmila, G., Ortega, J.-P., Cretin, E., Brunotte, G., et al., 2015. Substitute or complement? Defining the relative place of EEG and fMRI in the detection of voluntary brain reactions. Neuroscience 290, 435–444.
Galanaud, D., Perlbarg, V., Gupta, R., Stevens, R.D., Sanchez, P., Tollard, E.E., De Champfleur, N.M.M., et al., 2013. Assessment of white matter injury and outcome in severe brain trauma. A prospective multicenter cohort. Surv. Anesthesiol. 57 (4), 171–172.
Gardner, M.J., Altman, D.G., 2000. Estimating with confidence. In: Altman Douglas, G., David, M., Bryant Trevor, N., Gardner Martin, J. (Eds.), Statistics with confidence. Confidence intervals and statistical guidelines, second ed. British Medical Journal, London.
Giacino, J.T., Ashwal, S., Childs, N., Cranford, R., Jennett, B., Katz, D.I., Kelly, J.P., et al., 2002. the minimally conscious state definition and diagnostic criteria. Neurology 58 (3), 349–353.
Giacino, J.T., Kalmar, K., Whyte, J., 2004. The JFK coma recovery scale-revised: measurement characteristics and diagnostic utility. Arch. Phys. Med. Rehabil. 85 (12), 2020–2029.
Giacino, J.T., Fins, J.J., Laureys, S., Schiff, N.D., 2014. Disorders of consciousness after acquired brain injury: the state of the science. Nat. Rev. Neurol. 10 (2), 99–114.
Giacino, J.T., Schnakers, C., Rodriguez-Moreno, D., Kalmar, K., Schiff, N., Hirsch, J., 2009. Behavioral assessment in patients with disorders of consciousness: gold standard or fool's gold? Prog. Brain Res. 177, 33–48.
Gibson, R.M., Fernández-Espejo, D., Gonzalez-Lara, L.E., Kwan, B.Y., Lee, D.H., Owen, A.M., Cruse, D., 2014. Multiple tasks and neuroimaging modalities increase the likelihood of detecting covert awareness in patients with disorders of consciousness. Front. Hum. Neurosci. 8, 950. http://dx.doi.org/10.3389/fnhum.2014.00950.
Goldfine, A.M., Victor, J.D., Conte, M.M., Bardin, J.C., Schiff, N.D., 2011. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. Clin. Neurophysiol. 122 (11), 2157–2168.
Goldfine, A.M., Bardin, J.C., Noirhomme, Q., Fins, J.J., Schiff, N.D., Victor, J.D., 2013. Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study. Lancet 381 (9863), 289.
Gosseries, O., Schnakers, C., Ledoux, D., Vanhaudenhuyse, A., Bruno, M.M.-A.A., Demertzi, A.A., Noirhomme, Q., et al., 2011. automated EEG entropy measurements in coma, vegetative state/unresponsive wakefulness syndrome and minimally conscious state. Funct. Neurol.
Gosseries, O., Zasler, N.D., Laureys, S., 2014. Recent advances in disorders of consciousness: focus on the diagnosis. Brain Inj. 28 (9), 1141–1150. http://dx.doi.org/10.3109/02699052.2014.920522.
Guan, D., Yuan, W., Lee, Y.-K., Lee, S., 2011. Identifying mislabeled training data with the aid of unlabeled data. Appl. Intell. 35 (3), 345–358.
Guger, C., Edlinger, G., Harkam, W., Niedermayer, I., Pfurtscheller, G., 2003. How many people are able to operate an EEG-based brain-computer interface (BCI)? IEEE Trans. Neural Syst. Rehabil. Eng. 11 (2), 145–147.
Guger, C., Daban, S., Sellers, E., Holzner, C., Krausz, G., Carabalona, R., Gramatica, F., Edlinger, G., 2009. How many people are able to control a P300-based brain–computer interface (BCI)? Neurosci. Lett. 462 (1), 94–98.
Guger, C., Allison, B.Z., Großwindhager, B., Prückl, R., Hintermüller, C., Kapeller, C., Bruckner, M., Krausz, G., Edlinger, G., 2012. How many people could use an SSVEP BCI? Front. Neurosci. 6.
Habbal, D., Gosseries, O., Noirhomme, Q., Renaux, J., Lesenfants, D., Bekinschtein, T.A., Majerus, S., Laureys, S., Schnakers, C., 2014. Volitional electromyographic responses in disorders of consciousness. Brain Inj. 28 (9), 1171–1179.
Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110.
