Elsevier Editorial System(tm) for The Lancet
Manuscript Draft

Manuscript Number: THELANCET-D-12-03941R1

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

Article Type: Article

Corresponding Author: Dr Nicholas D Schiff, M.D.

Corresponding Author's Institution: Department of Neurology and Neuroscience, Weill Medical College of Cornell University

First Author: Andrew M Goldfine, MD

Order of Authors: Andrew M Goldfine, MD; Jonathan C Bardin, AB; Quentin Noirhomme, PhD; Joseph J Fins, MD; Nicholas D Schiff, M.D.; Jonathan D Victor, MD,PhD

Abstract: Background.
Cruse and colleagues (Lancet, 2011) described a new electroencephalographic (EEG)-based tool to detect fragments of consciousness, and using this tool, found evidence that 3 out of 16 vegetative state (VS) patients were able to perform a complex motor imagery task. Their analysis centered on applying a machine-learning algorithm, the support vector machine (SVM), to the EEG. Importantly, since there is no gold standard (e.g., behavioural measure) available for corroboration, their conclusions rest entirely on the statistical model used for validation.

Methods.
We first tested the statistical model used in Cruse et al. as applied to their dataset. We focused on its two key assumptions: independence of each trial, and lack of a special relationship between adjacent blocks. We then re-analyzed the data using a non-parametric approach to the SVM output that did not rely on these assumptions.

Findings.
Data from the 3 "positive" patients failed the test of trial independence, likely because of the presence of various artifacts. Data from 2 of these patients also failed the test concerning block relationships. The non-parametric reanalysis found no EEG evidence that patients performed the motor imagery task. In contrast, data from normal subjects typically met both model assumptions, and the non-parametric reanalysis, as expected, identified EEG evidence of motor imagery.

Interpretation
Unsupported assumptions in the statistical model used by Cruse et al. may account for the claimed evidence of an EEG signal of motor imagery in VS patients. This brings into question the findings of Cruse et al. (2011) and those of a follow-up paper using the same methodology (Cruse et al.,

2012). Application of machine-learning methods to large datasets such as EEG has great potential power, but requires a statistical model that is carefully vetted with data from the target population.

Funding
James S McDonnell Foundation, National Institutes of Health, Belgian Fonds de la Recherche Scientifique, Buster Foundation, Jerold B. Katz Foundation

*Reply to Reviewers Comments

CORNELL UNIVERSITY

NEW YORK PRESBYTERIAN HOSPITAL

Joan and Sanford I. Weill Medical College

Department of Neurology and Neuroscience
525 East 68th Street
New York, NY 10021

September 24, 2012

Zoë Mullan
Senior Editor
The Lancet
32 Jamestown Road
London NW1 7BY
UK

Dear Dr. Mullan,

We thank the editor and the reviewers for their detailed and careful review of our manuscript entitled Reanalysis of "Bedside detection of awareness in the vegetative state: a cohort study." We are pleased that they were strongly supportive of the importance of the re-analysis, and we have now undertaken the reorganization of the manuscript requested by the editor. Specifically, we have converted the primary manuscript into an approximately 1000-word, letter format document. The Webappendix includes the information removed from the manuscript, and is now written as an almost independent document with full details of the reanalysis, so that it retains the clarity of the original and allows for thorough evaluation of our work by interested parties. The original Figure 1 and Table 2 continue to be part of the manuscript, but the remainder of the figures and tables have been moved to the Webappendix. Only one figure was modified: the original Figure 4 (now Webappendix Figure 4) now has two additional components that replot the same data (an additional topographic headmap in section B and line graphs in section C) that we hope will clarify the points made in the original figure, and address issues raised in the review as noted below.

We have address all points raised in the reviews below and have modified the manuscript in response to the reviewer's comments. Specific responses to the reviewers are itemized below. Reviewer comments are in italic; our responses are in normal text; and specific changes from the text are in bold.

Thank you for your time and consideration.

Sincerely,

Nicholas D. Schiff, MD
Jerold B. Katz Professor of Neurology and Neuroscience
Professor of Neurology and Neuroscience (with Tenure)
Department of Neurology and Neuroscience
Weill Cornell Medical College

Reviewer #1: THELANCET-D-12-03941

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

Goldfine et al. reanalyse the data by Cruse et al and cast some doubt on their findings. In particular they correctly criticize the statistical independence assumption that is vital for the analysis of Cruse et al.

