Author's response to reviews

Title: Performance criteria for verbal autopsy-based systems to estimate national causes of death: development and application to the Indian Million Death Study

Authors:

Lukasz Aleksandrowicz (aleksandrowl@smh.ca)
Varun Malhotra (varunm@gmail.com)
Rajesh Dikshit (dixr24@hotmail.com)
Prakash C Gupta (guptapc@healis.org)
Rajesh Kumar (dr.rajeshkumar@gmail.com)
Jay K Sheth (jayksheth@yahoo.com)
Suresh K Rathi (RathiS@smh.ca)
Wilson Suraweera (SuraweeraW@smh.ca)
Pierre Miasnikof (MiasnikofP@smh.ca)
Raju Jotkar (rmjotkar@gmail.com)
Dhirendra N Sinha (sinhad@SEARO.who.int)
Shally Awasthi (rathis@smh.ca)
Prabhat Jha (Prabhat.jha@utoronto.ca)

Version: 2 Date: 4 October 2013

Author's response to reviews: see over
The paper presents an analysis of data quality of the Million Death Study conducted in India since 2002. This is from far the largest study on causes of death based on verbal autopsies (VAs) and on a large representative sample of deaths in a national population (some 170,000 deaths). This report will be highly valuable for researchers, international organizations, development agencies, and health professionals. However it suffers from a number of weaknesses and limitations, and could be improved before publication.

General comment 1. The paper has two parts, which could be separated: the first part deals with the quality of the VAs, the second with a comparison of re-classifications systems. This is not at all the same issue. Re-classification can be based either on good or bad data. I recommend extending the first part, the most important for readers. The second part deals with arbitrary re-classifications used by WHO and by GBD. It could be presented separately in the paper, or in an appendix, or in a separate paper.

R1. We agree to expand on the first section. However, we believe that the re-classification system is a critical feature of the performance of the MDS and should therefore remain in this manuscript.

2. The first part should deal specifically and in details with the specific points used for quality assessment: proportion ill-defined; CSMF compared with re-sampling; urban vs rural; total vs hospital based studies; age patterns. Each point could be addressed in a separate paragraph with a specific table or figure.

R2. For clarity, we have reorganized the first section in the sequence suggested.

3. Do authors have a direct comparison between VA’s and medically certified diagnoses?

R3. Not for individual deaths. As we now make more emphatically, medically-certified diagnoses in Indian hospitals are not reliable and should not serve as a robust comparison.

4. The grouping into 18, 19 or 21 categories is very arbitrary, and of little use for public health. Neither the WHO nor the GBD could be considered as proper references. The differences between the 3 classifications should be better investigated, since they use the same data and the same original diagnosis.

R4. We partially agree, as WHO and others have always collapsed the full ICD-10 list into a smaller set of ICD-10 codes. We have added a point in the discussion that further comparisons of these systems are needed, and the resquencing of the text now makes this point more clearly.

Some differences are hard to understand.

5. How were multiple causes dealt with?

R5. The MDS coders determine only the major underlying cause of death as assigned as a 3-digit ICD code. We have made this point of underlying cause of death clearer.

Detailed comments

6. The authors are unclear whether they focus on age 5-69, or on all ages. They should make a choice, and stick to it in the whole paper. I recommend including all ages.

R6. Agreed. All tables and graphs use ages 5-69. We have made this clearer in the titles.

7. Page 6. The urban / rural difference is not a matter of “over-estimating” or “under-estimating”. It is simply a different cause of death profile.

R7. Agreed. We have corrected to lower and higher- and made the point that extrapolation of deaths in urban areas or hospitals to other areas is problematic.

8. Page 10: what is ORGI?

R8. Office of Registrar General of India- now changed throughout the text.
9. The large differences in road traffic accidents should be discussed in detail. This is unexpected.  
**R9.** We have added a foot note that most RTIs in the MDS were classified as “other”, meaning death did not happen at home or hospital, but at scene of accident or en route elsewhere. The RTI results are discussed in more detail in Hsiao et al, BMJ Open 2013.

