Assessing the depth of language processing in patients with disorders of consciousness

Peng Gui[1,9], Yuwei Jiang[1,9], Di Zang[2,3,9], Zengxin Qi[2,3], Jiaxing Tan[2,3], Hiromi Tanigawa[2,3], Jian Jiang[1], Yunqing Wen[1], Long Xu[4], Jizong Zhao[4], Ying Mao[2,3,5], Mu-ming Poo[1], Nai Ding[6], Stanislas Dehaene[7,8], Xuehai Wu ●[2,3,5 ✉] and Liping Wang ●[1 ✉]

Assessing residual consciousness and cognitive abilities in unresponsive patients is a major clinical concern and a challenge for cognitive neuroscience. Although neuroimaging studies have demonstrated a potential for informing diagnosis and prognosis in unresponsive patients, these methods involve sophisticated brain imaging technologies, which limit their clinical application. In this study, we adopted a new language paradigm that elicited rhythmic brain responses tracking the single-word, phrase and sentence rhythms in speech, to examine whether bedside electroencephalography (EEG) recordings can help inform diagnosis and prognosis. EEG-derived neural signals, including both speech-tracking responses and temporal dynamics of global brain states, were associated with behavioral diagnosis of consciousness. Crucially, multiple EEG measures in the language paradigm were robust to predict future outcomes in individual patients. Thus, EEG-based language assessment provides a new and reliable approach to objectively characterize and predict states of consciousness and to longitudinally track individual patients' language processing abilities at the bedside.

Assessing residual consciousness and language abilities in unresponsive patients is a challenge for cognitive neuroscience and a major clinical concern. Every year, thousands of patients, due to severe brain injuries, lose their communication abilities and fall into different clinical conditions ranging from coma to unresponsive wakefulness syndrome (UWS) to a minimally conscious state (MCS). The clinical diagnostic assessment of patients' conditions is mainly based on motor and oro-motor non-reflex behavior at the bedside[1]. In particular, patients with UWS present moments of arousal, during which they open their eyes and produce complex behavior reflexes, but they show no clear signs of intentional behavior[2]. By contrast, patients in an MCS present some intentional behaviors but seem unable to establish any long-lasting functional communication[3]. The subtle difference between patients in an MCS and those with UWS, often compounded by the presence of additional deficits, can lead to a high rate of misdiagnosis[4].

Recent work has shown that the level of consciousness can be indicated by various dynamic features of EEG signals, such as the amplitude and latency of auditory evoked responses[5], spectral power[6] and signal complexity and functional connectivity, assessed using weighted symbolic mutual information[7]. Indeed, the major consciousness theories claim that consciousness is characterized by a dynamic process of self-sustained, coordinated brain activity that constantly evolves rather than being a static brain function[8]. Accordingly, a series of recent studies reported that the dynamics of resting-state activity in functional magnetic resonance imaging (fMRI) might provide a specific cortical signature of the loss of

consciousness[9]. Specifically, during loss of consciousness, brain activity is largely restricted to a dynamic pattern dominated by structural connectivity. In contrast, conscious states are characterized by a more complex pattern of brain activity with long-distance (e.g. frontal–parietal) interactions. Similar to those seen in fMRI signals, the dynamic patterns in EEG can be described as 'microstates', which are defined as global patterns of scalp potential topographies that dynamically vary over time in an organized manner[10]. That is, resting-state or task-related EEG can be described by a limited number of scalp potential topographies (maps) that remain stable for 60–120 ms before rapid transition to a different topography that remains stable again[11]. Given its temporal resolution, the pattern of EEG microstates seems likely to provide a better index of such fast dynamics and therefore better reflect the level of consciousness in patients with a disorder of consciousness (DOC), but this has yet to be experimentally tested.

Auditory oddball paradigms have commonly been used in EEG studies to detect the residual consciousness in patients with a DOC[12]. In these paradigms, although individuals might be instructed to count the number of times they hear a specific target sound[13] or a violation of temporal regularities, such paradigms rely primarily on an assessment of sensory responses at several hierarchical levels. Using active paradigms (e.g. mental imagery of playing tennis), some patients with a DOC were found to respond to commands, which requires greater cognitive abilities[14]. Rather than using pure tones as auditory stimuli, several studies attempted to develop reliable language paradigms to detect neural signatures of semantic

processing[15,16], as natural language stimuli might be easier for patients to attend to. Although cortical responses to natural speech in unresponsive patients in neuroimaging studies have provided evidence that natural language stimuli activated more auditory cortical regions than the scrambled auditory stimuli[15–17], EEG results have been variable. Most of these EEG studies examined the N400 component or inter-subject correlation of neural signals in response to the narrative content of natural speech and found no or weak differences between UWS and MCS patient groups[7,13,16,18,19]. The use of EEG in active linguistic paradigms that combine stimulus-evoked activity and dynamic brain states for assisting diagnosis and prognosis of DOCs remains unexplored.

In the present study, we adopted a new hierarchical auditory linguistic sequence paradigm that included three levels of processing, at the single-word, phrase and sentence levels[20], to compare the EEG features between activity in resting passive tasks and activity during active tasks (three language conditions). Our aim was to assess the depth of language processing in DOCs and to separate two distinct possibilities concerning this depth. First, the higher the consciousness level, the deeper the processing level of linguistic stimuli might be. This hypothesis might seem plausible given that the integration of multiple words into phrases and sentences calls for late, sustained and integrative brain activity that is typically found associated with conscious processing[8]. Second, alternatively, much language processing might remain possible in the absence of consciousness, as attested by a variety of masking and inattention paradigms in normal individuals, including the observation of brain responses to syntactic and semantic violations under non-conscious conditions[16,21,22]. Even in this case, it might still prove clinically useful to assess the depth of unconscious processing of linguistic stimuli in patients, as this might be predictive of their recovery.

Our paradigm allows us to combine both speech-tracking activity and the dynamic pattern of brain states. We first evaluated, in normal individuals, whether the detection of hierarchical structure in the sequences requires top-down cognitive resources and is modulated by attention. Moving to patients, we then employed a multivariate approach, integrating both speech-tracking activities and temporal dynamics of global brain states to evaluate the depth of linguistic processing in patients with DOCs and the value of EEG measures for diagnosis and prognosis of consciousness. We trained and validated classification algorithms that use EEG-derived metrics and clinical diagnosis as inputs and attempted to predict the clinical outcomes of individual patients.

Results

ITPC (speech-tracking activity). We constructed hierarchical linguistic structures using an isochronous, 4-Hz sequence of Chinese words that were independently synthesized (Fig. 1 and Methods). The auditory sequences included three linguistic levels: monosyllabic words, two-word phrases and four-word sentences. Monosyllabic words were presented at constant rates (250 ms per word), which means that the corresponding neural tracking of words, phrases and sentences could be tracked at distinct frequencies (4, 2, and 1 Hz, respectively; Fig. 1a).

We first tested whether top-down attention is required for the EEG responses to hierarchical language structures. We recruited 22 healthy human participants to perform an attentional task. The participants either attended to the auditory stimuli (the attend to condition), which were either word lists or sentence sequences, or performed a visual task while the auditory stimuli were simultaneously presented and ignored (the ignore condition; Fig. 1b and Methods). In both the attend to and ignore conditions, we found a significant and compatible 4-Hz response in the inter-trial phase coherence (ITPC) spectrum, for both the word list and sentence conditions, compared to baseline (Fig. 1c). However, the ITPC values at 1 and 2 Hz in the sentence conditions were significantly

weakened after the attention was shifted to the visual stimuli (attend to versus ignore, $P_{1\,Hz} = 0.015$ and $P_{2\,Hz} = 2.3 \times 10^{-5}$; paired-sample t-test; Fig. 1c,d). Thus, these results indicate the automaticity of processes underlying single-word tracking (4 Hz) and partially attentional modulation of neural processing of higher-level linguistic structures, i.e. phrases and sentences (at 2 and 1 Hz, respectively).

On the basis of these results, our hypothesis was that residual consciousness in patients with DOCs could be reflected by the strength of speech-tracking responses, especially neural tracking of higher-level linguistic structures, i.e. phrases and sentences (measured by ITPC). To test this hypothesis, we examined brain responses to sentence sequences in 42 patients with MCS, 36 patients with UWS and 47 healthy controls (see patient details in Supplementary Table 1 and patient selection in Extended Data Fig. 1). As shown in Fig. 2a, after a clinical diagnosis using the Coma Recovery Scale-Revised (CRS-R) and the Glasgow Coma Scale (GCS) (the categorization of MCS and UWS was done after the CRS-R diagnosis), a 5-min resting-state EEG was first recorded at the start of each recording session. After a 2-min rest period, three 8-min blocks containing Mandarin Chinese speech stimuli with different linguistic levels (word lists, phrase sequences and sentence sequences) were presented. The results showed a progressive increase in the strength of EEG ITPC that matched the increasing level of behavioral responsiveness as quantified by the CRS-R, from UWS to MCS, alongside the healthy control group for comparison. Specifically, word-level tracking, measured by the 4-Hz ITPC, was significant in the healthy control, MCS and UWS groups ($P_{4\,Hz\text{-Healthy}} = 1.3 \times 10^{-10}$; $P_{4\,Hz\text{-MCS}} = 2.1 \times 10^{-6}$; $P_{4\,Hz\text{-UWS}} = 5.8 \times 10^{-4}$; paired-sample t-test; Fig. 2b, left, and Extended Data Fig. 2). Phrase-level tracking, measured by the 2-Hz ITPC, was significant in the healthy control group, marginally significant in the MCS group and not significant in the UWS group ($P_{2\,Hz\text{-Healthy}} = 3.8 \times 10^{-9}$; $P_{2\,Hz\text{-MCS}} = 0.097$; $P_{2\,Hz\text{-UWS}} = 0.881$; paired-sample t-test; Fig. 2b, middle, and Extended Data Fig. 2). Sentence-level tracking measured by the 1-Hz ITPC was significant in the healthy control group but not significant in the MCS or UWS groups ($P_{1\,Hz\text{-Healthy}} = 4.9 \times 10^{-5}$; $P_{1\,Hz\text{-MCS}} = 0.567$; $P_{1\,Hz\text{-UWS}} = 0.546$; paired-sample t-test; Fig. 2b, right, and Extended Data Fig. 2).

It is worth noting that, although there were no group-level significant differences in 1- or 2-Hz ITPC between the MCS and UWS groups (Fig. 2b), some differences at the individual level were apparent (Extended Data Fig. 3). Namely, 11 patients with MCS and four patients with UWS exhibited significant ITPC at 1 or 2 Hz, which could indicate residual consciousness in these patients. Indeed, six (five MCS and one UWS) of these 15 (40.0 %) patients showed significant improvement of clinical diagnosis 100 d after the EEG recordings (outcome predictions; also see the classification results). We then applied multivariate pattern analysis on the 1-, 2-, and 4-Hz ITPC for all available electrodes to classify the patient groups (linear discriminant analysis (LDA); see Methods). Figure 2c shows successful decoding of patient groups. Even MCS and UWS groups were significantly distinguishable, particularly when listening to sentences.

Temporal dynamics of global brain states. The brain is inherently active in a regular manner both at rest and during cognitive tasks, and this dynamic pattern has been proposed to be the neural signature of consciousness[9,23]. Hence, we evaluated the second hypothesis that residual consciousness can be characterized by monitoring the dynamic patterns of brain states, in that these brain dynamics would be associated with different cognitive states. To this aim, we quantified the spatial and temporal dynamics of brain activity in healthy controls and patients by examining the properties (e.g. probability, occurrence, duration and transition) of the global pattern of scalp potential topographies (also referred as 'microstates')[11,24] in conditions with four increasing levels.
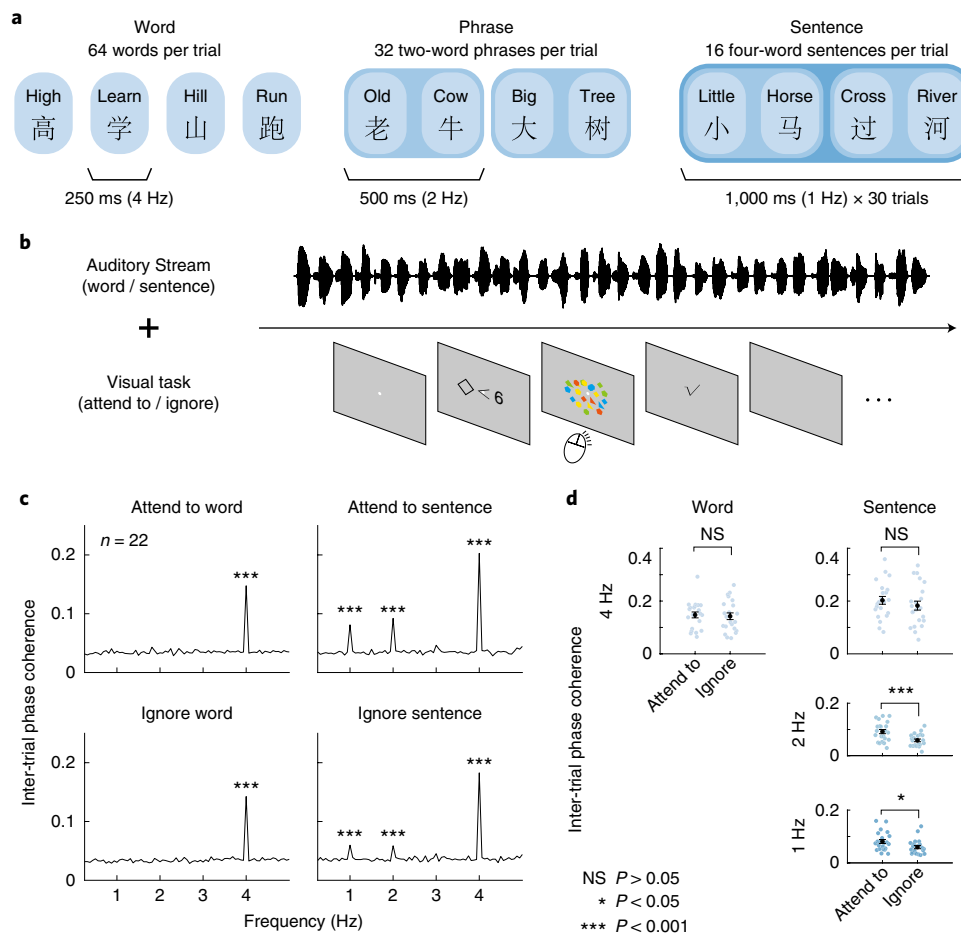
Fig. 1 | Paradigm and neural tracking of hierarchical linguistic structures. a, Illustration of stimuli presented in the word, phrase and sentence conditions. **b**, Schematic of the attentional experiment in normal individuals, which required attending either to the auditory sequence or to concurrent visual stimuli. **c**, Group-averaged ITPC under four conditions in the attentive study ($n = 22$ participants). Attend to: one-sided paired-sample t-test; $t_{\text{Word-4 Hz}}(21) = 10.11$, $P_{\text{Word-4 Hz}} = 8 \times 10^{-10}$; $t_{\text{Sentence-1 Hz}}(21) = 6.11$, $P_{\text{Sentence-1 Hz}} = 2.3 \times 10^{-6}$; $t_{\text{Sentence-2 Hz}}(21) = 7.26$, $P_{\text{Sentence-2 Hz}} = 1.9 \times 10^{-7}$; $t_{\text{Sentence-4 Hz}}(21) = 11.1$, $P_{\text{Sentence-4 Hz}} = 1.4 \times 10^{-10}$; ignore: one-sided paired-sample t-test; $t_{\text{Word-4 Hz}}(21) = 8.22$, $P_{\text{Word-4 Hz}} = 2.67 \times 10^{-8}$; $t_{\text{Sentence-1 Hz}}(21) = 4$, $P_{\text{Sentence-1 Hz}} = 3.2 \times 10^{-4}$; $t_{\text{Sentence-2 Hz}}(21) = 4.14$, $P_{\text{Sentence-2 Hz}} = 2.3 \times 10^{-4}$; $t_{\text{Sentence-4 Hz}}(21) = 8.58$, $P_{\text{Sentence-4 Hz}} = 1.3 \times 10^{-8}$. **d**, Comparison of ITPC values between attend to and ignore conditions ($n = 22$ participants). Attend to versus ignore: two-sided paired-sample t-test; $t_{\text{Word-4 Hz}}(21) = 0.42$, $P_{\text{Word-4 Hz}} = 0.678$; $t_{\text{Sentence-4 Hz}}(21) = 1.58$, $P_{\text{Sentence-4 Hz}} = 0.128$, $t_{\text{Sentence-2 Hz}}(21) = 5.41$, $P_{\text{Sentence-2 Hz}} = 2.3 \times 10^{-5}$, $t_{\text{Sentence-1 Hz}}(21) = 2.66$, $P_{\text{Sentence-1 Hz}} = 0.015$. Colored dots represent individual participants. Black dots represent mean values. Error bars represent s.e.m. All panels: not significant (NS), $P > 0.05$; *$P < 0.05$ and ***$P < 0.001$.

Group-level clustering identified an optimum of four clusters across groups and conditions, which reached the highest cross-validation criterion and explained approximately 80% of variance (Fig. 3a). The spatial configurations of the four maps in healthy controls (Fig. 3b) were highly consistent with the four maps described in previous studies[25,26]. We then labeled and sorted the four sets of maps according to the appearing probability of brain states at resting state in the healthy controls. Specifically, map A showed a fronto–central maximum, map B showed a symmetric frontal-to-occipital orientation, map C showed a left occipital-to-right frontal orientation and map D showed a right occipital-to-left frontal orientation (Fig. 3b; the group-averaged brain states of each group in each condition are shown in Extended Data Fig. 4).

Previous studies using simultaneous EEG and fMRI recordings suggested that brain states A and B are more closely related to the attention and saliency networks, as their corresponding blood-oxygen-level-dependent (BOLD) activations were located in the anterior cingulate cortex and parietal–frontal areas, and that states C and D are related to the auditory and visual sensory networks, as their corresponding BOLD signals were located in bilateral temporal and extrastriate visual areas[11,27]. We thus predicted

that higher levels of consciousness would be associated with a higher probability of the activation of high-level cognitive neural networks—that is, maps A and B (anterior–posterior maps, defined as the A–P map, Fig. 3b). In parallel, we predicted that reduced consciousness would involve a greater relative contribution of lower-level sensory areas, corresponding to maps C and D (left–right maps, defined as the L–R map, Fig. 3b). Multivariate analysis of variance (MANOVA) showed that, for the four conditions, the healthy controls demonstrated a pattern of a high probability of the A–P map and a low probability of the L–R map (Fig. 3c). The patient groups showed the opposite pattern, with a low probability of the A–P map and a high probability of the L–R map (Fig. 3c).