I essentially agree with this criticism (given that some missing info in the ms matches my prior) and also think that ultimately the comment should be published. However, at this point, I recommend a major revision, since (A) some information to make the analysis of Goldfine et al. transparent is missing. Without it the reanalysis in the submission would be irreproducible. (B) Furthermore some statements in the ms are a bit misleading and should be adjusted.

We have made a concerted effort in the transfer of the majority of the original manuscript into Webappendix form to provide a transparent and integrated document that should allow clarity and straightforward reproducibility of each step taken in this reanalysis.

Details

(A)

1. The model selection procedure for the SVM should be made very explicit. What regularization strength is being used. Also what kernel is being used (I assume it is a linear one).

We now give details in the Webappendix under "The Cruse et al. SVM approach", as we used exactly the same procedure as was used in the original study. Specifically:

> **An SVM classifier is then determined from a "training" component of the dataset (all of the data with one block of each type omitted), and its accuracy is determined by a "test" component (the two omitted blocks). Data are then normalized by subtracting off the mean and dividing by the standard deviation of the training features. A linear-kernel classifier is created with Matlab's 'svmtrain' with all default settings, except that 'autoscale' was disabled as data had already been normalized. (For further details on the default settings, see http://www.mathworks.com/help/toolbox/bioinfo/ref/svmtrain.html).**

Lemm et al 2011 is a good ref to compensate for overfitting during model selection while blocking effects are present. How exactly are the splits for cross-validation and testing made?

As mentioned above, the method used by Cruse et al. is now fully detailed in the Webappendix under "The Cruse et al. SVM approach". In brief, the cross-validation

approach used by Cruse et al. used adjacent block pairs; in our re-analysis we considered all block pairs as recommended by Lemm et al. (Webappendix Fig. 1).

2. From a ML perspective it is unclear how hard the classification problem presented in the ms is, so it would be important to have a baseline. The K-nearest neighbor algorithm is such a baseline. Note that it may also be rather practical in the present context, since both imagery classes and also rest can be contrasted. Also as described in Blankertzet al 2011 Neuroimage same volume as Lemm et al. shrinkage estimation maybe worth a try.

We appreciate the reviewer's comment for alternate ML approaches for classifying the subject datasets. However, our primary goal was to test the classification algorithm of Cruse et al. to determine its validity in normal and patient datasets. Whether there is an alternative ML approach that rigorously identifies task-related performance in patients is an interesting question, but far beyond the scope of our present re-analysis.

In terms of difficulty of the classification problem, this is of course hard to quantify, but we suspect that it is quite challenging. There are several reasons for this. First, there is no "gold standard" for patients, and we do not even know, a priori, whether any of the patients have a task-related signal. Second, the univariate analyses (see Webappendix Figures 3 and 4 and related text) show that the patient subjects who are positive in Cruse et al. have no evidence of a task-related change in their EEG, while the normals have the expected change in the sensorimotor rhythm (decreased EEG power between 7 and 30 Hz over sensorimotor cortex). Without a detectable signal on the standard univariate analysis, the classifier would have to rely on combinations of subthreshold signals – and without an a priori notion of where to look, this would appear to be difficult given the relatively limited amount of data and the very large number of possible features. Third, non-stationarity (or, at least, slow covariations) add further difficulty to the classification problem.

3. While the authors describe correctly that there is dependence in the trials, a further potential problem in the data incurred by the ML methods may be the underlying nonstationarity in the data. It is not at all uncommon that the underlying distributions of the first trials and the last ones (and the ones in the middle part) all are quite disjoint. Thus a reason for failure to decode the patients imagery maybe due to this nonstationarity. This point could be checked and should at least be discussed.

We agree: the underlying distributions of the trials from different parts of the experiment are in fact distinct, and that this could be taken as evidence of non-stationarity. We demonstrate this both at the block level (now Webappendix Figure 1 and related text) and at the trial level (Figure 1B and Webappendix Figure 2 and related text). We had not used the term "nonstationarity," though, because we are not able to distinguish between (a) a stationary process with multiple scales of temporal correlations and (b) one that is rigorously non-stationary. But we agree that "nonstationarity" is a useful notion, and we now include this term under the Webappendix sections "Testing the relationship between blocks" and "Testing independence of trials within blocks"

Because of nonstationarity, it is also theoretically possible that the patients performed the task in different ways in different blocks resulting in a negative result on the univariate analyses when all blocks are combined. We discuss this now in the Webappendix under "Potential for variation in task performance:"

