Author's response to reviews

Title: Classifying perinatal mortality using verbal autopsy: Is there a role for non-physicians?

Authors:

Cyril Engmann (cengmann@med.unc.edu)
John Ditekemena (josol70@yahoo.fr)
Imtiaz Jehan (imtiaz.jehan@aku.edu)
Ana Garces (anagarces@imsalud.org)
Mutinta Phiri (drtinta@yahoo.com)
Vanessa Thorsten (vthorsten@rti.org)
Manolo Mazariegos (mmazariegos@incap.int)
Elwyn Chomba (echomba@zamnet.zm)
Omran Pasha (omran.pasha@aku.edu)
Antoinette Teshefu (antotshe@yahoo.com)
Elizabeth McClure (mcclure@rti.org)
Dennis Wallace (dwallace@rti.org)
Robert Goldenberg (rgoldenb@drexelmed.edu)
Carl Bose (cbose@med.unc.edu)
Waldemar Carlo (Wcarlo@PEDS.UAB.EDU)
Linda Wright (wrightl@exchange.nih.gov)

Version: 3 Date: 20 May 2011

Author's response to reviews:

Dear Editors,

On behalf of my colleagues and myself, I would like to thank the reviewers for their insightful comments on this manuscript. We believe the manuscript will be strengthened by incorporating many of the reviewer's comments.

Please find a point-by-point response in italics.

Reviewer 1: Edward Fottrell

Minor Essential Revisions

1. The authors define physician consensus diagnoses as the reference standard and acknowledge that this is a limitation of the current study since the “true” cause of death is in fact unknown. It would be interesting to know how often physician consensus was reached immediately and how often there were discrepancies between physicians that required discussion in order to reach a consensus. If the authors were to compare the community coordinator diagnoses separately against each of the physicians, and indeed compare the physicians
against each other, one would better be able to assess whether agreement between physicians before discussion was substantially better than agreement between community coordinators, agreement between community coordinators and single physicians, and agreement between community coordinators and physician consensus.

In a previous publication (Engmann et al, An alternative strategy for perinatal verbal autopsy coding: single versus multiple coders Trop Med Int Health 2011 Jan; 16 (1): 18-29), we compared responses between individual physician coders for specific causes of stillbirth and early neonatal death (perinatal deaths), and responses of individual physician coders to the physician consensus for specific causes of perinatal deaths. We reported that physicians assigned the same COD for 75% of stillbirths (Kappa=0.69; 95% C.I. 0.61-0.78) and 82% early neonatal deaths (Kappa=0.75; 95% C.I. 0.65-0.84). In that manuscript, we also noted that the patterns and proportions of stillbirths and early neonatal deaths determined by the physician coders were very similar compared to causes individually assigned by each physician. Also, the rank order of the top 5 causes of stillbirth and early neonatal death were identical for each physician.

Although not shown in this manuscript, we completed analyses comparing the coordinators responses to physician consensus among those deaths which required further adjudication (i.e. more difficult cases where the physician consensus differed from one or both of the individual physician assessments of the COD) and among those deaths which did not require further discussion (i.e. easier cases where the physicians’ independent assessments of the COD agreed). These analyses did not change the results. Generally, there was high agreement among the physicians, and the community coordinators did not perform better for the easier cases than for the more difficult ones.

Given these data and results, comparisons of COD assigned by community coordinators to individual physicians would have been unlikely to change the overall findings of our study, and including those results in tables would have increased the density and complexity of those tables substantially for little benefit. Further, reporting these findings would not address the question of whether using community coordinators to assign COD is a reasonable alternative to the most common method of performing VA, that is, to use physician panels.

2. In relation to point 1, the methods on Page 6 should clarify that physician CONSENSUS causes COD responses were viewed as the reference standard.

Thank you. We have clarified this point on page 6. The sentence now reads" The underlying COD assigned by the community coordinator was compared to the consensus underlying COD assigned by the physician panel".

3. The Methods explain that the community coordinators administered the VA questionnaire in the community (page 5). The act of conducting the interview
may have provided the community coordinators with more explicit and implicit information from the respondent, environment etc. about the circumstances, signs and symptoms of the deceased before death than was recorded on the standardised VA questionnaire. A wide range of factors that were probably unknown to physician reviewers may have influenced the community coordinators’ ability to identify a cause. Some discussion of this would be appropriate in the Discussion section, perhaps in relation to the point about unknown validity.