Next, we examined the difference of the dynamic of brain states between the MCS and UWS groups. For each individual patient, we quantified the probability-weighted spatial correlation difference between the A–P and L–R maps ($\Delta C\rho$; see Methods). This difference reflected the spatial similarity between the maps of patient group and template maps derived from the healthy control group and was used as an index of residual consciousness. We found a progressive increase in the difference of $\Delta C\rho$ between the MCS and UWS groups as the level of the linguistic hierarchy increased
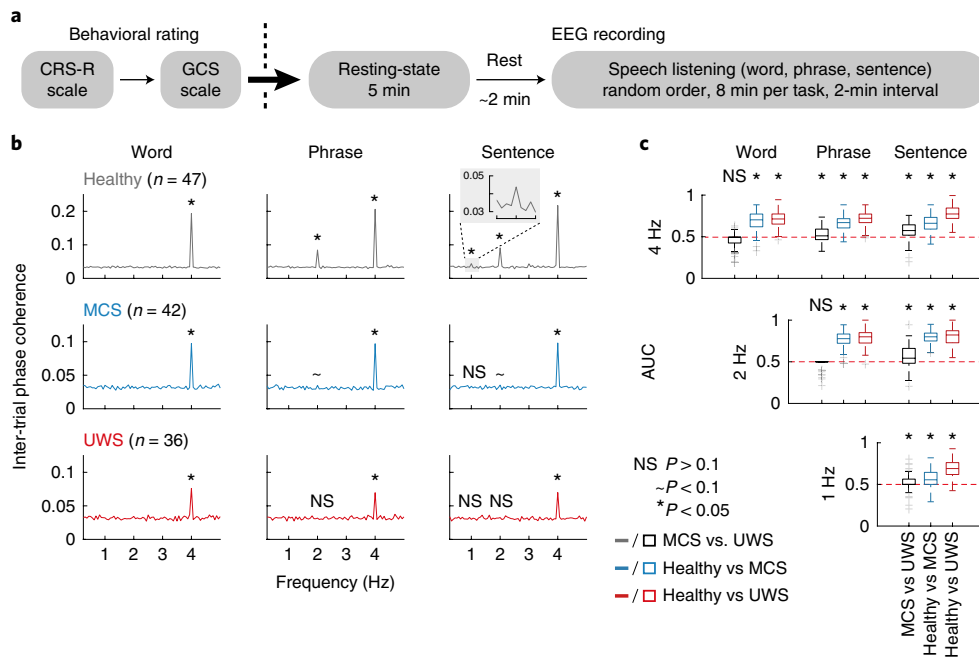
Fig. 2 | Procedure and auditory-evoked brain activity in the clinical study. a, Schematic procedure in patients with DOCs. In the same study day, after behavioral ratings, the EEG recording started with a 5-min resting-state, followed by a short-term rest and three linguistic auditory blocks in a random order across participants. **b**, Group-averaged ITPC in the word, phrase and sentence conditions. Data from the healthy control ($n = 47$), MCS ($n = 42$) and UWS ($n = 36$) groups are plotted from top to bottom. One-sided paired-sample t-test. Word condition: $t_{4\,Hz\text{-}Healthy}(46) = 8.24$, $P_{4\,Hz\text{-}Healthy} = 1.3 \times 10^{-10}$; $t_{4\,Hz\text{-}MCS}(41) = 5.52$, $P_{4\,Hz\text{-}MCS} = 2.1 \times 10^{-6}$; $t_{4\,Hz\text{-}UWS}(35) = 3.78$, $P_{4\,Hz\text{-}UWS} = 5.8 \times 10^{-4}$. Phrase condition: $t_{4\,Hz\text{-}Healthy}(46) = 8.78$, $P_{4\,Hz\text{-}Healthy} = 2.1 \times 10^{-11}$; $t_{4\,Hz\text{-}MCS}(41) = 5.36$, $P_{4\,Hz\text{-}MCS} = 3.5 \times 10^{-6}$; $t_{4\,Hz\text{-}UWS}(35) = 3.87$, $P_{4\,Hz\text{-}UWS} = 4.5 \times 10^{-4}$; $t_{2\,Hz\text{-}Healthy}(46) = 7.25$, $P_{2\,Hz\text{-}Healthy} = 3.8 \times 10^{-9}$; $t_{2\,Hz\text{-}MCS}(41) = 1.70$, $P_{2\,Hz\text{-}MCS} = 0.097$, $t_{2\,Hz\text{-}UWS}(35) = 0.15$, $P_{2\,Hz\text{-}UWS} = 0.881$. Sentence condition: $t_{4\,Hz\text{-}Healthy}(46) = 8.35$, $P_{4\,Hz\text{-}Healthy} = 9.1 \times 10^{-11}$; $t_{4\,Hz\text{-}MCS}(41) = 5.40$, $P_{4\,Hz\text{-}MCS} = 2.2 \times 10^{-6}$; $t_{4\,Hz\text{-}UWS}(35) = 3.93$, $P_{4\,Hz\text{-}UWS} = 3.8 \times 10^{-4}$; $t_{2\,Hz\text{-}Healthy}(46) = 6.62$, $P_{2\,Hz\text{-}Healthy} = 3.4 \times 10^{-8}$; $t_{2\,Hz\text{-}MCS}(41) = 1.88$, $P_{2\,Hz\text{-}MCS} = 0.068$, $t_{2\,Hz\text{-}UWS}(35) = 0.39$, $P_{2\,Hz\text{-}UWS} = 0.702$; $t_{1\,Hz\text{-}Healthy}(46) = 4.48$, $P_{1\,Hz\text{-}Healthy} = 4.9 \times 10^{-5}$, $t_{1\,Hz\text{-}MCS}(41) = 0.58$, $P_{1\,Hz\text{-}MCS} = 0.567$, $t_{1\,Hz\text{-}UWS}(35) = -0.61$, $P_{1\,Hz\text{-}UWS} = 0.546$. **c**, Decoding performance of the multivariate pattern analysis using phase coherence at 1, 2, and 4 Hz with LDA ($n = 100$ permutations). One-sided t-test, comparing with 0.5: MCS versus UWS, $t_{Word\text{-}4\,Hz}(99) = -4.31$, $P_{Word\text{-}4\,Hz} = 0.99998$; $t_{Phrase\text{-}4\,Hz}(99) = 2.75$, $P_{Phrase\text{-}4\,Hz} = 3.6 \times 10^{-3}$, $t_{Phrase\text{-}2\,Hz}(99) = -2.75$, $P_{Phrase\text{-}2\,Hz} = 0.996$, $t_{Sentence\text{-}4\,Hz}(99) = 6.52$, $P_{Sentence\text{-}4\,Hz} = 1.5 \times 10^{-9}$, $t_{Sentence\text{-}2\,Hz}(99) = 4.8$, $P_{Sentence\text{-}2\,Hz} = 2.4 \times 10^{-6}$, $t_{Sentence\text{-}1\,Hz}(99) = 2.02$, $P_{Sentence\text{-}1\,Hz} = 0.029$; healthy versus MCS, $t_{Word\text{-}4\,Hz}(99) = 17.63$, $P_{Word\text{-}4\,Hz} = 1.3 \times 10^{-32}$, $t_{Phrase\text{-}4\,Hz}(99) = 17.45$, $P_{Phrase\text{-}4\,Hz} = 2.8 \times 10^{-32}$, $t_{Phrase\text{-}2\,Hz}(99) = 30.04$, $P_{Phrase\text{-}2\,Hz} = 7.6 \times 10^{-52}$, $t_{Sentence\text{-}4\,Hz}(99) = 17.25$, $P_{Sentence\text{-}4\,Hz} = 6.7 \times 10^{-32}$, $t_{Sentence\text{-}2\,Hz}(99) = 38.61$, $P_{Sentence\text{-}2\,Hz} = 8.7 \times 10^{-62}$, $t_{Sentence\text{-}1\,Hz}(99) = 6.31$, $P_{Sentence\text{-}1\,Hz} = 4 \times 10^{-9}$; healthy versus UWS: $t_{Word\text{-}4\,Hz}(99) = 23.75$, $P_{Word\text{-}4\,Hz} = 5.6 \times 10^{-43}$, $t_{Phrase\text{-}4\,Hz}(99) = 25.79$, $P_{Phrase\text{-}4\,Hz} = 5 \times 10^{-46}$, $t_{Phrase\text{-}2\,Hz}(99) = 31.06$, $P_{Phrase\text{-}2\,Hz} = 3.8 \times 10^{-53}$, $t_{Sentence\text{-}4\,Hz}(99) = 31.25$, $P_{Sentence\text{-}4\,Hz} = 2.2 \times 10^{-53}$, $t_{Sentence\text{-}2\,Hz}(99) = 34.71$, $P_{Sentence\text{-}2\,Hz} = 1.6 \times 10^{-57}$, $t_{Sentence\text{-}1\,Hz}(99) = 17.36$, $P_{Sentence\text{-}1\,Hz} = 4.1 \times 10^{-32}$. Boxes represent interquartile range (IQR), central bars indicate the median, and whiskers indicate 1.5× IQR. "+" symbols indicate outliers. Horizontal dashed lines indicate the chance level. All panels: not significant (NS), $P > 0.1$; $^{\sim}P < 0.1$ and $^*P < 0.05$.

from resting to word, phrase and sentence conditions (Fig. 3d,e). Specifically, at the phrase and sentence levels (but not at the resting or word levels), the MCS group showed a significantly higher $\Delta C\rho$ than the UWS group ($P_{Phrase} = 0.037$; $P_{Sentence} = 0.014$; one-way ANOVA, Bonferroni corrected; Fig. 3e). This was indicative of an increased probability of the A–P map (frontal–parietal networks) and a decreased probability of the L–R map (sensory networks). Furthermore, the difference in $\Delta C\rho$ between MCS and UWS was significantly larger in the phrase and sentence condition than that in the resting and word condition (phrase versus word, $t_{76} = 2.29$, $P = 0.03$; sentence versus word, $t_{76} = 3.14$, $P = 0.002$; two-sided two-sample t-test). Thus, positive and high $\Delta C\rho$ potentially indicates residual consciousness.

We then investigated whether the probability difference of the maps was due to the duration (how long the map remained stable) or the frequency of occurrence (how many times the map occurred in 1 s) of each map. We found that the probability differences between the MCS and UWS groups in the phrase and sentence conditions could be attributed to a shorter duration of the L–R map, thought to reflect sensory networks ($P_{Phrase} = 0.016$; $P_{Sentence} = 0.017$; one-way ANOVA, Bonferroni corrected; Fig. 4a,b) and a higher occurrence of the A–P map, putatively associated with frontal–parietal

networks ($P_{Phrase} = 0.028$; $P_{Sentence} = 0.063$; one-way ANOVA, Bonferroni corrected; Fig. 4d,e) in the MCS group relative to the UWS group. Notably, the increase of the significant difference between the MCS and UWS groups matched the increasing linguistic level of conditions (Fig. 4b,e). No significant differences were found in the duration of the A–P map or in the occurrence of the L–R map (Extended Data Fig. 5).

If the duration and/or occurrence of maps indeed reflect the strength of residual consciousness in patients, we should observe corresponding changes before and after their recovery. In the subpopulation of participants who had multiple EEG recordings (12 out of 54 patients for resting-state recordings and 15 out of 60 patients for recordings during the linguistic tasks), the duration of the L–R map became shorter (Fig. 4a,c) and the occurrence of the A–P map increased along with recovery (Fig. 4d,f). Such changes were not observed in the non-recovery patients (Fig. 4a,c,d,f). The distribution of the A–P map duration and L–R map occurrence in healthy controls, recovery and non-recovery patients with MCS and UWS are shown in Extended Data Fig. 5.

To exclude the possibility that the different spatial maps in the two patient groups arose from a difference in brain injury, we then analyzed the volumes of brain damage in the 27 patients
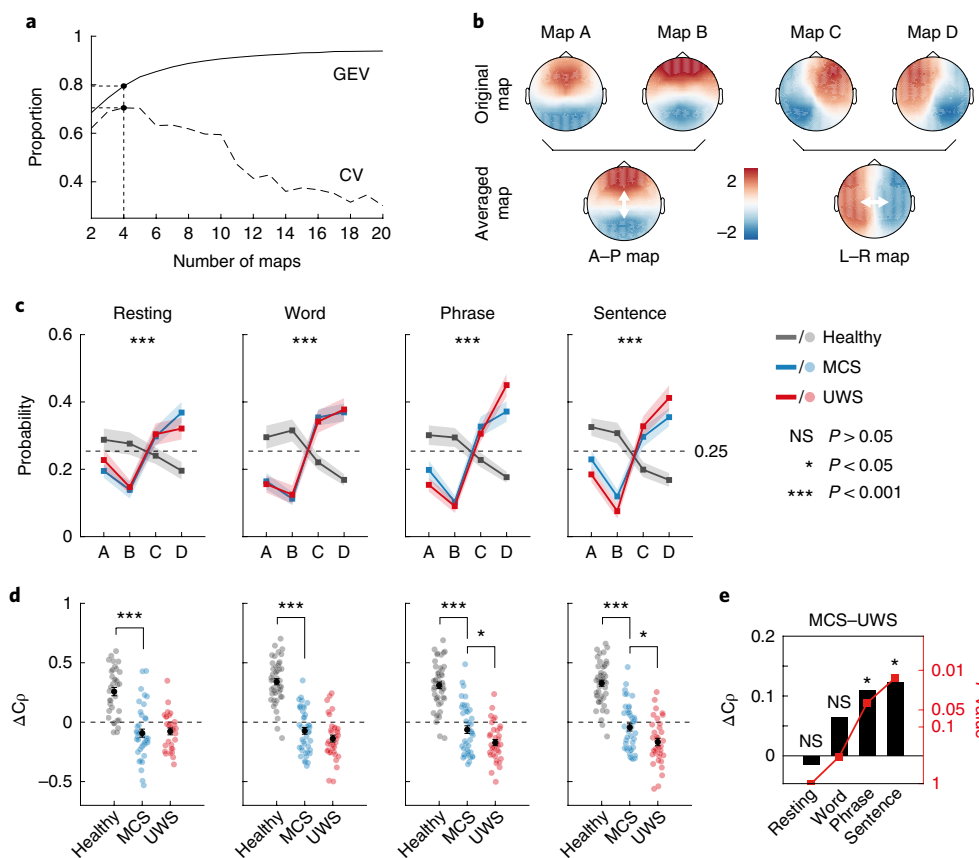
Fig. 3 | Global patterns of brain states. a, Mean accuracy of CV criterion and mean GEV varied with numbers of brain state maps, showing that the optimal number of clustering maps across participants was four. **b**, Group-averaged brain state maps of healthy controls in task conditions (averaged across all three task conditions, $n = 47$ participants), sorted by the brain state probability. Colors represent the relative potential distribution. **c**, The probability distribution of four maps in the three groups for each of the four conditions. MANOVA: resting, $F_{6, 200} = 5.176$, $P = 5.7 \times 10^{-5}$; word, $F_{6, 240} = 13.543$, $P = 3.1 \times 10^{-13}$; phrase, $F_{6, 240} = 14.258$, $P = 6.9 \times 10^{-14}$; sentence, $F_{6, 240} = 12.56$, $P = 2.6 \times 10^{-12}$. Lines indicate mean values; shaded areas represent s.e.m. **d**, $\Delta C\rho$ (probability-weighted spatial correlation difference) between A–P and L–R maps for each group and condition. One-way ANOVA, Bonferroni corrected: resting, $P_{\text{Healthy-MCS}} = 3.7 \times 10^{-11}$, $P_{\text{Healthy-UWS}} = 2.4 \times 10^{-9}$; word, $P_{\text{Healthy-MCS}} = 2.9 \times 10^{-19}$, $P_{\text{Healthy-UWS}} = 3.8 \times 10^{-22}$; phrase, $P_{\text{Healthy-MCS}} = 2.1 \times 10^{-15}$, $P_{\text{Healthy-UWS}} = 9.4 \times 10^{-21}$, $P_{\text{MCS-UWS}} = 0.037$; sentence, $P_{\text{Healthy-MCS}} = 1.7 \times 10^{-15}$, $P_{\text{Healthy-UWS}} = 9.9 \times 10^{-22}$, $P_{\text{MCS-UWS}} = 0.014$. Colored dots represent individual participants; black dots represent mean values; error bars represent s.e.m. **e**, Comparison of $\Delta C\rho$ between the MCS and UWS groups. The difference matched well along with the linguistic hierarchy from the resting state to sentence condition. The red line indicates statistical significance between patient groups in each condition. One-way ANOVA, Bonferroni corrected: $P_{\text{Resting}} = 1$, $P_{\text{Word}} = 0.33$, $P_{\text{Phrase}} = 0.037$, $P_{\text{sentence}} = 0.014$. **c–e**: $n_{\text{Healthy-Resting}} = 34$, $n_{\text{Healthy-Task}} = 47$, $n_{\text{MCS-Resting}} = 41$, $n_{\text{MCS-Task}} = 42$, $n_{\text{UWS-Resting}} = 30$, $n_{\text{UWS-Task}} = 36$. All panels: not significant (NS), $P > 0.05$; *$P < 0.05$ and ***$P < 0.001$. CV, cross-validation; GEV, global explained variance.

(17 MCS and ten UWS) from the database who had received a structural MRI scan on the same day as the EEG recordings. The results showed that there was no significant difference in the injured volume between the MCS and UWS patients (Extended Data Fig. 6a) and no significant correlation between the volume of brain injury and the $\Delta C\rho$ at all three levels of task (Extended Data Fig. 6b). For example, patient 7 with no brain damage had a low value of $\Delta C\rho$ (Extended Data Fig. 6c); by contrast, patient 17 with a high volume of brain injury showed a relatively higher $\Delta C\rho$ (Extended Data Fig. 6d). Furthermore, if we assume that the size of brain damage would not change dramatically during the recovery period, the observed changes in spatial maps before and after recovery (Fig. 4) also help rule out the possibility that the brain states measured with spatial maps purely reflect the underlying brain injury. The change of spatial maps of an example patient (ID: 2) is shown in Extended Data Fig. 6g–i.

