Reanalysis of “Bedside detection of awareness in the vegetative state: a cohort study”

In 2011, Damian Cruse and colleagues reported that a new EEG-based tool was able to show that three of 16 patients in a vegetative state performed a motor imagery task requiring language and short-term memory. This finding, if confirmed, has major implications for diagnosis and care of severely brain-injured patients. We were concerned about the method’s validity because of the difficulty of the task, and its critical reliance on certain statistical assumptions. To allow us to test the validity of the method, Cruse and colleagues graciously supplied their data and analysis software. Below we show that the patient data do not meet the statistical assumptions made in the original paper, probably because of the presence of various artifacts. We then show that when the data are reanalysed by methods that do not depend on these model assumptions, there is no evidence of task performance in the patients.

We begin by examining the EEG data themselves. The controls have findings typical of healthy adults (figure A, P13): rhythmicity in the α range (about 10 Hz) with minimal eye-blink and muscle artifact. By contrast, the patients’ EEGs (figure A, P13) are dominated by 1-4 Hz activity, as is typical of severe brain dysfunction, deep sleep, or anaesthesia. Frequency-domain representation (figure B) confirms these findings. It also reveals that patients’ EEGs have substantial muscle artifact that fluctuates block-to-block.

To determine whether individuals performed motor imagery tasks, Cruse and colleagues used a multivariate method (support vector machine [SVM]) to differentiate EEG signals recorded while individuals were asked to imagine moving their hand versus their toes. SVM is a powerful technique, but, without a gold-standard for task performance, the validity hinges on the appropriateness of the statistical model. As detailed below, the statistical model used in Cruse and colleagues’ study did not account for relations between adjacent blocks, or correlations between trials within a block.

For calculation of accuracy (how often the SVM correctly classified trials as “hand” vs “toe”), Cruse and colleagues’ methods did not take into account the possibility of slow variations across blocks, because their approach always classified pairs of neighbouring blocks (eg, hand and toe block 1, but never hand block 1 and toe block 4). We modified their analysis to use these alternative pairings for cross-validation (appendix). In two of the positive patients (webfigure 1), accuracy decreased to chance (P1), or worse-than-chance (P12) as the test-block-pairs were further apart. This drop in accuracy implies that idiosyncratic relations between adjacent blocks contributed substantially to SVM performance in these individuals.

For calculation of significance, Cruse and colleagues calculated p values by use of a binomial distribution for the number of correct trials—an approach that assumes that each trial is an independent assay. We found that this assumption does not hold in the patients. First, frequency domain representation of the EEG (figure B; appendix) reveals a lack of independence: data from individual trials are more closely matched within a block than across blocks. Second, we applied Cruse and colleagues’ analysis separately to all time points of the trials. For patients, we found that worse-than-chance classification occurred substantially more often than expected from binomial statistics. This excess of outliers implies that trials are correlated (appendix, webfigure 2).

We next show that when the SVM results are reanalysed with a statistical approach that takes into account the correlations mentioned above (see appendix and webtable 1 for full details), there is no statistical evidence of a task-related signal. To take into account correlations between blocks, we defined accuracy using all block-pairs as test components, rather than restricting consideration to adjacent block pairs. To account for dependence among trials, we determined significance via a permutation test that recognised the block design. With this approach, significance remained in positive controls, but in only one patient (P13) (p=0.0286; lowest possible p-value with four blocks). We further note that even for random data, a classifier would be expected to yield one in 20 positive individuals at p≤0.05.

Finally, we applied an independent approach that asked whether there was a significant difference between task and rest periods, by use of univariate statistics (ie, separate tests for each frequency and channel of the EEG; appendix, webfigures 3 and 4). Controls showed the expected task-related changes in motor imagery tasks (decreases in EEG power from 7 to 30 Hz, especially over the motor cortices contralateral to the imagined limb movement; p≤0.05 after FDR correction). None of the 16 patients had significant changes identified by this measure. This finding emphasises that, even if we were to accept Cruse and colleagues’ “positive” patient classifications as different from chance, the EEG signals lack the expected physiological changes associated with motor imagery (by contrast with the suggestion made by Cruse and colleagues in connection with their figure 2).

