**Reviewer's report**

**Title:** Direct Estimation of Cause-Specific Mortality Fractions from Verbal Autopsies: Multi-Site Validation Study Using Clinical Diagnostic Gold Standards

**Version:** 1  **Date:** 23 May 2011

**Reviewer:** Gail Williams

**Reviewer's report:**

The paper deals with a well-posed question and provides new information on the validity or otherwise of a recently proposed verbal autopsy process. The data used provide a gold-standard cause of death against which to test validity of verbal autopsy methods, which is very unusual. This information is important because accurate and cost-effective methods of establishing cause-of-death is vital in settings where death commonly occurs in the absence of medical attendance.

**Overview:**

The methods of the paper appear appropriate but there seem to be some gaps in the description of them. These are dealt with in the questions/comments below. The data are derived from several field sites within a major project. Unfortunately (and this remark applies to other aspects of the paper) detailed information on data sources are given in another paper under submission, and are not accessible to this reviewer. Some aspects of the reporting are unclear. Conclusions appear well supported by the results presented. The writing could be improved and appears in need of editing – several phrases are a little awkward or imprecise.

**Details:**

'Test' data are created in what appears to be a three-stage process: (i) drawing of 500 x 25% random subsamples of the original pooled set of deaths, then (ii) sampling from an uninformative Dirichlet distribution to obtain a set notional cause specific mortality fractions (CSMF), then (iii) sampling from the dataset (i), with replacement, to construct a set of final test datasets which have CSMFs corresponding to those obtained in (ii).

**Question/Comment 1:** Define specific ages for each of the age groups. 'Neonate' is usually used to refer to a liveborn child, so it is unclear how 'stillbirth' appears as a cause of death in Figure 5.

**Q/C 2:** How was the number of deaths determined at stage (iii) - reference was made to a 'prescribed total' but this does not seem to have been defined.

**Q/C 3:** If test data were literally generated from an uninformative Dirichlet, one would expect approximately equal fractions in each cause-specific group. This
needs to be clarified, as the abstract suggests that a wide variety of distributions was obtained. Why was an uninformative Dirichlet used (if it was) as this appears to be quite an artificial distribution compared to reality?

Some justification of the metrics used to assess validity of the KL and PCVA methods against the gold-standard cause specific mortality fraction (CSMF) is given in a paper which is in submission, which creates a difficulty in judging their choice.

Q/C 4: Could some further justification and explanation of the metrics be given in the methods section, unless the submitted paper is available, at least in press. Specifically could CSMF accuracy be explicitly defined? I assume it is the % of correct (according to the gold-standard) causes of death identified, either cause-specifically or over all cause by the method.

In 'Assessing the relationship between KL CSMF accuracy and the number of causes', up to 46 causes of death are randomly selected.

Q/C 5: Why 46?

OLS and corresponding RMS are used as metrics to perform overall comparisons.

Q/C 6: Did the authors consider forcing the regression line through the origin? This would give a better interpretation of RMS as directly related to the error in estimated vs true. It would however obviously remove the intercept as a measure of bias at low fraction. Unfortunately as the rationale for using the slope as a quality measure is referred to a paper in submission, it is not clear what aspect of quality the slope is measuring. By definition, the slope measures the increase in estimated mortality fraction per unit increase in 'true' mortality fraction and it is unclear why this is a measure of quality of the VA process, in terms of absolute accuracy or validity. Clearly a valid estimate would have slope close to unity and intercept close to zero, but how does one interpret slope-intercept combinations which are not close to those desired? While it is far from a perfect relationship, the Annex 1 data appear to demonstrate that slopes close to unity go with intercepts close to zero (stillbirths) and slopes well below unity go with larger positive intercepts (pneumonia in children) - is this bias or random measurement error in the estimated values, the latter which will technically contribute to attenuation of slopes and higher intercepts?

Q/C 7: No description is given for calculations involving PCVA - perhaps these are in papers also in submission.

Results

Q/C 8: How were the 95% CIs for median accuracy calculated?

Q/C 9: Why don't variances of some symptom profiles exist?

Examples of comments on writing
We repeated our assessment varying the number of symptoms in cluster from 8 to 18. We also explored varying the number of draws from 200 to 600.

Results section: ‘This method generates test/train datasets where there is no correlation between the CSMF composition of the training datasets and the test dataset’ – use of the concept of independence would be better than using the term ‘correlation’ which has specific statistical meaning.

Caption for table 1, second sentence – cannot understand what this means.

Use of ‘n.subset’ throughout seems unnecessarily obscure.

Articles are sometimes inappropriately included or omitted.

Revisions
Q/C 3 requires a compulsory response; other amendments should be made as indicated. Q/C 6 is discretionary.

**Level of interest:** An article of outstanding merit and interest in its field

**Quality of written English:** Needs some language corrections before being published

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

**Declaration of competing interests:**
I declare that I have no competing interests