Individual diagnosis and prediction. Our multiple measurements of brain activity allowed us to move from group-level analyses and attempt personalized diagnoses and predictions. We first

trained a three-class LDA classifier with leave-one-subject-out cross-validation for the diagnosis of individual control, MCS and UWS patients who had at least a 3-month duration of a DOC. The inputs to the classifier were the multiple EEG measurements, including the ITPC (three features: ITPC values at 1, 2, and 4 Hz) and global dynamic patterns of brain activity (six features: ΔProbability, $\Delta C\rho$, Occurrence$_{\text{A-P}}$, Duration$_{\text{L-R}}$, Transition$_{\text{A-P}}$ and Transition$_{\text{L-R}}$). A classifier with regularization first searched for the optimal feature combination within each task and calculated the classification probability for each individual participant (Fig. 5a; for detailed information of feature selection, see Methods). Then, to avoid model overfitting, only those selected feature combinations were entered in the final LDA. All steps were cross-validated (leave-one-subject-out). Figure 5b plots the confusion matrix generated by the LDA. The best classification was found in the sentence condition with $\Delta C\rho$, Transition$_{\text{A-P}}$, ITPC$_{\text{1 Hz}}$ and ITPC$_{\text{2 Hz}}$ as input EEG features. A chi-squared test was used to estimate the classifier's performance, which was highly significant ($\chi^2 = 95.84$, $P = 7.6 \times 10^{-20}$, accuracy $= 75\%$). The decoder categorized healthy control, MCS and UWS participants with 89%, 58% and 70%
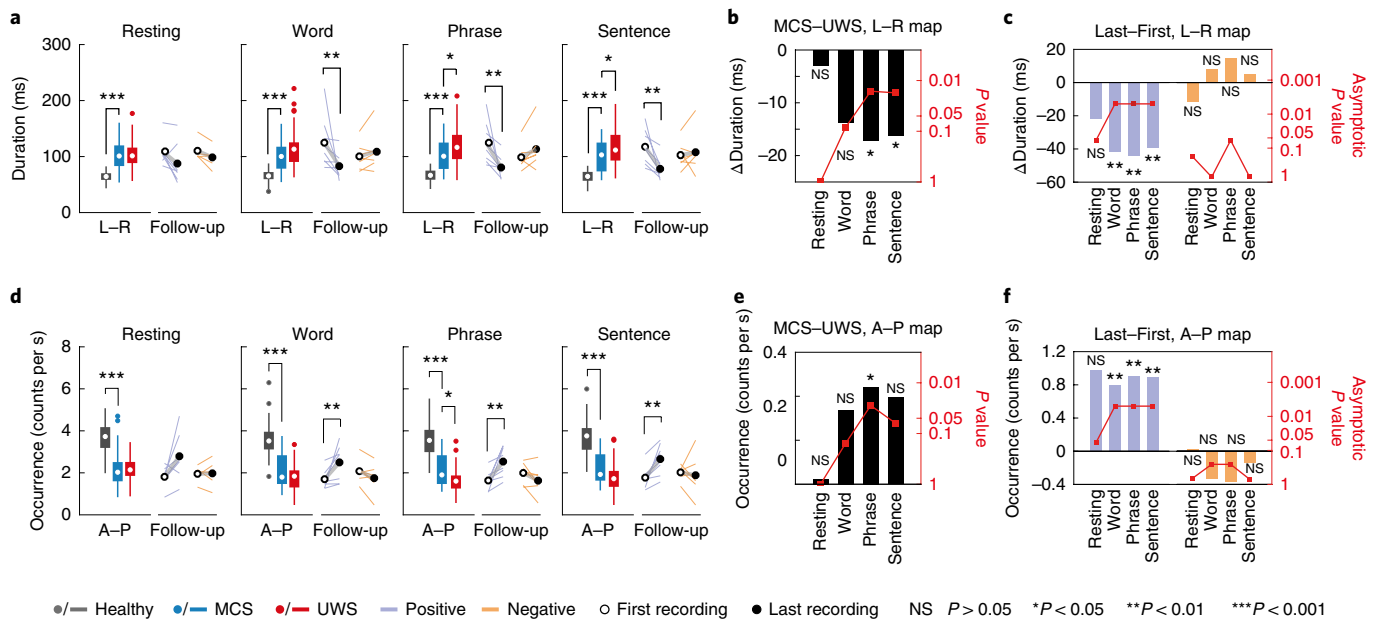
Fig. 4 | Duration and occurrence of brain state maps. a, Duration. Left: the duration of the L–R map for healthy controls ($n_{Resting} = 34$, $n_{Task} = 47$), patients with MCS ($n_{Resting} = 41$, $n_{Task} = 42$) and patients with UWS ($n_{Resting} = 30$, $n_{Task} = 36$). One-way ANOVA, Bonferroni corrected: resting, $P_{Healthy-MCS} = 5.6 \times 10^{-10}$, $P_{Healthy-UWS} = 5.8 \times 10^{-10}$, $P_{MCS-UWS} = 1$; word, $P_{Healthy-MCS} = 2.7 \times 10^{-8}$, $P_{Healthy-UWS} = 8.2 \times 10^{-13}$, $P_{MCS-UWS} = 0.083$; phrase, $P_{Healthy-MCS} = 1.1 \times 10^{-8}$, $P_{Healthy-UWS} = 1.05 \times 10^{-14}$, $P_{MCS-UWS} = 0.016$; sentence, $P_{Healthy-MCS} = 1.1 \times 10^{-9}$, $P_{Healthy-UWS} = 1.1 \times 10^{-15}$, $P_{MCS-UWS} = 0.017$. Boxes represent interquartile range (IQR), central dots indicate the median and whiskers indicate 1.5× IQR. Colored dots indicate outliers. Right: the duration of the L–R map on the first (circle) and last (solid) EEG recording for the recovered (outcome positive (Positive); $n_{Resting} = 7$, $n_{Task} = 8$) and unrecovered (outcome negative (Negative); $n_{Resting} = 5$, $n_{Task} = 7$) patients. Colored lines indicate the first and last EEG recording of individual participants. Friedman tests: resting, $\chi^2_{Positive} = 3.57$, $P_{Positive} = 0.059$, $\chi^2_{Negative} = 1.8$, $P_{Negative} = 0.18$; word, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 0.14$, $P_{Negative} = 0.705$; phrase, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 3.57$, $P_{Negative} = 0.059$; sentence, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 0.14$, $P_{Negative} = 0.705$. **b**, Comparison of the L–R map duration between MCS and UWS groups in the four conditions. The increase in the differences of L–R map duration between the MCS and UWS groups paralleled the increase in linguistic hierarchy. The red line indicates statistical significance between patient groups of L–R map duration in each condition. **c**, Comparison of the difference of L–R map duration between the first and last recording in the four conditions in the recovered and non-recovered patients. **d–f**, Occurrence of the A–P map, with the same format as **a–c**. For A–P differences in **d** and **e**, one-way ANOVA, Bonferroni corrected: resting, $P_{Healthy-MCS} = 1.1 \times 10^{-12}$, $P_{Healthy-UWS} = 1.2 \times 10^{-11}$, $P_{MCS-UWS} = 1$; word, $P_{Healthy-MCS} = 6.1 \times 10^{-15}$, $P_{Healthy-UWS} = 7.1 \times 10^{-19}$, $P_{MCS-UWS} = 0.156$; phrase, $P_{Healthy-MCS} = 3.4 \times 10^{-15}$, $P_{Healthy-UWS} = 8.4 \times 10^{-21}$, $P_{MCS-UWS} = 0.028$; sentence, $P_{Healthy-MCS} = 3.6 \times 10^{-16}$, $P_{Healthy-UWS} = 5.4 \times 10^{-21}$, $P_{MCS-UWS} = 0.063$. For follow-up differences in **d** and **f**, Friedman tests: resting, $\chi^2_{Positive} = 3.57$, $P_{Positive} = 0.059$, $\chi^2_{Negative} = 0.2$, $P_{Negative} = 0.655$; word, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 1.29$, $P_{Negative} = 0.257$; phrase, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 1.29$, $P_{Negative} = 0.257$; sentence, $\chi^2_{Positive} = 8.0$, $P_{Positive} = 0.005$, $\chi^2_{Negative} = 0.14$, $P_{Negative} = 0.705$. All panels: not significant (NS), $P > 0.05$; $*P < 0.05$, $**P < 0.01$ and $***P < 0.001$.

accuracy, respectively, all well above the chance level of 33%. The high decoding accuracy was confirmed by another discriminative classifier, support vector machine (SVM), with 96%, 65% and 73% accuracy for healthy control, MCS and UWS participants, respectively (Extended Data Fig. 7a).

Although a proportion of patients with UWS were classified as patients with MCS (30%, 9/30), it is possible that these patients had some degree of consciousness that was not detected by the CRS-R in the behavioral assessment. Interestingly, a greater proportion (33.3%, 3/9) of such potentially misdiagnosed patients with UWS had positive outcomes (fully awakened or exhibited improved behavior after the EEG recording), as compared to patients for whom both the CRS-R and EEG classifier agreed on a diagnosis of UWS (9.5%, 2/21). Conversely, the classifier also diagnosed some patients with MCS as having UWS (12/31). A lower proportion (25%, 3/12) of those potentially misdiagnosed MCS patients had positive outcomes as compared to patients diagnosed with MCS by both the CRS-R and EEG measurements (44.4%, 8/18). However, we should note that the number of patients in these groups might be too small to generate sufficient power for statistical analysis of between-group differences in these proportions.

We also examined whether EEG recordings could predict the subsequent recovery of consciousness in individual patients.

Thirty-eight patients with multiple measurements were included in this analysis. On the basis of the CRS-R total score 6 months or more after DOC onset[28], 15 patients had a positive outcome, denoted as positive (for patient selection, see Extended Data Fig. 1). We used the same cross-validated method as above to construct an LDA classifier, this time aiming to separate the 15 outcome-positive and 23 outcome-negative patients. The results showed that, whereas the CRS-R total score could partially predict outcomes (AUC = 70%, $\chi^2 = 7.2$, $P = 0.016$, chi-squared test; Fig. 5c), the prediction using the EEG measurements was better (AUC = 77%, $\chi^2 = 11.5$, $P = 9.2 \times 10^{-4}$, chi-squared test; Fig. 5e, left; sensitivity: EEG versus CRS-R, $P = 0.07$, McNemar's test). The best EEG predictive ability was achieved by the mean performance (Methods) of the word, phrase and sentence conditions with 87% sensitivity and 70% specificity, resulting in correct prediction for 13 of 15 outcome-positive and 16 of 23 outcome-negative patients. The best feature combinations for the prediction at each task level were ΔProbability + Occurrence$_{A-P}$ for the word condition, Occurrence$_{A-P}$ + ITPC$_{4\,Hz}$ for the phrase condition and $\Delta C\rho$ + Occurrence$_{A-P}$ for the sentence condition (Supplementary Table 3). We then examined the prognostic ability within individual patients by calculating the predictive scores within patients with UWS and MCS who had CRS-R-based clinical outcomes. The prediction results demonstrated a high prediction
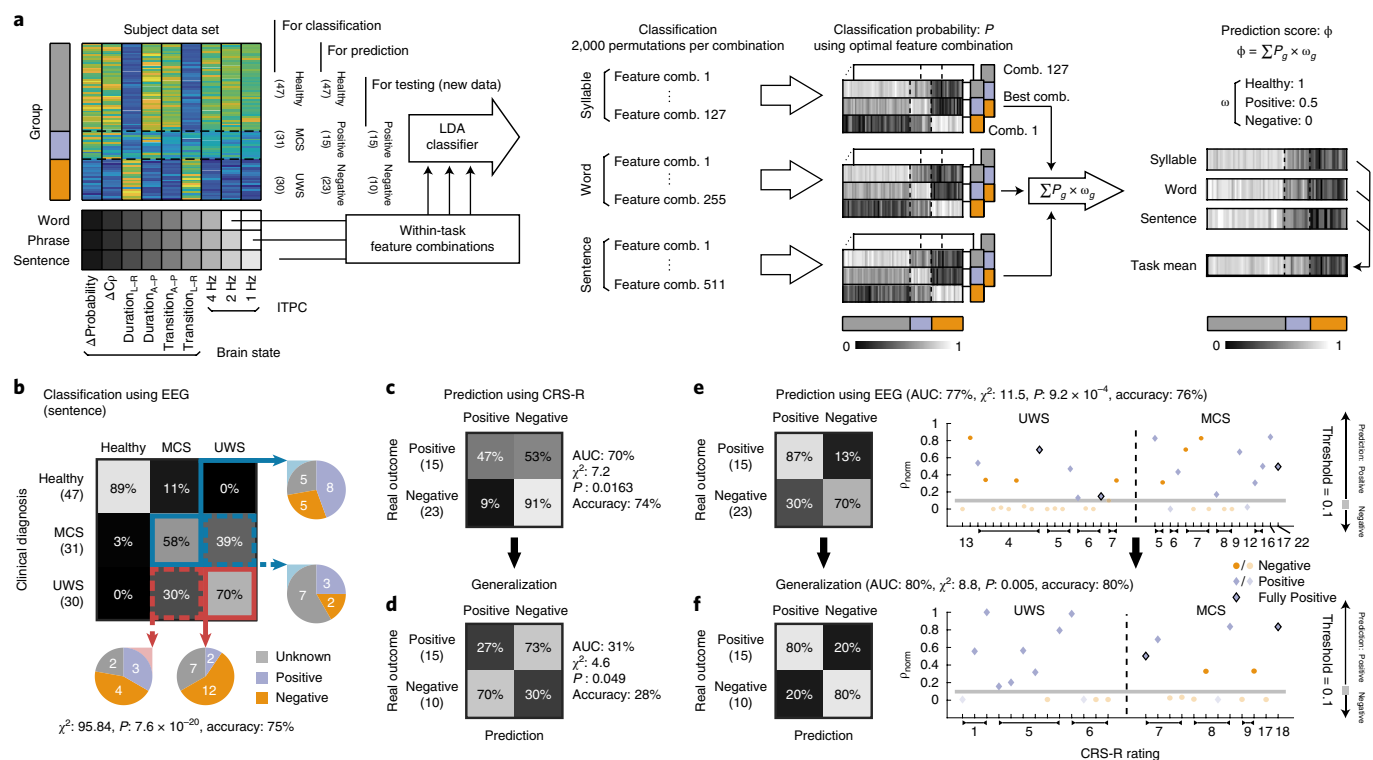
Fig. 5 | Diagnosis and outcome prediction. a, Classification and prediction procedures using the multiple EEG features (for details, see Methods). **b,** The confusion matrix of diagnosed consciousness classification generated by the LDA. The classifier was trained on the EEG metrics derived from the sentence task. The pie chart shows the mismatch between clinical diagnosis of patients and outcomes. **c,** The performance of outcome prediction using CRS-R total scores. Chi-squared test: AUC = 70%, $\chi^2 = 7.2$, $P = 0.016$, accuracy = 74%. **d,** The prognostic validity of the model using CRS-R total scores. The classifier was trained on the dataset of 38 patients from **c** with cross-validation and then tested the generalization ability on the new dataset of 25 patients. Chi-squared test: AUC = 31%, $\chi^2 = 4.6$, $P = 0.049$, accuracy = 28%. **e,** The performance of outcome prediction using EEG (left, chi-squared test: AUC = 77%, $\chi^2 = 11.5$, $P = 9.2 \times 10^{-4}$, accuracy = 76%) and the comparison of prediction and actual outcome of individual patients (right). The dots and diamonds above the threshold (gray line, prediction score = 0.1) represent the patients with predicted positive outcomes, whereas the others represent those with predicted negative outcomes. **f,** The prognostic validity of the model using EEG. The classifier was trained on the old dataset **e** and generalized to the new dataset. Chi-squared test: AUC = 80%, $\chi^2 = 8.8$, $P = 0.005$, accuracy = 80%. **f** shows the same format as **e**. g, group; ω, normalization coefficient.

accuracy in both groups (81% UWS and 71% MCS; individual prediction scores are shown in Fig. 5e, right). Thus, prediction using EEG measures was better than using behavioral observations alone, using either a single CRS-R measure in our study or multiple CRS-R measurements in another study[29].

It is worth noting that, although the number of participants in the dataset for the prediction model was relatively small (38 patients), the best feature combination for the classification consisted only of two or three features, suggesting that it was unlikely that the model was overfitted. Nevertheless, to ensure the reliability of our results, we validated the same dataset using a different classifier, SVM, with cross-validation. This additional analysis confirmed our results, showing a significant predictive accuracy from EEG (Extended Data Fig. 7b, $\chi^2 = 15.4$, $P = 1.6 \times 10^{-4}$, accuracy = 82%, chi-squared test).

Most importantly, to further test the external validity and the generalization ability of our models, we tested them (without retraining) on a new dataset consisting of 25 additional patients (12 MCS and 13 UWS, Supplementary Table 2). The classifier (LDA) trained with the previous dataset using the same feature combinations showed a high predictive accuracy in both outcome-positive and outcome-negative groups across the two sample sets (Fig. 5f, left, $\chi^2 = 8.8$, $P = 0.005$, accuracy = 80%, chi-squared test; individual prediction scores in Fig. 5f, right). However, similar generalization using the classifier trained with the CRS-R total score showed a much lower predictive accuracy (Fig. 5d; $\chi^2 = 4.6$, $P = 0.049$,

accuracy = 28%, chi-squared test). As a control, the classification accuracy was 50% when we shuffled the outcome label of the testing dataset; and the classifier for diagnosis (i.e. MCS versus UWS), once trained with CRS-R scores on the first 38 patients, successfully generalized to the new dataset (Extended Data Fig. 8a; $\chi^2 = 21.3$, $P = 2.7 \times 10^{-6}$, accuracy = 96%, chi-squared test).

Could the higher predictive accuracy with EEG measurements be due to the larger number of EEG features used? To evaluate this, we expanded the behavioral measurements by including the total score and six subscores (auditory, visual, motor, oromotor, communication and arousal) and used the same model to search for the optimal CRS-R feature combination for prognosis. With cross-validation on the first 38 patients, the best prediction was obtained using the visual subscale alone (Extended Data Fig. 8b, left), but again that classifier failed to generalize to the new dataset of 25 patients (Extended Data Fig. 8b, right). Direct comparison of prediction performance between CRS-R or the EEG recording under the word condition (both had seven features) yielded a similar result (Extended Data Figs. 8c,d). Furthermore, we examined the outcome prediction using a standard LDA model without feature selection before classification. The results confirmed the superior generalization ability using EEG metrics than using CRS-R features (Extended Data Fig. 8e).

Finally, the confusion matrix generated by the CRS-R scores showed that an important proportion of outcome-positive patients

(53%, 8/15; three MCS and five UWS) was mis-predicted as outcome negative (Fig. 5c). Compared to the CRS-R, our EEG task-based assessments could potentially contribute to more accurate diagnoses, as the classifier that we constructed using EEG measurements was able to significantly predict future CRS-R dichotomized outcomes (Fig. 5e): six of those eight patients were classified as outcome positive by the EEG-based classifier. Furthermore, we combined the features from both CRS-R and EEG and submitted them to the model. The results showed that, at the word and phrase levels, the classification accuracies by using both EEG and CRS-R were slightly higher than those using only EEG signals (for classification accuracies at all levels, see Extended Data Fig. 9 and Supplementary Table 3). The best feature combination in the classification indeed contained the CRS-R, which might indicate that the prediction model could benefit from the use of both EEG and CRS-R. Taken together, these results confirm the prognostic ability of EEG metrics recorded during language listening and suggest that these EEG measures might complement demographic and behavioral diagnoses in the clinic.

The outcome of the patients in the prediction was based on a single CRS-R, which could lead to misdiagnosis owing to the variability of the level of consciousness over time (e.g. morning versus afternoon within a day and across different days; see ref. [30]). Although we examined a subset of patients ($n = 15$) who had multiple diagnoses before and after EEG recordings and found relatively stable CRS-R scores across days at both group and individual levels (Extended Data Fig. 10), repetitive behavioral measurements (CRS-R) are likely to be important for accurate diagnosis and prediction, and a systematic comparison of the predictive value with multiple EEG measurements and multiple CRS-R scores remains to be explored.