This is an important point which we have included in our discussion. We have stated “It is conceivable that direct interactions between the community coordinator and the respondents (mothers and birth attendants) may have provided the community coordinators with more information from the respondents and the environment about the circumstances, signs and symptoms of the deceased before death than was recorded on the standardized VA questionnaire”.

4. All data are pooled, analysed and presented together in the current study. It would be interesting to know whether there was any difference in community coordinators’ performance between the different study settings. Perhaps the role of non-physicians in classifying perinatal causes of death is context specific, influenced by regional differences in literacy, recognition of symptoms etc. and thus differs between settings. In addition to the pooled results already presented, results presented by country might be interesting and some comment on this would be appropriate.

While we agree that issue may be an important one to evaluate, this study was neither designed, nor powered to do this type of analysis. Without sufficient power to complete this analysis and sample sizes that are relatively small by country, we are concerned that presenting these data (even in a strictly descriptive manner) may be misleading and result in erroneous conclusions.

5. Robustness criteria used in the study are based on previous work by Setel et al and provide a useful measure of agreement derived from previous work on VA. However, alternative measures of agreement without the implicit designation of any one method as being “true” do exist e.g. Kappa. Some explanation as to why the authors chose the robustness measures described in the manuscript would be helpful.

Since the most commonly used method for coding COD uses physician panels (Garenne & Fauveau. Bull WHO 2006; 84: 164, Fotrell & Byass. Epid Rev 2010; 32: 38-55) and these panels represent the reference standard, other measures
such as kappa, which are used to compare two alternatives considered of equal weight and not to compare an alternative approach against a reference/gold standard, would be less appropriate in this case. We have used kappa’s in previous publications where we did not have a reference standard per se and where each response had equivalent value (Engmann et al, Trop Med Int Health 2009:14:1496-504, and Engmann et al Trop Med Int Health 2011: 16:18-29).

6. In relation to point 5, the authors describe VA as an epidemiological tool (page 3), implying a public health utility of the results. However, assessment and discussion of relative differences in CSMFs is perhaps not the most useful measure of the epidemiological/public health utility of VA. The extent to which large relative differences matter in public health terms depends on the magnitude of the true value. For causes that account for a fairly small proportion of all causes (e.g. tetanus in Table 2, APH, maternal accident, malpresentation in Table 4), even a large relative difference can be somewhat meaningless in absolute terms. Conversely, absolute differences can be large and of public health importance if relative differences are small and the true CSMF is large. Therefore, presentation and discussion of absolute differences would be appropriate and may enable comparisons of community coordinator and PR results in terms of public health utility of the data.

The reviewer identifies an import consideration in the interpretation of relative differences. We did not include absolute differences in the tables because we felt that the volume of data would become somewhat overwhelming. However, the data to calculate absolute differences is in the tables and therefore available to the reader.

7. Page 11 – the comment about non-physician coders perhaps not being a “reliable” alternative to physician coders seems inappropriate since reliability is not actually assessed in the current study.

Thank you. We have changed the sentence to read “Although the use of non-physician coders to assign COD may not be a suitable alternative to the use of physician coders, other alternatives may have a role.

8. Recommendations on the use or not of non-physicians to determine perinatal causes of death is difficult based on the current study since we do not know the true cause of death – discrepancies between community coordinators and physician consensus might be because the community coordinators identified the correct cause. Therefore conclusions and recommendations need to be carefully considered and limited to the evidence of the current and previous published studies, focussing on agreement rather than reliability or validity. The authors
have generally done this but I am not convinced that, in the absence of “true”
cause-of-death gold-standards, the call for further research is justified. Can the
authors explain the nature of this further research and how it will bring us further
than where their work has taken us?

We agree that the use of physician panels as a reference standard to determine
cause of death has its limitations, some of which we have outlined on pages 11
and 12. This study focused on concordance between community coordinators
and physician panels, rather than on reliability or validity, after community
coordinators and physician panels had both been taught a standardized
package. Since the determination of underlying COD requires a deep
understanding of pathophysiology, it is possible that more intense, prolonged or
varied training of community coordinators may improve their ability to determine
the single underlying cause of death comparable to physicians. In addition,
should multiple causes of death (underlying, final and contributing) be
considered, it is possible the concordance of community coordinators responses
to physicians might improve.