10. It there any way to provide more details on infectious & parasitic disease (e.g. measles, pertussis, meningitis, etc.)  
**R10.** Some of these results are detailed in separate MDS papers so we do not repeat their key results.

11. Is there any way to provide more details on injuries: road traffic accident, domestic accident, snake bite, other injuries; homicide; suicide; other violence.  
**R11:** See R10.

12. The section on snake bite could be expanded if illustrated by a figure. However, is it necessary in this paper, since there is already a publication on this issue?  
**R12.** Despite being already published, we prefer to retain as this illustrates plausibility of seasonality for a clear cause of death.

Tables & Figures  
Table 1 is of little use. Its information could be presented in a few sentences in the Data & Methods section.  
**R13.** We would disagree with the reviewer. We prefer to highlight each of the specific methodological features in the table.

Figure 1 is of little use. It could be put in words in the Data & Methods section.  
**R14.** We believe that this flowchart of MDS field work/process is useful to the general audience and prefer to retain.

Figure 2 is of little use. Numbers could be explained in text.  
**R15.** Agreed- deleted and now added to the text.

Figure 3, 4, 5, 6 are not readable in black and white. Add markers for each series.  
**R16.** Agreed. Now labeled clearly.

Additional files:  
It would be better to explain how and why the classifications differ, and to provide only the proper numbers. The figures are hard to read and of little use.  
**R17.** Agreed. We have retained the classification and additional file 2 has detailed spreadsheet which should be useful for specific technical uses. We have deleted Additional figures in File 3 and 4.

References  
Author could add a link to the questionnaire used in the MDS.  
**R18.** We have added additional file (no 4) with the MDS Forms, as well as additional file 5 with screenshots to show the training materials.

**REVIEWER 2**  
The intended contribution of this study is the metrics of evaluation. I think the authors have this completely wrong. The key issue here is whether the CSMF’s are estimated correctly. And with the MDS, from the evidence they provide, there isn’t much evidence that their goal has or has not been achieved.  
**R19.** We believe that the reviewer has very much confused the questions of the paper. First, there is no “gold standard” against to which to compare CSMFs. As identified consistently in the literature (see references 1-6), in this paper (and
indeed in the other papers in this series), the use of hospital—based patients as “gold standard” is not valid as hospitalized patients do not yield the same diseases or symptom patterns (and may suffer from other biases) as home deaths.

The closest “correct or accurate” CSMF is that expected from two completely independent sampling of the same deaths—which we provide in Table 2. This shows the CSMF distribution for the original MDS and the re-sample are very similar. This gives us the best confidence in the results of the MDS.

In particular, the inter-coder reliability for this study is remarkably low — 66%. Even if these deaths were all correctly coded, which is unlikely, adding the unobserved 33% to one or more of the CSMF rates — i.e., where they might belong — could completely change the results of the study. R20. Physician agreement does not equate to accuracy, but more to precision. Even with 100% agreement by physicians, the underlying CSMF would not automatically be used to predict the population distribution. The classic example is our findings for malaria, where slide or test–confirmed malaria deaths would both be confirmed by two doctors as being deaths from some other condition, but these could not be used to estimate that fever deaths in rural areas were from causes other than malaria. We have however, re-written this section for greater clarity- pointing out 1/3 disagreement at the outset, which fell to 15% after reconciliation.

If the MDS is going to be used as the authors describe, there needs to be serious validation. (1) When physicians agree, how often do they agree on the correct COD? This is a common issue, but needs to be studied. More to the point, the pattern of selection bias that would result from omitting all the missing data must be shown to be uncorrelated with the quantites of interest in the study. If not, the bias could be huge. R21. See R20. This goes back to lack of a reference standard. There was no change in the MDS CSMFs even when we incorporated the ICD codes that produced disagreement in physician coding (rather than just taking the final, adjudicated code). The important issue is not rate of agreement (which can be influenced by quality of the narrative and other variables in the field work), but whether the CSMFs generated are reliable. Plausibility, age patterns, ill-defined deaths and other support to the CSMF fractions are plausible in the MDS. We also point out that CSMFs based only hospital studies or urban areas (which were the VAST majority of VA studies to date) are misleading. A true random sample of deaths is far less. Moreover, what we offer here are simple, transparent criteria to extend VA studies to national levels to generate national COD estimates (including CSMFs). These build on performance criteria used for systems which rely on medically certified causes of death.