Discussion

We adopted a hierarchical linguistic processing paradigm to test residual consciousness in patients with DOCs. We demonstrated that two EEG-derived neural signals—speech-tracking activity and global dynamic pattern—were associated with the behavioral diagnosis of consciousness and clinical outcomes. This correlation significantly increased along with the increase in language hierarchy. Furthermore, the multiple EEG measurements were robust enough for the prediction of behavioral diagnosis and future outcomes in individual patients. This therefore represents a new approach for clinical use of EEG measures in the diagnosis and prognosis of consciousness in patients with DOCs.

In the past decade, neuroimaging and electrophysiological approaches, including fMRI[14,23,31,32], positron emission tomography[28], EEG[5–7,18] and multimodal imaging[33,34], have been used to examine states of consciousness in unresponsive patients. fMRI has several limitations, including high cost, lack of portability and impossibility of bedside clinical testing, whereas high-density EEG is more feasible to deploy at the patient's bedside and help track individual patients longitudinally.

Previous fMRI experiments showed that the brain spontaneously generates a dynamic series of constantly changing activity and functional connectivity between brain regions[35,36], which contributes to efficient exchanges between neural populations, which suggests that the neural correlates of consciousness could be found in temporally evolving dynamic processes[9,37]. Our study investigated the similar discrete and dynamic states in EEG (referred to as 'microstates') and proposed that the dynamic pattern of scalp potentials reflects the momentary state of global neural activity, which might correspond to the changes in consciousness over time[9,23]. Specifically, we showed that the global brain activity, in particular the duration of L–R maps and the occurrence of A–P maps during language tasks, can significantly differentiate between patients with UWS and patients with MCS at both

the group and individual levels. Furthermore, the capacity to discriminate patients with UWS and patients with MCS by examining the dynamics of brain states increased with task hierarchy, from resting, word and phrase to sentence conditions. The results lend support to the idea that those EEG states are not meaningless recurrent patterns but might indeed separate sensory versus higher-level cognition functions and are associated with the level of consciousness and task performance (e.g. the hierarchical level of language processing)[38,39].

Despite some promising neuroimaging studies[14,32], there have been few reliable active EEG paradigms for assisting diagnosis and prognosis of DOCs[19,28,40]. Here we present the first EEG evidence that speech-tracking neural responses and cortical dynamic patterns are directly associated with multiple levels of speech processing in patients with DOCs. We found that phrase- and sentence-level responses disappear in patients with UWS, suggesting no preservation of deeper-level linguistic processing once consciousness is lost (in agreement with previous studies of anesthesia[41] and sleep[42]). Although these responses drastically reduce in patients with MCS, multivariate and brain state analyses indicated that linguistic structures continue to modulate neural processing, suggesting some degree of deeper-level processing in patients with MCS. Increased variability of cortical dynamics in patients with MCS (Figs. 3d and 4) could explain why phrase- and sentence-rate responses diminished in the phase coherence spectrum in this group (Fig. 2b). Recently, it has been debated whether phrase- and sentence-rate responses reflect neural entrainment to mentally constructed syntactic structures[43] or semantic properties of individual words[44]. Our study, however, did not intend to distinguish semantic and syntactic processing and employed the phrase- and sentence-rate responses as general measures of higher-level linguistic processing. The current result that word-rate responses remain in patients with UWS and MCS, whereas phrase- and sentence-rate responses diminish, actually further confirms that the phrase- and sentence-rate responses reflect deeper levels of speech processing than the word-rate response. In the future, a combination of multiple EEG paradigms, including the present paradigm as well as syntactic and semantic violation paradigms[16,17,45], could facilitate the assessment of language comprehension abilities in individual patients.

Our results demonstrate the diagnostic potential of EEG speech responses. The diagnosed state classification and future outcome prediction models showed that both speech-tracking responses (e.g. ITPC at 2 and 4 Hz) and global dynamic patterns (e.g. Occurrence$_{A–P}$ and ΔC_ρ) were required for the best decoding and prediction accuracy. This suggests that combining different analytical methods could deliver better diagnostic and predictive capabilities. Note that, different from most previous studies showing the prediction of recovery from coma[29], our study demonstrated the recovery prediction of patients from different stages (e.g. UWS, MCS and emergence from MCS (EMCS)). However, we should also note that we lacked a detailed set of consecutive behavioral measurements for each patient during the recovery period. Thus, although we selected the patients whose CRS-R outcome was obtained more than 100 d after EEG assessments, our research does not indicate that EEG signals can precisely predict clinical outcome 100 d ahead of behavior. On the basis of previous studies[46,47], 6 months after DOC onset, 17% of patients with non-traumatic UWS will recover consciousness, and 67% of patients with post-traumatic UWS will recover consciousness. A systematic evaluation of whether and by how many days the EEG model can anticipate behavioral recovery requires further investigation. Nevertheless, there are several reasons why our results engender confidence in the clinical use of EEG at the bedside. First, the current paradigm has been demonstrated to be useful for multiple languages, including Chinese, English and Hebrew[20,42], indicating that it can be applied clinically in different language environments. Second, EEG ITPC and global brain states can

theoretically be measured using fewer electrodes (e.g. 16 channels[48]) and with less expensive EEG systems, while still preserving discriminative power and clinical utility. Third, the timeframe of the paradigm is also feasible for daily bedside examinations at the hospital or at home, as the experiment lasted for less than 20 min when including only two linguistic conditions (word and sentence).

However, there are also several limitations of the current study that should be noted. First, the speech stimuli were synthesized using software; prior studies have demonstrated that personal and meaningful stimuli elicit more robust and reliable responses in patients with brain injuries[49]. Future work could therefore use more personalized speech stimuli, e.g. on topics that are familiar to the patient. Second, the phrase- and sentence-level processing could potentially be enhanced when the speech rate is slowed down, especially for patients with DOCs. Also note that the ITPC signals at 1, 2 and 4 Hz of the healthy controls recorded in the noisy hospital environment were significantly reduced compared to laboratory recordings. Improvements in data acquisition systems and recording conditions might provide higher EEG data quality. Third, although some patients were diagnosed multiple times using EEG recordings, and the diagnostic accuracy of the model seems very high, it is possible that the model misdiagnosed some patients owing to fluctuations in levels of consciousness over time. Indeed, clinicians are required to conduct multiple behavioral assessments of consciousness before they achieve a stable and accurate diagnosis. For example, there is evidence that the number of assessments has a significant effect on clinical diagnosis within 2 weeks, and even within a day[50], and a higher responsiveness during behavioral assessment was found in the morning than in the afternoon[30]. Similarly, the level of consciousness that is evident from a patient's EEG is likely to fluctuate both within and across days. Future work might therefore require multiple EEG sessions from the early stage of coma to the follow-up recovery.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-020-0639-1.

References

1. Giacino, J. T., Fins, J. J., Laureys, S. & Schiff, N. D. Disorders of consciousness after acquired brain injury: the state of the science. *Nat. Rev. Neurol.* **10**, 99–114 (2014).
2. Bruno, M. A., Gosseries, O., Ledoux, D., Hustinx, R. & Laureys, S. Assessment of consciousness with electrophysiological and neurological imaging techniques. *Curr. Opin. Crit. Care* **17**, 146–151 (2011).
3. Giacino, J. T. et al. The minimally conscious state: definition and diagnostic criteria. *Neurology* **58**, 349–353 (2002).
4. Owen, A. M. The search for consciousness. *Neuron* **102**, 526–528 (2019).
5. Strauss, M. et al. Disruption of hierarchical predictive coding during sleep. *Proc. Natl Acad. Sci. USA* **112**, E1353–E1362 (2015).
6. Goldfine, A. M., Victor, J. D., Conte, M. M., Bardin, J. C. & Schiff, N. D. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. *Clin. Neurophysiol.* **122**, 2157–2168 (2011).
7. Sitt, J. D. et al. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain* **137**, 2258–2270 (2014).
8. Dehaene, S. & Changeux, J. P. Experimental and theoretical approaches to conscious processing. *Neuron* **70**, 200–227 (2011).
9. Barttfeld, P. et al. Signature of consciousness in the dynamics of resting-state brain activity. *Proc. Natl Acad. Sci. USA* **112**, 887–892 (2015).
10. Lehmann, D., Ozaki, H. & Pal, I. EEG alpha map series: brain micro-states by space-oriented adaptive segmentation. *Electroencephalogr. Clin. Neurophysiol.* **67**, 271–288 (1987).
11. Michel, C. M. & Koenig, T. EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review. *Neuroimage* **180**, 577–593 (2018).
12. Laureys, S. & Schiff, N. D. Coma and consciousness: paradigms (re)framed by neuroimaging. *Neuroimage* **61**, 478–491 (2012).
13. Schnakers, C. et al. Voluntary brain processing in disorders of consciousness. *Neurology* **71**, 1614–1620 (2008).
14. Owen, A. M. et al. Detecting awareness in the vegetative state. *Science* **313**, 1402 (2006).
15. Braiman, C. et al. Cortical response to the natural speech envelope correlates with neuroimaging evidence of cognition in severe brain injury. *Curr. Biol.* **28**, 3833–3839 (2018).
16. Rohaut, B. et al. Probing ERP correlates of verbal semantic processing in patients with impaired consciousness. *Neuropsychologia* **66**, 279–292 (2015).
17. Coleman, M. R. et al. Do vegetative patients retain aspects of language comprehension? Evidence from fMRI. *Brain* **130**, 2494–2507 (2007).
18. Faugeras, F. et al. Event related potentials elicited by violations of auditory regularities in patients with impaired consciousness. *Neuropsychologia* **50**, 403–418 (2012).
19. Engemann, D. A. et al. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain* **141**, 3179–3192 (2018).
20. Ding, N., Melloni, L., Zhang, H., Tian, X. & Poeppel, D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat. Neurosci.* **19**, 158–164 (2016).
21. Kiefer, M. The N400 is modulated by unconsciously perceived masked words: further evidence for an automatic spreading activation account of N400 priming effects. *Brain Res. Cogn. Brain Res.* **13**, 27–39 (2002).
22. Batterink, L. & Neville, H. J. The human brain processes syntax in the absence of conscious awareness. *J. Neurosci.* **33**, 8528–8533 (2013).
23. Demertzi, A. et al. Human consciousness is supported by dynamic complex patterns of brain signal coordination. *Sci. Adv.* **5**, eaat7603 (2019).
24. Khanna, A., Pascual-Leone, A., Michel, C. M. & Farzan, F. Microstates in resting-state EEG: current status and future directions. *Neurosci. Biobehav/Rev.* **49**, 105–113 (2015).
25. Koenig, T. et al. Millisecond by millisecond, year by year: normative EEG microstates and developmental stages. *Neuroimage* **16**, 41–48 (2002).
26. Brodbeck, V. et al. EEG microstates of wakefulness and NREM sleep. *Neuroimage* **62**, 2129–2139 (2012).
27. Britz, J., Van De Ville, D. & Michel, C. M. BOLD correlates of EEG topography reveal rapid resting-state network dynamics. *Neuroimage* **52**, 1162–1170 (2010).
28. Stender, J. et al. Diagnostic precision of PET imaging and functional MRI in disorders of consciousness: a clinical validation study. *Lancet* **384**, 514–522 (2014).
29. Scarpino, M. et al. Prognostic value of post-acute EEG in severe disorders of consciousness, using American Clinical Neurophysiology Society terminology. *Neurophysiol. Clin.* **49**, 317–327 (2019).
30. Wannez, S. et al. The repetition of behavioral assessments in diagnosis of disorders of consciousness. *Ann. Neurol.* **81**, 883–889 (2017).
31. Wu, X. et al. Intrinsic functional connectivity patterns predict consciousness level and recovery outcome in acquired brain injury. *J. Neurosci.* **35**, 12932–12946 (2015).
32. Monti, M. M. et al. Willful modulation of brain activity in disorders of consciousness. *N. Engl. J. Med.* **362**, 579–589 (2010).
33. Chennu, S. et al. Brain networks predict metabolism, diagnosis and prognosis at the bedside in disorders of consciousness. *Brain* **140**, 2120–2132 (2017).
34. Di Perri, C. et al. Neural correlates of consciousness in patients who have emerged from a minimally conscious state: a cross-sectional multimodal imaging study. *Lancet Neurol.* **15**, 830–842 (2016).
35. Zalesky, A., Fornito, A., Cocchi, L., Gollo, L. L. & Breakspear, M. Time-resolved resting-state brain networks. *Proc. Natl Acad. Sci. USA* **111**, 10341–10346 (2014).
36. Hutchison, R. M. et al. Dynamic functional connectivity: promise, issues, and interpretations. *Neuroimage* **80**, 360–378 (2013).
37. Tagliazucchi, E. et al. Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J. R. Soc. Interface* **13**, 20151027 (2016).
38. Cojan, Y. et al. The brain under self-control: modulation of inhibitory and monitoring cortical networks during hypnotic paralysis. *Neuron* **62**, 862–875 (2009).
39. Barttfeld, P. et al. Distinct patterns of functional brain connectivity correlate with objective performance and subjective beliefs. *Proc. Natl Acad. Sci. USA* **110**, 11577–11582 (2013).
40. Claassen, J. et al. Detection of brain activation in unresponsive patients with acute brain injury. *N. Engl. J. Med.* **380**, 2497–2505 (2019).
41. Davis, M. H. et al. Dissociating speech perception and comprehension at reduced levels of awareness. *Proc. Natl Acad. Sci. USA* **104**, 16032–16037 (2007).

42. Makov, S. et al. Sleep disrupts high-level speech parsing despite significant basic auditory processing. *J. Neurosci.* **37**, 7772–7781 (2017).
43. Jin, P. et al. Low-frequency neural activity reflects rule-based chunking during speech listening. *eLife* **9**, e55613 (2020).
44. Frank, S. L. & Yang, J. Lexical representation explains cortical entrainment during speech comprehension. *PLoS ONE* **13**, e0197304 (2018).
45. Steppacher, I. et al. N400 predicts recovery from disorders of consciousness. *Ann. Neurol.* **73**, 594–602 (2013).
46. Giacino, J. T. et al. Practice guideline update recommendations summary: disorders of consciousness: report of the Guideline Development, Dissemination, and Implementation Subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National Institute on Disability, Independent Living, and Rehabilitation Research. *Neurology* **91**, 450–460 (2018).
47. Giacino, J. T. et al. Comprehensive systematic review update summary: disorders of consciousness: report of the Guideline Development,
Dissemination, and Implementation Subcommittee of the American Academy of Neurology; the American Congress of Rehabilitation Medicine; and the National Institute on Disability, Independent Living, and Rehabilitation Research. *Neurology* **91**, 461–470 (2018).
48. Lehmann, D. et al. EEG microstate duration and syntax in acute, medication-naive, first-episode schizophrenia: a multi-center study. *Psychiatry Res.* **138**, 141–156 (2005).
49. Perrin, F., Castro, M., Tillmann, B. & Luaute, J. Promoting the use of personally relevant stimuli for investigating patients with disorders of consciousness. *Front. Psychol.* **6**, 1102 (2015).
50. Cortese, M. D. et al. Coma recovery scale-r: variability in the disorder of consciousness. *BMC Neurol.* **15**, 186 (2015).

Methods

Participants. The study protocol was approved by the Ethical Committee of Huashan Hospital of Fudan University (approval no. HIRB-2014-281), and informed consent was obtained from all healthy participants and caregivers of all patients. All patients were native Mandarin Chinese speakers. No statistical methods were used to predetermine sample sizes, but our sample sizes were similar to those reported in previous publications[18,34,51].

Twenty-seven healthy individuals participated in the first EEG study investigating how top-down attention modulates speech-tracking activity (experimental details shown in the below description; 15 males; mean age = 23.9 years; range = 20–30 years). Five participants were excluded from the final analyses owing to poor quality of EEG data (for 22 participants used in the final analysis: 12 males; mean age = 23.73 years; range = 20–30 years).

For the clinical study, the analyses were based on usable EEG data acquired from patients with DOCs during July 2016 and June 2019. All patients had been diagnosed with MCS or UWS according to the Chinese versions of the CRS-R[52,53] and GCS[54]. EEG was recorded in patients who had not received sedation (mostly midazolam) within the previous 24 h, to minimize the influence of drugs on spontaneous brain activity and arousal levels.

First, a total of 93 patients were recruited from July 2016 to October 2018. Because some patients were followed-up several times, the final dataset included 133 resting-state recordings from 89 patients and 132 task-related recordings from 92 patients. Owing to contamination of EEG data from environmental noise and extreme body movements, 62 resting-state recordings from 50 patients and 54 task-related recordings from 43 patients were discarded (Extended Data Fig. 1). Data from a final total of 70 patients were used in the present study, and the etiologies of these patients were stroke (36, 51.43%), traumatic brain injury (31, 44.29%) and anoxia (3, 4.29%) (see Supplementary Table 1 for details). Thus, the final dataset included 71 resting-state recordings from 54 patients with DOCs (48 males; mean age = 49.3 years; range = 17–75 years, 41 MCS recordings and 30 UWS recordings) and 78 linguistic task-related recordings from 60 patients with DOCs (52 males; mean age = 47.8 years; range = 9–68 years). The number of recordings varied between patients.

To test the external validity and generalization ability of the prediction model, another EEG dataset acquired from October 2018 to June 2019 was included (Extended Data Fig. 1; 25 recordings, 12 MCS and 13 UWS, 13 males, mean age = 39.9 years, range = 18–69 years; detailed information in Supplementary Table 2).

We also recruited 61 healthy volunteers from local communities as a control group (20 males, mean age = 31.3 years, range = 22–65 years). After checking the data quality, there were 34 usable resting-state recordings (eight males, mean age = 25.8 years, range = 22–50 years) and 47 usable task-related recordings (17 males, mean age = 31.1 years, range = 22–58 years).

Blinding. Both the attentional and clinical studies were double blind (both participant and experimenter blinding was instituted). Neither the individuals who participated in the studies nor the experimenters who collected data knew the real purpose of this research. Both the interviewees and interviewers obtaining the follow-up assessment were blinded to the behavioral and EEG-based diagnosis in the hospital. In addition, the experimenters who collected EEG data were not involved in data analyses.

Specifically, in the clinical study, the behavioral measurement and EEG and MRI data analyses were conducted by two separated groups of experimenters in a completely blinded fashion. P.G. and Y.J., from the Institute of Neuroscience, analyzed and modeled the EEG and MRI data and were blinded to the measurements of patient outcomes. X.W., D.Z. and their colleagues from Huashan Hospital performed the clinical evaluation of patient behavior and were not involved in the EEG and MRI data analyses.

Behavioral classification and evaluation of consciousness. Patients were categorized into an MCS group and an UWS group according to CRS-R-based behavioral evaluations[55,56].