> **One potential concern present for both the univariate and SVM approaches is that they use data from all blocks, assuming that the task was performed in the same way each time. If the changes in the EEG were not the same each time, then both techniques would have difficulty detecting them. This is not directly relevant to this manuscript since our goal was to test the Cruse et al. approach, which is run on multiple blocks at once. Nevertheless, to test this possibility, we ran the univariate analyses on P13 (best patient subject) separately for each block and still saw no evidence for task performance. When we ran the analyses for N2 (normal with similar classification rate), we see evidence for task performance on each block, though with slightly different patterns. This is akin to what we found in our previous work (Goldfine et al., Clinical Neurophysiology 2011) where evidence for task performance on individual blocks had slightly different patterns, but with sufficient commonality so that there was a stronger signal when all blocks were combined.**

(B) The discussion univariate vs SVMs is occasionally a bit misleading (I am sure the authors can pinpoint these parts of the ms easily). It is clear: if univariate features are able to distinguish between brain states then this is great (but also quite lucky). If not no further conclusion can be drawn. See for a recent discussion Biessmann et al Neuroimage April 2012.

Here too, we agree: multivariate techniques can be more sensitive than univariate ones, and the lack of a univariate signal does not imply the lack of a signal via multivariate techniques. This is why the crux of our analysis is the testing of the SVM statistical model, and not comparing univariate to multivariate results. We further clarified this emphasis in the reorganized manuscript, in which a discussion of the Cruse et al. statistical model is immediately followed by a re-analysis of the same SVM results via an alternative model.

Additionally, we now have a section in the Webappendix under "Details of the univariate approach" where we stress that multivariate and univariate approaches each have advantages and disadvantages.

Regarding the issue of "no further conclusion can be drawn" in the setting of a negative univariate signal – we agree that in general, this cannot be considered as predictive of whether a multivariate signal is likely to be present. But in this particular study, there is another ingredient, namely, that in normal subjects, a univariate signal is readily identified. Therefore, even if a signal can ultimately can be identified in patients via multivariate techniques, the signal would be qualitatively different than the easily-identified signal that is present in normals. We think this is important to mention this, since the biological meaning of a "positive" result in Cruse et al. relies not only on statistical significance, but also on making a connection with normal physiology. To make this point, we state in the manuscript:

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 conclusion, I think it is important to ultimately publish the ms, including the data and the code. It is an interesting and important discussion of an very much debated topic namely potential awareness in the vegetative state. However, for the moment, I suggest to supply further information as outlined above during a major revision.

We appreciate these comments and agree.

Reviewer #2: MAJOR COMMENTS
This is an unusual article, in that it is a re-analysis of a previously published paper, which comes up with a completely different conclusion. The importance of misdiagnosis of the vegetative state might warrant publication of such a paper.

Previous work (in relatively small number of patients) has suggested that fMRI might identify patients who are in a vegetative state by clinical criteria, but actually show signs of 'minimal consciousness' in their fMRI responses. The original paper by Cruse used simpler and more practical EEG-based techniques. They concluded that 3/16 patients, who had been diagnosed as being in a vegetative state, showed some ability to alter their (motor cortex) EEG in responses to command. This was interpreted as showing covert awareness. However the technique seems to be very difficult - both in the accuracy of the data collection, and in its statistical analysis. The very modest success of the control group seems to indicate a large component of chance/error in the use of the technique. (ie numerous false positives and false negatives)

Goldfine's re-analysis paper is primarily a statistical critique of the original paper. They suggest that the conclusions of the original are incorrect because of two well-known statistical problems:
1) Cruse did not sufficiently allow for slowly fluctuating correlations between/within trials - probably largely the result of EMG noise.
2) Cruse did not allow for multiple comparisons

They also suggest that it is better to use univariate (rather than multivariate) analysis in these cases where data numbers are limited.

1) The methods of Goldfine seem to be more robust - as they can distinguish controls vs patients better. They also are measuring a well validated real neurobiological phenomenon - namely the task-related suppression of high frequency EEG activity.

2) Both papers are the output of high profile, experienced, research teams. Both papers indicate that the use of EEG might be useful in the diagnosis of the vegetative state (they both showed differences between the patients and controls). But just visual examination of the raw EEG signal (or its FFT) might be just as good (see fig 1).