Henriques, J., Gabriel, D., Grigoryeva, L., Haffen, E., Moulin, T., Aubry, R., Pazart, L., Ortega, J.-P., 2014. Protocol design challenges in the detection of awareness in aware subjects using EEG signals. Clin. EEG Neurosci. (1550059414560397).
Hoffmann, U., Vesin, J.-M., Ebrahimi, T., Diserens, K., 2008. An efficient P300-based brain–computer interface for disabled subjects. J. Neurosci. Methods 167 (1), 115–125.
Höller, Yvonne, Jürgen Bergmann, Aljoscha Thomschewski, Martin Kronbichler, Peter Höller, Julia S. Crone, Elisabeth V. Schmid, Kevin Butz, Raffaele Nardone, Eugen Trinka. "Comparison of EEG-features and classification methods for motor imagery in patients with disorders of consciousness." (2013): e80479.
Höller, Y., Thomschewski, A., Bergmann, J., Kronbichler, M., Crone, J.S., Schmid, E.V., Butz, K., Höller, P., Nardone, R., Trinka, E., 2014. Connectivity biomarkers can differentiate patients with different levels of consciousness. Clin. Neurophysiol. 125 (8), 1545–1555.
Howell, D.C., 2012. Statistical Methods for Psychology. Wadsworth.
King, J.R., Faugeras, F., Gramfort, A., Schurger, A., El Karoui, I., Sitt, J.D., ... Dehaene, S., 2013a. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. NeuroImage 83, 726–738.
King, J.-R., Sitt, J.D., Faugeras, F.F., Rohaut, B., El Karoui, I.I., Cohen, L., Naccache, L., Dehaene, S.S., 2013b. information sharing in the brain indexes consciousness in noncommunicative patients. Curr. Biol. 23 (19), 1914–1919.
Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Ashburner, J., Frackowiak, R.S.J., 2009. A plea for confidence intervals and consideration of generalizability in diagnostic studies. Brain 132 (4), e102.
Kriegeskorte, N., Kyle Simmons, W., Bellgowan, P.S.F., Baker, C.I., 2009. circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.
Kübler, A., Birbaumer, N., 2008. Brain–computer interfaces and communication in paralysis: extinction of goal directed thinking in completely paralysed patients? Clin. Neurophysiol. 119 (11), 2658–2666.
Landsness, E., Bruno, M.-A., Noirhomme, Q., Riedner, B., Gosseries, O., Schnakers, C., Massimini, M., Laureys, S., Tononi, G., Boly, M., 2011. Electrophysiological correlates of behavioural changes in vigilance in vegetative state and minimally conscious state. Brain 134 (8), 2222–2232.
Laureys, S., Celesia, G.G., Cohadon, F., Lavrijsen, J., León-Carrión, J., Sannita, W.G., Sazbon, L., Schmutzhard, E., von Wild, K.R., Zeman, A., Dolce, G., 2010. European Task Force on Disorders of Consciousness, Unresponsive wakefulness syndrome: a new name for the vegetative state or apallic syndrome. BMC Med. 8, 68. http://dx.doi.org/10.1186/1741-7015-8-68 (Nov 1).
Lehembre, R., Bruno, M.-A., Vanhaudenhuyse, A., Chatelle, C., Cologan, V., Leclercq, Y., Soddu, A., Macq, B., Laureys, S., Noirhomme, Q., 2012. Resting-state EEG study of comatose patients: a connectivity and frequency analysis to find differences between vegetative and minimally conscious states. Funct. Neurol. 27 (1).
Lemm, S., Blankertz, B., Dickhaus, T., Müller, K.-R., 2011. Introduction to machine learning for brain imaging. NeuroImage 56 (2), 387–399.
Lesenfants, D., Habbal, D., Lugo, Z., Lebeau, M., Horki, P., Amico, E., Pokorny, C., et al., 2014. an independent SSVEP-based brain–computer interface in locked-in syndrome. J. Neural Eng. 11 (3), 035002.
Lugo, Z.R., Rodriguez, J., Lechner, A., Ortner, R., Gantner, I.S., Laureys, S., Noirhomme, Q., Guger, C., 2014. A vibrotactile p300-based brain–computer interface for consciousness detection and communication. Clin. EEG Neurosci. (1550059413505533).
Lulé, D., Noirhomme, Q., Kleih, S.C., Chatelle, C., Halder, S., Demertzi, A., Bruno, M.M.-A.A., et al., 2013. probing command following in patients with disorders of consciousness using a brain–computer interface. Clin. Neurophysiol. 124 (1), 101–106.