Patients diagnosed with UWS were awake but showed no behavioral signs of consciousness. They could open their eyes, had basic reflexes and woke up or fell asleep at various intervals. Patients diagnosed with MCS had partial preservation of consciousness, with the presence of subtle but reproducible signs of consciousness. Moreover, patients with MCS can be subcategorized into two distinct subgroups based on the complexity of their behaviors. MCS+ describes high-level behavioral responses and is determined by the presence of command following, intelligible verbalization or nonfunctional communication. MCS− describes low-level behavioral responses and is determined by the presence of visual pursuit, localization of noxious stimulation or contingent behavior related to environmental stimuli (such as smiling or crying in response to the linguistic or visual content of emotional stimuli). Finally, patients who were able to functionally communicate and/or use different objects were diagnosed with EMCS. The classification and behavioral ratings were performed by experienced doctors on the same day of testing, usually before EEG recording.

Chinese materials. The Chinese materials in the auditory stream in the two studies were adapted from previous work[20,53]; auditory stimuli had one (word),

two (word and phrasal) or three (word, phrasal and sentential) linguistic levels. The speech materials were synthesized using a free online text-to-speech engine (http://ai.baidu.com/tech/speech/tts).

Word stimuli. First, 50 four-word sentences with an np–vp structure were predefined (Supplementary Table 4). The 200 words were submitted to the online text-to-speech engine individually to generate their pronunciations. Subsequently, 20 64-word sequences were generated using these words. For each sequence, 64 unique words were randomly selected from the 200-word speech library, concatenated in a row and adjusted manually for their relative temporal spacing. Finally, the duration of each word was adjusted to 250 ms. Thus, the duration of each sequence was 16 s.

Phrase stimuli. For the 50 four-word sentences used in the word condition, 50 noun phrases were chosen to form the phrase library (Supplementary Table 4). Thirty-two phrases were randomly selected from the library and connected for each 16-s phrase sequence. A total of 20 sequences were generated. To avoid liaison in phrasal pronunciation, the speech of every phrase sequence was synthesized at the word level but not at the phrasal level.

Sentence stimuli. Similarly to the sequences used in the word condition, 16 sentences were randomly chosen from the 50 four-word sentences (np–vp) and concatenated together to form a 16-s sentence sequence.

Organization of stimuli in the tasks. In the sentence condition, three levels of semantic hierarchies were used, including single-word frequency at 4 Hz, phrasal frequency at 2 Hz and sentential frequency at 1 Hz. The phrase condition included only word and phrasal levels. The word condition included only the 4-Hz word frequency.

For each condition of every recording, 30 sequences were randomly selected from respective 20 pre-synthesized candidates and connected to form a 480-s speech stream, without any additional blanks between them.

Experimental design of the attentional study. The tasks were conducted in a sound-attenuated chamber and performed using the Psychtoolbox in MATLAB (R2015b, MathWorks).

The attentional experiment involved a full factorial design with two factors: attention (two levels, attend to or ignore) and linguistic condition (two levels, word or sentential stimuli). Thus, there were four blocks in total, with different task conditions (attend to word: attend to word audio while ignoring the simultaneous visual task; attend to sentence: attend to sentential audio while ignoring the simultaneous visual task; ignore word: attend to visual task while ignoring simultaneous word audio; ignore sentence: attend to visual while ignoring simultaneous sentential audio). Participants were asked to either attend to or ignore a visual attention task in separate blocks, with the simultaneous presentation of 8 min of Chinese speech material (Fig. 1b). The auditory stream was adapted from previous work[20,53] and consisted of Chinese monosyllabic words, which had either one (word) or three (word, phrase and sentence) linguistic levels (Fig. 1a and Supplementary Table 4).

The auditory stream started 20 s after the onset of the first visual trial in each block and ended before the offset of the last visual trial and was delivered through two loudspeakers next to the monitor ~80 cm away from the participants' ears at ~65 dB SPL. In general, the auditory stream in each condition was composed of 30 16-s-long Chinese sequences with no noticeable gap between them. The audio in each block was played for 8 min without a break.

Visual stimuli were presented on a 23-inch LCD monitor ~60 cm from the participants. A fixation cross was presented at the start of visual trials for 1.5 s, followed by a statement composed of a shape, a greater-than-or-smaller-than sign and a number. After 4 s, a shape matrix was presented. The shape matrix consisted of a random number (24 ± 2) of five shapes (isosceles right triangle, equilateral triangle, square, pentagon and hexagon) in four colors (blue, green, yellow and magenta). In the visual attention condition, the participants needed to respond within 12 s of the presentation of this matrix by pressing the left or right arrow key to indicate whether the prospective statement was correct (the number of a specific shape in the matrix was greater or smaller than the given number). The assignments of the keys (agree or disagree with the cue) were counterbalanced across participants. A 2.5-s visual feedback was given as soon as the response was made to indicate whether the response was correct or incorrect. The next trial began after a 3–6.1-s inter-trial interval. There were 32 trials in each block, which lasted ~10 min in total.

In the visual–ignore condition, the shape matrix was always presented for 7 s in each trial because the participants did not need to respond. The participants were asked to attend to the audio while ignoring the visual trials. After each block, participants were asked to decide whether the words or sentences in a testing list had been played or not.

The order of the four task blocks was randomized and counterbalanced across participants. EEG data were collected continuously and segmented into 16-s epochs. To obtain clean data, we excluded trials with noise, extreme movement and eye blinks. The mean trials used in attend to word, attend to sentence, ignore word and ignore sentence conditions were 28.7, 28.2, 28.8 and 27.4, respectively.

Note that the analysis of behavioral performance of the visual task showed no significant difference between the attend to sentence and attend to word conditions (accuracy: 75.71% ± 1.79% versus 76.56% ± 2.52%, $P = 0.796$; reaction time: $6.99 ± 0.24$ s versus $6.84 ± 0.33$ s, $P = 0.518$; paired-sample t-test).

Experimental design of the clinical study. Tests were conducted in hospital wards or similar places and performed using the Psychtoolbox in MATLAB (R2015b, MathWorks).

First, a 5-min resting EEG was measured at the beginning of each recording session. After a 2-min rest period, three blocks were then presented. These blocks corresponded to three 8-min Mandarin Chinese audio sequences with different semantic levels: word, phrase and sentence conditions (Fig. 2a). Before each task block, a brief introduction was played to instruct the participant to be quiet and listen carefully, which was also synthesized using the same online text-to-speech engine. To reduce environmental noise, the acoustic stimuli were delivered through headphones at ~65 dB SPL, over which participants wore an additional pair of sound-shielding earmuffs.

The order of task conditions was randomized and counterbalanced across participants, controlled by a random function in MATLAB. In addition, the order of stimuli in each task condition was also shuffled across participants.

EEG recording. In the attentional study, data were collected using a 64-channel EEG recording system (actiCHamp, Brain Products). In the clinical study, a 257-channel system (GES 300 or GES 400, Electrical Geodesics) and a 257-channel electrode cap (HCGSN 257-channel net cap, Electrical Geodesics) were used. EEG signals were referenced online to the FCz (the attentional study) or Cz (the clinical study) electrode. The impedance of all electrodes was kept below 5 kΩ (the attentional study) or 20 kΩ (the clinical study). The EEG signals were sampled at 1,000 Hz.

Criteria of EEG data quality. Considering the noisy recording environment and abundant artifacts caused by patients' involuntary movements, we first examined the proportion of bad duration during each recording by applying the following criteria on each data channel (for the ITPC analysis, the number of data channels was 257; for the brain state analysis, the number was 204, because electrodes placed on the cheeks and neck were excluded firstly):

1. Gradient criterion: the instant voltage change exceeds the maximal allowed step, which is 30 μV ms⁻¹;
2. Max–Min criterion: the absolute difference between the maximal and minimal voltage within every 200-ms sliding window exceeds 120 μV, and the sliding step is 10 ms;
3. Amplitude criterion: the absolute voltage value exceeds 100 μV;
4. Low-activity criterion: the absolute difference between the maximal and minimal voltage within every 100-ms sliding window is smaller than 1 μV, and the sliding step is 10 ms.

Data points that met any of these criteria were marked as artifacts, and the 200-ms periods both before and after each artifact were marked as bad intervals. Channels with a bad duration longer than 20% of the total recording length were marked as bad channels. Recordings in which the number of bad channels exceeded 70 were discarded.

Data pre-processing. For the ITPC analysis, EEG data were pre-processed using BrainVision Analyzer (2.0.1, Brain Products) as follows: data were band-pass filtered (0.1–40 Hz) with a notch filter (50 Hz) firstly. Then, channels were semi-automatically inspected and bad ones were interpolated. Next, data were re-referenced to the common average of signals from all EEG channels, and an independent component analysis (ICA) was performed to remove blinks and eye movements. Finally, data were segmented to 16-s epochs and downsampled to 50 Hz.

For the brain state analysis, EEG data were pre-processed in the EEGLAB toolbox (version 14.1.1) as follows: the electrodes placed on the cheeks and on the neck were removed firstly and data of the maintained 204 electrodes were band-pass filtered (0.2–40 Hz). Then, channels were semi-automatically inspected and bad channels were interpolated before and after ICA. Next, ICA was performed to remove blinks and eye movements, and data were segmented into 2-s epochs and bad epochs were manually removed. Finally, data were re-referenced and band-pass filtered again (2–20 Hz).

Phase coherence analysis and multivariate pattern analysis. The single-trial EEG data were transformed into the frequency domain using discrete Fourier transform (DFT) without additional smoothing windows. The DFT coefficient was denoted as $X_k(f)$ for the kth trial ($k = 1, 2, …, n$), and the phase information was $A_k(f) = \angle X_k(f)$. The ITPC was defined as:

$$ITPC(f) = 1/n \sum_{k=1}^{n} (\cos(A_k(f))^2 + 1/n \sum_{k=1}^{n} (\sin(A_k(f))^2$$

Binary classifiers were used to discriminate different participant groups. Because there were three groups (healthy controls, patients with MCS and

patients with UWS), the LDA was trained for pairwise classifications at each target frequency under each task. The decoding was implemented as follows: 1) the input features were the 257 ITPC values at all EEG channels; 2) for each comparison (ITPC values of two groups at one frequency in one task condition), 4/5 participants were randomly chosen as the training set and the other 1/5 was the testing set; 3) a five-fold cross-validation was applied on the training set—that is, for each fold, the classifier was fit on 4/5 participants and validated on 1/5 of the training set; 4) the classification performance was computed as the sum of the area under the receiver operating characteristic curve (AUC), based on the probabilistic classification of the independent testing set; and 5) steps 2–4 were repeated 100 times to produce the mean classification AUC for these two groups at each frequency for each condition.

Brain state analysis. Brain state analysis was performed using MicrostateAnalysis (version 0.3, software free at http://www.thomaskoenig.ch/index.php/software/microstates-in-eeglab/). For each condition, we computed EEG map topographies at the time of global field power peaks at the individual level, disregarding map polarity, and identified the predominant brain state maps using k-means clustering. Four maps were selected as the optimal number of brain states, which was determined using the cross-validation criterion and global explained variance. According to the best assessments of global explained variance and stability, we defined the group-averaged maps using the healthy controls as template maps of each condition.

With the template maps, we analyzed brain state probability, mean duration, mean occurrence and mean transition probability of the healthy controls and patient groups. To summarize the spatial information of the four predominant brain states in a single participant, we calculated a probability-weighted spatial correlation difference, $\Delta\mathbf{C}_\rho$. The template maps were further classified into the two following categories: the A–P map, which was created by averaging template maps 'A' and 'B', and the L–R map, which was created by averaging template maps 'C' and 'D'. For each participant, the spatial correlation of each given map corresponds to the spatial Pearson's correlation between the given tested map and the template maps (A–P and L–R maps) averaged from healthy participants. The difference in the spatial correlation with the two template maps (ΔC) indicates the similarity of the four maps in each patient compared to the healthy controls[26].

Each difference of spatial correlation corresponds to the spatial Pearson's correlation, which was calculated as follows:

$$\Delta C = \frac{\sum_i^n (\mathbf{I}_i \cdot \mathbf{V}_{AP,i})}{\sqrt{\sum_i^n \mathbf{I}_i^2} \cdot \sqrt{\sum_i^n \mathbf{V}_{AP,i}^2}} - \frac{\sum_i^n (\mathbf{I}_i \cdot \mathbf{V}_{LR,i})}{\sqrt{\sum_i^n \mathbf{I}_i^2} \cdot \sqrt{\sum_i^n \mathbf{V}_{LR,i}^2}}$$

ΔC is the difference of spatial correlation of the two template maps and n is the number of electrodes. $\mathbf{I}$ is the measured voltage of the individual map and .. is the measured voltage of the A–P template map. $\mathbf{V}_{LR}$ is the measured voltage of the L–R template map and i is the electrode i.

Accordingly, we calculated the probability-weighted spatial correlation difference as follows:

$$\Delta\mathbf{C}_\rho = \sum_{k=1}^{4} \Delta\mathbf{C}_k \cdot \rho_k$$

where ρ is the probability of a given map (Fig. 3c, Maps A, B, C and D) and k denotes the map k.

Diagnosis and prediction analysis. We first used classification analysis to identify the consciousness states of individuals. The exclusion criteria for the classification dataset were as follows: 1) patients with a DOC duration shorter than 3 months; 2) patients who had received deep brain stimulation in the last 120 d; and 3) patients with an unstable level of consciousness caused by an unexpected disease. After exclusion, data from a final total of 47 healthy controls, 31 patients with MCS and 30 patients with UWS were included. These feature combinations were used to train three-class LDA classifiers to discriminate between healthy controls, patients with MCS and patients with UWS. There were, in total, 893 possible feature combinations when using the EEG metrics from all three levels of the language task (the sentence condition contains three ITPC metrics and six brain state metrics, which produces $2^9 - 1 = 511$ feature combinations; the phrase and word condition produces 255 and 127 feature combinations, respectively). For each classification, to avoid model overfitting, only one out of 893 feature combinations was selected and submitted to the model. All steps were cross-validated (leave-one-subject-out). A classifier with regularization first searched for the optimal feature combination within each task and calculated the classification probability for each individual participant. To avoid model overfitting, only those selected feature combinations were entered in the final LDA. The regularized version of LDA was used by estimating covariance matrices. Cross-validation relied on the leave-one-subject-out method with 108 permutations. Considering the bias effect of unequal class sizes in the LDA classification, we did not rely on uneven prior probabilities for class sizes but assumed that all classes had the same number of samples[57,58]. We thus randomly chose 29 samples from the healthy controls and MCS group individually to match the sample number of the UWS group.

This sampling process was repeated 2,000 times. For the model fitting of each feature combination, the accuracy of the classification was averaged over the 2,000 permutations. The mean accuracies of classifications allowed us to determine the optimal feature combination. For individual particpants, under the optimal feature combination, the maximum probability during the 2,000 permutations decided which group a given participant was classified to.

Subsequently, to predict patient outcome, we selected the patients for whom we had behavioral measurements more than 100 d after EEG assessments. Each patient was first labeled as showing a positive or negative outcome. The clinical diagnoses of patients could be subcategorized into four different subclasses with proposed ascending levels of consciousness, namely UWS/vegetative state, MCS−, MCS+ and EMCS[12,55]. Here, positive outcome was defined as any advance in the transition of clinical categorization during follow-up, whereas negative outcome was defined as stasis or retrogress in the transition. Data from a total of 38 patients were used in the prediction analysis, including 15 patients with positive outcomes (ten MCS and five UWS) who became fully awakened or exhibited significantly improved behavioral signs in the follow-up measurements, and 23 patients with negative outcomes (seven MCS and 16 UWS). Similarly, the process of consciousness state classification was used to classify the outcomes of individuals. The classifier was built using three classes—healthy controls, patients with positive outcomes and patients with negative outcomes—>that corresponded to normalized coefficients (ω) of 1 (healthy control), 0.5 (positive outcome) and 0 (negative outcome). Under the optimal feature combination, we applied the normalized coefficients to classification probabilities (P) and then defined the weighted sum as the predicted score ($\phi = \sum P \times \omega$) of task-single prediction. Because the task-single predictive scores varied across task conditions for individual participants, the task-mean prediction was used and defined as the average of three scores from the three task conditions.

The external validation (generalization ability) of the classifier was examined on the new dataset (25 patients), which contained 15 patients with positive outcomes (five MCS and ten UWS) and ten patients with negative outcomes (seven MCS and three UWS) (Fig. 5d,f and Extended Data Fig. 8). The classifier (LDA) for outcome prediction using EEG metrics was first trained on the dataset of 38 patients with the cross-validation procedure within the dataset and then tested on the new dataset of 25 patients. The similar analysis procedures for outcome prediction and generalization were performed using behavioral features: CRS-R total scores and six subscales (1, Auditory (0–4); 2, Visual (0–5); 3, Motor (0–6); 4, Oromotor (0–3); 5, Communication (0–2); 6, Arousal (0–3)). For the CRS-R total score classifier, we also computed chance performance by repeating the same generalization 100 times using shuffled outcome labels of the testing dataset.

The direct comparisons of outcome prediction and its generalization between EEG and CRS-R scores were also examined by using LDA without searching for the optimal feature combinations (Extended Data Fig. 8e). The input features for training the two-class classifier were values of the EEG or CRS-R (total score and six subscales) metrics. The labels corresponding to each participant (samples) were either outcome positive or outcome negative. We used five-fold cross-validation in all the task conditions, with random samples allocated to folds stratified by labeled class.

We plotted the receiver operating characteristic curves of predicted scores to carry out AUC measurements, which were used to estimate the abilities of task-single and task-mean to prognosticate outcomes. The optimal threshold for prognosticating outcomes was determined by the point with maximal sum of sensitivity and specificity on the receiver operating characteristic curve. The corresponding predictive threshold was equal to 0.1 after normalization (Fig. 5e,f). Patients with predicted scores that were higher than the threshold were identified as having a positive outcome. The prediction accuracy was calculated by comparing the predicted labels of patients and their actual outcome in the follow-up diagnosis.

MRI data acquisition and extracting the volume of brain injure. Structural MRI images were acquired from the 27 patients (16 males; mean age = 44.6 years; range = 9–68 years) on the same day as the EEG recordings. MRI data were collected by a 3-Tesla MRI scanner ($n = 21$, Siemens Magnetom Verio, using turbo spin-echo sequence) or a 1.5-Tesla MRI scanner ($n = 6$, GE Signa EXCITE Twinspeed Zoom, using a fast spin-echo sequence). By checking the structural contrast, head motion of images and reliability of lesion detection[59], we identified hyperintensity of brain lesions in T2-weighted image data from 21 patients, hypointensity of brain lesions in fluid-attenuated inversion recovery data from four patients and hypointensity of brain lesions in T1-weighted image data from two patients. The areas of the lesions were drawn on each image manually under doctors' instruction. The volume of the lesions was calculated by summing all voxels within the lesion areas across all slices and multiplying by the voxel size[60,61]. We then evaluated the correlation between $\Delta C\rho$ and lesion volume using Pearson's correlation. The analyses of MRI data were performed using MATLAB (2017b, MathWorks) and ITK-SNAP (3.8).

