

Reviewer's report

Title: Verbal Autopsy interpretation: a comparative analysis of the InterVA model versus physicians' review in determining causes of death in the Nairobi DSS.

Version: 1 **Date:** 13 February 2010

Reviewer: Gary King

Reviewer's report:

I don't see how the author can claim that "the validity of PR is well established" when the inter-coder reliability of different physicians is usually very low. If you don't have reliability, you can't have validity.

The inter-coder reliability for the three physicians used to code COD in this sample needs to be reported prior to them meeting to resolve differences.

Is there evidence that having the physicians meet to resolve differences is more likely to get to the right answer than other approaches. In most situations like this, keeping the experts apart is a better approach, since asking them to average is asking them to do the research. Averaging can be done by the analyst (with ties going to the modal category from the data), with uncertainty appropriately represented.

The equation on p.11 that is supposedly the key to the author's method has four unique quantities on the right side, but the text says that 100 are elicited and 60 categories of outcome variables are included. In a revision, the method needs explication so that readers can understand what's going on. I'm afraid I can't figure it out and so can't evaluate it. If standard mathematical or statistical notation is used, it should be easy.

Statements like this need to be explained, even if they are explained in prior literature: "Additionally, the InterVA model has provisions to adjust for the prevalence of malaria and HIV/AIDS in any setting such that before running the model, the prevalence of HIV/AIDS and malaria in the study population can be set as high or low." What

exactly is the method being used here? I infer that the method is fairly straightforward, and so I don't see any reason why it can't be presented in sufficient detail so that readers can understand what it is.

Why should the analysis be done on 60 causes of death and then grouped into 14 rather than just running the method on 14? Is there any advantage to one way vs the other? If the method is being run on 60, then it is essential to see the results from that analysis prior to grouping. The accuracy will be lower, but that's fine; we just need to see what it is.

p.13. using only the best estimate of the cause of death from the model can be highly biased unless the predictions are all perfect. For example, suppose the probability of 1000 deaths falling in COD category 1 is 0.51 and the rest in category 2. Then the procedure used in the paper is to assign ALL 1000 deaths to category 1 even though in reality we'd expect to only see 51% of the deaths in that category. This should be fixed -- certainly for the the statistical method, but it should also be fixed if possible for the physician judgments by eliciting uncertainty estimates from them and using them appropriately too.

The comparison between the statistical model predictions and the physician classifications is done as if both are deterministic. To ascertain whether the statistical model is accurate, readers need more than point predictions. Uncertainty estimates must be presented so that policy makers will not be misled by the analysis, and in the paper these uncertainty estimates should also be evaluated. The model is Bayesian and so credible intervals ought to be easy to compute.

Level of interest: An article of limited interest

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'