Additional comments concerning:
Bedside detection of awareness in the vegetative state: a cohort study (Cruse et al., 2011)

It is not surprising, given the importance of the issue, that Cruse et al. have responded vigorously to the WCMC re-analysis of their 2011 study. But we find it disappointing that their response (2013) fails to to address the crucial issues.

Below we go through the Cruse et al. response letter, point by point. We encourage the reader to read the Webappendix of our re-analysis letter, along with the original paper of Cruse et al. and their response letter, as these provide the foundations for the discussion below. (Our re-analysis letter was peer-reviewed by at least six experts. As we understand from the Journal, the response letter of Cruse et al. was not peer-reviewed.) We also mention that, although their response letter defends the validity of the procedures in the original 2011 study, Cruse et al. have subsequently modified their procedures in a way that appears to be directly responsive to our concerns – as elaborated in the next-to-the-last paragraph below, and the last paragraph of their response letter.

Comments

Paragraphs 1 and 2 of the response letter concern how the significance of the correctly-classified fraction of trials was determined. Our point was that Cruse et al. used a binomial test, which amounted to assuming that trials within the same block were independent. We showed that trials within a block were not independent, that therefore the binomial approach used by Cruse et al. was not valid. We then re-analyzed the data via another approach that does not suffer from this pitfall – a permutation test that respected the block structure. With the permutation test, we found that at least two of the three patients of Cruse et al. would not be considered "positive." Importantly, Cruse's letter does not defend their original approach (i.e., the use of binomial statistics and a blocked design), but instead offers reasons that alternative approaches would suffer from other drawbacks.

The first paragraph ("The primary suggestion…") of their response states that our use of the permutation test was invalid because there was not enough data. They are correct that permutation tests are data-hungry, but their comments are misleading, since the limited-data concern about permutation tests actually strengthens our interpretation. That is, the only patient who might be considered significant by the permutation test (P13: raw p-value 0.03) was the patient with the *least* amount of data (4 blocks), and it is for this patient that the permutation test should be considered the most suspect. For the three patients with 8 or more blocks of data (for which one would anticipate that the permutation test is fully reliable, as there were more than 1000 permutations), raw p-values were 0.09, 0.64, and 0.25; the first of these (P1) was also one of the patients that Cruse et al. considered positive. We also included a control: when our test was applied to the three normals supplied by Cruse et al. who were found to be positive in their hands, we found raw p-values of 0.05, <0.01, and <0.01. These normal-subject datasets had only 6 blocks of data. So the permutation test, even with datasets of the size available, is reasonably sensitive. These data are summarized in Webappendix Table 1. (Also, we do not understand how they arrived at the figure of 36 permutations, as this is neither a factorial quantity nor a relevant number of pairings.)

The second paragraph ("One could argue…") of the response offers a biological justification for the use of a block design. We agree with the stated merit of a block design, but we feel strongly that the statistical analysis must take into account the pitfalls that are inherent in it, and again emphasize that the Cruse et al. analysis did not do so.

In paragraph 3 ("Moreover, Goldfine…"), Cruse et al. take issue with our inference that there were correlations between trials within blocks, based on the U-shaped distribution of p-values. They raise the point that the patient group may not be homogeneous, as it might contain patients who are able to carry out the task, and

thus, skew the distribution of p-values. This is a red herring, for two reasons. First, our argument does not assume homogeneity among the patients: if individual subjects have p-value distributions that are either flat or biased towards the low end, then it cannot be U-shaped when the individual distributions are combined. Moreover, our argument was based on the excess of *large* p-values (i.e., p-values close to 1); the presence of patients who might carry out the task would only lead to an excess of *small* p-values (i.e., p-values close to 0). They also state that if our test is applied to patients individually, there is no evidence for correlations. This is also beside the point: applying the test of a flat p-value distribution one patient at a time reduces its power drastically, so the observation that data from individual patients pass the test does not provide the necessary reassurance that the statistics are properly calibrated. Note also that we included a control that shows that our test was not overly stringent when applied to a group: when applied to the data from all normal subjects, we find no evidence of a U-shaped distribution, and hence, no evidence for correlations. We encourage the reader to view the section of our Webappendix related to Figure 2 for further details, and for the rationale for pooling across subjects for this analysis. Finally, we carried out an entirely independent test for correlations between trials within blocks, applicable to data from individual patients. This also showed (main Figure panel B) that such correlations were present in patient data, but not in that of normal subjects. It also suggests that the source of these correlations, at least in part, is muscle artifact.

Paragraph 4 ("Although there are…") concerns the vetting of an analysis procedure by its ability to detect command-following in normals. However, once again, their response is one of mis-direction. They state that our analysis approach only detects command-following in "40%" of normals. This is not what we showed. We showed that statistical analysis of *their* multivariate approach via the permutation test detects command-following in two of the three normals that *they* found to be positive (they provided us with data from five normals: these three normals, and two others that *they* found to be negative.) We also showed that a straightforward univariate approach (Goldfine et al., 2011) applied to the same data detected evidence of task performance in *all 5* normal subjects and *none* of the patients. But more fundamentally, our message is that the performance of the analysis procedure in normals is of surprisingly little value as a vetting procedure, because the statistical properties of the EEG in patients are so different from that of normals.