3) As a clinician, I would not strongly base my diagnosis/prognostication on the results of the Cruse paper, because the numbers are so small, and there are too many issues of what is the gold standard etc. the true test of any diagnostic test is its ability to reliably predict clinically relevant outcomes. This always trumps p-value chicanery.

4) Both papers demonstrate the deficiencies of applying the arbitrary Neyman-Pearson approach to diagnostic problems. Perhaps we are artificially forcing continuous data into a binary mould.

MINOR COMMENT
1) pg 4 - end. I agree that there are big (probably insoluble) problems with the EEG data acquisition; because the broadband EMG signal is too large to be ignored or even filtered out successfully.

We agree with all of reviewer 2's' comments. In particular, point 1 that using the univariate approach of task versus rest allows us to show the well-validated physiological signal in the normal subjects but is missing from the patient subjects. While SVM and other multivariate techniques do have their strengths, it is essential to examine the data with standard univariate approaches to connect the findings with known physiological changes. Length limitations prevent doing justice to all of these points, but we have made a strong attempt to address items 1, 2, and 3 within the confines of the reorganized Letter.

In regards to the minor comment, we now clearly mention in the manuscript that fluctuating muscle artifact is a likely cause of the false positive classification in the patients. "**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)"** and **"Specifically, the model does not allow for correlations between nearby trials and blocks, which are likely induced by fluctuating artifact and arousal state.**."

Reviewer #3: Reanalysis of bedside detection_ cohort study
Introduction:
** In general the introduction is fair. Try to elaborate further about the vegetative state. How much of the vegetative brain can perform conscious activities.*

Several studies have looked at elements of preserved cognitive function in vegetative state, but given the restriction in the new format, a broader review of the literature cannot be undertaken. Had the space been available, we would be eager to make the point that each patient with a severe structural brain injury significant enough to induce enduring conditions consistent with the clinical criteria for VS is unique, and can have any of a wide range of residual cognitive capacities. Published case reports of individual VS patients

demonstrating evidence of consciousness through brain imaging, however, do not imply that typical patients in VS have cognitive capabilities. Therefore, functional imaging tools such as EEG need to be interpreted on their own, with statistical testing that makes no assumption about individual patient capabilities.

* *You need to further rationalize the need for reanalysis of the data.*

We now include a sentence in the first paragraph regarding the primary rationalization for taking on the reanalysis. "**We were concerned about the method's validity because of the difficulty of the task in these subjects, and because of its critical reliance on certain statistical assumptions."**

**Important thing relate to the write up of the introduction, do not include the methodology and conclusion in this section.*

The paper is now completely reorganized as a letter with no demarcated sections.

* *Explicitly mention what were your objectives for this reanalysis.*

We now mention in the first paragraph that the objective is: "**To allow us to test the validity of the method**"

Methods:
* *I have certain reservations against your analysis. I think the method of cruse et al foe entering mid point data was quite justified. The method results in 30,000 features were well conducted.*

We are unclear of the reviewer's criticism here. We don't question Cruse et al.'s reasons for their choice of data segments or features; rather, our critique of the methods aims at whether the data satisfy the assumptions of the statistical models. In other words, we do not claim that the use of a large number of features is wrong, only that it makes the model assumptions a critical issue. We find that the patient data do not meet the assumptions of the models, whereas the normal data do. While this methodology is valid and accepted in the brain computer interface field using subjects with normal cognition and without severe brain injury, it needs modification to study non-communicative, severely structurally brain-injured patients. In these patients there is no gold-standard for performance, or expectation of normal EEG signal characteristics, and there is an associated higher likelihood of various artifacts. In the manuscript we state: "**Importantly, the model generally suffices for normals, where there is minimal artifact contamination**."

* *Secondly the method of analysis for test component by hand block 1 and toe block 1 was also appropriate. The method that you are using is very much prone to biased estimates.*

The reviewer makes an assertion but does not provide a basis for it, and we respectfully disagree. As we state in the manuscript, restricting the classification to adjacent blocks leaves open the possibility that the classifier is identifying idiosyncratic relationships that

are not robust. Put another way, if slow variations just happen to be on the timescale of block alternation, the power to reject overfit models can be substantially reduced. This is a well-recognized phenomenon and the literature provides a basis for our approach (Lemm et al., 2011, ref. 6 in the manuscript.)

