Reviewer’s report

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

Version: 0 Date: 11 Nov 2018

Reviewer: Peter van de Ven

Reviewer's report:

Review of "Development of an algorithm for evaluating the impact of measurement variability on response categorization in oncology trials" submitted to BMC Medical Research Methodology.

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.

More specific comments:

Background
* 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
* 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.
* 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.

Methods: Data
* 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?
* 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?)

* 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.

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

* 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.

Methods: Modeling

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

* 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.

* 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.

* Page 8 (line 2): Now a subject effect is mentioned, but a lesion (within subject) effect is missing.

Methods: Simulation

* Page 9-10: Shouldn't the true percent change c not be applied to μ2 rather than Y?

* 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.

* 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?

Methods: Evaluation algorithm

* 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. 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.

Methods: Validation

* 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.

* On which data was the hierarchical model estimated? On (part) of this dataset or the dataset with 75 patients?

Results
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.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

No

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Not suitable for publication unless extensively edited

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the
4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal