

MEMORANDUM

22 August 2006

To: Philip W. Lavori, PhD
From: Mark A. Demitrack, MD 
Re: Request for Summary Statistical Opinion for Panel Meeting Packet Materials

[Dr. Philip W. Lavori, Professor of Biostatistics, Stanford University, has served as a statistical consultant to Neuronetics since 2003. He provided principal input in the design of the statistical plan for Neuronetics' clinical studies, and also consulted in the review and interpretation of the statistical analyses that were conducted and submitted in Neuronetics regulatory submission.]

Dr. Lavori is unable to attend the FDA Advisory Panel Meeting that will review Neuronetics' regulatory submission on 31 October 2006. In lieu of his participation at the Panel Meeting, Neuronetics asked Dr. Lavori to provide his expert opinion on two key statistical questions that were raised by FDA regarding Neuronetics clinical study data and that may be raised during the Panel Meeting. Dr. Lavori's written responses to these two questions are contained in the attached letter that was kindly provided by Dr. Lavori in response to our request.]

Request to Dr. Lavori:

Dear Dr. Lavori,

On behalf of Neuronetics, I am requesting that you provide your expert opinion on two key questions pertaining to our clinical studies. These questions have been raised by the FDA during their review of Neuronetics' clinical data. Here, I have paraphrased these questions in the form of the following two questions stated below. We ask that you consider the evidence obtained from Neuronetics' clinical studies and to offer your expert opinion on these statistical concerns:

- 1) The sponsor has proposed that, in the face of a P value of 0.057 on the primary outcome measure, it is nonetheless legitimate to consider the statistical evidence of effect on the secondary outcome measures, i.e., to appraise the outcome of study based on the totality of the observed data. Is this appropriate from a statistical view?
- 2) On the key question of the failure of the primary outcome measure to fall below the conventional threshold of $P < 0.05$, the Sponsor has argued that the additional subset analysis that removes from the evaluable population the patients whose baseline MADRS total scores fell below 20 reveals a P value = 0.038. Can you comment on the appropriateness of this subset analysis, and in so doing, also comment as to why the ANCOVA performed by the sponsor did not address the issue of the imbalance in baseline scores on the MADRS?



Memorandum

4 September 2006

To: Mark Demitrack, MD
From: Philip W. Lavori, PhD

1. In response to the first question in your memorandum of 22 August 2006, I have formed a statistical opinion about the overall strength of evidence contained in the Neuronetics database, specifically with regard to the issue of the failure to reach nominal statistical significance in the primary analysis, while several secondary analyses have small nominal P-values. I conceptualized this as a test of an overall null hypothesis that TMS had no effect on any measure of depressive symptoms. To do so, I performed the following steps:

a. From Table 3 (A Priori-Defined Primary and Secondary Outcome Measures Observed in Study 44-01101) of your 'AI Response to FDA letter of 24 May 2006.doc' I extracted the P-values for the treatment comparisons of the primary outcome and the secondary outcomes (at week 4) that involved 'scored' variables measuring the symptoms of depression. The final list of 13 variables is provided at the end of this memorandum. My rationale for excluding the 'remission' and 'response' variables was that the study was not powered to detect differences in binary variables, and my rationale for omitting the scores corresponding to outcomes other than depressive symptomatology follows from my intent to focus on the antidepressant effect of TMS. I did not include the 'baseline imbalance corrected' MADRS score P-value, because I wanted to stick with pre-specified analyses. I did not use the week 6 values because they are heavily imputed, due to protocol-driven dropout from Study 44-01101, and the LOCF (or any other imputation method) is inappropriate for such a 'missingness mechanism'.