To test this possibility, we followed the recommendations of Lemm et al., 2011 (and also see Reviewer 1) for use of non-adjacent blocks as test datasets. This ensures that features that slowly vary through the experiment are not used for the classification. We now explain our choice of methodology in detail in the Webappendix.

* *The reason for performing SUM classifier by binomial distribution was quite appropriate considering the distribution of the data. Look at the table and the values they have reported. Binomial distribution is an appropriate method in this scenario.*

Again, we respectfully disagree, and fail to see the logic in the reviewer's comment: the values reported in the manuscript do not provide a justification for the use of binomial statistics; the crucial issue in this regard is whether each trial can be considered an independent assay. Using binomial distribution statistics assumes that the each trial represents an independent assay of the classifier. This is unlikely to be the case if there are slow variations (e.g., changes in the level of muscle artifact) that run through the trials, as is the case for the patient data (Figure 1). As a demonstration of the inadequacy of the binomial test, we show that it yields too many outliers *in either direction* (both better-than-chance and worse-than-chance) when applied to the patient data (see Webappendix Figure 2, right).

Interestingly, Cruse and colleagues recognize the within-block dependence of trials in carrying out cross-validation (under "Classification and statistical analysis" they state: "This blockwise cross- validation procedure, in addition to the pseudorandomised block order, ensured that task-irrelevant intrablock and interblock correlations in the EEG did not significantly account for the classification results"), but they ignore these same correlations when they choose the binomial test as a way to assess the significance of the accuracy data.

Finally, we note that we do not imply that the binomial method for statistical significance is invalid in general, as the data from the normals meets the assumptions for the model; this point is made in the text of the Letter.

Reviewer #4: The article of Goldfine and colleagues is a re-analysis of data from a past study focused on detection of awareness in patients diagnosed with vegetative state. The authors test the statistical assumptions of the original Cruse et al. article and find them to be invalid. They furthermore apply a statistical test that does not depend on the invalid assumptions and demonstrate that the findings of Cruse et al. are not robust. As such, it is concluded that the original interpretation of awareness in the 3 patients with a diagnosis of vegetative state was not supported.

From my perspective as a clinician, I found the article clearly written, cogently argued, and ultimately compelling. Of course, the statistical nature of this article requires a thorough statistical review. However, the authors demonstrate at the very least the fundamentally important point that the interpretation of states of consciousness based solely on electrophysiological analysis is crucially dependent on the model. The publication of this article will serve as an important lesson for all such future investigations.

We thank the reviewer for the supportive comments.

Here are several relatively minor points to enhance the manuscript.

1) The manuscript is very clearly written. However, most of the readers of the Lancet will be unfamiliar with many statistical concepts. An additional table to help your clinical colleagues with the concepts involved will enhance impact.

We have enhanced the previous Table 2 (now simply the Table accompanying the letter) to summarize the tests and results from the paper. We have also modified the text of the manuscript to ensure that statistical concepts are clarified, but left in full details of the analyses in the Webappendix. The Webappendix also includes multiple expository to clarify our tests of the assumptions.

2) Avoid the term "fragments of consciousness", which is imprecise. The original authors were assessing awareness of the environment.

We have now removed this term from the manuscript and simply mention the cognitive processes required by the task (e.g., motor imagery, language and short-term memory).

3) Cruse et al did not include the basic spectral information. It seems that the spectral data could potentially support your statistical arguments regarding states of consciousness. If all patients indeed had dominant delta in their baseline state, shouldn't it be mentioned that this reflects a state of thalamocortical disconnection (such as slow-wave sleep or anesthesia) that is thought to be inconsistent with environmental awareness? Of course, this does not mean that the patient is incapable of being aware, but it does potentially provide some neurobiological support to your statistical claim.

We agree with the reviewer that the spectral features of the patient data make the positive results in the patients even more surprising (beyond that of the behavioral difficulty of the task). We therefore show a typical patient spectral dataset in Figure 1B and mention that these spectra are **typical of severe brain dysfunction, deep sleep or anesthesia**. Importantly, though, we do not make this the crux of our argument, because it is theoretically possible that a patient with a very abnormal EEG could perform a cognitively difficult task (such as reported cases of patients with apparent continuous absence epilepticus yet normal cognition – see Gökyiğit and Çalişkan, 2005).

4) A style point: I would carefully re-evaluate your article for tone. Although you do make several explicit statements about data sharing, the readers should feel that this is a collegial

evolution of ideas rather than a public refutation. I know that authors from both manuscripts are colleagues, but you want the reader walk away really feeling the importance of sharing primary data.