About 12% of the MDS records could not be surveyed due mostly to outmigration or locked houses during the field work. As published in various earlier papers, the broad age, sex and regional distribution of these deaths did not differ from the 88% of deaths we included (see Cancer/Suicide mortality papers in the Lancet- all on www.cghr.org).

(3) The authors explain that half of the 1/3rd of the observations where physicians disagreed where adjudicated until agreement. This is a procedure, but the authors need to provide a rigorous validation, with external information such as medical verification from a subsample, that this procedure works to reduce the bias. There is no evidence (or even clearly stated assumptions) that would prevent this procedure from increasing the bias. R22. See above. It would be misleading to compare results to hospital-based deaths (see also review by Leitao in this series). The assumptions in this paper are that independent coding my dual physicians provides a quality check on any diagnostic error by one physician and the knowledge that their work will be checked means that physicians are far more likely to adhere to the guidelines (in particular as payment is made only after reconciliation). We disagree with the idea that dual coding would increase bias- usually- half the time in disagreements, one doctor yields to the opinion of the other and the other half of the time, they stick with their diagnosis, leading to adjudication. There are no clear physician correlates (such as seniority, etc of those who yield or those who don’t at reconciliation). Thus, the main effect is that we are pushing physician coders to adhere more to the guidelines, which is not a bias, but a desirable outcome.
Overall, to refer to these as “good quality results” is unjustified.

**R23.** We completely disagree. Strong CSMF agreement on re-sampling, plausible CSMF patterns on using hospital or urban deaths, and age-sex plausibility all are clear evidence of good quality results.

Re the next section: It is inappropriate to use the results of the study to evaluate the methods. If you want to do that, there’s no reason to collect the data in the first place. The methods must be justified ex ante, without looking at the results. I see no justification for them here.

**R24.** The methods are extension of commonly used criteria for COD statistics. Our proposed metrics (ill-defined before old age, and other criteria), are similar to those have been applied to physician-based COD systems (see Mahapatra et al, ref 47, other studies by Mathers et al, ref 3. Lopez). If VA studies are going to be used to generate national COD estimates (which we support), our metrics are clear, transparent and replicable. The methods of simple measurement of COD have been historically justified and used for over one hundred years in high-income countries. The key issue of medical attention at death cannot be wished away.

Sophisticated statistical methods exist with which to accurately -- with known degrees of bias and uncertainty -- estimate the quantities of interest in this study at lower cost. I don’t see why the authors are pursing such outdated and inaccurate methods.

**R25.** We strongly disagree with the criticism that these are outdated or inaccurate. As pointed out in Paper 1 in the series, sophisticated statistical methods applied form high income data sets to LMICs have shown huge inconsistencies (such as huge variation in estimation of TB deaths in India, or even diametrically opposed conclusions on trends in all-cause child mortality in South Africa).

Several automated methods exist, however are not yet ready for widespread use (see paper 4 by Desai et al). The VA software developed by King and Lu performs well at the population-level (for community and hospital-based deaths), but does not offer individual coding, which is a priority for death registration. Many automated methods require labeled data to train on which has so far come from physician-assigned CODs from community and hospital-based deaths.

A few other points: Define “SRS” before it is used.

**R26.** Done

This needs to be made precise, so that a reader could follow the same rules. as is, I can’t tell what it means. “Odds ratios to compare CSMFs in various sub-groups were adjusted for age (linear year), sex, religion (Hindu versus other), education (illiterate versus literate), poorer or richer state, and as needed, hospital versus home and rural versus urban status”

**R27.** We have clarified that logistic regression was used to generate the odds ratios, and that the model included the variables listed.