Statistics. For the ITPC analyses, the significance tests were applied to individual participants and group participants. At the individual level, the one-sided exact test was recruited. For ITPC between 0.2 and 5 Hz, 77 frequencies were used in total (1/16 Hz for each bin). The null hypothesis was that the response phase is not synchronized to the stimulus and the ITPC at the target frequency is not significantly larger than those in neighboring frequencies. Thus, the statistical

significance (exact P) of the response at a target frequency is the probability that the target frequency response differs from the null distribution (nontarget frequencies; numbers of nontarget frequencies within participants under the three conditions: 76 frequencies for word, 75 for phrase and 74 for sentence). At the group level, the chance-level phase coherence for each target frequency was the average of its neighboring nontarget frequencies (four bins on each side of each target frequency, which was equivalent to 0.25 Hz). The statistical significance is the difference between the response at a target frequency and the response at its neighbors (one-sided paired-sample t-test). For classification results, one-sided one-sample t-tests were applied to examine the significances of decoding performance, comparing with the chance level of 0.5.

In the brain state analysis, for the probability of four maps, the main effect of group under each condition was examined using MANOVA. A repeated-measures ANOVA was used to evaluate the group effect in each task condition, in which EEG metrics and task (three levels: word, phrase and sentence) were treated as repeated measures, whereas group (three levels: healthy control, MCS and UWS) was the between-participants factor. For pairwise comparisons between the three groups, one-way ANOVA tests (Bonferroni corrected) were applied to all EEG metrics in each condition. In addition, for follow-up patients, Friedman tests were used to test the changes in brain state parameters between their first and last recordings. Chi-squared tests (Fisher's exact test) were used to estimate the statistical significance of the match between the classified/predicted labels and the diagnosed labels in the classification and prediction analyses. All data distributions were assumed to be normal, but this was not formally tested.

Data availability
The data that support the findings of this study are available from the corresponding authors upon reasonable request.

Code availability
EEG data analyses were performed in BrainVision Analyzer, the freely available toolbox EEGLAB and MicrostateAnalysis in combination with custom MATLAB scripts. The software code that support the findings of this study is available from the corresponding authors upon reasonable request.

References

51. Edlow, B. L. et al. Early detection of consciousness in patients with acute severe traumatic brain injury. *Brain* **140**, 2399–2414 (2017).
52. Giacino, J. T., Kalmar, K. & Whyte, J. The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. *Arch. Phys. Med. Rehabil.* **85**, 2020–2029 (2004).
53. Ding, N. et al. Characterizing neural entrainment to hierarchical linguistic units using electroencephalography (EEG). *Front. Hum. Neurosci.* **11**, 481 (2017).
54. Teasdale, G. & Jennett, B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* **2**, 81–84 (1974).
55. Bruno, M. A., Vanhaudenhuyse, A., Thibaut, A., Moonen, G. & Laureys, S. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. *J. Neurol.* **258**, 1373–1384 (2011).
56. Wannez, S. et al. Prevalence of coma-recovery scale-revised signs of consciousness in patients in minimally conscious state. *Neuropsychol. Rehabil.* **28**, 1350–1359 (2018).
57. Büyüköztürk, Ş. & Çokluk Bökeoğlu, Ö. Discriminant function analysis: concept and application. *Eurasian J. Educ. Res.* **8**, 73–92 (2008).
58. Dixon, S. J. et al. Application of classification methods when group sizes are unequal by incorporation of prior probabilities to three common approaches: application to simulations and mouse urinary chemosignals. *Chemometrics Intell. Lab. Syst.* **99**, 111–120 (2009).
59. Geurts, B. H., Andriessen, T. M., Goraj, B. M. & Vos, P. E. The reliability of magnetic resonance imaging in traumatic brain injury lesion detection. *Brain Inj.* **26**, 1439–1450 (2012).
60. Pialat, J. B. et al. Evolution of lesion volume in acute stroke treated by intravenous t-PA. *J. Magn. Reson. Imaging* **22**, 23–28 (2005).
61. Rosso, I. M. et al. Reduced amygdala volumes in first-episode bipolar disorder and correlation with cerebral white matter. *Biol. Psychiatry* **61**, 743–749 (2007).

Acknowledgements

Author contributions

L.W., X.W., S.D., J.Z. M.P. and N.D. conceptualized the study. L.W. and P.G. designed the experiments. D.Z., Z.Q., J.T., H.T., J.J., Y.W. and L.X. collected the data. D.Z., Z.Q., J.T., H.T., X.W. and Y.M. performed clinical ratings on patients. P.G. and Y.J. analyzed the data. L.W., X.W., S.D., M.P., P.G., Y.J. and N.D. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-020-0639-1.

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-020-0639-1.

Correspondence and requests for materials should be addressed to X.W. or L.W.

Peer review information *Nature Neuroscience* thanks Damian Cruse and Ziv Williams for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Dataset 2016.7 - 2018.10

132 recordings from **92** patients
72 from **47** MCS
60 from **50** UWS

↓

Data quality check — *Rejected* → **54** recordings from **43** patients
30 from **22** MCS
24 from **22** UWS

Passed

↓

78 recordings from **60** patients
42 from **30** MCS
36 from **33** UWS → **Brain state & ITPC analyses**

Yes

↓

DOC > 3 months — *No* / *Rejected* → **17** recordings from **16** patients
11 from **10** MCS
6 from **6** UWS

Yes

↓

61 recordings from **51** patients
31 from **23** MCS
30 from **28** UWS → **Diagnosis analysis**

Yes

↓

Follow-up after 3 months — *No* / *Rejected* → **23** recordings from **22** patients
14 from **13** MCS
9 from **9** UWS

Yes

↓

38 recordings from **38** patients
15 outcome-positive
23 outcome-negative → **Outcome prediction analysis**

Dataset 2018.10 - 2019.6

36 recordings from **21** patients
17 from **12** MCS
19 from **13** UWS

↓

Data quality check

↓

27 recordings from **19** patients
13 from **10** MCS
14 from **12** UWS

Yes

↓

DOC > 3 months

Yes

↓

27 recordings from **19** patients
13 from **10** MCS
14 from **12** UWS

Yes

↓

Follow-up after 3 months

Yes

↓

Generalization →

25 recordings from **17** patients
15 outcome-positive
10 outcome-negative

Extended Data Fig. 1 | The flowchart showing patients selection in data analysis.
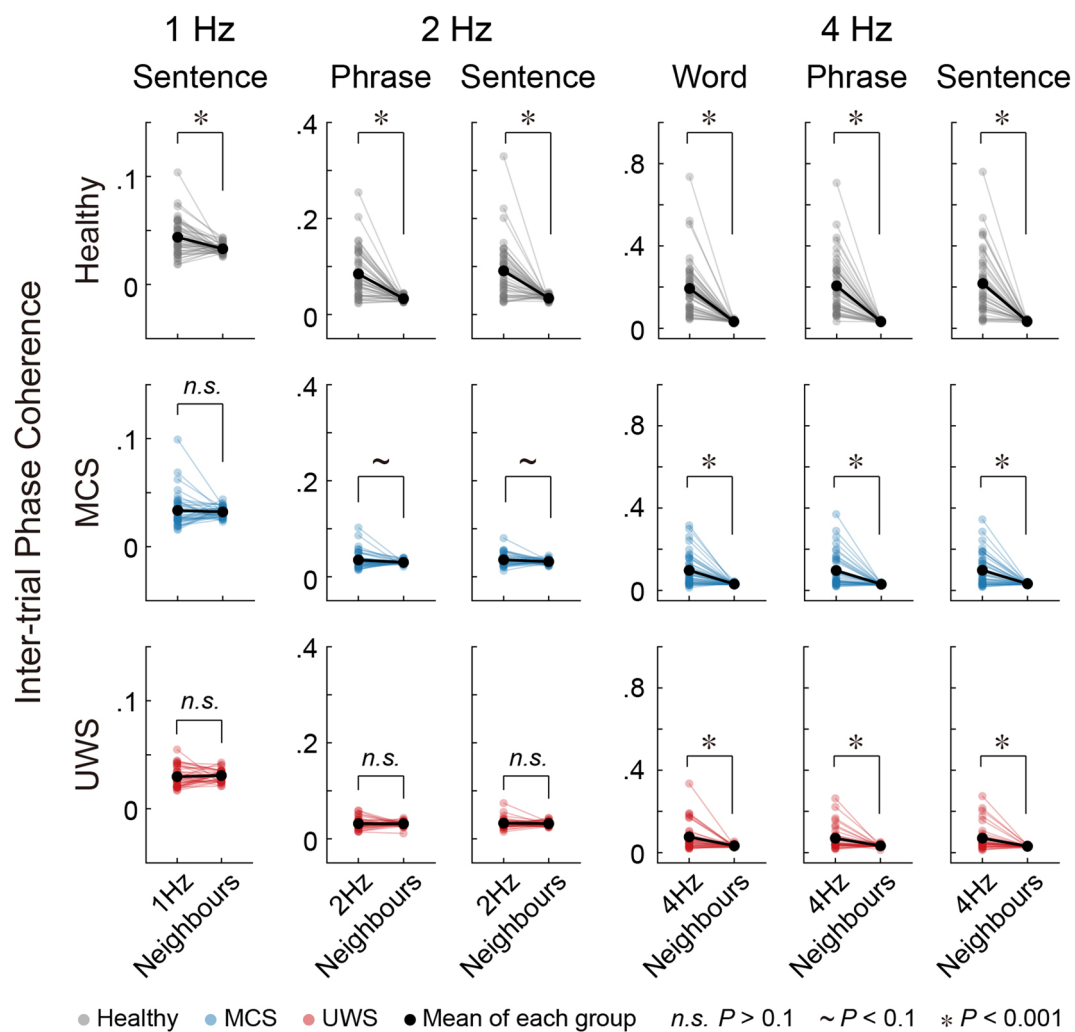
Extended Data Fig. 2 | Individual ITPC responses to hierarchical linguistic structures in Healthy controls (n = 47), MCS (n = 42) and UWS (n = 36) patients at three task levels. In each inset, the dots in the left represent the ITPC values from individual subject at target frequencies (1/2/4 Hz), while the dots in the right represent the individual mean value at its respective neighbours. Solid black dots represent grand mean values. *n.s.*, $P > 0.1$; ~, $P < 0.1$; *, $P < 0.001$; one-sided paired-sample *t*-test: see legend of Fig. 2b for precise statistical values.
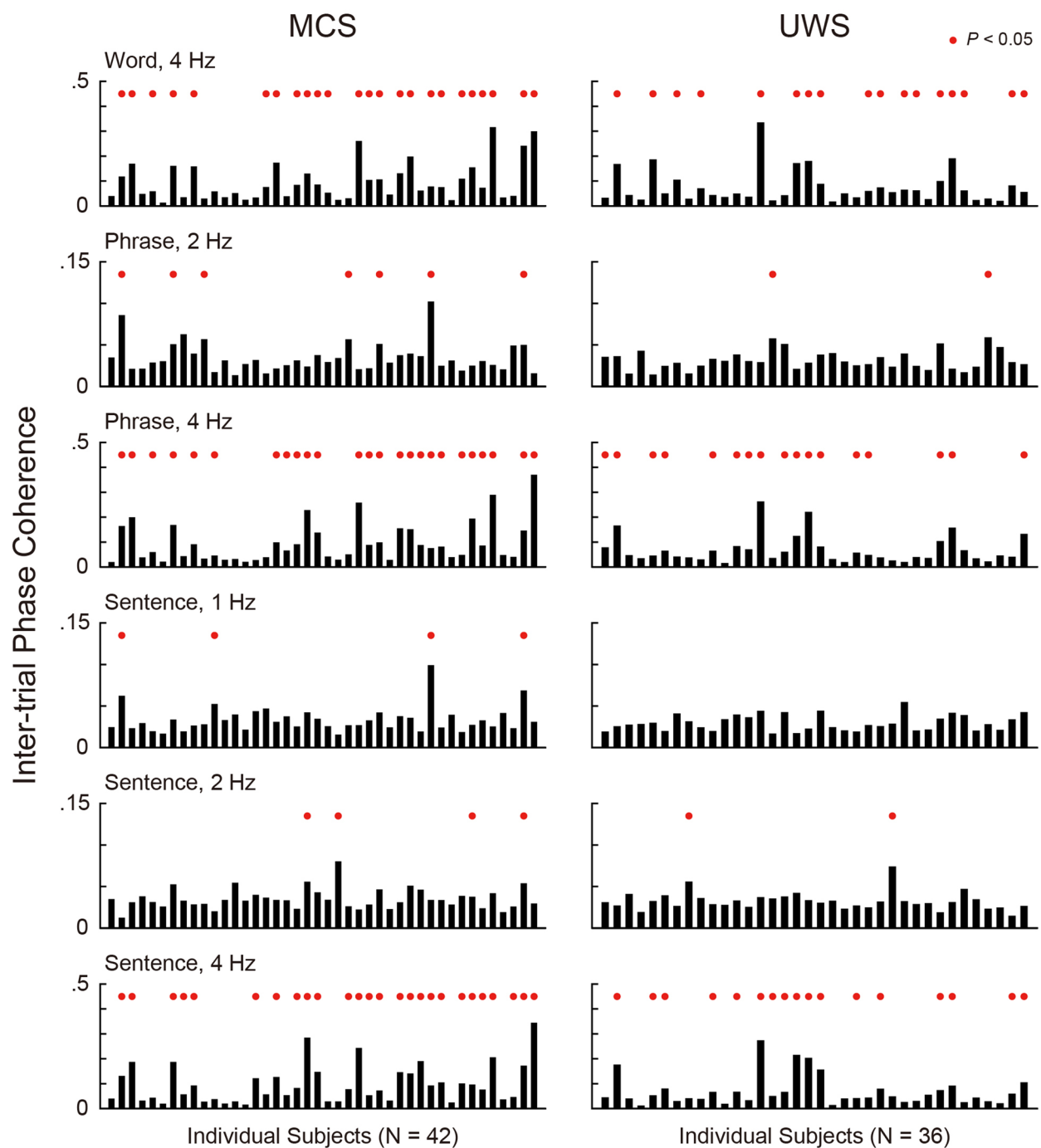
Extended Data Fig. 3 | ITPC responses to hierarchical linguistic structures in individual patients at three task levels. Each bar denotes the response from one subject. Red dots indicate the significance (exact $P < 0.05$; one-sided exact test, the statistical significance (exact P) of the ITPC response at a target frequency is the probability that the target frequency response differs from the null distribution, which consisted of responses at all non-target frequencies, see *Methods*).
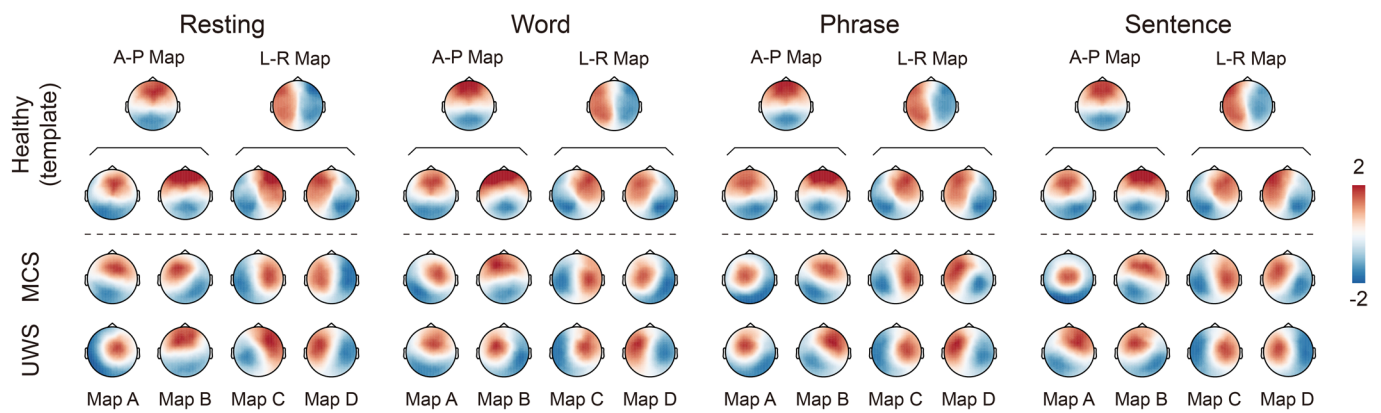
Extended Data Fig. 4 | Brain state maps of healthy controls and patients in all four task conditions. Number of subjects: $n_{Healthy-Resting} = 34$, $n_{Healthy-Task} = 47$, $n_{MCS-Resting} = 41$, $n_{MCS-Task} = 42$, $n_{UWS-Resting} = 30$, $n_{UWS-Task} = 36$. *Top row*: Template Anterior-Posterior (A-P) and Left-Right (L-R) maps obtained from healthy controls. *Bottom three rows*: Original four maps (Maps A, B, C, and D) of each group in each condition.
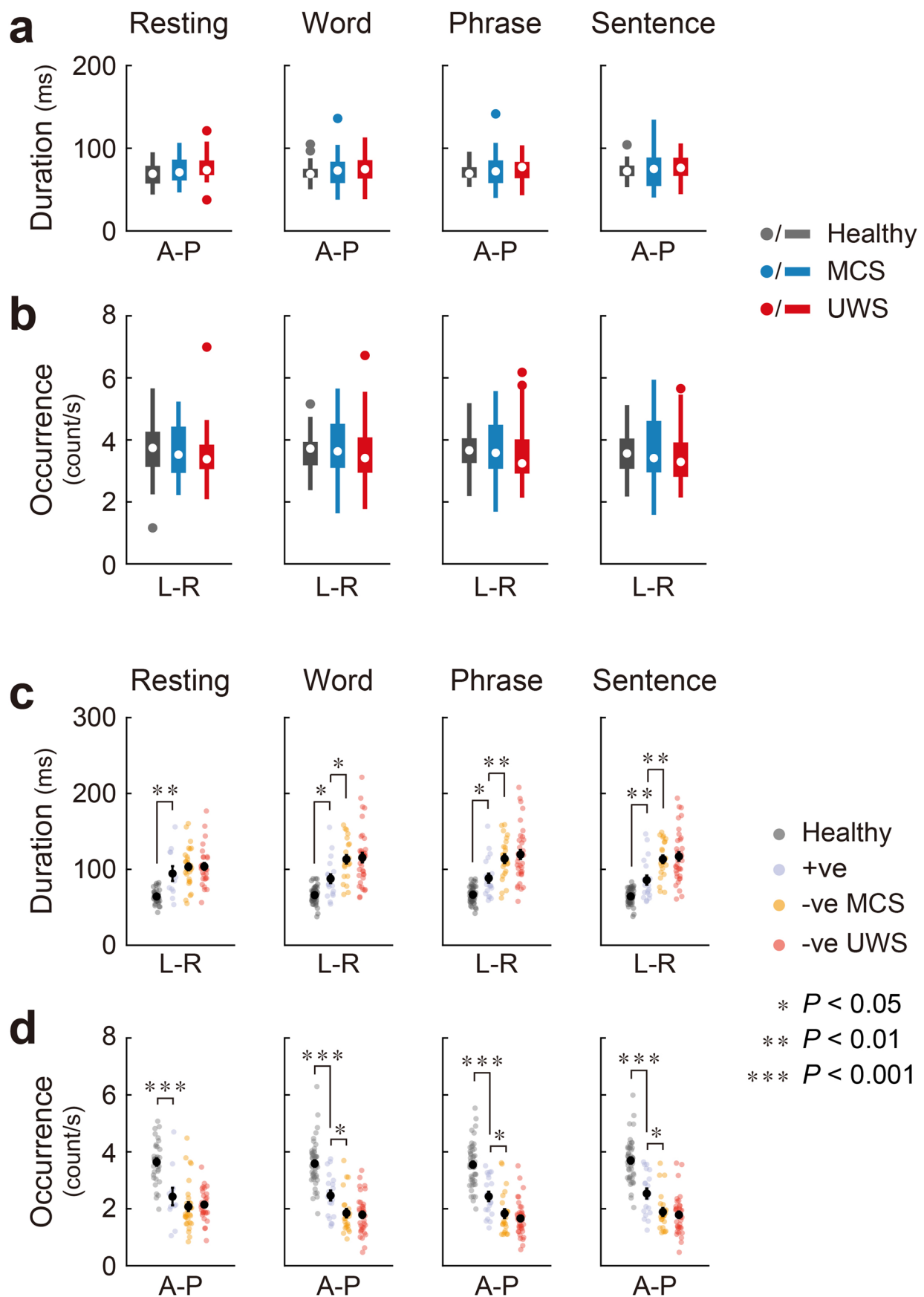
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Duration and occurrence of brain state maps. a, The duration of the A-P Map for the healthy control ($n_{Resting}$ = 34, n_{Task} = 47), MCS ($n_{Resting}$ = 41, n_{Task} = 42), and UWS (n_{UWS} = 30, n_{Task} =36) groups in all four task conditions. **b**, The occurrence of the L-R Map for healthy control, MCS, and UWS groups. Note that there were no differences between the three groups. Boxes represent IQR, central dots indicate the median, and whiskers indicate 1.5 × IQR. Colored dots indicate outliers. **c**, The duration of the L-R map for healthy controls (gray; $n_{Resting}$ = 34, n_{Task} = 47), recovery patients (+ve, green; $n_{Resting}$ = 11, n_{Task} = 19), and non-recovery MCS (-ve MCS, blue; $n_{Resting}$ = 30, n_{Task} = 23) and non-recovered UWS (-ve UWS, red; $n_{Resting}$ = 30, n_{Task} = 36) patients in all four conditions. One-way ANOVA, Bonferroni corrected: Healthy *vs.* +ve, $P_{Resting}$ = 0.001, P_{Word} = 0.018, P_{Phrase} = 0.013, $P_{Sentence}$ = 0.008; +ve *vs.* -ve MCS, P_{Word} = 0.012, P_{Phrase} = 0.009, $P_{Sentence}$ = 0.002. **d**, The occurrence of the A-P map for healthy controls, recovery patients, non-recovery MCS patients, and non-recovery UWS patients. One-way ANOVA, Bonferroni corrected: Healthy *vs.* +ve, $P_{Resting}$ = 6.8×10^{-5}, P_{Word} = 8.1×10^{-7}, P_{Phrase} = 5.7×10^{-7}, $P_{Sentence}$ = 1.8×10^{-7}; +ve *vs.* -ve MCS, P_{Word} = 0.046, P_{Phrase} = 0.048, $P_{Sentence}$ = 0.026. Panel c and d: colored dots represent individual subjects. Black dots represent mean values. Error bars represent S.E.M. All panels: *, $P < 0.05$; **, $P < 0.01$; ***, $p < 0.001$.
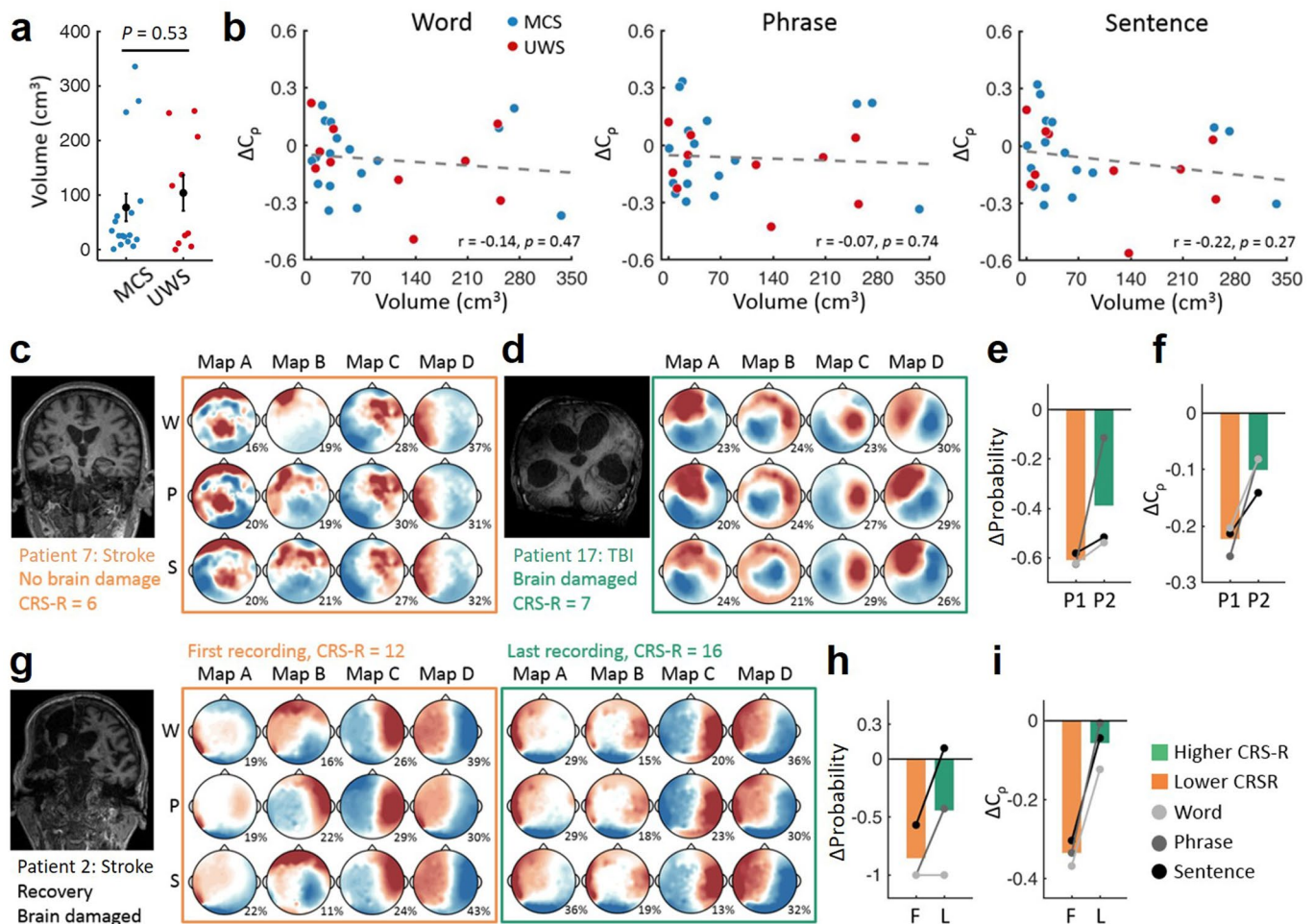
Extended Data Fig. 6 | Correlation between the volumes of brain injury and the ΔCρ. a, The comparison of volumes of brain injury between MCS (n = 17) and UWS (n = 10) patients. The black dots and error bars denote the mean value and S.E.M. $t_{25} = 0.64$, $P = 0.53$, two-tailed two-sample t-test. **b**, The correlation between ΔCρ and the volumes of brain injury in three task conditions. Pearson's correlation test (two-tailed), $n_{MCS} = 17$, $n_{UWS} = 10$. **c**, An example patient (Patient 7): the MRI data and maps of the stroke patient without brain damage. **d**, An example patient (Patient 17): the MRI data and maps of the TBI patient with large brain damage. **e**, The comparison of Δprobability between the stroke patient without brain damage and the TBI patient with brain damage, as shown in **c** and **d**. P1: Patient 7, P2: Patient 17. **f**, ΔCρ, the same format as e. **g**, The MRI data and brain states of a stroke patient with brain damage (an example patient, Patient 2). The orange box indicates the first EEG recording in unrecovered state. The green box indicates the last EEG recording in recovery state. **h**, The comparison of Δprobability in Patient 2 between the first EEG recording in unrecovered state and the last EEG recording in recovery state. **i**, ΔCρ, the same format as **h**. **W**: Word, **P**: Phrase, **S**: Sentence. **F**: First recording, **L**: Last recording. The percentage under each spatial map indicates the probability of each map.
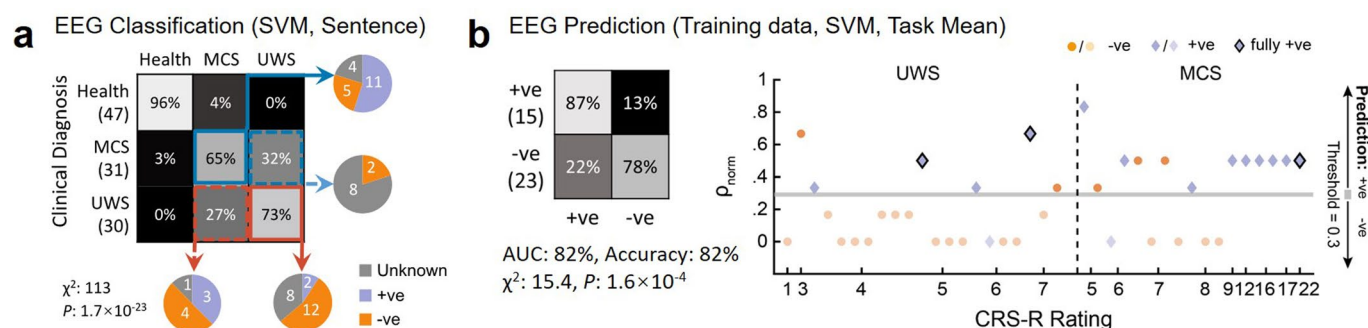
Extended Data Fig. 7 | Diagnosis and outcome prediction using SVM. a, The confusion matrix of diagnosed consciousness classification generated by the cross-validated SVM. The feature combinations we used were [$\Delta C\rho$ + Duration$_{L-R}$ + Occurrence$_{A-P}$ + ITPC$_{1Hz}$ + ITPC$_{2Hz}$ + ITPC$_{4Hz}$] for Sentence task. **b**, The performance of outcome prediction on training data using SVM classifier with the best feature combinations. *Left*: Outcome prediction accuracies by EEG on 38 EEG recordings (15 outcome-positive patients). *Right*: Comparison of individual predictions and actual outcomes. The patients with UWS are shown to the left of dashed line, and the patients with MCS are shown to the right. The dots above the threshold (gray line, prediction score = 0.3) represent the patients with predicted positive outcomes, while the others represent those with predicted negative outcomes. The actual outcome-negative patients are marked by orange dots, and the actual outcome-positive patients are marked by green diamonds. Solid green diamonds represent the outcome in patients that regained wakefulness. The feature combinations we used were: [ΔProbability + Duration$_{L-R}$ + Transition$_{A-P}$] for Word condition, [ΔProbability + Occurrence$_{A-P}$ + Duration$_{L-R}$ + Transition$_{L-R}$ + ITPC$_{4Hz}$] for Phrase condition, [Occurrence$_{A-P}$ + Duration$_{L-R}$ + Transition$_{L-R}$ + ITPC$_{1Hz}$ + ITPC$_{2Hz}$] for Sentence condition.

