Author's response to reviews

Title: How much do the physician review and InterVA model agree in assigning probable causes of death? A Comparative analysis from rural Ethiopia

Authors:

Berhe Weldearegawi (berheph@gmail.com)
Yohannes Adama Melaku (yelidya@yahoo.com)
Geert Jan Dinant (beletejon@gmail.com)
Mark Spigt (berhe.sahle@monash.com)

Version: 2 Date: 11 June 2015

Author's response to reviews: see over
Title: How much do the physician review and InterVA model agree in assigning probable causes of death? A Comparative analysis from rural Ethiopia

Authors: Weldearegawi et al.

Dear Dr. Saeid Shahraz

Editor-in-Chief: *BMC Public Health*

**Re: Submission of revised manuscript**

Thank you for your letter dated 14th of May 2015, in which we have received feedback on the above-mentioned manuscript. We appreciate the thoughtful comments which were very helpful for improving the manuscript. We agree with almost all of the comments and we have revised our manuscript accordingly. Some of the paragraphs are rephrased to remove ambiguities, and additional data are provided where required.

Appended to this letter is our point-by-point response to the comments raised by the reviewers. We would like to take this opportunity to express our sincere thanks to the reviewers who identified areas of our manuscript that needed revision.

Kind regards

Berhe Weldearegawi (on behalf-of the authors)
Comments from the editor(s):

1-the comment from the editor(s) requests including the “Authors’ Contributions” section.

Response: The authors are grateful for the suggestion, and this now been included after the “Competing Interests section”.

Authors’ contributions

BW and YAM involved in design of the surveillance, data collection and supervision. BW was involved in study conception, data processing and analysis. BW wrote the manuscript and interpreted the results. BW, YAM, GJD, and MS made substantial revision to the manuscript. MS and GJD mentored the process of paper writing from preparation to manuscript write up. All authors read and approved the final manuscript.

Reviewer 1

Comment-1:

The authors appropriately use kappa statistics as a means of making their comparisons, but they seem to be a bit confused in the interpretation of the results. The implied null hypothesis is that the two approaches to VA interpretation are equivalent, in the absence of any source of "truth". In other words, if there are differences, nothing can be inferred about which approach is "better" or "more accurate". In some instances this clarity is not entirely reflected in the text. For example, in line 81 "the reliability of the diagnosis that can be reached by using these methods" is not something that can be addressed in this type of comparison. The final conclusion "further refinement to the InterVA model" (line 266) may be true, but is not supported by the results presented here, in the same way that a conclusion "retrain Tigrean physicians" would not be supported by these findings.

Response: The authors agree that neither physician review (PR) nor the InterVA are gold standard in defining probable COD; and that is why the study was comparative rather than a validation. The purpose of the study was to answer the question: “how similar diagnosis can be reached if InterVA were used in place of the PR”. In terms of this, the findings tried to answer this specific question rather than validating either of the methods. Yet, as the reviewer rightly mentioned, we have modified the terms in a way that explain “comparison”, throughout the text. Particularly, the conclusion and recommendations are also modified accordingly.
The conclusion in the abstract is now modified as:

“If the InterVA were used in place of the existing PR process, the overall diagnosis would be fairly similar. The methods had better agreement in important public health diseases like; TB, perinatal causes, and pneumonia/sepsis; and lower in cardiovascular diseases and neoplasms. Therefore, both methods need to be validated against a gold-standard diagnosis of death.”

And the conclusion of the study as:

“In summary this study reported an overall low chance corrected agreement in probable cause of death between PR and InterVA. The level of agreement varies across different categories of causes of death, and age of the deceased. It ranged from moderate to substantial for important public health diseases like; TB, perinatal causes, pneumonia/sepsis, and accidents and injuries; while, the agreement for NCDs, especially for cardiovascular causes and neoplasms was low. Both methods had relatively better agreement in under-five children and adults aged 15-45, while they least agreed for cases aged 45 and above years. Therefore, if the InterVA were used in place of the PR process, the overall diagnosis would be fairly similar.”