Lutkenhoff, E.S., Chiang, J., Tshibanda, L., Kamau, E., Kirsch, M., Pickard, J.D., Laureys, S., Owen, A.M., Monti, M.M., 2015. Thalamic and extrathalamic mechanisms of consciousness after severe brain injury. Ann. Neurol. 78 (1), 68–76. http://dx.doi.org/10.1002/ana.24423 (Jul).
Luyt, C.-E., Galanaud, D., Perlbarg, V., Vanhaudenhuyse, A., Stevens, R.D., Gupta, R., Besancenot, H., et al., 2012. Diffusion Tensor Imaging to Predict Long-term Outcome after Cardiac ArrestA Bicentric Pilot Study. Anesthesiology 117 (6), 1311–1321.
Malinowska, U., Chatelle, C., Bruno, M.-A., Noirhomme, Q., Laureys, S., Durka, P.J., 2013. Electroencephalographic profiles for differentiation of disorders of consciousness. Biomed. Eng. Online 12 (1), 109.
Mallett, S., Halligan, S., Thompson, M., Collins, G.S., Altman, D.G., 2012. Interpreting diagnostic accuracy studies for patient care. BMJ 345, e3999.
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG-and MEG-data. J. Neurosci. Methods 164 (1), 177–190.
Martin, J.K., Hirschberg, D.S., 1996. Small sample statistics for classification error rates II: Confidence intervals and significance tests. Department of Information and Computer Science, University of California, Irvine, CA.
Monti, M.M., Vanhaudenhuyse, A., Coleman, M.R., Boly, M., Pickard, J.D., Tshibanda, L., Owen, A.M., Laureys, S., 2010. Willful modulation of brain activity in disorders of consciousness. N. Engl. J. Med. 362 (7), 579–589.
Mueller-Putz, G., Scherer, R., Brunner, C., Leeb, R., Pfurtscheller, G., 2008. Better than random: A closer look on BCI results. Int. J. Bioelectromagn. 10, 52–55 no. EPFL-ARTICLE-164768.
Näätänen, R., Paavilainen, P., Rinne, T., Alho, K., 2007. The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin. Neurophysiol. 118 (12), 2544–2590.
Naci, L., Owen, A.M., 2013a. Making Every Word Count for Nonresponsive Patients. JAMA Neurol. 70 (10), 1235–1241. http://dx.doi.org/10.1001/jamaneurol.2013.3686.
Naci, L., Cusack, R., Jia, V.Z., Owen, A.M., 2013b. The brain's silent messenger: using selective attention to decode human thought for brain-based communication. J. Neurosci. 33 (22), 9385–9393.
Nakase-Richardson, R., Yablon, S.A., Sherer, M.M., Nick, T.G., Evans, C.C., 2009. emergence from minimally conscious state insights from evaluation of posttraumatic confusion. Neurology 73 (14), 1120–1126.
Newcombe, R.G., 1998. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat. Med. 17 (8), 857–872.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.
Noirhomme, Q., Lehembre, R., Lugo, Z., Lesenfants, D., Luxen, A., Laureys, S., Oddo, M., Rossetti, A.O., 2014a. Automated analysis of background EEG and reactivity during therapeutic hypothermia in comatose patients after cardiac arrest. Clin. EEG Neurosci. 45 (1), 6–13. http://dx.doi.org/10.1177/1550059413509616 (Jan).
Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., Laureys, S., 2014b. Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. NeuroImage 4, 687–694.
Owen, A.M., Coleman, M.R., Boly, M., Davis, M.H., Laureys, S., Pickard, J.D., 2006. Detecting awareness in the vegetative state. Science 313 (5792), 1402.
Pan, J., Xie, Q., He, Y., Wang, F., Di, H., Laureys, S., Yu, R., Li, Y., 2014. Detecting awareness in patients with disorders of consciousness using a hybrid brain–computer interface. J. Neural Eng. 11 (5), 056007.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45 (1), S199–S209.
Perlbarg, V., Puybasset, L., Tollard, E., Lehericy, S., Benali, H., Galanaud, D., 2009. Relation between brain lesion location and clinical outcome in patients with severe traumatic brain injury: a diffusion tensor imaging study using voxel-based approaches. Hum. Brain Mapp. 30 (12), 3924–3933.
Peterson, A., Cruse, D., Naci, L., Weijer, C., Owen, A.M., 2015. Risk, diagnostic error, and the clinical science of consciousness. NeuroImage 7, 588–597.

