Author’s response to reviews

Title: Development of an algorithm for evaluating the impact of measurement variability on response categorization in oncology trials

Authors:
Jeong-Hwa Yoon (yjh891114@snu.ac.kr)
Soon Ho Yoon (yshoka@gmail.com)
Seokkyung Hahn (hahns@snu.ac.kr)

Version: 1  Date: 03 Mar 2019

Author’s response to reviews:

Carel F.W. Peeters
Editor
BMC Medical Research Methodology

3 Mar 2019

Re: BMRM-D-18-00418

Dear Dr. Peeters,

Thank you for giving us the opportunity to revise our manuscript entitled “Development of an algorithm for evaluating the impact of measurement variability on response categorization in oncology trials” for BMC Medical Research Methodology.

We appreciate the constructive comments from both the editor and the reviewers. The feedback motivated us to expand our data and to improve our analyses. We have amended our manuscript in light of all the comments that we received, and we hope that we have addressed the points raised satisfactorily. We hereby enclose our revised manuscript, together with point-by-point responses to the detailed comments provided.

Thank you for your consideration and we look forward to hearing a positive response from you.

Sincerely,

Seokkyung Hahn
Responses to the comments from editors and reviewers

Associate Editor
I have reread the manuscript as well as the referee reports closely. Both referees raise valid points of concern and I feel that, ultimately, Major Revisions are required. If you decide to take the opportunity to embark on a revision then I suggest that you follow the suggestions made by both referees.

We would like to thank the editor and reviewers for their constructive comments. We have amended our manuscript in the light of the comments provided, as described below. All changes in the manuscript are highlighted.

In particular:
- Both referees indicate that the method likely underestimates the variability in ORRs (Points 6 and 7 by Referee 1 and the Main Concern by Referee 2). In light of this concern and subsequent points raised by Referee 2 it seems appropriate to expand the simulation and validation exercises. In particular, one would like to see (a) a larger number of re-samplings in the real-data validation and (b) a simulation study in which the true ORR is known. Moreover, in this latter simulation study it indeed seems appropriate to (as Referee 2 suggests) assess the impact of heterogeneity in baseline tumor size and percent change.

AUTHORS’ RESPONSE:
We have now expanded the exercise using 100 resampled datasets in the real-data validation and provided the results.

We have also performed the suggested simulation study in which the “true ORR” was hypothesized to be known. We compared the 95% CIs for the observed ORRs, which were based on the simulated datasets from the known ORR, and the 95% central ranges of ORRs obtained by reassessment. We explored trends in whether the observed ORR was covered by the 95% central range in a series of simulations. The simulation results will help understand the patterns of discrepancy of these two different measures. However, it should also be noted that it is difficult to interpret the results of the simulation in terms of the “true ORR” for validation of the evaluation algorithm. Since the algorithm tool should be considered as a pragmatic alternative to a reassessment exercise by the same reader or another reader, the algorithm validation can only be meaningful when it is performed using a real dataset that contains actual measurements obtained by readers re-reading the tumor size. Our base model is not relevant to the “true ORR,” but to disagreement between the results of the original assessment and a reassessment. In other words, close agreement between an original assessment and its reassessment could be achieved, even if the original assessment itself was inaccurate, depending on the condition of data for generating the baseline tumor burden and percent change. What we essentially tried to highlight through our work was the pattern of such reproducibility, and we tried to present this point more clearly in this revision.
We presented and discussed the results in the text and the Supplementary Materials.

- Referee 2 raises several points regarding the modeling (under heading “Methods: Modeling”) that may affect the reliability of ORR determinations. For a full appreciation of the proposed methodology these points should be addressed.

AUTHORS’ RESPONSE:
We appreciate the referee’s comments with respect to the modeling. We tried to clarify the modeling step better in the revised manuscript, and we also explained in detail how it was dealt with in the relevant responses to the referee.

- Both referees seem to hint that, in clinical trials, uncertainty regarding the ORR is driven by small sample sizes rather than measurement error. In this light and given that most trials are small, please (a) indicate more clearly the relative merit of the proposed approach and (b) provide guidelines on how researchers should assess the results stemming from your tool.

AUTHORS’ RESPONSE:
In this revision, we described the role of a 95% central range of ORRs determined in reassessments through a comparison with the 95% confidence interval. We highlighted that those two measures shed light on different aspects of uncertainty of an observed result obtained from a given clinical trial. We have provided guidance on how researchers should interpret the 95% central range along with the 95% CI.

