TITLE PAGE

Title: Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study."

Authors:

Andrew M. Goldfine, MD
Burke Medical Research Institute and
Department of Neurology and Neuroscience
Weill Cornell Medical College
White Plains, NY, USA

Jonathan C. Bardin, AB
Department of Neuroscience
Weill Cornell Graduate School of Medical Sciences
New York, NY, USA

Quentin Noirhomme, PhD
Coma Science Group
Cyclotron Research Centre and Neurology Department
University and University Hospital of Liège
Liège, Belgium

Joseph J. Fins, MD
Division of Medical Ethics
Weill Cornell Medical College
New York, NY, USA

Nicholas D. Schiff, MD
Department of Neurology and Neuroscience
Weill Cornell Medical College
New York, NY, USA

Jonathan D. Victor, MD, PhD
Department of Neurology and Neuroscience
Weill Cornell Medical College
New York, NY, USA

Corresponding Author:

Nicholas D. Schiff, MD
Department of Neurology and Neuroscience
Weill Cornell Medical College
LC-803
1300 York Ave
New York, NY 10065, USA
Phone: 212-746-2372
email: ndschiff@med.cornell.edu

Cruse and colleagues reported[1] that a new electroencephalography (EEG)-based tool was able to show that 3 out of 16 vegetative state (VS) patients 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 Cruse et al., likely because of the presence of various artifacts (Table). We then show that when the data are re-analyzed by methods that do not depend on these model assumptions, there is no evidence for task performance in the patients.

To begin, we examine the EEG data itself. The normals have findings typical of healthy adults (Figure 1A, left): rhythmicity in the alpha range (~10 Hz) with minimal eye-blink and muscle artifact. In contrast, the patients' EEG (Figure 1A, right) is dominated by 1-4 Hz activity, as is typical of severe brain dysfunction, deep sleep or anesthesia[2]. Frequency-domain representation (Figure 1B) confirms these findings. It also reveals that the patient's EEG has significant muscle artifact[3] that fluctuates block-to-block.

To determine whether subjects performed motor imagery, Cruse and colleagues used a multivariate method (Support Vector Machine; SVM) [4,5] to differentiate EEG signals recorded while subjects were asked to imagine moving their hand, vs. 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.[6] As detailed below, the statistical model used in Cruse et al. did not account for relationships 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"), the Cruse et al. methods did not take into account the possibility of slow variations across blocks, as their approach always classified pairs of *neighbouring* blocks (e.g., 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[6] (Webappendix). In two of the positive patients (Webappendix Figure 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 relationships between adjacent blocks contributed substantially to SVM performance in these subjects.

For calculation of significance, Cruse and colleagues calculated p-values using 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 1B; Webappendix) reveals a lack of independence: data from individual trials are more nearly matched within a block than across blocks. Second, we applied the Cruse et al. 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 (Webappendix and Webappendix Figure 2).

We next show that when the SVM results are re-analyzed with a statistical approach that takes into account the correlations mentioned above (Webappendix and Webappendix Table 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[6], rather than

restricting consideration to adjacent block pairs. To account for *dependence among trials*, we determined significance via a permutation test that recognized the block design. With this approach, positive normals remained significant, but only one patient (P13) remained significant ($p=0{\cdot}0286$; lowest possible p-value with 4 blocks). We further note that even for random data, a classifier would be expected to yield 1 in 20 positive subjects at $p\leq0.05$. We therefore corrected for multiple comparisons via the False-Discovery Rate (FDR)[7]; normals remained significant but none of the patients were significant at $p\leq0{\cdot}05$.

Finally, we applied an independent approach that asked whether there was a significant difference between task and rest periods, using univariate statistics (i.e., separate tests for each frequency and channel of the EEG; methods in Webappendix and [8]; Webappendix Figures 3 and 4). Normals showed the expected task-related changes in motor imagery tasks (decreases in EEG power from 7-30 Hz, especially over the motor cortices contralateral to the imagined limb movement; $p\leq0.05$ after FDR correction)[9,10]. None of the 16 patients had significant changes identified by this measure. This emphasizes that even if we were to accept the 'positive' patient classifications of Cruse et al. as different from chance, the EEG signals lack the expected physiological changes associated with motor imagery (in contrast to the suggestion made by Cruse and colleagues in connection with their Figure 2).

In sum, we found that the method of Cruse et al. 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 induced by fluctuating artifact and arousal state; when these factors are taken into account, there is no statistical evidence for task performance in patients. Importantly, the model of Cruse et al. generally suffices for normals, where there is minimal artifact contamination. These findings cast doubt about conclusions drawn from this method, both in Cruse et al., and a more recent study[11].