Diagnosis

a

CRS-R
Total-score

CRS-R (Total-score)
Cross-validated diagnosis on data (n=38)

	MCS	UWS	
MCS (17)	53%	47%	**AUC: 75%**
UWS (21)	4%	96%	χ^2: 11.2
	MCS	UWS	P: 0.001
			Accuracy: 76%

Clinical Diagnosis

Generalization →

Generalization on the new data (n=25)

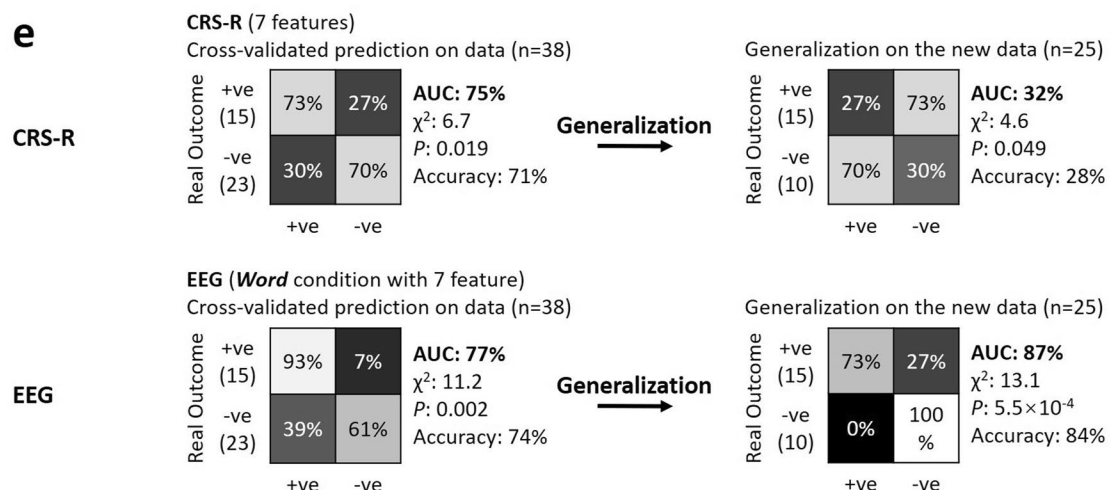|	MCS	UWS	
MCS (12)	92%	8%	**AUC: 96%**
UWS (13)	0%	100%	χ^2: 21.3
	MCS	UWS	P: 2.7×10^{-6}
			Accuracy: 96%

Clinical Diagnosis
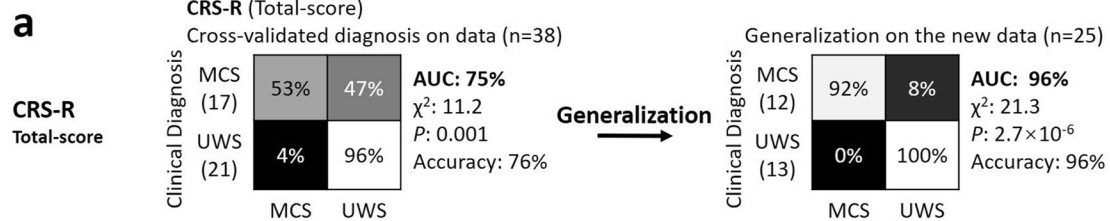
Outcome prediction with feature selection

b

CRS-R
With optimal feature combination

CRS-R (Visual subscale)
Cross-validated prediction on data (n=38)

Real Outcome

	+ve	-ve	
+ve (15)	67%	33%	**AUC: 75%**
-ve (23)	13%	87%	χ^2: 11.6
	+ve	-ve	P: 0.001
			Accuracy: 79%

Generalization →

Generalization on the new data (n=25)

Real Outcome

	+ve	-ve	
+ve (15)	27%	73%	**AUC: 32%**
-ve (10)	70%	30%	χ^2: 4.6
	+ve	-ve	P: 0.049
			Accuracy: 28%

c

EEG
With optimal feature combination

EEG (*Word* condition, ΔProbability + Occurrence$_{A-P}$)
Cross-validated prediction on data (n=38)

Real Outcome

	+ve	-ve	
+ve (15)	73%	27%	**AUC: 74%**
-ve (23)	26%	74%	χ^2: 8.2
	+ve	-ve	P: 0.007
			Accuracy: 74%

Generalization →

Generalization on the new data (n=25)

Real Outcome

	+ve	-ve	
+ve (15)	60%	40%	**AUC: 79%**
-ve (10)	0	100%	χ^2: 9.4
	+ve	-ve	P: 0.003
			Accuracy: 76%

d

CRS-R
Using the features with the first two highest weights

CRS-R (Visual + Arousal subscales)
Cross-validated prediction on data (n=38)

Real Outcome

	+ve	-ve	
+ve (15)	67%	33%	**AUC: 74%**
-ve (23)	17%	83%	χ^2: 9.5
	+ve	-ve	p: 0.005
			Accuracy: 76%

Generalization →

Generalization on the new data (n=25)

Real Outcome

	+ve	-ve	
+ve (15)	27%	73%	**AUC: 31%**
-ve (10)	70%	30%	χ^2: 4.6
	+ve	-ve	P: 0.049
			Accuracy: 28%

Outcome prediction without feature selection

e

CRS-R

CRS-R (7 features)
Cross-validated prediction on data (n=38)

Real Outcome

	+ve	-ve	
+ve (15)	73%	27%	**AUC: 75%**
-ve (23)	30%	70%	χ^2: 6.7
	+ve	-ve	P: 0.019
			Accuracy: 71%

Generalization →

Generalization on the new data (n=25)

Real Outcome

	+ve	-ve	
+ve (15)	27%	73%	**AUC: 32%**
-ve (10)	70%	30%	χ^2: 4.6
	+ve	-ve	P: 0.049
			Accuracy: 28%

EEG

EEG (*Word* condition with 7 feature)
Cross-validated prediction on data (n=38)

Real Outcome

	+ve	-ve	
+ve (15)	93%	7%	**AUC: 77%**
-ve (23)	39%	61%	χ^2: 11.2
	+ve	-ve	P: 0.002
			Accuracy: 74%

Generalization →

Generalization on the new data (n=25)

Real Outcome

	+ve	-ve	
+ve (15)	73%	27%	**AUC: 87%**
-ve (10)	0%	100%	χ^2: 13.1
	+ve	-ve	P: 5.5×10^{-4}
			Accuracy: 84%