Paragraph 5 ("Goldfine and colleagues…") concerns our comments about the differences in the spatial and spectral characteristics of the EEG changes seen patients and controls. In their letter, they appear to question why we consider these differences to raise questions about the meaning of the patient "responses," and point out that in a previous paper (Goldfine et al., 2011) we showed that patient responses could be highly significant yet still different from those of normal subjects. But they miss two crucial points. First (see Webappendix Figure 4 and the discussion of it), the spectral changes seen in the Cruse et al. patients do not reach statistical significance, while those in Goldfine et al. (2011) did. Second, in contrast to the changes seen in the Cruse et al. patients, those seen in the Goldfine et al. (2011) patients have a clear spatiotemporal structure.

But there is also a troubling element of misdirection in this paragraph: in the response letter, Cruse et al. state that differences between responses in normal subjects and patients are unsurprising, but in the original (2011) paper, they stated that the positive responses they identified in all three patients were "formally identical" to those in normals (see caption to their Figure 2.)

Paragraph 6 *("These methodological concerns…")* questions our use of a multiple-comparisons correction, claiming that Cruse et al. 2011 had a specific a priori hypothesis, so no multiple-comparisons correction was necessary. But again, the reader is misdirected by the response letter. The problem is that the specific hypothesis of Cruse et al. (2011) is at a population level, not at the level of individual patients. They wrote, "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." Therefore, a multiple-comparisons problem remains, because the crux of the issue is to determine if the presence of a few low p-values within the population represents an occurrence that is not likely to be due to chance. To put it another way, in completely random datasets, one expects 5% of individual

subjects to be "significant" at $p<0.05$ (just by chance), so having a small fraction of patients with low p-values does not necessarily indicate that those patients were able to carry out the task.

To address this issue, we used two approaches, both standard. First, we used the false-discovery rate correction, which controls the probability that a declared "positive" might be due to chance. At $p<0.05$, none of the patients reached this threshold. Second, we asked whether the entire distribution of p-values was consistent with a flat distribution, which is the null hypothesis. At $p<0.2$, this could not be ruled out (Kolmogorov-Smirnov). Finally, we showed that our approach is not too stringent, as the same tests – which were consistent with the null hypothesis in patients – clearly ruled out the null hypothesis in normals.

Paragraph 7 ("Finally…") states that the results of the Cruse et al. analysis were corroborated by fMRI obtained during the same week. We do not know the level of detail at which the fMRI study was scrutinized, and Cruse et al. do not report on the fMRI studies in the negative patients, so it is difficult to comment on the utility of the fMRI data as a corroboration. But more fundamentally, it is difficult to see how it has implications for the intrinsic validity of the statistical approach used in the EEG analysis.

Paragraph 8 ("In conclusion…") states that we used an "unconventional" cross-validation approach that indicated that "suggest[ed] that the EEG responses of two of our three positive patients became less consistent across time." Our approach was not "unconventional"; in fact, it is the recommended approach for a block-design study, as we cite in our letter (Lemm et al., ref. 6). It is also inaccurate to say that the responses merely became "less consistent over time." As seen in Webappendix Figure 1, they fell to chance levels in these two patients (P1 and P12).

Perhaps the most telling and constructive part of the response is the mention of their 2012 PLoS-ONE paper, in which Cruse and colleagues use a different behavioral paradigm and a different analysis approach to identify command-following in a single patient (not one of the original 16 patients of the 2011 paper). There were four significant changes in the 2012 approach, compared to the 2011 approach: (i) a randomized design rather than a block design, (ii) a permutation test rather than a binomial test, (iii) randomized cross-validation rather than cross-validation by deletion of adjacent trials, and (iv) classification based on 8 physiologically-motivated features driving a simple Bayes classifier (as we understand it, a quadratic or linear discriminant), rather than based on thousands of features and a support vector machine. Together, these changes appear to address all of our concerns. We had presented our re-analysis to the Owen group in early 2012, and the PLoS-ONE paper was received by the Journal on August 22, 2012. This is a very constructive outcome.

In sum, we showed (i) that the statistical assumptions underlying the Cruse et al. analysis were invalid, (ii) that analyses of their classifier via tests that do not make use of these assumptions yield results consistent with chance, and (iii) that the same procedures detect task-dependent signals in normals. But we emphasize that we have not shown whether the patients in Cruse et al. 2011 can or cannot follow commands – we only showed that their analysis methods are invalid.

Jonathan D. Victor
Andrew M. Goldfine
Joseph J. Fins
Nicholas D. Schiff

28 January 2013