Daniel F. Heitjan (Reviewer 1)

1. 4:21 — I would say that many trials "designate the OR as a primary or secondary outcome". That is, "designate" rather than "provide", and "OR" rather than "ORR". OR is the outcome, as blood pressure is the outcome in a hypertension study. ORR is the estimand, as mean BP is the estimands in a hypertension study.

AUTHORS’ RESPONSE:
Thank you for the comments. We have taken more care in the use of these terms.

2. 6:11 — I would say "a phase III randomized trial of chemotherapy in advanced SCLC".

AUTHORS’ RESPONSE:
This phrase was modified as suggested. In the revised paper, the acronym for small-cell lung cancer was used.

3. It seems to me that an important potential use of this work would be to provide some indication of how measurement error potentially affects uncertainty about the ORR value. You do not emphasize this element as much as you could.
AUTHORS’ RESPONSE:
We thank you for your constructive feedback. In this revision, we more clearly emphasized the role of the 95% central range of ORRs determined in reassessments through comparisons with the 95% confidence interval. We highlighted that those two measures shed light on different aspects of uncertainty of an observed result from a given clinical trial. It is therefore advised that researchers should more carefully consider the robustness of their usual inferences based on the CI when considering potential reassessments of the same tumor burdens by themselves or by other observers. We also provided more practical guidance on how researchers should interpret the results stemming from this tool along with the 95% CI.

4. Figures 1A and 1B should use the same vertical scale.

AUTHORS’ RESPONSE:
Thank you for noting this. The scale has now been modified as suggested (Figure 1).

5. Figure 4 — Some readers will find it difficult to understand what you are presenting in this figure. For example, what is the axis in 4A?

AUTHORS’ RESPONSE:
We have added more specific descriptions in the legends (Figure 4).

6. Phase II cancer trials are often of modest size, and therefore the confidence intervals for estimated ORRs, even with substantial measurement error, would be much larger than the intervals that your method would create. This gets back to my point #3 above — I think you need to provide a more cogent explanation for how people are supposed to use this program.

AUTHORS’ RESPONSE:
Our suggested 95% central range of reproduced ORRs deals with the uncertainty of the observed ORR from the original assessment due to measurement variability, whereas the 95% confidence interval deals with the uncertainty against sampling errors in the estimation of the ORR. Our aim was to highlight that those two intervals shed light on different aspects of uncertainty of an observed result from a given clinical trial, rather than to combine those different sources of variability to estimate the true ORR. We also demonstrated how the 95% central range of reproduced ORRs could be different from the 95% CI of the originally observed ORR depending on the condition of data for generating the baseline tumor burden and percent change. Through this exercise, we have tried to provide more understandable explanations for how clinical trialists should explore their observed results.

7. Figure 5B suggests that your method is conservative, giving 100% coverage rather than the intended 95%. Of course this is likely to be an artifact of the small number of "resamplings". How hard would it be to do a larger study, allowing you to acquire a better estimate of the coverage rate?

AUTHORS’ RESPONSE:
Thank you for your comments. We have extended the validation using 100 resampled datasets and provided the results.
Peter van de Ven (Reviewer 2)

The paper describes an algorithm/webtool to evaluate the reliability of the objective response rate (ORR)/progression rate. I found the paper difficult to read. Details of both clinical context (definition RECIST) are lacking and various assumptions/choices regarding statistics/methodology are not communicated in a clear manner.

My main concern with the algorithm is that the proposed method will likely underestimate the variability in ORRs as uncertainty is not taken correctly into account. In particular, in the first step of the evaluation algorithm a probability of progression is estimated for each patient (using the hierarchical model involving parameter uncertainty and a simulation procedure involving simulation uncertainty), however when estimating uncertainty of ORR these probabilities are then assumed fixed. This may not have been picked up in the validation study, which is very limited and uses only a real dataset with unknown true ORR.

AUTHORS’ RESPONSE:
Thank you for your constructive comments. We realized that we were not very clear about what we really tried to accomplish through this study. Taking your valuable comments into account to the extent possible, we have responded to each of the specific comments point-by-point as below and revised the manuscript accordingly.