Comment-2:

The Kilite Awlaelo site has actually been using both physician and InterVA-4 methods since 2009, as can be seen from several INDEPTH publications in which Kilite participated using InterVA-4 that are not referenced here (Global Health Action 2014; 7:25362 and following papers). Therefore the statement in lines 87-89 is not entirely correct.

Response:

The Authors would like to provide the following clarifications on this. The KA-HDSS is using PR method since its establishment in Sept 2009. As described in the methods section, the PR approach was applied longitudinally; verbal autopsy questionnaires were filled at the end of mourning period and reviewed by physicians (in less than a month time). Still the KA-HDSS (and other centers in Ethiopia) are using this method- as a “standard method of determining possible COD”.

In 2013, there was a multi-centre cause of death analysis organized by the INDEPTH Network, where the KA-HDSS was a participant. In this multi-centre analysis there was a need to aggregate data from member centers, and the InterVA was a preferred method to use, as the outputs must be “comparable”. To this effect, the already existing VA dataset-established based on a PR- were migrated into the InterVA-4 format. So, the first time the KA-HDSS piloted the InterVA was in 2013, and the paper mentioned by the reviewer (Global
Health Action 2014; 7:25362) was produced during this trial. To sum up, the KA-HDSS is still using a physician review approach.

**Comment 3:**

Line 116 onwards describes the process by which two independent physicians reviewed each case. It is important to also present kappa statistics for the agreement between the first and second physicians to understand the effect of inter-physician variation.

**Response:** Of the total VA cases reviewed to date, only less than 10% of the VA cases required a tie-breaker physician. However, this data were not part of the existing VA database. As result, it was not possible to compute the kappa statistics for the agreement between the first and second physicians.

**Comment 4:**

Line 125 onwards - it is not clear how the fractional likelihoods of different causes produced by the InterVA-4 model were handled, as well as the residual uncertainty associated with cases have a total likelihood of less than 100%. These methods have been clearly described in previous papers using InterVA and need to be stated and followed more carefully here. The InterVA-4 findings should align with the Kilite Awlaelo data presented in the INDEPTH series of papers and this should be checked.

**Response:** The authors agree that the InterVA produce up to three likely COD per case, and corresponding likelihood ratios. The most probable COD have the highest likelihood; while, the subsequently ranked probable causes (secondary, tertiary), if any, will have a lower likelihood. Physicians also coded up to three underlying COD, ranked according to the chance to be a cause in that specific subject. But, in most cases, as is also true in the InterVA, physicians list one probable COD if the evidence presented in the VA questionnaire is sufficient to pick up single cause (such as in accidents/injuries). So, the comparison was made between the most probable COD listed by both methods, as is used in many other studies (*Population Health Metrics* 2010, 8:21, and, *Population Health Metrics* 2011, 9:49).

Still we agree that this should be clearly explained. Accordingly, we have added the following sentence in line-152.

“In the present study, the analysis was made by taking the most probable COD assigned by both methods rather than all three possible causes.”
Comment-5:
Line 156 - a complex process has been used to reconcile the different sets of causes of death assigned by InterVA-4 and the physicians. What is not made clear is that InterVA-4 classifies causes according to the WHO VA 2012 standard (described in detail in Global Health Action 2013; 6:21518), and so it would be more informative for an international audience to conform to this standard categorisation for both the physician and InterVA-4 results.

Response: The two methods use different coding to represent similar probable causes. The PR process use ICD-10 coding, while the “WHO VA cause of death code” is used in the InterVA. Thus, to facilitate the comparison, outputs of the InterVA model were re-coded in to equivalent ICD-10 code, which is also given in the InterVA manual (http://www.interva-4.net/).

As described in the methods section (line 152-153), there was some difference in the categories of COD between the two methods. For example, the InterVA model has only one category of “maternity-related deaths”; while the physicians coded a detailed classifications such as “eclampsia” and “antepartum and post-partum haemorrhage”. Therefore, in an effort to have comparable categories of COD, such causes were grouped into one broad category such as: “maternal deaths”. The causes of death listed from each method and the corresponding re-categorisations are given in the attachment file.

To improve the clarity, we have modified the paragraphs starting from line 172 onwards.

