Author's response to reviews

Title: Performance of InterVA for Assigning Causes of Death to Verbal Autopsies: Multi-Site Validation Study using Clinical Diagnostic Gold Standards

Authors:

Rafael Lozano (rlozano@uw.edu)
Michael Freeman (mikefree@uw.edu)
Spencer L James (spenci@uw.edu)
Benjamin Campbell (bcamp10@uw.edu)
Alan D Lopez (a.lopez@sph.uq.edu.au)
Abraham D Flaxman (abie@uw.edu)
Christopher JL Murray (cjlm@uw.edu)

Version: 2 Date: 15 June 2011

Author's response to reviews: see over
Reviewer's report

**Title:** Performance of InterVA for Assigning Causes of Death to Verbal Autopsies: Multi-Site Validation Study using Clinical Diagnostic Gold Standards

**Version:** 1  **Date:** 30 May 2011

**Reviewer:** Daniel Chandramohan

I think this is the first time the accuracy of interVA has been tested against a gold standard. The previous “validation” of interVA was done by assessing the correlation between PCVA and interVA. Given that the accuracy of PCVA itself is debatable assessing the concordance between PCVA and interVA is not the appropriate way to determine the accuracy of interVA.

It is not surprising that the chance corrected concordance rate of interVA was lower compared to PCVA and SPmethod. The expert opinion based algorithms developed previously performed as good as or lower than PCVA. However data driven algorithms particularly artificial neural network performed better than expert opinion based algorithms. Loronzo et al demonstrate well the key reason for the lower performance of interVA by comparing the probabilities of interVA versus SPmethod for selected causes given the symptom acute cough. It is not uncommon that expert opinion based probabilities are often wrong as demonstrated in Figure 6 that the probability of dying from chronic respiratory disease, other acute infection, maternity related death, suicide and drowning is the same given that some had acute cough according to interVA.

The comparison of chance corrected concordance rates of interVA, PCVA and SPmethod is useful to determine which of these methods perform well. However, this is not comparable to the previously reported “validation studies” of interVA. An analysis of the concordance between PCVA and interVA (individual causes of death and CSMF) consistent with the previous studies would be informative to understand whether the poor performance of interVA is due to the fact that the comparator in this study is clinically confirmed goldstandard.

To address this point from the reviewer, we performed an analysis of InterVA in which we used the Physicians Certification of Verbal Autopsy (PCVA) results as the comparison, instead of the Gold-Standard (GS). The table below compares the GS analysis to PCVA results for adults and children. We have excluded Neonates from this analysis because physician coding of causes of death made a direct comparison of the models impossible. For both adults and children, the chance corrected concordance and accuracy were quite similar. It is likely not a change in reference category, but a change in evaluation method/metric which has caused our numbers to be different from prior studies. For chance corrected concordance, both varying the cause composition and correcting for chance have provided a more rigorous evaluation of InterVA. In terms of population level
metrics, we’ve found that varying the cause composition is a stricter test of accuracy because it simulates applications of the program in various settings.

<table>
<thead>
<tr>
<th></th>
<th>Chance Corrected Concordance</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>InterVA to Gold Standard</td>
<td>InterVA to Gold Standard</td>
</tr>
<tr>
<td>Adults</td>
<td>25.2</td>
<td>26.3</td>
</tr>
<tr>
<td></td>
<td>InterVA to PCVA</td>
<td>InterVA to PCVA</td>
</tr>
<tr>
<td></td>
<td>0.549</td>
<td>0.56</td>
</tr>
<tr>
<td>Children</td>
<td>25.0</td>
<td>25.6</td>
</tr>
<tr>
<td></td>
<td>0.546</td>
<td>0.507</td>
</tr>
</tbody>
</table>

Previous studies do not use the same standardized metrics for their analysis, and are, therefore, difficult to compare. These studies generally report that the cause fractions are roughly the same between InterVA and PCVA as their primary metric. We believe our paper offers the following improvements 1) it corrects for chance, 2) it scrambles the cause fractions 3) it utilizes standard metrics of performance to draw more accurate comparisons.

Proponents of interVA would argue that there was only one African site in the PHMRC validation study and the expert opinion based probabilities of interVA are primarily developed and tested against PCVA in datasets from Africa. However, the accuracy of any algorithm based VA should not be depended on the cultural and epidemiological context. An analysis of the concordance between PCVA and interVA in the PHMRC dataset from Tanzania could be carried out to examine whether interVA performs as good as PCVA as claimed by previous studies.