Background
1.* Page 4 (lines 17-18): Talking about percentages change it is not clear whether the changes of 10% and 30% are absolute or relative increases in percentage change

AUTHORS’ RESPONSE:
We acknowledge that we may have confused readers and we have made the definitions and use of terminology clearer in the text.

2.* Page 5 (lines 2-10): The ORR itself is an ordinal measure. Many methods have been developed for assessing agreement for ordinal measures. These methods are not mentioned in the introduction. It is not motivated why a complex algorithm using the underlying continuous measures is needed to assess variability in the ORR or why this would be better.

AUTHORS’ RESPONSE:
We agree that the RECIST-based response is a type of ordinal measure. However, response categorization is essentially an ordinal dichotomization of the measured change of tumor burden, which is also a composite of individual measurements of target lesions in each patient. The tumor burden is calculated according to the RECIST guideline by summing the measurements of the various target lesions that can adequately represent the overall tumor burden in a patient. Thus, measurement error fundamentally occurs when observing lesion sizes. It then contributes to measurement variability of tumor burden size, percent change, and response categorization in succession. In this study, the uncertainty of the ORR (or progression rate) derived sequentially from measurement error at the lesion level was quantified. We have added a paragraph to the background section to make this context clearer.
In fact, tumor burdens that are categorized as showing the same response can be differently reproducible depending on whether or how the measured change of tumor burden is close to the cut-off values for complete or partial response (-30%) or progressive disease (20%). This point actually motivated us to conduct this study to understand the primary behavior of measurement variability through modeling the measurements of lesion size and then elaborating those differences by measuring the uncertainty and reproducibility of the eventual RECIST-based response categorization, which would not be sufficiently evaluated by assessing agreement simply at the level of the ordinal RECIST-based response.

3.* Page 5 (lines 11 -18): Clearly, you need repeated evaluations for each before and after assessment on the lesions in order to enable estimation of the within-lesion component of the measurement variability. This should be explicitly stated.

AUTHORS’ RESPONSE:
We have added a brief description of the data structure to the Background section.

Methods: Data
4.* Page 6 (line 15): Author's state "According to the RECIST guidelines, ...". It might be good to briefly mention the relevant guidelines here. Was 249 the maximum number according to guidelines?

AUTHORS’ RESPONSE:
We briefly described the relevant guideline in the Methods section. The RECIST guideline recommends that a reader should select up to five target lesions per patient to ensure representativeness of the tumor burden. The guideline focuses on how to choose and to measure lesions for each patient, not on recommending an adequate sample size, which should be determined depending on the primary statistical inference designed in each specific clinical trial. The number 249 was the total number of the selected target lesions from all 75 patients with a median number of 3 target lesions per patient that were determined in a particular clinical trial already undertaken. We clarified the selection process of the target lesions in the Methods section.

5.* Page 6 (line 16): Shouldn't variability due to variation in the images/timing be taken into account when assessing variability in ORR (rather than just between- and within-rater variability?)

AUTHORS’ RESPONSE:
For each lesion, image review comprised four separate sessions as follows: first reading of the baseline CT scan, first reading of the post-treatment CT scan, second reading of the baseline CT scan, and second reading of the post-treatment CT scan. An identical image of each lesion at a specific time point was used for all repeated measurements. Therefore, the image is not considered to have been an issue here. The timing is only related to whether the image was from the baseline or at post-treatment, which was considered in our framework.

6.* Page 6 (line 16): It should be mentioned somewhere that RECIST defines the sum of the patient's target lesions as the reference for ORR. It might be more clear to stated that "for each target lesion you get (a maximum) of 24 sums of longest diameters (12 at baseline and 12 at follow up)", rather than the total of 5820.

AUTHORS’ RESPONSE:
We have described the data structure more clearly in the Methods section.

7.* Page 6 (line 20): Why are additional/different measures considered for the lymph nodes (it this also
in the RECIST guidelines)?

AUTHORS’ RESPONSE:
We developed a separate additional model based on the short-axis diameter for lymph nodes because the RECIST guideline (version 1.1) recommends measuring the short-axis diameter for lymph nodes. However, lymph nodes are the most common type of organ where target lesions are chosen, and the long-axis diameter of lymph nodes was also used for representing the tumor burden in an earlier version of the RECIST guideline (version 1.0). The change to the newer version was required on a clinical basis. For this reason, we included the long-axis diameter measurements of lymph nodes together with other long-diameter measurements from various organs for a general long-diameter model, and we also constructed a separate model for lymph nodes using their short-axis diameter measurements. We described this process in the Discussion section.

