Reviewer's report

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

Version: 1 Date: 17 December 2010

Reviewer: Sébastien Dejean

Reviewer's report:

- Major Compulsory Revisions

1) It is not clear for me whether the method is not appropriate to address the problem or whether the problem is not well defined. Indeed, the background highlights the challenge lying in 'distinguishing encephalitis from other disguising conditions', but as every patients in the study are confirmed encephalitis cases, this question cannot be addressed. As mentioned in the section 'method', CCA can 'assess association between a set of exposure and a set of symptom and diagnostic variables' but I cannot see how it could also identify potential risk factors

*associated with human encephalitis*. Other sentences in the paper ('Age is a potential risk factor in developing encephalitis', 'Other variables that have to be found to be strongly associated with encephalitis exposure'...) cast doubt on the real purpose of the paper. At the end of section 3, when discussing results over 263 patients, it is written '...identified the same set of exposure variables as potential risk factors associated with symptom, clinical and diagnostic variables', which is more in agreement with the actual results of CCA. An exposure variable can be very important to distinguish encephalitis, but will not be identified by CCA if

no significant correlations exist with at least one symptom, clinical or diagnostic variable.

2) Using Gini's measure to build a covariance matrix enable to perform computation for CCA on qualitative variables, but what about the interpretation of the results. For instance, how to interpret 'linear combinations of X set variables' mentioned page 7?

3) Qualitative variables are usually explored directly using Multiple Correspondence Analysis. Why not to use this method, potentially as a preliminary step? Relations between variables would also appear in the plots provided by MCA.

- Minor Essential Revisions

4) "Duration of illness" & "Length of Hospital stay" are quite obviously highly correlated and this may hide others potentially relevant correlations. Only one variable would be enough in the study.
5) Concerning the heatmap: i) labels on the figure will facilitate interpretation; ii) 'large size diamonds': does it refer to patterns along the diagonal corresponding to sets of highly correlated variables?

6) Some paragraphs are not justified.

7) Figure 3 is mentioned at the end of the section 3 but does not exist in the pdf file I have downloaded.

8) Page 7: a reference is missing in the text just after CCA : "[reference]"

9) End of page 7 : "vitiate" -> "variate"

... - Discretionary Revisions

10) Patients were recruited in England and only 14.9% traveled within UK. Is it OK? How a travel is defined?

11) Table 2 contains many figures and it is quite difficult to have a comprehensive overview of the data. It could be replaced by a "Mosaic plot" (function mosaicplot in R).

**Level of interest:** An article of importance in its field

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

I declare that I have no competing interests.