Level of interest: An article whose findings are important to those with closely
related research interests
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.

Reviewer 2: Timothy Colbourn
Major Compulsory Revisions

1. The calculation of Relative Difference (RD) is based on the rounded figures of
CSMF for PC and CC. Using unrounded figures alters the reported RD
sometimes above the 20% threshold (maternal accidents goes to 29%) or
below/level with it (Tetanus goes to 20% and ‘other (SB) goes to 17%). Should
redo with unrounded figures (see attached Excel spreadsheet ‘Tables 2 and 4
calcs.xlsx’).

We thank the reviewer for this suggestion. In an earlier version of the manuscript
we reported the unrounded absolute relative differences and a coauthor wanted
the rounded numbers since these could be calculated from the values presented
in the table. However, we agree that the unrounded numbers are more
appropriate and have updated the tables accordingly. Additionally, we have
clarified the calculation in the table footnotes. For early neonatal deaths, tetanus
now fulfils criteria for robustness; for stillbirth, cord prolapse fulfill criteria for
robustness.

2. Also, regarding the causes with small RD: The reasons these causes are in
fairly good agreement (small RD) may be because they are very small CSMFs - if
you look at the sensitivity of CC diagnosis for these causes it is actually pretty low (53% for maternal accidents and 50% for cord prolapse). Worth mentioning given that these small causes are less relevant overall.

We agree and have included the statement “It is worth pointing out that where relative difference values are small, it may be because there are very small CSMFs”, on page 8.

3. Methods: The exclusion criteria seem overly strict and need to be explained to the reader. Why exclude hospital deaths (are they already covered by VA in all cases?) Why exclude those where a birth attendant was not present - especially if you are not interviewing the birth attendant? Also, 7 days seems too short a period to be worried about recall bias, especially given that the death of a baby is such a major event that it is bound to stick in the mothers mind. Given your sample is relatively small it could have served you better to have a more inclusive inclusion criteria. Not including 145 cases because the mother was not available in the first 7 days lost you 33% (145/437) of your sample.

This study was a community-based, non-facility study because it is among these births that identification of COD remains the greatest challenge. Thus by definition hospital-based deaths were not included. We chose seven days to minimize the risk of recall bias during the conduct of the study, since other authors have commented that recall bias may influence results obtained from VA respondents, (Soleman et al Bull WHO 2006;84:239-45, Lee AC et al Pediatrics 2008: 121:e1372-80, Fottrell and Byass, Epid Rev 2010; 32:38-55). As part of the study design, we compared responses of birth attendants with those of conventional VA respondents, mothers, thus birth attendants needed to be at the delivery.

4. Methods: Training and VA methodology: ‘Train the trainer’: Please expand. It is not clear to me why a ‘train the trainer’ method was required when only 13 physicians and 40 non-physicians were trained in total. Also what is the breakdown of the 53 trained people (and subsequent VA done) by country given that this study was done in 4 countries?

We have outlined why and how we used the train-the-trainer method in a previous publication, to which we draw the reader’s attention (reference 19, Engmann et al Trop Med Int Health 2009; 14: 1496-504). Briefly, an expert in ICD-10 rules, regulations and classification trained 9 individuals in the USA. These individuals then returned to their countries and trained 53 additional physicians and non-physicians. This model avoided the need for a master trainer to visit each country.

5. Methods: Why not compare Community Coordinator PANELS vs. Physician
panels? What happens if you analyse by single Physicians rather than panels as the gold standard? Also, it would be useful to the reader if you present the data on the % agreement between members of each physician panel (inter-rater reliability). This would give an indication of the reliability of individual physician diagnosis and would give the reader a sense of the reliability of the 'gold standard'. Comparing panels of Physicians with single community coordinators seems unfair too. I realise you are after a cost-effective intervention (the community co-ordinators) - but it wouldn't cost much more to have panels of coordinators rather than single coordinators as the comparison group.

This is an interesting question, and perhaps one our group may pursue later in another study. We considered comparing physician panels versus community coordinator panels, however decided on the eventual experimental design in the interests of cost and logistics. Also we have compared physician concordance in a previous publication (please see response 1 to reviewer A.

6. Discussion: strength of study being that it is from a variety of countries: But you have not broke down the results to show how they differ by each of the 4 countries.

