Author's response to reviews

**Title:** Simplified Symptom Pattern Method for Verbal Autopsy Analysis: Multi-Site Validation Study Using Clinical Diagnostic Gold Standards

**Authors:**

Chistopher JL Murray (cjlm@uw.edu)
Spencer L James (spencj@uw.edu)
Jeanette K Birnbaum (kurian@uw.edu)
Michael Freeman (mikefree@uw.edu)
Rafael Lozano (rlozano@uw.edu)
Alan D Lopez (a.lopez@sph.uq.edu.au)

**Version:** 2  **Date:** 9 June 2011

**Author's response to reviews:** see over
Reviewer’s report
Title: Simplified Symptom Pattern Method for Verbal Autopsy Analysis: Multi-Site Validation Study Using Clinical Diagnostic Gold Standards
Version: 1 Date: 10 May 2011
Reviewer: D. R. Mani

Reviewer's report:
The manuscript "Simplified Symptom Pattern Method for Verbal Autopsy Analysis: Multi-Site Validation Study Using Clinical Diagnostic Gold Standards" by Christopher JL Murray, Spencer L James, Jeanette K Birnbaum, Michael Freeman, Rafael Lozano and Alan D Lopez proposes a simplified and more effective implementation of the (previously published) Symptom Pattern method for computer coded verbal autopsy. The new Simplified Symptom Pattern (SSP) approach is rigorously validated using accepted train/test methodology based on gold standard VA data.

Major Compulsory Revisions

1. In its current incarnation, the manuscript is excessively dependent on companion papers that are all referenced as being "In submission". Unless many of these are co-published as part of a special issue, it would greatly benefit the reader to replicate relevant aspects from these papers. Some key aspects that fall under this category include the design of the datasets, the tariff calculation, and chance-corrected concordance. Having these definitions and details available in this manuscript will make this document more self-contained, and enable interested readers to replicate the methodology laid out here.

   We understand the reviewers concern. However, we have submitted this paper along with the papers referred to as part of a set of papers that are being considered for publication in a special initiative of Population Health Metrics. It is our understanding that the editors would prefer not to repeat so much of the text and documentation as proposed to make the paper more self-contained. We would, however, be happy to follow the suggestions of the editors in this regard.

2. There are several questions and comments related to the Section titled "Validation Using the PHMRC Gold Standard Train-Test Datasets".
   i. Paragraph 2 in this section states that 100 data subsets were analyzed. The next paragraph claims to assess performance on 500 test datasets. This is confusing to the reader and should be explained in more detail

   We fixed this mistake and added clarification.

   ii. Paragraph 2, sentence 2. Shouldn't the last two words be "test dataset"?
This section has been re-written to provide clarification for the previous comment.

**iii. The authors also state that a set of 500 train/test splits was developed. Are these a FIXED set of splits? Or, are the splits created randomly as and when needed? If a fixed set of splits was created, an explanation of the rationale would be very helpful.**

These splits are fixed in order to allow for direct performance comparison between methods. We added this explanation to the first paragraph.

**iv. 12 SP variants are evaluated using the train/test datasets, and the best combination is chosen. It is possible that the performance of the best variant is biased by the set of train/test splits used, especially if a fixed set of splits was used for all the variants. In such a case, the choice of the variant can be based on the train/test splits, but the final performance should be evaluated using an unseen validation dataset(s) that was set aside exclusively for this purpose. At the very least, random subsets of the 500 train/test splits should be used for each variant.**

We thank the reviewer for this important point. To explore this concern, we have conducted two types of tests. First, we repeated the analysis using splits 101-200 as the original assessment in the submitted paper used only the first 100 train-test splits. Further, following the reviewer’s recommendations we have also created random samples of size 50 and tested which methods perform better. Following the recommendations of the second reviewer, we have also repeated this assessment for children and neonates. In this assessment, we found the results from the first 100 splits, the second 100 splits and the random samples were highly consistent. We have found that for adults the ordering of methods was invariant across the 12 methods for adults. However, the ordering of the 12 methods for children and neonates was somewhat different. For adults, the best performing method was the cluster 10, single cause, top 40 symptoms as assessed by tariff, uniform prior model. By summing the rank order of methods for chance-corrected concordance, and CSMF accuracy for adults, children and neonates, we identified that the best overall method is actually cluster 10, single cause, all symptom, uniform prior model. The variant used in the submitted paper ranked second overall. The difference between the two variants lies only in which set of symptoms are included in the model. Given a substantial improvement for children and neonates and small reduction in performance for adults, we have chosen to use the variant that performs best across the three age-groups.

3. **What are the implications for deployment of the SSP model? Other than computational cost, would the model be applicable to all regions that the PHMRC data originated from? Would tuning/retraining be necessary? Will the code and/or model be publicly available?**
The SSP model performs well and currently could conceivably be implemented to analyze other VA datasets. Training of the model on the full PHMRC dataset would be required. The code and model will be publically available on the internet. We have added more to the discussion section to outline our plans in this regard. If a country or location has gold standard data, training on that dataset or the combination of the PHMRC data set and local validation dataset would probably be the most useful. Given the cost and challenge of obtaining a rigorously defined gold standard dataset, many users will likely need to use the PHMRC trained SSP.

Minor Compulsory Revisions

1. Byass et al (2003) should be included in the list of references.

We added this reference.