In sum, we found that Cruse and colleagues’ method is not valid because the patient data do not meet the assumptions of their statistical model. Specifically, the model does not allow for correlations between nearby trials and blocks, which are likely to be induced by fluctuating
Correspondence

artifact and arousal state; when these factors are taken into account, there is no statistical evidence for task performance in patients. Importantly, Cruse and colleagues' model generally suffices for controls, in whom there is minimal artifact contamination. These findings cast doubt about conclusions drawn from this method, both in Cruse and colleagues’ Lancet study and a more recent one.11

In BCI applications, participants can confirm task performance and the consequences of classifier failure are limited to reduced device performance. But in the diagnostic setting (eg, determination of consciousness, genomic diagnosis of cancer13,14), classifier failure can misinform clinical decision making, with major consequences for patients and families. Given this problem, and the ease of dissemination of EEG technology, standards of demonstration of validity need to be high. Our analysis suggests that Cruse and colleagues’ approach falls short of this standard.

Finally, we wish to emphasise the importance of data sharing. This analysis would not have been possible without full access to the original data and code.15

Figure: Time and frequency domain representations of the EEG of a typical control (N2) and patient (P13) who had similar classification rates in the paper by Cruse and colleagues (75% and 78%, respectively).

(A) Laplacian-montaged EEG of the first trial of hand and toe block 1. The 25 channels used by Cruse and colleagues are shown. Note high frequency activity in P13 that differs between the trials.

(B) Spectra of the EEG calculated from each block, colour-coded by block type, for the same individuals as (A). Rest period is data 1·5–0·0 s pretone, and task period is data 0·5–2·0 s post-tone. Channels displayed include extreme left, midline, and extreme right of the 25 channels shown in (A). I-bar symbol in each plot of (B) represents average 95% CIs for the spectra (by jackknife). If trials were independent, the spectral estimates from each block should agree with each other, up to the CIs of each estimate. This holds for the data from controls (left) but not patients (right).

SVM and related methods are useful, particularly in EEG analysis for brain-computer interface (BCI).10,12

AMG, NDS, and JDV designed the overall structure of the study. AMG did the analysis. JCB and QN also designed the study. AMG, JCB, QN, NDS, and JDV interpreted the results and contributed to the writing of the paper. JJF contributed to the writing of the paper. We receive grant support from the same James S McDonnell Foundation grant as Cruse and colleagues.
Andrew M Goldfine, Jonathan C Bardin, Quentin Noirhomme, Joseph J Fins, 
“Nicholas D Schiff, Jonathan D Victor nds2001@med.cornell.edu

Department of Neurology and Neuroscience, Weill Cornell Medical College, New York, NY 10065, USA (AMG, NDS, JDV); Burke Medical Research Institute, White Plains, NY, USA (AMG); Department of Neuroscience, Weill Cornell Graduate School of Medical Sciences, New York, NY, USA (JCB); Coma Science Group, Cyclotron Research Centre and Neurology Department, University Hospital of Liège, Liège, Belgium (QN); and Division of Medical Ethics, Weill Cornell Medical College, New York, NY, USA (JF)


Authors’ reply

The primary suggestion put forward by Andrew Goldfine and colleagues is that a permutation test should have been used to infer command-following in our 2011 publication in The Lancet.1 One obvious problem with this argument is that, if a permutation test were used for all of the patients, half of them would only produce 36 permutations that could contribute to the test. It is accepted statistical practice that at least 1000 permutations are required to draw valid conclusions.2 As such, the outcome of their suggested approach would be statistically invalid for half the patients in our original study—ie, no one can know whether the answer generated by their approach is right or wrong because it is not an appropriate test to use given the data available.

One could argue that a task requiring more frequent switches between commands could be used to generate the requisite number of permutations. However, such a task would inevitably increase the cognitive load substantially and would probably be impossible for severely brain-injured patients to do. Indeed, the task that we chose for our study, with its blocked structure, has cognitive demands that are more similar to the mental imagery tasks that have previously been shown to detect awareness in a significant proportion of vegetative state patients using functional magnetic resonance imaging (fMRI).4

Moreover, Goldfine and colleagues’ suggestion that our patient data violate the independence requirement of the binomial test is based on an assumption that the patient group should be treated as homogeneous. To make their point, they show that, across the patient group, there seems to be a violation of independence—ie, a U-shaped histogram of p values. Although this might be the case across the group as a whole, it is certainly not the case when the data are inspected on an individual patient basis. It is widely accepted, even by Goldfine and colleagues,2,3 that a significant minority of patients (about 17%) who are diagnosed as being in the vegetative state nevertheless retain some level of conscious awareness. By extension then, this group is clearly not all homogeneous—that is to say, some are likely to be truly vegetative, whereas others might appear to be vegetative behaviourally, but are in fact covertly aware. It makes little sense, therefore, to group all of our vegetative state patients together in the way suggested by Goldfine and colleagues, because the (known) majority of truly vegetative patients will water down the covertly aware subgroup, rendering the latter more difficult to detect using any statistical method. Indeed, when we applied the same test for independence used by Goldfine and colleagues to each patient dataset individually, rather than as a group (ie, using the standard working hypothesis that all patients are different), we found that all three of our positive patients pass the assumption of independence—ie, one-tailed histograms. By Goldfine and colleagues’ own test, therefore, our use of the binomial method is validated in these positive individuals.