We agree with the reviewer that the accuracy of any algorithm based VA should not be depended on the cultural and epidemiological context. Because gold standard deaths were collected in six locations, this study provided an evaluation of how InterVA performs overall. We completely agree that it would be important to test the concordance of the Tanzania datasets alone (data was collected at 2 sites within Tanzania - Pemba and Dar es Salaam), however, the site-specific analyses will be done by the lead investigators at the sites. The Tanzania-specific analysis is outside of the scope of this publication, though the results will be published by these lead investigators. We have added to the discussion that this is an important analysis that will be published separately.

The limitations of using clinically confirmed causes of death occurring at hospital setting as a gold standard to assess the validity of VA have been cited by some investigators to justify the concordance between PCVA and a second method of VA interpretation such as interVA. Lozano et al have shown the limitation of validation studies not having a goldstandard. A brief discussion of the limitations of methodology of validation of VA without a goldstandard should be included in the discussion section.
Following this suggestion of the reviewer, we have included the following text in the paper:

"...The contribution of this study is the use of gold-standard cases for the validation of InterVA. The aforementioned studies only provide information on the relation between InterVA and hospital assigned or physician reviewed cause of death. This study provides a direct comparison of InterVA to gold-standard verified causes of death..."

Reviewer's report

Title: Performance of InterVA for Assigning Causes of Death to Verbal Autopsies: Multi-Site Validation Study using Clinical Diagnostic Gold Standards

Version: 1 Date: 10 June 2011

Reviewer: Ayaga Agula AA Bawah

Reviewer's report:

Performance of InterVA for Assigning Causes of Death to Verbal Autopsies: Multi-Site Validation Study using Clinical Diagnostic Gold Standards

Authors: Rafael Lozano, Michael Freeman, Spencer L James, Benjamin Campbell, Alan D Lopez, Abraham D Flaxman, Christopher JL Murray

As verbal autopsies (VA) have become widely recognized as an important source of data for obtaining causes of death information in low- and middle-income countries, there have been intense methodological developments to find ways to efficiently analyze such VA data. Traditionally, VAs data have been analyzed using physicians to assign causes of death based on accounts of relatives. Physicians would normally plough through stacks of completed VA forms reading accounts of relatives who served as caregivers of the deceased when he/she was sick until death inferring from such accounts what in their judgment would have been the probable cause of death. Apart from being laborious and often takes lots of time to code the VA forms, physician accounts have often are often deemed inconsistent and not replicable. In other words one physician could offer two different judgments to the same account at different points in time.

The InterVA model developed by Peter Byass and colleagues aims to address the inadequacies of physician coding. InterVA is a Bayesian probabilistic technique for interpreting causes of death data obtained from verbal autopsies. Designed to handle up to 106 items yielding predictions for 35 causes of death across all ages, the InterVA is able at any point in time able to predict up to three causes of death. Given the importance of generating in a timely fashion information on causes of death both for public health and policy intervention, the development of the InterVA represents an important development in the efforts to mechanize the analysis of verbal autopsy data. It averts the dual problems of delays in physician coding and the inconsistencies in interpretation associated with physician
interpretation. Since the development of InterVA, a number of validation studies have been conducted by both the developers themselves and other researchers. However, many of those validation studies have been based on comparison with hospital data which may not necessarily be validated clinically.

The paper by Lozano and colleagues evaluates the performance of the InterVA tool. This is an important analytical effort since almost all the validation studies that have been conducted to evaluate the InterVA are either by the developers of the tool themselves or their colleagues or collaborators. Secondly, many of the validation studies are based on hospital records, which Lozano et al argue are themselves often suspects. Using data based on strict clinical diagnostic criteria from six sites in four countries in Mexico, Tanzania, India and the Philippines, Lozano et al. applied a complex Bayesian method known as the simplified symptom pattern (SP) method to determine causes of death.

**Level of interest:** An article of importance in its field  
**Quality of written English:** Acceptable  
**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

We agree with the reviewer that hospital information and access could potentially introduce bias in a verbal autopsy study. One way in which we have addressed this was that we ran our analyses with and without medical record review that we call in the paper Health Care Experience (HCE). This deliberately excluded information which could be obtained through an interaction with the health system (i.e. “Was the deceased diagnosed with HIV/AIDS?”).