Author’s response to reviews

Title: Automated versus Physician Assignment of Cause of Death for Verbal Autopsies: Randomized Trial of 9374 Deaths in 117 Villages in India

Authors:

Prabhat Jha (prabhat.jha@utoronto.ca)
Dinesh Kumar (drdineshk@charutarhealth.org)
Rajesh Dikshit (dixr24@hotmail.com)
Atul Budukh (atul.budukh@gmail.com)
Rehana Begum (BegumR@smh.ca)
Prabha Sati (SatiP@smh.ca)
Patrycja Kolpak (patrycja.kolpak@mail.utoronto.ca)
Richard Wen (WenR@smh.ca)
Shyamsundar Raithatha (shyamjr@charutarhealth.org)
Utkarsh Shah (drutkarsh2537@gmail.com)
Zehang Li (lizehang@gmail.com)
Lukasz Aleksandrowicz (lk.aleksandrowicz@gmail.com)
Prakash Shah (ShahPr@smh.ca)
Kapila Piyasena (delpepiyasena@gmail.com)
Tyler McCormick (tylermc@uw.edu)
Hellen Gelband (hellengelband@gmail.com)
Samuel Clark (sinafala@gmail.com)

Version: 2 Date: 05 May 2019

Author’s response to reviews:
Reviewer #1: The authors have satisfactorily addressed my comments on the original version. Given the strong correlation between two independent physicians from this study as well as other previous studies, I recommend that this manuscript could revise its recommendation of compulsory dual physician review for all VAs. Dual physician review adds to costs and logistical challenges, and does not seem to offer that much of an advantage, as shown here. Instead, the manuscript could recommend analysis of specific causes which demonstrate divergence of cause of death assignment, and develop guidelines to promote greater consistency. Further, the recommendation could be revised to single physician review with a random sample review to check for agreement. This could start with a higher percentage - say 25% random re checking, then gradually brought down as agreement is systematically proven. Such strategies are essential to reduce workload and costs as VA implementation expands to cover larger populations in the developing world.

REPLY 1: There is much evidence that dual review is worthwhile (references 16-18, 21-23). An additional benefit of dual review is the ability to de-bias physician codes (reference 11). Nonetheless, we have toned down the recommendation on dual physician coding, stating that this is desirable (abstract, discussion p 17).

Reviewer #3: Overall comment: I would like to thank the authors for addressing most of my comments, and clarifying several things in the manuscript. However, I do not think the main issue has been fully addressed. The study design the authors used to compare COD between two arms has introduced a problem, as the ascertainment of the outcome and the method for classifying the outcome differ between the two arms. So they cannot disentangle which of these two would lead to discordance. In addition, the true distribution of the outcome could be different, and therefore discordance would be expected - and should not lead to the conclusion that either physician review or algorithm assignment are flawed (and appendix 10 and 11 support this, as physician versus physician did not show the same distribution of causes). For example, in the results section, I do not agree with the statement that the symptom patterns were the same between the two trial arms. In Table 1, a difference between 50% fever in the physician arm and 39% in the algorithm arm for children, with the numbers of participants, I would assume would be statistically different (as with jaundice in both neonates and children). Therefore the finding that concordance between the two arms for the child and neonate groups was low is unsurprising. Based on the primary analysis approach, I don't think you can conclude that it is low because of poor classification of the algorithms versus physicians, because they may have actually died of different causes, or because the respondents reported different symptoms during narratives versus closed questions. So it is still unclear to me how the study design used is an advantage over comparing COD using different methods within the same sample of deaths and I think this needs to be much more explicitly explained and addressed.
Reply 2. We thank the reviewer for this point, but believe s/he has fundamentally misunderstood the paper and disagree with their points.

The trial outcome is the cause of death (COD) distribution using two different approaches, and comparing algorithms to standard physician assignment, so naturally the ascertainment and method for classifying the outcome will differ. That’s the whole point of a randomised assessment!

Moreover, the main goal of verbal autopsies is to establish the population distribution of CODs, and not to replace individual death certification by a physician. So the important public health question is to know if in any particular population, say among 1000 deaths, how many arise from TB, malaria, vascular disease, injury etc. (references 1,5,7). We know that individual comparisons (based on past reviews-reference 18) show much lower agreement than at the population level (formally CSMF%), so this RCT adopts a lower threshold to show that algorithms could be used for VAs. (Our conclusion is that they cannot yet be so used).

We chose not to randomize individual deaths to both instruments (which are nearly identical except for the narrative) as obviously the first interview would influence the data captured during the second from household recall. Trying to tackle this possible bias (by for example randomizing the sequence of interviews) would make for a complex trial design that might well be difficult to interpret (see Pocock https://doi.org/10.1016/j.jacc.2015.10.051). The reviewer misses the design simplicity of randomization- randomize A and B and count dead bodies.

Moreover, the reviewer confuses the origin of the symptoms. In both arms, symptom lists shown in Table 1 come from a checklist of questions, NOT from narratives in one arm. The text already states clearly: “Completion of all questions, including negative symptoms, was mandatory [in each arm]. The main difference in the forms between the arms was the inclusion of a narrative of the symptoms and events leading up to death at the end of the interview in the physician assignment arm.” We have further clarified this to state that the symptom patterns were drawn from the checklist of questions.

Finally the reviewer makes the mistake (made still by too many clinicians) of inappropriate post-hoc sub-group analyses (see https://www.nytimes.com/2017/11/28/magazine/a-failure-to-heal.html). There is little doubt that the overall balance between arms worked well (particularly for adults with larger sample sizes). Cherry picking some differences (such as jaundice and fever which don’t differ in adults, only in children, and that too only in one of the two states of the study) that likely arise simply from the play of chance is misleading. Table 1 has 35 comparisons, so just the play of chance would mean at the 5% confidence level, 1 to 2 would be different.

To make the point further, we show below that among adults, those whose name started with the letters B, K or P were more often found in the automated arm whereas those whose names began
with M or N were more often in the physician arm. Naturally these differences are specious, as
the overall distribution by letter is no different.

<table>
<thead>
<tr>
<th></th>
<th>Overall (n=9374)</th>
<th>Standard assignment (n=4651)</th>
<th>(physician assignment (n=4723)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult’s first letter of name</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B, K and P</td>
<td>1872 (22%)</td>
<td>882 (20%)</td>
<td>990 (22%)</td>
</tr>
<tr>
<td>M and N</td>
<td>1287 (15%)</td>
<td>687 (16%)</td>
<td>600 (13%)</td>
</tr>
<tr>
<td>Remaining Letters</td>
<td>5,545 (63%)</td>
<td>2742 (64%)</td>
<td>2803 (65%)</td>
</tr>
</tbody>
</table>

We do not expect that the reviewer will change his/her views, but believe strongly that it is a misleading to do post-hoc analyses of randomised evidence, as most journals now recognize. Nonetheless, we have incorporated some of the points raised by the reviewer into the introduction and results (page 10).

Minor Comments:

Methods (Participants): thank you for responding on the sampling in the reply. I still think the random selection of villages and that all households in those villages were included needs to be explicitly stated in the methods, rather than implied.

REPLY 3: We don’t make the claim that villages were selected randomly. All households in the village were included, wherever possible.

Methods (Trial Procedure): I am now unclear, the automated arm did not collect narrative data, so did SmartVA use this?

REPLY 4: As per SmartVA procedures, we entered key symptoms from the checklist into the list of words that SmartVA requires.