Although there are few known truths when attempting to detect covert awareness, the one thing we can assume to know is that when healthy volunteers are asked to do the imagery tasks described in our original paper, they are doing them. Equally, when asked to not do the imagery tasks, it is reasonable to assume that they are not doing them. It is reassuring then, that our task and analyses identified significant command-following in 75% of the healthy participants who contributed to the original Lancet article (and correctly detected the absence of command-following in 100% of cases). Although not perfect, this is, on balance, a reasonable approximation of the only known truth. By stark contrast, the method expounded by Goldfine and colleagues only detects command-following in 40% of the healthy participants they analysed. In short,
because their method fails to detect command-following in 60% of healthy volunteers, it is equally likely to fail to detect command-following (where it exists) in most patients. Goldfine and colleagues also point to differences in the spatial and spectral characteristics of the neural command-following response seen in our three positive patients, relative to healthy controls. We would certainly have to agree that there are differences (as one would expect after serious brain injury), but question their relevance here. Indeed, in their own recent EEG study, Goldfine and colleagues highlight the “variability in healthy control results, along with the fact that those with severe brain injury have differences in neuroanatomy and connectivity due to injury and the recovery process”, yet go on to accept as evidence of command-following a broad range of EEG responses that varied widely in terms of their spatial and spectral characteristics. They also state that “It is not possible to determine whether the reason for the difference in this patient’s [EEG] spectral pattern [when compared with healthy controls] reflects variation in the way the task was performed, or an injury-induced reorganization in cerebral networks supporting the behavior”.9 Our interpretation of these spatial and spectral differences, therefore, concurs fully with their own and does nothing to undermine the key results reported in our Lancet paper.

These methodological concerns about Goldfine and colleagues’ assumptions notwithstanding, their reanalysis only pushes two of our three positive patients to just beyond the widely accepted p<0.05 threshold for significance—i.e., to p=0.06 and p=0.09, respectively. To dismiss the third patient, whose data remain significant, they state that the statistical threshold for accepting command-following should be adjusted to account for the number of patients who have been assessed (a so-called multiple comparisons correction). We know of no groups in this field who routinely use such a conservative correction with patient data, including Goldfine and colleagues.10-11 In this particular case, the only reason for doing so would be if we had no a-priori hypothesis. In the Introduction to our Lancet paper, we reviewed several previous papers,4,5,11 and concluded that “these findings confirm that a population of patients exist who meet all the behavioural criteria for the vegetative state, but nevertheless retain a level of covert awareness that cannot be detected by thorough behavioural assessment”. Our a-priori hypothesis could hardly have been clearer.

Finally, it is reassuring to note that corroborative data using independent methods, including a previously validated fMRI test of command-following,12 is available for two of our three positive patients. These data confirm that these patients were aware during the same week in which the EEG data in question was acquired.

In conclusion, Goldfine and colleagues make some interesting points about the choice of statistical model when seeking to identify covert command-following in severely brain-injured patients. Their unconventional cross-validation approach does suggest that the EEG responses of two of our three positive patients became less consistent across time, and argues for future iterations of the task structure to be altered to accommodate this. Indeed, our goal, like that of Goldfine and colleagues, is to develop increasingly sensitive tools to identify covert command-following and, in that spirit, we have recently published a method that more formally addresses many of their current concerns.12 Clearly, it is only through the continuing improvement of our complementary approaches that we will converge on the optimum methods for accurately identifying covert awareness, where it exists, in every severely brain-injured patient.

We declare that we have no conflicts of interest.

“Damian Cruse, Sinivas Chennu, Camille Chatelle, Tristan A Bekinschtein, Davinia Fernández-Espejo, John D Pickard, Steven Laureys, Adrian M Owen dcruise@uwo.ca

Brain and Mind Institute, University of Western Ontario, London, ON N6A 5B7, Canada (DC, DF-E, AMO); Department of Clinical Neurosciences, University of Cambridge, Cambridge, UK (SC); Coma Science Group, Cyclotron Research Centre and Neurology Department, University Hospital of Liège, Liège, Belgium (CC, SL); Medical Research Council Cognition and Brain Sciences Unit, Cambridge, UK (TAB); and Division of Academic Neurosurgery, Addenbrooke’s Hospital, Cambridge, UK (JDP)