We have sought to achieve a neutral tone in the rewriting of the manuscript for the required reorganization and thank the reviewer for pointing out this concern. We now mention in both the first and final paragraphs of the manuscript that the data were shared and the importance of this data sharing in this type of work.

I congratulate the authors for this critical re-appraisal.

We thank the reviewer for this supportive comment.

Reviewer #5: This is an important paper doing something which is done too rarely: Examining data behind other researchers' conclusions. It presents a reasonable re-analysis, and is cautious in not suggesting that this analysis is a priori "better" than the original. However, for a journal like Lancet, the same point could be made in a shorter paper.

We thank the reviewer for the supportive comments and have undertaken shortening of the presentation as suggested by reviewer and required by the editors.

In the introduction, the Cruse et al result is discussed as presenting "fragments of consciousness". This concept seems rather undefined, and it is debatable whether these methods studying sustained cognitive processes are indicative of sustained conscious experience (see e.g. discussion in Overgaard, Lancet, 2011 or Overgaard, Progress in Brain Research, 2009). Even though this is obviously not the point of this paper, a cautious theoretical interpretation goes naturally with this "reanalysis agenda".

We have now removed the term “fragments of consciousness” from the manuscript.

Reviewer #6: [No comments were provided]

Reviewer #7: I like this study. You have shown that the findings of Cruse et al. are biased due to improper statistical analysis of SVM output of EEG. Moreover, you suggest the appropriate amendments of the method.
Major comment:
1. I still have one major comment that relates both to your study and to the study of Cruse et al. I do not agree that the borderline p-value should be the only criterion of positivity of EEG in detecting motor imagery in VS patients. The problem arises due to varying, non-standardized number of trials in different patients. The results from patients with larger number of trials are more likely to be positive solely because of smaller estimated variance of EEG classification accuracy. Compare, for example, the results of the Patient 1 (number of trials = 202; EEG classification accuracy 61.38%, p<0.01), that was considered positive, and the Patient 2 (number of trials= 113; EEG classification accuracy 61.90%, p= NS), that was considered negative in the report of Cruse et al. Clearly, there is about 50% chance that EEG

classification accuracy of the Patient 2 is as good as or better than EEG classification accuracy of the Patient 1. Therefore, the results of the Patient 2should not be considered negative, but inconclusive.

If we understand the concern properly, the reviewer's point is that the failure of a p-value to be significant should not be taken as sole evidence for the absence of a signal. We entirely agree, and we are now explicit about this with respect to the findings in Patient 13 (see Webappendix under "Interpretation of findings from P13"). But the goal of the manuscript is not to demonstrate that signals are absent (in fact, it is unclear whether one could ever demonstrate that signals are absent), but rather, that the method of Cruse et al. does not provide convincing evidence that the signals are present. We have attempted to convey this by expressing the findings of our re-analysis in terms of failure to reach significance, and refraining from the use of the term "negative." Interestingly, in our re-analysis, the "positive" patient with the greatest amount of data (Patient 1) no longer appears "positive," while the patient with the least amount of data (patient 13) has the lowest p-value. This suggests – but of course does not prove – that "failure" to obtain significance is not simply due to the limited amount of data.

Minor comments:
2. Shouldn't 'inter-trial' be 'intra-trial' at page 3, 11-th line of the third paragraph.

To clarify,' inter-trial' is the intended term here as we are referring to dependencies between trials, but within blocks (and therefore not within a single trial).

3. I suggest you attribute 'the tone' (command tone?) at page 4, line 5.

This sentence has now been moved to the Webappendix in the section "Data provided to us". We now precede the first use of the word "tone" with the word command. **All EEG data provided had already been preprocessed including: segmentation of the continuous data into trial epochs spanning 1·5 seconds before the command tone to 4·0 seconds after the tone, filtering (1 to 40 Hz), application of a Laplacian montage, and manual removal of trials with significant artifact.**

4. I suggest 'correctly classified' instead of 'correct' at page 4, 5-th line of the 4-th paragraph.

This sentence is now in the Webappendix under "The Cruse et al. SVM approach" and we made the recommended change.

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 $p\leq 0.05$. We therefore corrected for multiple comparisons via the False-Discovery Rate (FDR)[7]; normals remained significant but none of the patients were significant at $p\leq 0{\cdot}05$.