SVM and related methods are useful tools, particularly in EEG analysis for Brain-Computer Interface (BCI)[10,12]. In BCI applications, subjects can confirm task performance and the consequences of classifier failure are limited to reduced device performance. But in the diagnostic setting (e.g., determination of consciousness, genomic diagnosis of cancer[13,14]), classifier failure can misinform clinical decision making, with major consequences for patients and families. Given this, and the ease of dissemination of EEG technology, standards of demonstration of validity need to be high. Our analysis suggests that the approach of Cruse et al. falls short of this standard.

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

ROLE OF THE FUNDING SOURCE

The funding source had no role in study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the paper for publication.

CONFLICTS OF INTEREST

The authors of the Cruse et al. report receive grant support from the same James S. McDonnell Foundation grant.

AUTHOR'S CONTRIBUTIONS

Andrew Goldfine, Nicholas Schiff and Jonathan Victor designed the overall structure of the study. Andrew Goldfine conducted the analysis. Jonathan Bardin, Quentin Noirhomme also designed the study. All above authors interpreted the results and contributed to the writing of the paper. Joseph Fins contributed to the writing of the paper.

TABLE

Assumption of Cruse et al.	Relevance to Analysis	Test(s) of the Assumption	Outcome
no special relationship between adjacent blocks	calculation of accuracy and significance	dependence of classification accuracy on temporal separation of hand and toe blocks	invalid in two positive patients
independence of trials within blocks	calculation of significance	1. consistency of spectra from different blocks of same task type	invalid in all positive patients
		2. distribution of p-values with classification tested at all time points	invalid in patients as a group

Table – Overview of analyses and findings.

FIGURE LEGENDS

Figure 1: Time and frequency domain representations of the EEG of a typical normal (N2) and patient (P13) who had similar classification rates in Cruse et al. (75% and 78%, respectively; Webappendix for methods). A. Laplacian-montaged EEG of the first trial of hand and toe block 1. The 25 channels used in Cruse et al. are shown. Note high frequency activity in P13 that differs between the trials. B. Spectra of the EEG calculated from each block, color-coded by block type, for the same subjects as Panel A. Rest period is data 1·5 to 0 seconds pre-tone, and task period is data 0·5 to 2·0 seconds post-tone. Channels displayed include extreme left, midline and extreme right of the 25 channels shown in Panel A. I-bar symbol in each plot of Panel B represents average 95% confidence limits for the spectra (by jackknife). If trials were independent, the spectral estimates from each block should agree with each other, up to the confidence limits of each estimate. This holds for the data from normals (left) but not patients (right).

REFERENCES:

1 Cruse D, Chennu S, Chatelle C, *et al.* Bedside detection of awareness in the vegetative state: a cohort study. *Lancet* 2011; **378**: 2088–94.

2 Schomer DL, da Silva FL. Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields, Sixth. Philadelphia, PA, Lippincott Williams & Wilkins, 2010.

3 Whitham EM, Pope KJ, Fitzgibbon SP, *et al.* Scalp electrical recording during paralysis: Quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG. *Clin Neurophysiol* 2007; **118**: 1877–88.

4 Bishop CM. Pattern Recognition and Machine Learning, 1st ed. 2006. Corr. 2nd printing. New York, NY, Springer, 2007.

5 Noble WS. What is a support vector machine? *Nature Biotechnology* 2006; **24**: 1565–7.

6 Lemm S, Blankertz B, Dickhaus T, Müller K-R. Introduction to machine learning for brain imaging. *Neuroimage* 2011; **56**: 387–99.

7 Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Statist Soc B* 1995; **57**: 289–300.

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

9 Pfurtscheller G, Lopes da Silva FH. Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clin Neurophysiol* 1999; **110**: 1842–57.

10 Bai O, Lin P, Vorbach S, Li J, Furlani S, Hallett M. Exploration of computational methods for classification of movement intention during human voluntary movement from single trial EEG. *Clinical Neurophysiology* 2007; **118**: 2637–55.