Phillips, C.L., Bruno, M.-A.A., Maquet, P., Boly, M.M., Noirhomme, Q., Schnakers, C., Vanhaudenhuyse, A., et al., 2011. relevance vector machine" consciousness classifier applied to cerebral metabolism of vegetative and locked-in patients. NeuroImage 56 (2), 797–808.

Plum, F., Posner, J.B., 1971. The diagnosis of stupor and coma. Contemp. Neurol. Ser. 10, 1–286.

Pokorny, C., Klobassa, D.S., Pichler, G., Erlbeck, H., Real, R.G.L., Kübler, A., Lesenfants, D., et al., 2013. the auditory P300-based single-switch brain–computer interface: paradigm transition from healthy subjects to minimally conscious patients. Artif. Intell. Med. 59 (2), 81–90.

Rasmussen, C.E., 2004. Gaussian processes in machine learning". Advanced Lectures on Machine Learning. Lect. Notes Comput. Sci 3176, 63–71.

Rosanova, M., Gosseries, O., Casarotto, S., Boly, M., Casali, A.G., Bruno, M.-A., Mariotti, M., et al., 2012. Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. Brain, awr340.

Ruf, C.A., De Massari, D., Wagner-Podmaniczky, F., Matuz, T., Birbaumer, N., 2013. semantic conditioning of salivary pH for communication. Artif. Intell. Med. 59 (2), 91–98.

Sackett, D.L., Straus, S., Richardson, W.S., Rosenberg, W., Haynes, R.B., 2000. Evidence based medicine. How to practise and teach EBM, second ed. Churchill Livingstone, Edinburgh, pp. 67–93.

Sanders, R.D., Tononi, G., Laureys, S., Sleigh, J., 2012. Unresponsiveness ≠ unconsciousness. Anesthesiology 116 (4), 946.

Schiff, N.D., Rodriguez-Moreno, D., Kamal, A.A., Kim, K.H.S., Giacino, J.J.T., Plum, F., Hirsch, J.J., 2005. fMRI reveals large-scale network activation in minimally conscious patients. Neurology 64 (3), 514–523.

Schnakers, C., Perrin, F., Schabus, M., Majerus, S., Ledoux, D., Damas, P., Boly, M.M., et al., 2008. voluntary brain processing in disorders of consciousness. Neurology 71 (20), 1614–1620.

Schnakers, C., Vanhaudenhuyse, A., Giacino, J., Ventura, M., Boly, M., Majerus, S., Moonen, G., Laureys, S., 2009a. Diagnostic accuracy of the vegetative and minimally conscious state: clinical consensus versus standardized neurobehavioral assessment. BMC Neurol. 9 (1), 35.

Schnakers, C., Perrin, F., Schabus, M., Hustinx, R., Majerus, S., Moonen, G., Boly, M., Vanhaudenhuyse, A., Bruno, M.-A., Laureys, S., 2009b. Detecting consciousness in a total locked-in syndrome: an active event-related paradigm. Neurocase 15 (4), 271–277.

Sitt, J.D., King, J.-R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., Naccache, L., 2014. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. Brain 137 (8), 2258–2270.

Stender, J., Gosseries, O., Bruno, M.M.-A.A., Charland-Verville, V., Vanhaudenhuyse, A., Demertzi, A., Chatelle, C., et al., 2014. diagnostic precision of PET imaging and functional MRI in disorders of consciousness: a clinical validation study. Lancet 384 (9942), 514–522.

Steppacher, I., Kaps, M., Kissler, J., 2014. Will time heal? A long-term follow-up of severe disorders of consciousness. Ann. Clin. Transl. Neurol. 1 (6), 401–408. http://dx.doi.org/10.1002/acn3.63 (Jun).

Stoll, J., Chatelle, C., Carter, O., Koch, C., Laureys, S., Einhäuser, W., 2013. Pupil responses allow communication in locked-in syndrome patients. Curr. Biol. 23 (15), R647–R648.

The Multi-Society Task Force on PVS, 1994. Medical aspects of the persistent vegetative state (2). N. Engl. J. Med. 330 (22), 1572–1579.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Tzovara, A., Rossetti, A.O., Spierer, L., Grivel, J., Murray, M.M., Oddo, M., De Lucia, M., 2013. Progression of auditory discrimination based on neural decoding predicts awakening from coma. Brain 136, 81–89.

Vanhaudenhuyse, A., Noirhomme, Q., Tshibanda, L.J.-F., Bruno, M.-A., Boveroux, P., Schnakers, C., Soddu, A., et al., 2010. Default network connectivity reflects the level of consciousness in non-communicative brain-damaged patients. Brain 133 (1), 161–171.

Wannez, S., Annen, J., Aubinet, C., Thonnard, M., Charland-Verville, V., Heine, L., Habbal, D., Martial, C., Bodart, O., Vanhaudenhuyse, A., Chatelle, C., Thibaut, A., Schnakers, C., Demertzi, A., Gosseries, O., Laureys, S., 2016. The Eleventh World Congress on Brain Injury, The Hague, The Nederlands, 2-5 March.

Wilhelm, B., Jordan, M., Birbaumer, N., 2006. Communication in locked-in syndrome: effects of imagery on salivary pH. Neurology 67 (3), 534–535.

Wolfers, T., Buitelaar, J.K., Beckmann, C., Franke, B., Marquand, A., 2015. From estimating activation locality to predicting disorder: a review of pattern recognition for neuroimaging-based psychiatric diagnostics. Neurosci. Biobehav. Rev. 57, 328–349. http://dx.doi.org/10.1016/j.neubiorev.2015.08.001.