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; $p\leq 0.05$ 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.

***Manuscript with revisions highlighted**

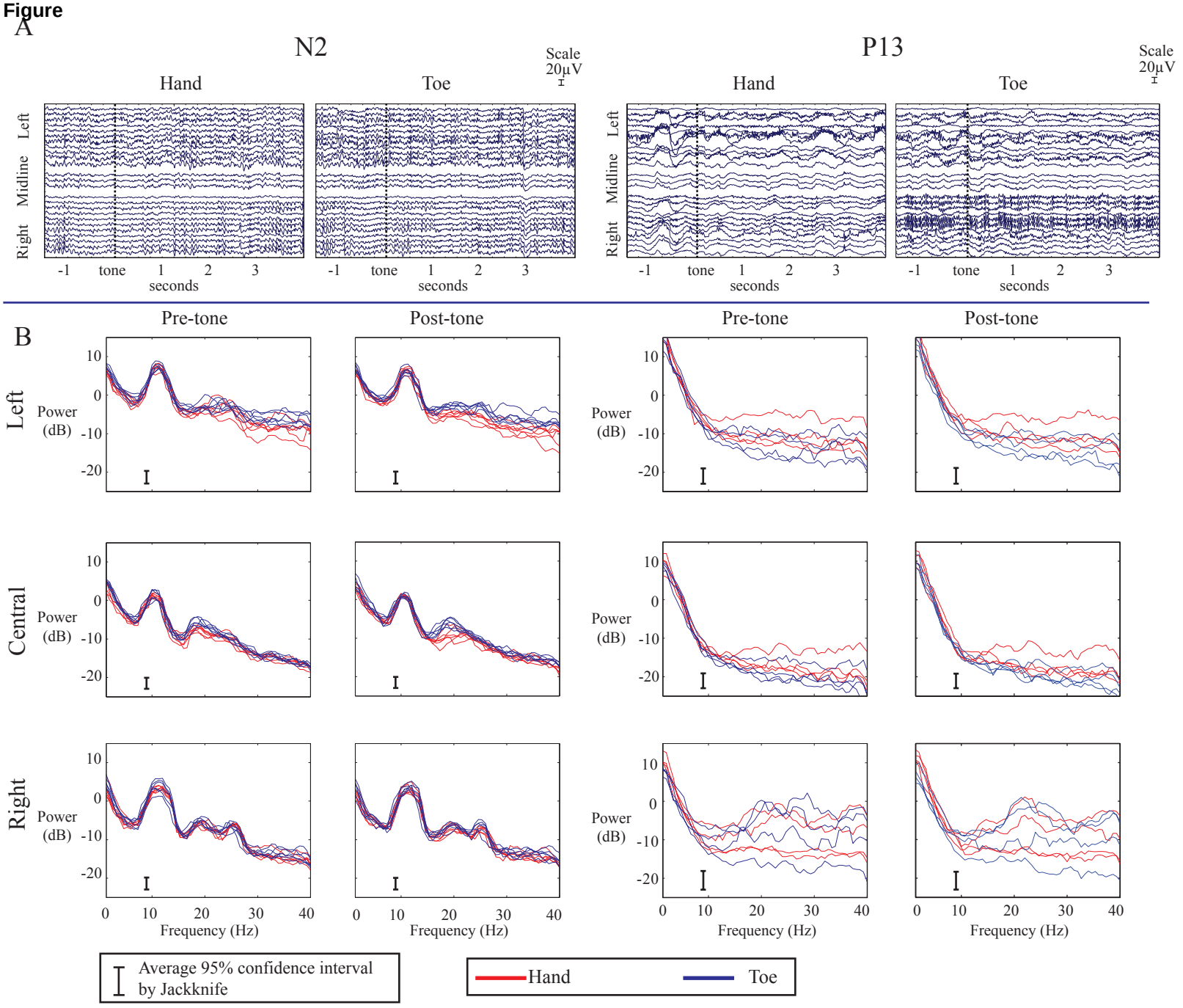
To the Lancet editorial staff,

Due to the change in manuscript format from a 3000 word manuscript to a 1000 word letter, there is no longer any text identical to the original manuscript. Therefore we are not submitting a separate manuscript with revisions highlighted (as the entire manuscript is revised).

Thank you,

Nicholas Schiff

Figure


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

Web Appendix

Click here to download Web Appendix: WebappendixAll.pdf