Reviewer's report

Title: Usefulness of the Population Health Metrics Research Consortium Gold Standard Verbal Autopsy Data for General Verbal Autopsy Methods

Version: 1
Date: 16 December 2013

Reviewer: Alan Lopez

Reviewer's report:

1. My main concern with this paper is that I could not determine what exactly the author is trying to establish. Is it that InterVA performs as well in diagnosing causes of death as hospitals? Is it that the PHMRC dataset is or is not usable to evaluate diagnostic methods for VAs, including InterVA and physician coded VA (PCVA)? Is it to advertise the new WHO VA short form questionnaire which conveniently maps to InterVA (or the converse), but not to other possible analytical approaches? Or all of the above, or something else? The paper requires a clear and unambiguous statement about aims.

2. It is not clear why you would want to compare the PHMRC dataset with the WHO 2012 standard and InterVA. The author does not make clear that the PHMRC dataset consists of two sub data sets: 12500 cases where the true cause of death has been established ex-ante using the PHMRC clinical definitions: and full VA questionnaire responses for each of these 12500 cases which were then used as the basis for testing various diagnostic methods, including InterVA. I could not find this distinction anywhere in the paper nor what part of the PHMRC dataset was being used, or for what purpose.

3. The author claims that methods built on the PHMRC dataset (which one?) should be independently evaluated, but does not say why? Nor does he offer any cogent critique of the test-train procedure used on the PHMRC dataset by the PHMRC collaboration to establish model performance.

4. The author appears to think that the PHMRC dataset was used to estimate community cause of death patterns rather than using VA on all-community deaths (see penultimate sentence of Abstract). This was never claimed by the PHMRC collaboration.

5. Background, para 2: you cannot test models with "high-quality verbal autopsy data". There is no such thing!

6. I do not understand the interest in comparing the PHMRC dataset (presumably the VA questionnaire used in the PHMRC Study?) with the WHO Verbal Autopsy questionnaire. Why would you do this?

7. Results, last para: the author compares CSMF (cause-specific mortality fractions) from application of InterVA to those in the PHMRC hospital dataset. I have several problems with this. First, the concordance of CSMF as shown in Table 4 is terrible with very little agreement between the two, and with NO
confidence intervals to aid interpretation. On this basis, I would conclude that InterVA4 does very poorly in relocating the true cause of death distribution in the PHMRC dataset. Second, is a visual comparison of CSMFs appropriate? Why not use a summary metric such as CSMF Accuracy as defined in the PHMRC Study which captures agreement over all causes of interest? Third, this comparison has already been published (see Lozano et al, PHM, 2011). What more does this comparison add?

8. The author seems to place great store in the WHO 2012 VA questionnaire as the "standard". But why should we be concerned whether the PHMRC questionnaire (which is a long form, while the WHO 2012 effort is a short form) does or does not agree with what WHO proposes? What were the criteria for choosing the questions by WHO and have they been empirically evaluated? We do NOT know the drop in performance in going from the WHO long form (2007) to the WHO short form, so how can it be confidently used to monitor causes of death in populations?

9. Why is the high proportion of respiratory symptoms reported in the PHMRC VA dataset "surprising"?

10. If you cannot use tertiary facility deaths as the evidence base for VA methods, then what can you convincingly use?

11. I do not understand why differences between the WHO 2012 VA questionnaire and the PHMRC one are labelled as "omissions". Surely they simply reflect different value judgements between the two processes?

12. I see no convincing evidence to invalidate the use of the PHMRC dataset to build and investigate the comparative performance of different VA diagnostic methods. What are the authors reasons for claiming that this is the case?

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

I should reveal what I suspect are conflicts of interest. I have recently submitted a paper where I was a co-author with Prof Byass. It has not been accepted but is in review.

On the other hand, I have been involved in verbal autopsy research (as reflected in a paper on which I am a co-author currently in review at BMC Medicine) which suggests that a method developed by Prof Byass (InterVA4) performs substantially worse than other methods in diagnosing causes of death.