Comment-6:
Line 172 - the study population here is very small, at least when it comes to comparing subgroups or rare causes, and this is not well discussed as a limitation of the study. No assessment of the statistical power of the study is presented. The authors also do not refer a recent large international study that made similar comparisons on a much larger dataset (Journal of Global Health 2015; 5:010402).

Response: The authors have addressed this comment, and the following paragraph was added after line-146.

“Based on previous studies, we assumed that physicians would identify infectious diseases (most common outcome) in 58% of the cases and the InterVA model would do it for 61% of the cases. Besides, existing κ estimates between the two methods range from 0.27 to 0.8. Thus, in the present study we would expect a kappa of 0.5 on average. With an absolute precision of 0.1 and α of 0.05, the minimum sample size required for this study is 332 participants. However, there were 434 cases that have both PR and InterVA based COD, and all of them were included.”

We also thank for indicating us the recent large international study, published after this manuscript was prepared, which is now referred in this manuscript.
Comment-7: Minor essential revisions

There is huge inconsistency in the way that kappa is referred to in the paper - Kappa, kappa, K, k, etc. It would be more correct to use the Greek character kappa throughout.

Response: we agree, and the Greek character has been used throughout the text.

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

Response: We have made a rigorous proofread to improve the clarity and level of English command. Besides, it has been proof read by an expert whose native language is English. The revisions made are highlighted in track change.

Reviewer 2
Comment-1:

Although there is reasonable agreement at the broad causes of death categories the agreement between the two methods for several specific causes of death is poor. The authors suggest that one of the reasons for this discrepancy between the two methods is the difference in the number of causes of death used by the two methods. The effect of the difference in the number of cause of death used is unlikely to explain the large disagreement between the two methods for assigning causes of death such as neoplasm and respiratory infections. The key factor that explains this inconsistency between the two methods is the main source of information – InterVA uses the data from the closed questions only while the PR uses mostly the open ended narrative part of the VA questionnaire. Furthermore the probability matrix of InterVA is not consistent with the probability applied by the local physicians. The authors have mentioned these issues but they need to be well discussed.

Response: this is very valuable suggestion. We agree on modifying the possible explanation, and the following paragraph has been added in line 291 onwards, in effect.

“Variation in the diagnosis between these two methods may not be unexpected, though further investigation is needed to explain the variation. Nevertheless, findings from previous studies have attributed the variation to how the two methods process and use the verbal autopsy data. The InterVA uses the data from the closed ended questions only, while, the PR involve extensive use of the open ended narrative part of the VA data. In addition, the InterVA process uses a probability matrix to process the indicators in the verbal autopsy data, while, in PR it is entirely based on expert judgment.

Comment-2:
The authors should discuss the limitations of comparing the performance of two methods with unknown validity without a gold standard to compare. Although there is plenty of data on the validity PR compared to hospital diagnosis as gold standard there is very little data on
the validity of InterVA. The uncertain in the validity of InterVA needs to highlighted in the discussion.

**Response:** We have addressed this suggestion in the discussion and limitations sections. The following paragraph has been added to the discussion in line 274 onwards.

“In the present study, inference about validity of either of the methods cannot be made in the absence of a gold-standard diagnosis. However, validation studies which simultaneously evaluated both methods against hospital certified deaths, showed that the PR performs better than the InterVA model.\textsuperscript{13,14} A validation study which compared both the InterVA and PR methods against hospital causes of death revealed that the level of agreement between InterVA and hospital CoD (k =0.32) was lower than the agreement between physician review and hospital CoD (k =0.52). In addition, in another study which evaluated the PR and InterVA using clinical diagnostic gold standards in a sample of 12,542 verbal autopsy cases, the PR has shown a better performance than the InterVA, across all age-groups.\textsuperscript{13}”

And the sentence below in the limitations section:

“However, our study has the following limitations. The two methods were compared in the absence of a gold-standard diagnosis. As a result, it was possible to compare the diagnosis between the widely used PR and the InterVA model, but not to investigate the validity of the methods. Although the study included more cases than the minimum sample size required, it was not sufficient when it comes to comparing sub-groups or rare causes of death such as maternal deaths.”