Reviewer's report
Title: Simplified Symptom Pattern Method for Verbal Autopsy Analysis: Multi-Site Validation Study Using Clinical Diagnostic Gold Standards
Version: 1 Date: 16 May 2011
Reviewer: Soeharsono Soemantri
Reviewer's report:
I. Major Compulsary Revisions
1. SSP vs PPVA
a. Individual cause assignment. The paper described that selection of a simplified symptom pattern (SSP) from 12 possible modifications of SP was basically based on adult causes of death and the selected variants have been implemented for children and neonates. Could the procedure may effect the performance of SSP for children and neonates in terms of chance corrected concordance?

We thank the reviewer for this important observation. We ran the 12 Simplified Symptom Pattern variants on the child and neonate datasets and observed in fact the top performing model for each age module is different. Moreover, the top performing model is different when considering CSMF accuracy versus chance-corrected concordance. To choose one variant that will have the best overall performance on CSMF accuracy and chance-corrected concordance, we ranked each variant within each age group, for chance-corrected concordance and CSMF accuracy. The method with the lowest sum of ranks was cluster 10, single cause, all symptoms, uniform prior model. This differs from the variant selected solely on the basis of adults by using all symptoms and not a subset of symptoms based on tariff. As a result of this finding, we have changed our main model for all modules to the cluster size 10, all symptoms, uniform prior, and single cause specifications. This change has slightly lowered adult performance but substantially improved performance for children and neonates. We have explained this in the text and added detail on these tests.

As shown in Table 3, SSP for adults outperform PVCA better than
SSP for children and neonates. Even Table 3 shows that PCVA outperforms SSP for children and neonates for group with HCE (Note the authors wrongly stated SSP does better than PCVA both with and without HCE).

We fixed these mistakes and added clarification.

Chance corrected concordance shown in Table 3 can be used to indicate performance difference between SSP and PCVA by contrasting 95% CI values, but it can also be used to evaluate how strong the relationship between the methods (both SSP and PCVA) to assign individual cause with the true one. All chance corrected concordance coefficients shown in Table 3 are below 50% indicating a weak relationship.

Assessment of what is or is not a strong relationship depends in this case on what are expectations for medical certification of causes of death, and VA as an alternative. For individual cause assignment, in medical certification of causes of death in Mexico in the best national hospitals with extensive resources, chance-corrected concordance is 66.5% for adults, 38.5% for children, and 54.3% for neonates. We expect that chance-corrected concordance to substantially lower for deaths in less sophisticated facilities and especially for deaths outside of hospital. Given these benchmarks, we believe that chance-corrected concordances above 35% are actually quite high.

b. CSMF Estimation. CSMF accuracy achieved by SSP in comparison to PCVA by age group is shown by Table 4. Not all cases SSP performs better than PCVA. For group of neonates PCVA produces better accuracy of CSMF (although SSP accuracy results are not comparable). Although median CSMF accuracy for adult and child based on SSP are statistically better than PCVA, the increment of CSMF is only within the range 4% to 7%, it can not be claimed as substantial.

The reviewer is correct PCVA outperforms SSP for neonates in specific cases. Although the comparison is not truly correct as PCVA is only assigning causes to 5 neonatal causes and stillbirths which is an easier task. With regards to the claim that differences in CSMF accuracy are not substantial, we disagree with the reviewer. First the range is from 0.624 to 0.751 across all the cases. A 10 percentage point difference in accuracy is actually quite dramatic. Two methods would differ in CSMF accuracy by 10 percentage points if on average over 500 tests, one cause was mis-estimated to be 10 CSMF percentage points higher on average. Such a difference would be rather large and dramatic. Even small differences in CSMF accuracy represent substantial gains in performance. We have added to the paper to clarify this point.

2. Estimated of CSMF as a function of true CSMF
Annex 2 shows relationship of estimated CSMF derived from SSP model and true CSMF from 500 different test splits in the form Estimated CSMF = True CSMF * slope + intercept and Figures 4A thru 4G highlight characteristics of SSP’s prediction. Slope and intercept indicate how well SSP’s method will predict true CSMF. Slope close to one and intercept close to zero indicate that estimated CSMF will be perfect to predict true CSMF. SSP only gives slope above 0.8 for four adult causes, three child causes and two neonates causes (see column HCE of Annex 2). Many slopes derived from SSP are below 0.5 (20 adult causes, 14 child causes and 7 neonates causes). For the 9 causes with slope above 0.8 their intercepts ranging from 0.010 to 0.051. Many low values of slope coupled with relatively high values of intercept and RMSE predicted by SSP indicate that the method is still unpredictable and imprecise for many causes.

The key issue is the comparison between SSP and PCVA. SSP does better than PCVA on average in terms of higher slopes and lower intercepts indicating a better estimation of the true CSMFs. We do not claim that VA methods are perfect, only that the SSP method performs equal to or better than the current standard in the field, namely PCVA.

3. In conclusion, the Simplified Symptom Pattern, in comparison with PCVA indeed improved the individual cause of death assignment as well as cause specific mortality fraction estimates. However SSP is still far from expectation to predict true CSMF for many causes of death.

We agree with the reviewer and the text of the paper makes this quite clear. It is worth noting that in the future, it might be possible to consider correcting VA based estimates of cause fractions based on the relationships between true and estimated CSMFs. Such corrected VA cause fractions could be more useful. We believe however that this will require the collection of more data and more experience using new methods before choosing to use this approach. At the very least, the relationships shown in the Annex can be used to assess the sensitivity of results for any given cause.

II. Minor Essential Revisions

Table and Figure heading should follow standard format. Inconsistency in numbering Figure within text body and in attachment.

We have checked and corrected inconsistencies.