b. Having defined the 'family' of comparisons, I used four methods that adjust P-values for multiple comparisons, generally known as Holm, Hochberg, Hommel, and Benjamini-Hochberg, as implemented in the R function 'p.adjust'. All four methods (which improve on the Bonferroni inequality method, which is highly conservative) yield a significant ($P < 0.05$) rejection of the overall null hypothesis (that none of the 'depression' scores are lowered by TMS compared to sham). They differ in the number of specific hypotheses that would be rejected, reflecting their varying conservatism in the context of positively correlated tests (Holm and Hochberg would reject 1 of the hypotheses, while Hommel, a more powerful test, would reject 4 of them, and Benjamini-Hochberg would propose that a 'false discovery rate (FDR)' criterion would reject 9 of the 13. My own view is that the Benjamini-Hochberg criterion is the most sensible one for this context, given that the underlying construct of depressive symptomatology is common to all these measures, so the number of true nulls is either 13, in which case the false discovery rate is just the Type 1 error rate, or close to 0, in which case the FDR is close to 0. In light of these analyses, my opinion is that the overall pattern of P-values for primary and secondary depressive symptom scores at the pre-specified week 4 endpoint, is not consistent with the absence of an effect of TMS vs. sham on at least one of the scores. Indeed, the FDR analysis leads me to conclude that there is statistical evidence for an effect on at least 9 of the 13 scores, including almost the entire 'family' of scores based on the HAMD.

2. In response to the second question in the memorandum of 22 August 2006, I offer the following opinions concerning the subset analysis, opinions that are informed by and dependent on the analysis described in (1) above. Having decided that the overall null hypothesis can be rejected, it is useful to try to understand 'what went wrong' with the primary analysis. The 4 methods for adjusting for multiple comparisons described above agree that the MADRS (primary) null cannot be rejected at the 5% level.

This could of course be a consequence of random noise, but the observation that patients with MADRS <20 at baseline entered the study (while the HAMD baseline score was the basis for exclusion) and the finding that excluding such patients improves the effect to the extent that the P-value falls below 5%, suggests another explanation, that patients with low MADRS at baseline experience a 'floor effect', limiting the degree of improvement (and the difference between TMS and sham) at week 4. Furthermore, the imbalance in the count of such patients (more of them in the group randomized to TMS) would make the TMS group more subject to the floor effect. If this were true, the ANCOVA (adjusting for baseline) would not address the issue, since it should be characterized as an interaction effect rather than a main effect of baseline score. Thus, the subgroup analysis and the ANCOVA analysis are directed at different needs for adjustment. I do not believe that the subgroup analysis would stand on its own; rather, it is my opinion that the results of the analysis provide a possible explanation for the failure of the primary analysis, in the light of the relatively convincing secondary analyses (as described in 2 above).

I hope these opinions are useful to the Panel.

Sincerely,

A handwritten signature in black ink, appearing to read "Philip W. Lavori". The signature is fluid and cursive, with a long horizontal stroke at the end.

Philip W. Lavori, PhD
Professor of Biostatistics
Stanford University

Appendix: Results of multiple comparisons analysis:

Note: 'pval' is the raw p-value, from Table 3 as cited above

k is the order of the raw p-value, from highest (1) to lowest (13) 'kxp' is the product $k \cdot pval$
the last 4 columns are the p-values from the Holm, Hochberg, Hommel, and Benjamini-Hochberg methods, respectively.

Vaname	pval	k	kxp	holm	hoch	homm	bh
1 MADRS Total Score	0.057	4.0	0.2280	0.228	0.174	0.171	0.069
2 HAMD 24 Total Score	0.012	6.5	0.0780	0.084	0.072	0.072	0.020
3 HAMD17 Total Score	0.006	11.5	0.0690	0.072	0.063	0.042	0.018
4 SF-36 Mental Health Score	0.006	11.5	0.0690	0.072	0.063	0.042	0.018
5 HAMD Anxiety/Som Score	0.025	5.0	0.1250	0.125	0.125	0.100	0.036
6 HAMD Core Depr Score	0.012	6.5	0.0780	0.084	0.072	0.072	0.020
7 HAMD Maier Score	0.003	13.0	0.0390	0.039	0.039	0.027	0.018
8 HAMD Gibbons Score	0.007	9.5	0.0665	0.072	0.063	0.042	0.018
9 HAMD Retardation Score	0.007	9.5	0.0665	0.072	0.063	0.042	0.018
10 HAMD Sleep Score	0.211	1.0	0.2110	0.362	0.211	0.211	0.211
11 IDS-SR Score	0.058	3.0	0.1740	0.228	0.174	0.174	0.069
12 CGI-S Score	0.009	8.0	0.0720	0.072	0.072	0.054	0.019
13 PGI-I Score	0.181	2.0	0.3620	0.362	0.211	0.211	0.196