8.* Page 7 (line 5): Were longest diameters in the validation set only obtained for lymph node lesions? The description of the development dataset suggests longest diameter is obtained for all target lesions.

AUTHORS’ RESPONSE:
As we mentioned in the previous response, the long-axis diameter was obtained for all target lesions, and additionally the short-axis diameter was obtained for lymph nodes for the training dataset. For validation, one other actual external clinical trial dataset was used where only the short-axis diameter was obtained for lymph nodes as per the RECIST 1.1 guideline, which is why long-axis diameters from lymph nodes were not obtained in this validation set.

Methods: Modeling
9.* Page 7 (line 9): A level for patient is missing from the description. Lesions are nested with patients.

AUTHORS’ RESPONSE:
Our model was constructed with the goal of estimating variance components capturing measurement variability as the remaining variation after accounting for other specific sources of variation. Although the data structure suggested that lesions were nested with patients, because the variation among the patient-specific measurements was in fact part of the variation of the lesion-specific measurements, it was unnecessary to separate these in the model because this did not influence the remaining variations. The estimated distributions of intra-reader and inter-reader measurement errors did not change depending on whether or not the patient effect was added to the model. As we understand that this point could be questioned, we have added a paragraph to the Discussion section regarding this issue.

10.* Page 7 (line 17): Why is 5% trimmed off? It might be exactly these cases that cause imperfect reliability of ORR determination. In addition, variance components may be underestimated as a result.

AUTHORS’ RESPONSE:
From an examination of the outlying data outside the 95% range, it was determined that those outlying results originated solely from misperceptions of the boundaries of target lesions due to abutting non-malignant pathology such as atelectasis or the misunderstanding of target lesions, and such cases exceeded the range of measurement error, which rarely occurs in oncology practice. For this reason, we decided to trim off the outlying 5% of the data, and we have described this decision in the text.

11.* Page 19 (line 19): At this point it is not yet clear phase is (and will be) used to refer to baseline and after treatment measurement.

AUTHORS’ RESPONSE:
To avoid this source of confusion, we have revised to no longer use the term ‘phase’ in the sentence.

12.* Page 8 (line 2): Now a subject effect is mentioned, but a lesion (within subject) effect is missing.
AUTHORS’ RESPONSE:
Thank you for pointing this out. It was actually a typo for “lesion effect.” As we mentioned in a previous response, we chose not to include the subject effect term in the model. We have now corrected the text.

Methods: Simulation
13.* Page 9-10: Shouldn't the true percent change c not be applied to \( \mu_2 \) rather than Y?
AUTHORS’ RESPONSE:
We simulated a situation in which the baseline and post-treatment diameters observed from a tumor burden in the first reading were re-measured 100 times by the original reader or by other readers in order to obtain the probability of the change being classified as a certain tumor response based on the re-measured tumor burden sizes. In this context, c is just a constant value that plays a role in denoting the post-treatment observation as distinct from its baseline observation. In other words, c expresses the observed relationship between the baseline and post-treatment measurements in the first reading, to facilitate the estimated probability function depending on this value. We have clarified the definition better in the text.

14.* Page 9-10 (line 22): How are patients with complete response included in the algorithm/simulation? The probability defined here seems to be a probability of PR or CR.
AUTHORS’ RESPONSE:
The ORR is defined as the percentage of patients who achieve complete or partial response. In accordance with this definition, you are right that the definition of the probability contained in the evaluation algorithm would be the probability of PR or CR (i.e., the probability of achieving at least partial response). We have clarified the definition throughout the text to avoid confusion.

15.* Page 11 (line 11): Is c = -1.0 complete response? The simulation set-up will likely result in negative realizations of measures for the extreme negative c. How was this dealt with?
AUTHORS’ RESPONSE:
Our primary interest was to investigate the degree to which the original conclusion of a specific clinical trial would be stable against measurement variability if the data of the given trial were reassessed. We used a framework focusing on early-phase clinical trials that draw conclusions based on the ORR and the percentage of patients who achieve complete or partial response. Complete response was therefore dealt with as part of the definition of ORR rather than by evaluating it separately. This has been clearly addressed in the text.