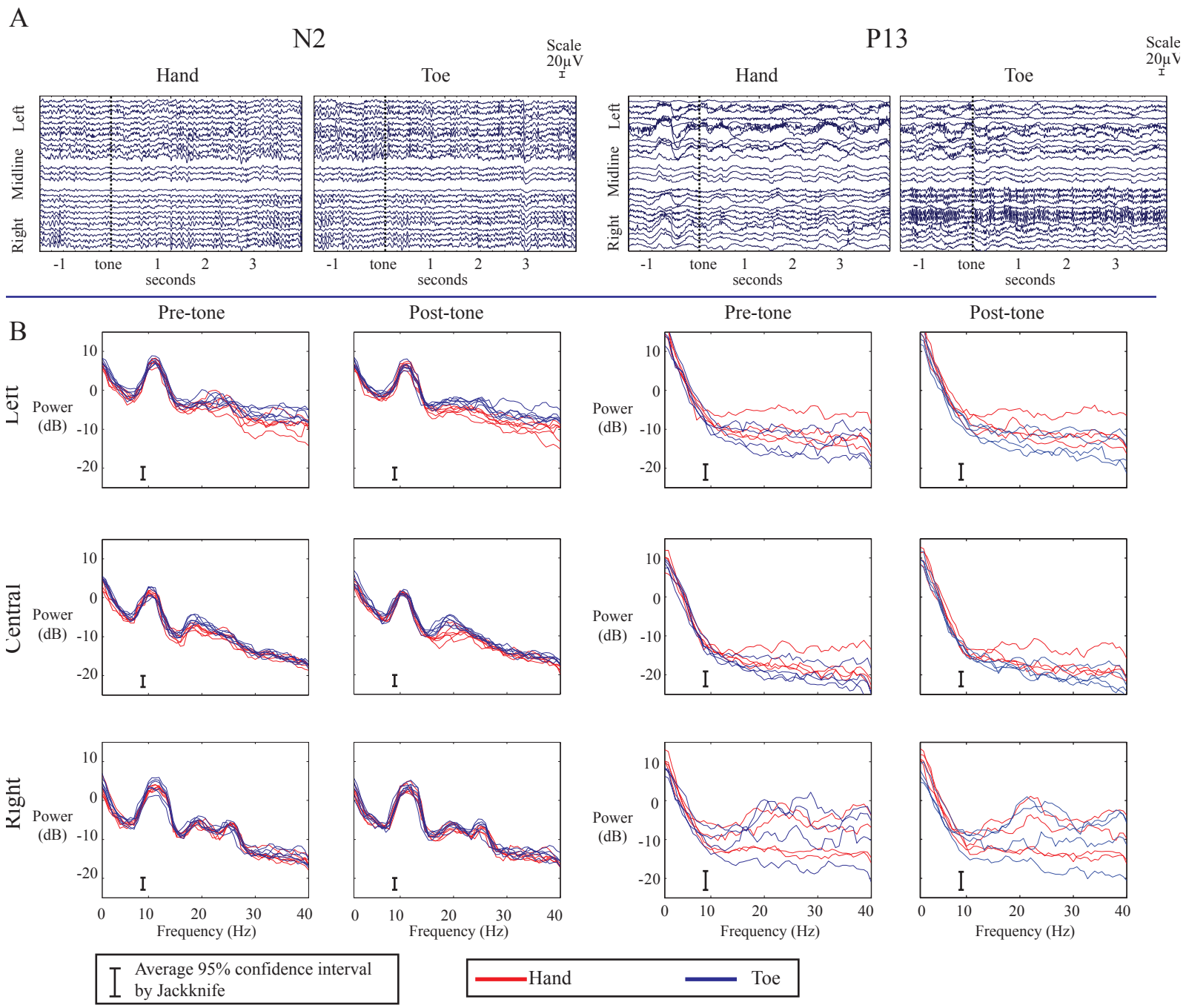
11 Cruse D, Chennu S, Chatelle C, *et al.* Relationship between etiology and covert cognition in the minimally conscious state. *Neurology* 2012; **78**: 816–22.

12 Daly JJ, Wolpaw JR. Brain-computer interfaces in neurological rehabilitation. *The Lancet Neurology* 2008; **7**: 1032–43.

13 McCarthy JF, Marx KA, Hoffman PE, *et al.* Applications of machine learning and high-dimensional visualization in cancer detection, diagnosis, and management. *Ann N Y Acad Sci* 2004; **1020**: 239–62.

14 Zhang F, Chen JY. Data mining methods in Omics-based biomarker discovery. *Methods Mol Biol* 2011; **719**: 511–26.

15 Hauser SL, Johnston SC. Extraordinary claims require extraordinary evidence. *Annals of Neurology* 2011; **69**: A9–A10.


A
N2
P13
Hand
Toe
Scale
20μV
Left
Midline
Right
-1
tone
1
2
3
seconds
B
Pre-tone
Post-tone
Left
Central
Right
Power
(dB)
10
0
-10
-20
0
10
20
30
40
Frequency (Hz)
Average 95% confidence interval
by Jackknife
Hand
Toe