Please see response for Reviewer 1, Item 4, explaining why we have not reported results by country.

7. Discussion: comparison with Chowdhury et al: you need to calculate Kappa so as to compare with their results.

Please see response to Reviewer 1, Item 5 for why we elected not to use kappas.

Minor Essential Revisions
8. Abstract: Methods: It is not clear who the Physician panels are. You could be referring to the 13 physicians or a separate panel of experts? From the background it is apparent that a panel of 2-3 physicians (presumably from the 13) are used - need to make this clearer here in the abstract.

Thank you. We have amended the abstract to read “Subsequently, panels of two physicians and individual non-physicians from this trained cohort independently reviewed verbal autopsy data from a sample of 118 early neonatal deaths and 134 stillbirths”


We have done this.

10. Methods: calculation of RD: Should mention that this is always on a positive
scale i.e. negative differences are reported as positive Relative Differences.

The reviewer is correct. We calculated the absolute relative difference. During the revisions of the manuscript, this distinction was inadvertently removed. We have added back to the methods that we took the absolute value of the relative difference.

11. Methods: calculation of overall concordance in COD across all causes of SB (57%) and END (47%): Would be good to show this calculation in the Tables.

We have added this calculation as a footnote to the tables.

12. Author contributions: If all 16 authors had significant input to the data analysis and draft writing, how come you all missed the rounding errors for the RD, no-one calculated Kappa statistics to compare with Chowdhury et al; no-one looked at inter-rater reliability of physicians within panels, no-one looked at the differences between countries etc? I’m not convinced that all of the 16 authors participated as stated.

Please see previous responses which address these comments.

13. Figure 1: split of 518 PD to 229 END and 289 SB: How was this split determined (given that no VA was done) - maybe worth adding a footnote here and also a footnote for the bottom 134 SB and 118 END split saying that that is according to physician diagnosis following VA.

We have stated on page 7 that determination of enrolled deaths was based on the physician consensus. The determination of END/SB status among those not enrolled in the study was determined during the FIRST BREATH trial by the birth attendant.

14. Table 1 and 2 footnotes: should change the 'other' within the 'other' category to spell out what these causes were or state that they were unknown. Also, what is the difference between the main category 'Trauma' and the causes 'birth trauma' and 'neonatal accident' which are categorised under 'other'. Should make this clear as currently the categorisation seems inconsistent.

We agree and have amended the footnotes as suggested.

We used the following definitions commonly accepted by WHO and published in our previous manuscripts.

Birth trauma - injuries to the infant during the process of birth.
Neonatal accident – any injury or trauma affecting the infant after it is born.
Fetal trauma – injuries affecting the fetus before it is born.
Discretionary Revisions

15. Abstract: Methods: would be better if you briefly described the robustness criteria in the abstract.

We are constrained by word numbers in the abstract so are unable to incorporate a description of robustness criteria.

16. RD calculations: Is it possible for you to calculate 95% CI for the RD? As a conservative estimate you could simply base it on the RD between the upper 95% CI for CSMF(PC) and the lower 95% CI of CSMF(CC) as the lower limit, and the lower 95% CI for CSMF(PC) and the upper 95% CI of CSMF(CC) as the upper limit.

These tables currently provide a variety of measures. In order to make room for the relative difference confidence intervals we would need to cut something. What would the reviewer suggest we remove?

17. Conclusion: I think the way forward is actually computer programs like InterVA. They are more accurate, are consistently reliable, valid (based on state of the art expert medical opinion) don't rely on one cause of death and are far quicker and far cheaper and thus more cost-effective. Would be good if you could expand upon the sentence in the discussion on the advantages and disadvantages of computer programs (or even simple algorithms for the SB / END split).

We agree that the way forward may utilize computer simulation techniques such as InterVA or the King-Lu probabilistic method. Since the focus of this manuscript was not on these alternative simulation techniques, we elected to make brief mention of such techniques, rather than elaborate in detail.

Level of interest: An article of importance in its field
Quality of written English: Acceptable
Statistical review: No, the manuscript does not need to be seen by a statistician.
Declaration of competing interests:
I declare that I have no competing interests Reviewer 2:
Cyril Engmann MD FAAP
Attending Neonatologist
Associate Professor of Pediatrics and Public Health University of North Carolina Schools of Medicine and Public Health
CB# 7596, 4th Floor UNC Hospitals