Methods: Evaluation algorithm
16.* (a) The evaluation algorithm seems a very indirect way to quantify the uncertainty in ORR. Follow-up measurements could be directly simulated for the percentual change (c in text, x in Figure 3!) from which the outcome (Partial Response, Stable disease or Progression) can be directly read. This can be repeated for several simulated dataset to assess uncertainty in ORR. (b) The applied approach where a probability p for partial response is first estimated based on x and then quantifying uncertainty by drawing a yes/no for occurrence of Partial Response from a beta distribution moreover assumes that the model-based estimate for the proportion is not uncertain.
AUTHORS’ RESPONSE:
As we mentioned in the response to your second comment, the tumor burden according to the RECIST guideline is calculated by summing the measurements of the various target lesions that can adequately represent the overall tumor burden of a patient. Thus, measurement error fundamentally occurs when observing lesion sizes. It then manifests as measurement variability of the tumor burden size, percent change, and response categorization. In this study, the uncertainty of ORR (or the progression rate)
derived sequentially from measurement error at the lesion level was quantified. In the simulation process to obtain the probability of a complete or partial response, the probability distribution was considered through an iterative procedure by using randomly extracted values from the posterior distributions of the variance-covariance component parameters. Since the posterior distributions were leptokurtic and the variance of those probabilities obtained from the iterative processes was very small, we determined that using a fixed probability based on the median value would be sufficient to ensure simplicity of the evaluation algorithm. In the Discussion section, we added a paragraph to clarify this.

Methods: Validation
17.* The method is only validated on a real dataset. The actual ORRs that are used as a reference are unlikely to be the true ORRs. A simulation study in which the true ORR is known is needed to properly assess coverage level.

AUTHORS’ RESPONSE:
We performed the suggested simulation study in which the “true ORR” was hypothesized to be known. We compared the 95% CIs for the observed ORRs, which were based on the simulated datasets from the known ORR, and the 95% central ranges of ORRs obtained by reassessments. These simulation results will help understand the patterns of discrepancies between these two different measures. In this revision, we also described the role of a 95% central range of ORRs determined in reassessments through this comparison with the usual 95% confidence intervals. We highlighted that those two measures shed light on different aspects of uncertainty of an observed result obtained from a given clinical trial. We also explored trends in coverage of the observed ORR by the 95% central ranges of ORRs obtained by reassessments.

We have presented this simulation study in the text and the Supplementary Materials. However, it should also be noted that it is difficult to interpret the results of the simulation in terms of the “true ORR” for validation of the evaluation algorithm. Our base model is not relevant to the “true ORR,” but to disagreement between the results of the original assessment and a reassessment. Since this algorithmic tool should be considered as a pragmatic alternative to reassessment exercises by the original reader or another reader, the algorithm validation can only be meaningful when it is performed using a real dataset that contains actual measurements obtained by readers re-reading the tumor size.

18.* On which data was the hierarchical model estimated? On (part) of this dataset or the dataset with 75 patients?
AUTHORS’ RESPONSE:
We developed the hierarchical model using a dataset based on 249 lesions from a clinical trial of 75 patients. We used another external clinical trial dataset based on 56 lesions from 22 patients for the algorithm validation. This has now been clearly presented in the Methods section.

Results
19.* Validation is limited to a single (small) dataset. A simulation study in which the true ORR is known is needed to properly assess coverage level. Impact of heterogeneity in tumour size at baseline and progression rate should be considered as factors possibly affecting reliability of the method.

AUTHORS’ RESPONSE:
As also mentioned in the response to your valuable comment, we performed the suggested simulation study in which the “true ORR” was hypothesized to be known. We explored the trend of coverage of the observed ORR by 95% central range in a series of simulations where the data vary in terms of baseline tumor size and percent change. This exercise should be helpful to understand how variation in the data impacted the behaviors of the 95% central range of ORRs. However, it should also be noted that it is difficult to interpret the results of the simulation in terms of the “true ORR” for validation of the evaluation algorithm.

Our base model is not relevant to the “true ORR,” but to disagreement between the results of the original assessment and a reassessment. In other words, close agreement between an original assessment and its reassessment could be achieved, even if the original assessment itself was inaccurate, depending on the baseline tumor size and percent change. What we essentially tried to highlight through our work was the pattern of such reproducibility, and we tried to present this point more clearly in this revision.