Author’s response to reviews

Title: Data extraction from machine-translated versus original language randomized trials: a comparative study

Authors:

Ethan M Balk (ebalk@tuftsmedicalcenter.org)
Mei Chung (mei_chun.chung@tufts.edu)
Minghua L Chen (lilychen71@gmail.com)
Lina Kong Win Chang (linakongwinchang@gmail.com)
Thomas A Trikalinos (thomas_trikalinos@brown.edu)

Version: 2 Date: 23 August 2013

Author’s response to reviews: see over
Dear Systematic Reviews Editorial Team,

Thank you for your continuing interest in our manuscript “Data extraction from machine-translated versus original language randomized trials: a comparative study.” We found the reviewers’ comments to be very insightful and helpful. Below our their comments with our point-by-point responses.

We are happy to discuss any additional revisions that may be necessary.

Sincerely,

Ethan Balk, MD MPH
Tufts Medical Center, Boston, MA, USA

Reviewer: Guy Tsafnat

Major Compulsory Revisions
1. The data collected by the authors is substantial and they are to be commented for it. The methods are sound and presented clearly. However, the results are not presented in more detail and mostly the results section reads like a discussion section. In particular, there are very few details in the results. For example, "the average time to translate was about 30 minutes" would better be described as the average time and standard deviation. Similarly the following statements should be changed to include statistical results: "articles were estimated to take between 6 and 30 additional minutes to extract" (consider showing average and confidence interval), "the accuracy was about the same regardless of extractors’ confidence level (76 or 79%)." (not sure what this means so can’t suggest how to better present it). Table 1 and Table 2 are

We have made revisions. We corrected the sentence about “average” time to translation to read that the median time to translate…. We do not have the exact times it took to translate, just rounded estimates, so calculation of a mean and SD are not possible. As noted in the Methods “a research assistant … estimated the approximate time she required to translate each article.” We have also added Table 3 from the full report as an appendix table (Appendix 2, Table 2).
Likewise for time to extract, “Extractors provided a rough estimate of how much extra time they believe they spent with the article compared to an extraction of a similar English article.” We believe the lack of specificity is made clear by the phrasing “extractors estimated they took <5 additional minutes to extract... took <30 additional
minutes to extract... were estimated to take between 6 and 30 additional minutes to extract.” We have also added Table 4 from the full report as an appendix table (Appendix 2, Table 3).

We corrected and rephrased our sentences about extractor confidence. We moved these to the section “Association between extractor confidence and accuracy” and have also added Table 5 from the full report as an appendix table (Appendix 2, Table 4).

We did not think that SDs of proportions (percentages) was of added value, but we are happy to add these if the editors prefer. As noted in Table 2, Appendix 6(originally Appendix 5) has the confidence intervals about the odds ratio estimates.

We do not know what the reviewers thoughts were about Tables 1 and 2. We are happy to work with the editors to make them clearer, if necessary.

Minor Essential Revisions
2. This might be an issue with the submission system rather than the author’s prerogative but supplementary file names do not match their file name. E.g. Appendix 2 is in sup4.docx.

We believe this is an issue with the submission system. We will work with the journal editors to ensure that files are called out by their proper names.

3. Appendix 3 starts with a long sentence I did not actually understand

We have removed some unnecessary detail and simplified the sentence.

Discretionary Revisions
4. In addition to 1 above, I strongly suggest to the authors to present to their results not just more accurately but also graphically to help the reader.

We would be happy to take suggestions for any specific graphics that would add value and information to the article. There was no figure we have thought of that would do so beyond the tables and text already included.

Reviewer: Alexis Llewellyn

The authors addressed an interesting and relevant question. The methods were well described and it appears that sufficient details were provided to replicate the work (taking into account the original AHRQ report to which the paper refers). Attempts were made to distinguish between issues associated with poor translating, poor reporting and errors unrelated to translation in
extraction. Relevant limitations of the study were acknowledged, and cautious attempts were made to explain variations in findings relating to accuracy across languages.

Thank you.

I have the following suggestions for revision:
A) Major compulsory revisions
1) Methods, “Data extraction form and comparison”, first sentence: I would at least nuance the statement about which domains are the most important for systematic review. I think judgement about importance and detail given to each domain depends amongst other things on the review question, and one could argue that extracting at least some basic data on participant characteristics and comparator is often equally important.

   We have changed this sentence to: We focused data extraction on common and important study domains for systematic review: study design and methods, interventions (and comparators),...
   In Appendix 5 (old Appendix 4), we have also clarified that by “Intervention” we actually meant “Intervention / Comparator”, ie, each study arm.

2) Conclusions: The authors clearly attempted to provide nuanced and balanced conclusions. However, I do not think that the conclusions sufficiently reflect the evidence presented: concerns about the inaccuracy of extraction across all foreign languages regarding results and/or outcomes data do not suggest, in my opinion, that basic translation software is sufficiently accurate for data extraction in systematic reviews of health interventions. I am not convinced that the better results in other domains, or the relatively limited resources and effort needed for using the tool may sufficiently outweigh the inaccuracies found in the analyses.

   We have reworked parts of the Conclusions to downplay our conclusion favoring use of machine translation, in particular adding a phrase to the first Conclusions sentence that the accuracy might not be sufficiently high and adding a final sentence saying the same.

3) Conclusions: 3rd sentence: the threshold according to which extraction is considered “fairly accurate” is unclear. Some clarification is required to strengthen this statement.

   We have changed the sentence to: “Specifically, extractions of Spanish articles were most accurate, followed by somewhat less accurate extractions from German, Japanese, and French articles.” We do not repeat the results here since, but we are happy to repeat some data if the editors desire.
4) Discussion: Despite the fact that a power calculation was performed, I still have some concerns about power/precision of estimates, notably given the wide confidence intervals of ORs presented. I think a brief comment on this (similar to the one main in p22 of your original AHRQ report) would be appropriate.

We had omitted this sentence from the original report for the sake of brevity, but we agree that it is a useful statement to include: “However, despite our power calculation, the confidence intervals of the adjusted odds ratios between translated and English articles were generally wide, possibly resulting in either an overestimation of the number of items with “trends” toward large differences in accuracy (i.e., small but nonsignificant odds ratios) or an underestimation of the number of true effects (due to frequent nonsignificance).”

B) Minor essential revisions
5) Results, “Comparison of extractions from translated to original articles”: For consistency, and as for Chinese articles, the % of items that are each #98% correct for Spanish papers should be reported.

We agree. To the description of Spanish articles, we have added: “but only 30% of items were extracted correctly by >98% of the extractors”.

C) Discretionary revisions
6) Abstract: The results section of the abstract currently reports the percentage of items that were correctly extracted more than half the time. I think this could be seen as a low threshold for medical research. I would prefer to see the % of items that are each #91% (or #98%).

We have kept the 50% and added the 98% thresholds to the abstract. The word count is still well under the limit (287 words).

7) Background: For readability, I suggest breaking the last sentence of the background into two.

We agree. Done.

8) Methods: Figure 1 in the original AHRQ report may make the translation/extraction process more explicit. I would consider adding it to the appendices.

We agree and have added this as Appendix 4, along with a callout to it in the Data extraction process section of the Methods (page 6).