**Extended Data Fig. 8 | See next page for caption.**

Extended Data Fig. 8 | Comparison of EEG-based and CRS-based classifiers for diagnosis and outcome prediction. a, Performance of clinical diagnosis using the CRS-R total-score. The classification model (LDA) was trained on the first 38 patients with cross validation (*Left*) and then tested (without retraining) on a novel dataset of 25 patients (*Right*). **b**, *Left*: Performance of outcome prediction using the optimal CRS-R sub-score (Visual subscale). *Right*: The prognostic validity of this model. The classifier was trained with cross-validation on the first dataset of 38 patients, then tested for generalization on a new dataset of 25 patients. **c**, Comparison of the prediction performance for models with the same number of features, based either on CRS-R (7 features, as in **b**) or the EEG recording under the word condition (7 features: 1 ITPC and 6 microstates). The optimal feature combination was ΔProbability + Occurrence$_{A-P}$. **d**, To test whether the superior EEG generalization ability was due to the larger number of EEG features used in the model, we then ran another way of selecting features by merely using the features with the first two highest weights in the model, and compared their performance of generalization. For the model using EEG see **c**, and for CRS-R, the best two features were Visual and Arousal subscales. **e**, Comparison of the performance of outcome prediction, using a standard LDA without feature selection, using all 7 features under the word condition of EEG versus all 7 features of the CRS-R scores (1 total-score and 6 sub-scores: auditory, visual, motor, oromotor, communication and arousal). Generalizations using EEG recorded during the other two task conditions (phrase: AUC = 89%, χ^2 = 13.1, P = 5.5×10-4, accuracy = 84%; sentence: AUC = 93%, χ^2 = 13.1, P = 5.5×10-4, accuracy = 84%; chi-squared test) showed similar results as that during the word condition.
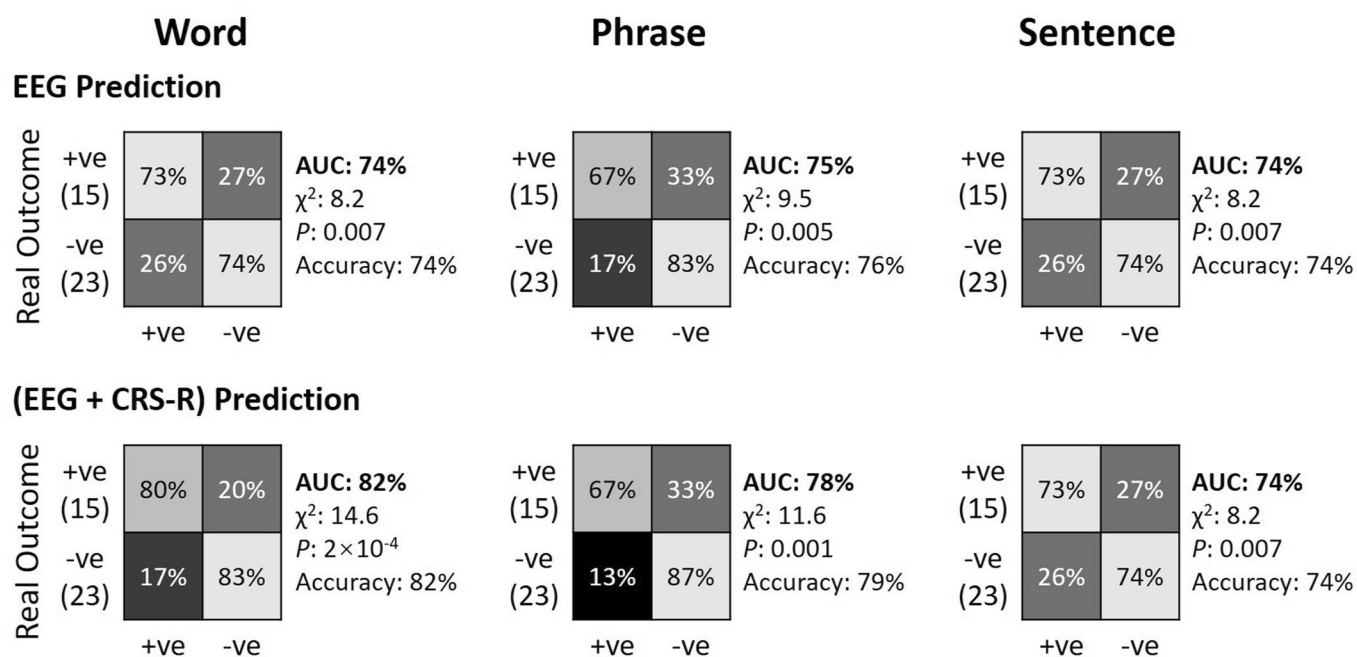
Extended Data Fig. 9 | Comparisons of performance of outcome prediction using EEG versus EEG plus CRS-R scores. *Upper*: The confusion matrix of outcome prediction by EEG scores. *Lower*: The confusion matrix of outcome prediction by the combination of EEG and CRS-R scores.
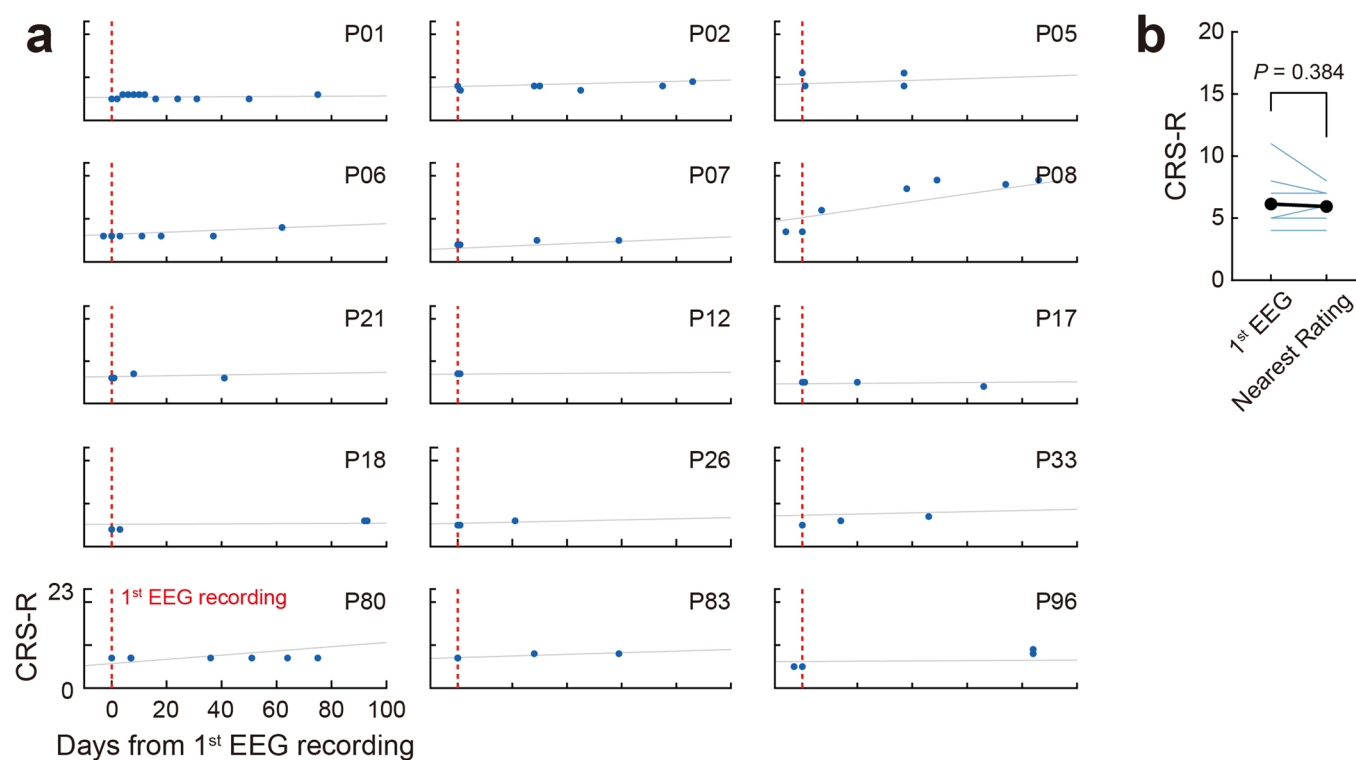
Extended Data Fig. 10 | Multiple CRS-R ratings across time. Each inset indicates one patient. a, Individual patients (n = 15). Within each inset, every blue dot indicates one CRS-R rating, and the gray line indicates the GLM fitting of all ratings in the entire period. Day 0 and red vertical dashed lines indicate the day of first EEG recording. **b**, The comparison of CRS-R scores between the EEG recording day and the day within a week (on average within 2.67 days). Colored lines indicate the ratings of individual patients. Black line indicates the mean. No significant difference was found between the two ratings (n = 15, $t_{14} = 0.899$, $P = 0.384$; two-sided paired-sample t-test).

nature research

Corresponding author(s):	Liping Wang, liping.wang@ion.ac.cn
Last updated by author(s):	Mar 25, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	In the attentional study, behavioral data were collected using Psychtoolbox (Brainard, 1997; RRID: SCR_002881) within MATLAB (2015b, Mathworks; RRID: SCR_001622), while EEG data were collected using BrainVison Recorder (1.20.0801, Brain Products GmbH, Germany; RRID:SCR_016331). In the clinical study, Net Station software (5.42, Electrical Geodesics, Inc., USA; RRID: SCR_002453) were used to collect EEG data. The MRI data were collected by a 3 Tesla MRI scanner (Siemens Magnetom Verio, Germany), or a 1.5 Tesla MRI scanner (GE Signa EXCITE Twinspeed Zoom, USA).
Data analysis	Data pre-processing: for the phase coherence analysis, EEG data were pre-processed using BrainVision Analyzer (2.0.1, Brain Products, GmbH, Germany; RRID: SCR_002356); for brain state analysis, EEG data were pre-processed using the EEGLAB toolbox (Version 14.1.1; RRID: SCR_007292) within MATLAB (2017b, MathWorks, USA; RRID: SCR_001622). Phase coherence analysis and multivariate pattern analysis were performed in MATLAB (2018a MathWorks, USA; RRID: SCR_001622). Brain state analysis was performed using MicrostateAnalysis toolbox (0.3, http://www.thomaskoenig.ch/index.php/software/microstates-in-eeglab/) within MATLAB (2017b, MathWorks, USA; RRID: SCR_001622). The analyses of MRI data were performed using MATLAB (2017b, MathWorks, USA; RRID: SCR_001622), and ITK-SNAP (3.8, RRID: SCR_002010).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see <u>nature.com/documents/nr-reporting-summary-flat.pdf</u>

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to pre-determine sample sizes but our sample sizes are similar to those reported in previous publications (Faugeras et al, 2012 Neuropsychologia; Di Perri et al, 2016 The Lancet Neurology; Edlow et al, 2017 Brain).
Data exclusions	Data exclusion criteria were pre-established based on previous literatures. In the attentional study, 5 out of 27 healthy subjects were excluded due to poor quality of EEG data (highly contaminated by ocular or muscular activities; Picton et al, 2000 Psychophysiology). In the clinical study, considering the noisy recording environment and abundant artifacts caused by patients' involuntary movements, we first examined the proportion of bad duration during each recording by applying the following pre-defined criteria on each data channel (204 channels, electrodes placed on the cheeks and neck were excluded), based on a previous paper (Faugeras et al, 2012 Neuropsychologia): 1) gradient criterion: the instant voltage change exceeds the maximal allowed step, which is 30 μV/ms; 2) max–min criterion: the absolute difference between the maximal and minimal voltage within every 200 ms sliding window exceeds 120 μV, and the sliding step is 10 ms; 3) amplitude criterion: the absolute voltage value exceeds 100 μV; 4) low activity criterion: the absolute difference between the maximal and minimal voltage within every 100 ms sliding window is smaller than 1 μV, and the sliding step is 10 ms. Data points that met any of these criteria were marked as artifacts, and the 200 ms period both before and after each artifact were marked as bad intervals. Channels with bad duration longer than 20% of the total recording length were marked as bad channels. Recordings in which the number of bad channels exceeded 70 were discarded. As a result, for patients' data collected during 2016-2018, 62 out of 133 resting-state recordings and 54 out of 132 task-related recordings were discarded. For patients' data collected during 2018-2019, 15 out of 40 resting-state recordings and 13 out of 38 task-related recordings were discarded. For the healthy controls, 8 out of 42 resting-state recordings and 14 out of 61 task-related recordings were discarded.
Replication	Given the longitudinal nature of the patients population, no attempts were made at replication.
Randomization	The order of task conditions (four conditions in the attentional study and three conditions in the clinical study) was randomized across subjects/patients in both studies, controlled by a random function in MATLAB. In addition, the order of stimuli in each task condition was also shuffled across subjects/patients.
Blinding	Both the attentional and clinical studies were double-blind (both subject and experimenter blinding was instituted). Both the subjects who participated in the studies and experimenters who collected data didn't know the real purpose of this research. Both the interviewee and interviewer obtaining the follow-up assessment were blinded to the behavioral and EEG-based diagnosis in the hospital. In addition, the experimenters who collected EEG data were not involved in data analyses. Specifically in the clinical study, the behavioral measurement and EEG & MRI data analyses were conducted by two separated groups of experimenters with a completely blinded fashion. The first and second author (Peng Gui and Yuwei Jiang, from the Institute of Neuroscience, CAS) analyzed and modeled the EEG & MRI data were blinded with the measurements of patient outcomes. The other authors (Dr. Xuehai Wu, Di Zang and their colleagues, from the Huashan Hospital) performed the clinical evaluation of patients' behavior did not involve in the EEG and MRI data analyses.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒ ☐	Antibodies
☒ ☐	Eukaryotic cell lines
☒ ☐	Palaeontology
☒ ☐	Animals and other organisms
☐ ☒	Human research participants
☒ ☐	Clinical data

Methods

n/a	Involved in the study
☒ ☐	ChIP-seq
☒ ☐	Flow cytometry
☐ ☒	MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics

Twenty-seven healthy subjects participated in the attentional EEG study; mean age = 23.9 years; range = 20 to 30 years. EEG data from twenty-two out of these subjects were used in final analyses (mean age = 23.73 years; range = 20 to 30 years).

For the clinical study, the usable datasets of patients collected during 2016-2018 included 71 resting-state recordings from 54 patients with DOC (48 male; mean age = 49.3 years; range = 17 to 75 years), and 78 linguistic task-related recordings from 60 patients with DOC (52 male; mean age = 47.8 years; range = 9 to 68 years).
To test the validity of the prediction model, a new group of patients were recruited during year 2018-2019, consisted of 12 MCS and 13 UWS recordings (13 male; mean age = 39.9 years; range = 18 to 69 years).
We also recruited 61 healthy volunteers from local communities as a control group (20 male; mean age = 31.3 years; range = 22 to 65 years). The usable datasets of healthy controls included 34 resting-state recordings (8 males; mean age = 25.8 years; range = 22 to 50 years) and 47 task-related recordings (17 males; mean age = 31.1 years; range = 22 to 58 years).

For the analysis of correlation between brain damage and brain states, the usable MRI data of 27 patients (16 males; mean age = 44.6 years; range = 9 to 68 years) were included, which were collected on the same day as the EEG recordings.

Recruitment

In the attentional study, subjects were recruited from the Institute of Neuroscience via electronic advertisement and posters, most of which were graduate students or staffs. Thus no potential self-selection bias was introduced. The inclusion criteria were: native Mandarin Chinese speaker, no metal implants in head, no history of psychological/neurological diseases or head injury, not being pregnant, no current use of psychoactive medication, normal hearing ability.

In the clinical study, health subjects were recruited from the Huashan Hospital via posting announcements on local and electronic bulletin boards, with the same requirement in the attentional study. No potential self-selection bias was introduced. All patients were recruited from the Huashan Hospital and the affiliated rehabituation hospitals. The inclusion criterion was all patients with a prolonged (≥28 days) DOC after brain damage. The exclusion criteria were: 1) when there was a fluctuation of consciousness or a severe medical condition, which is judged by experienced clinicians, 2) patients with other neurological and psychiatric disorders.

Ethics oversight

Ethical Committee of the Huashan Hospital of Fudan University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type

The data were scanned for structural diagnosis. It is not a task or resting-state design.

Design specifications

The experiments were structural scanning, neither task nor resting-state design was used.

Behavioral performance measures

The experiments were structural scanning, no behavior data was acquired.

Acquisition

Imaging type(s)

Structural MRI

Field strength

3 Tesla for Siemens Magnetom Verio; 1.5 Tesla for GE Signa EXCITE Twinspeed Zoom.

Sequence & imaging parameters

Siemens Magnetom Verio 3 T MRI system:
Two T2WI data were acquired using a turbo spin-echo (TSE) sequence: field of view = 240 × 240, matrix size = 640 × 640, slice thickness = 2 mm, orientation: transverse position, TR/TE = 7360/98 ms, flip angle = 150°.
Twelve T2WI data were acquired using a TSE sequence: field of view = 220 × 220, matrix size = 320 × 320, slice thickness = 4 mm, orientation: transverse position, TR/TE = 7120/101 ms, flip angle = 90°.
One T2WI data were acquired using a TSE sequence: field of view = 210 × 210, matrix size = 320 × 320, slice thickness = 4 mm, orientation: transverse position, TR/TE = 7120/101 ms, flip angle = 90°.
Four T2WI data were acquired using a TSE sequence: field of view = 230 × 230, matrix size = 256 × 256, slice thickness =

6 mm, orientation: transverse position, TR/TE = 3300/95 ms, flip angle = 90°.
Two T1WI data were acquired using a magnetization-prepared rapid gradient-echo (MP-RAGE) sequence: field of view = 240 × 256, matrix size = 240 × 256, slice thickness = 1 mm, orientation: sagittal position, TR/TE/TI = 2300/2.98/900 ms, flip angle = 9°.

GE Signa EXCITE Twinspeed Zoom 1.5 T MRI system:
Two T2WI data were acquired using a fast spin-echo (FSE) sequence: field of view = 240 × 240, matrix size = 512 × 512, slice thickness = 6 mm, orientation: transverse position, TR/TE = 4820/124.74 ms, flip angle = 90°.
Four T1WI data were acquired using a Fluid-attenuated inversion recovery (FLAIR) sequence: field of view = 240 × 240, matrix size = 256 × 256, slice thickness = 6 mm, orientation: transverse position, TR/TE/TI = 1708/20.48/720 ms, flip angle = 90°.

Area of acquisition

All MRI data were collected using a whole brain scan.

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software

The analysis of MRI data were performed using MATLAB (2017b, MathWorks, USA; RRID: SCR_001622), and ITK-SNAP (3.8, RRID: SCR_002010).

Normalization

We extracted the lesion areas manually, normalization was not needed in the data processing.

Normalization template

Normalization was not performed in the data processing.

Noise and artifact removal

Only data with good quality were collected in the dataset. So we didn't conduct the noise and artifact removal.

Volume censoring

We checked the volume visually. The criteria for volume censoring is high signal-to-noise ratio and no motion artifact.

Statistical modeling & inference

Model type and settings

Two-sample t-test was used to compare the injured volumes between the MCS and UWS patients. Pearson's correlation tests were used to test the correlation between ΔC_P and the volumes of brain injury in three task conditions.

Effect(s) tested

No ANOVA or factorial design was used. No effect test was performed.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference
(See Eklund et al. 2016)

The sizes of brain injured volumes were defined using voxel-wise calculation.

Correction

The comparison of brain lesion volumes was performed between two groups. No correction for multiple comparison was performed.

Models & analysis

n/a	Involved in the study
☒ ☐	Functional and/or effective connectivity
☒ ☐	Graph analysis
☒ ☐	Multivariate modeling or predictive analysis