Reviewer's report

Title: Potential risk factors associated with human encephalitis: application of canonical correlation analysis

Version: 1 Date: 25 December 2010

Reviewer: Aeilko Having Zwinderman

Reviewer's report:

The manuscript assesses the associations between exposure variables and symptom and diagnostic variables in a sample of 263 patients with confirmed or suspected encephalitis. Fifteen exposures are associated with 24 symptoms using canonical correlation analysis. Because all variables were dichotomous (or were dichotomized) an extension of Gini’s index was used to quantify the pairwise correlations. Main results was a cluster of 10 symptoms that was significantly related to a cluster of 10 outcome variables.

The manuscript is well written and the analyses seem appropriate, but I have a number of remarks.

1. Because the problem is presented as trying to "define" or diagnose encephalitis I wonder whether it is better to use a multivariate technique that maximizes the explained variance of the outcome variables instead of the canonical correlation (f.i. redundancy analysis). This issue could be mentioned at least in the discussion or perhaps the authors could do the redundancy analysis and report on any difference with cca.

2. The inclusion of duration of hospital stay (and perhaps illness too) seems odd if the primary objective is diagnosis. Or is it customary to make the diagnosis at the end of the hospitalization (illness) and if so what is the consequence of having this diagnosis? I would suggest no to include these two variables in the outcome set, but their removal will have little consequence.

3. It is unclear to me what the extension of Gini’s coefficient entails and some elaboration is necessary. What is the vector x(ia)x(ib)? What makes this extension useful for binary data? Are the results robust in the sense that similar results are obtained with other coefficients? For instance, tetrachoric or even phi correlations. (If the psd-requirement of X or Y is important there are many Gramian coefficients: Hamann, Yule, ...)

4. The interpretation of the components using the structural coefficients is not clear. Are these the canonical weights? But weights are difficult when it comes to 'importance' of variables if the variances of the variables are different. These will vary since the "prevalences" vary quite a lot. In that case the 0.3 threshold is not really useful.

5. Also unclear is how the missing data were handled. Were only complete cases
6. On page 7, in the fourth paragraph the results of the CCA are described as suggesting the variables are "risk factors" for encephalitis. This seems odd because all 263 (or 208) patients had encephalitis. So which risk is meant here? I think it is much better to describe the cca results in the sense of the association of exposures with particular symptoms: like "younger patients with water, tick, ... exposure have abnormal wcc, headache, abnormal eeg ,...".

7. The results are also formulated in terms of importance. That is a daunting aim and requires data on the stochastic variability of the weights (and also requires the variables to have the same variance). So how significant is the difference of the EEG weight of 0.7 and the WCC weight of 0.9 or the coma weight of 0.6?

8. On page 8 there is a reference to "[reference]"; this should be detailed.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** Yes, and I have assessed the statistics in my report.

**Declaration of competing interests:**

I have no competing interests.