The section on ill-defined codes is ill-defined; it doesn’t explain what they are or how to recognize them if we see them, other than that a different study defined them. I can’t even tell what the odds ratios are trying to estimate. There are different ways of estimating odds ratios; this must be clarified.

**R28.** We have further defined on ill-defined deaths.

adjustments in the tables must be described in sufficient detail so that they can be replicable. I can’t figure out what they are, except in the most vague sense.

**R29.** See R27.

**REVIEWER 3**
This is an important and well-written paper relating developments in physician-based VA CoD determination in the long-standing Indian Million Deaths Study. It details the procedures related to this study, which aims to cover a representative sample of about 0.5% of the Indian population. At the time when the MDS started, the state of the art in terms of VA was at a very different position compared with today. Consequently the fairly elaborate procedure, as described in the paper, to have large numbers of physicians assessing largely narrative-based VA interview material, was probably the only viable option at the time, apart from doing nothing. It would be helpful to the reader to make this point clear.

R30. We would slightly disagree with the reviewer. The paper provides metrics for VA-studies used to generate national COD statistics, and this is independent of the type of coding method (physician or CCVA). Other papers in this series point out that CCVA and PCVA results do not match well at the individual level, and CCVA methods are therefore not yet an obvious replacement for physicians. We have expanded in the paper that physician coding is NOT a rate-limiting step, indeed with use of electronic training and certification and coding. We were able to identify and train 300 doctors within 8 months, and they coded about 50,000 records from 2004-5 (not analyzed here) within 3 months. Additional files have been added to show the training methods.

In addition, the considerable effort involved here in covering just 0.5% of the population raises questions about scalability, in terms of moving towards full civil registration of deaths with cause, integrating VA strategies, which also needs some additional discussion.

R31. Agreed. We have expanded the discussion in the context of civil registration. See also Paper 1 which discusses the tradeoffs of civil versus sample registration to obtain CODs. The key point is that random sample of deaths are needed for sample surveys. Eventually, civil registration will occur (as is happening in China) but could take decades. For example, 100% coverage of CODs in poor black men in Louisiana, USA was not achieved until 1975.

If one were starting such a venture now, there would be a more complex array of possible strategies. The major question would be whether to use a physician-based approach as MDS continues to do, or to consider completely automated approaches which are now coming into much wider use. This is not an argument for now changing the MDS strategy mid-stream, which would almost certainly be a bad idea; but as time passes, the MDS approach will appear to be an increasingly “legacy” solution to the very real problem of documenting cause of death in India.

R32. We thank the reviewer for this perspective. However, it is premature to move to CCVA, for reasons discussed in the other papers. We do add that future development should emphasize a random sample of deaths (versus small non-random studies which have been the norm) and need to test combinations of PCVA and CCVA. Finally, we add that even with limitations, PCVA approximates the accepted clinical standard of having a physician review (and is akin to a medical student interview but with diagnosis by a team of junior and senior doctors).

Thus the authors here face something of a dilemma; they usefully document the methodology behind a huge on-going VA survey, with relatively positive results; but, at least in my view, the conclusion should not be that their approach represents an optimal future strategy to fill the global gaps in cause of death data. I would therefore like to see a more nuanced discussion along these lines, to accompany the very interesting detailed documentation given here.

R32. We agree. Paper 1 describes other options besides SRS to enable countries to obtain a reliable representative surveys. We encourage other VA systems (INDEPTH, etc) to also apply these metrics and perhaps to improve on them. The discussion is re-written to emphasize these points.

Page 7 – important to make clear that the WHO 2012 VA cause categories are simply groupings of ICD-10 codes, not a separate classification system like GBD (which in particular treats poorly defined codes in a totally different way).

R33. Yes. Point now added.

However, in the related table at the start of the supplementary material, I do not recognise the WHO 2012 codes as being those specified in the WHO 2012 documentation. All those WHO VA cause codes have a format of e.g. 01.04 (Diarrhoeal
Diseases). This needs clarifying in the supplementary material. It may be helpful to refer to the newly published paper by Leitao et al. (Global Health Action 6:21518) for this.

**R34. Yes, we have now corrected to the WHO VA cause codes.**