Reviewer’s report

Title: Defining responders to therapies by a statistical modelling approach applied to randomized clinical trial data

Version: 1 Date: 24 Apr 2019

Reviewer: Kit Roes

Reviewer's report:

The paper presents an interesting case study of post-hoc analyses and modelling of randomised clinical trials to identify a subgroup of patients which is likely to have a better response to treatment. The case study presented is in multiple sclerosis, with a drug that ultimately failed to gain EMA approval due to a negative benefit - risk ratio. It uses three trials, ALLEGRO, BRAVO and CONCERTO. The first two are used to build and validate the model, the latter to independently verify. The methodology closely follows Zhao et al. (2013), with an unsupervised model building approach (selection among "all possible" models) based on initial separate (Cox) models per treatment arm. This is a good approach to do in essence an exploratory analysis in such a way that the strongest possible conclusions be drawn. The paper is well written, clear and focused. The approaches described are important and welcome to better define treatment effects (if they are there). The conclusion in the abstract is somewhat overstating what can be concluded on one specific case study, with in the end at best moderate evidence for a differential treatment effect in the CONCERTO trial. The analysis may be relatively easy to do, but needs (biologically, medically) well understood and validated predictors and even then, translation to e.g. decision making on drugs or personalizing treatment are not that easy. Points in my review below will point to that, and a more modest conclusion would be more helpful to stimulate follow-up of similar exercises. The conclusion at the end of the paper is more balanced in that respect.

General comments

The BRAVO and ALLEGRO studies were "negative" on their primary endpoint (which was different from disability progression used for this analysis) and CONCERTO was negative on disability progression as used in this paper, while it was prospectively planned to confirm the disability progression results of BRAVO and ALLEGRO. Conventional wisdom is then that no confirmatory conclusion can be drawn from a (any) subgroup analysis of such a trial (CONCERTO), as the type 1 error associated with the trial is exhausted (see 1). Do the authors consider that their analysis has shown that the treatment is effective in the score defined subgroup (in the CONCERTO study). If so, this has to be stronger motivated (see below). If not, what would be required/sensible to reach that point and how does benefit - risk enter that
equation (i.e. should then not also safety be studied in the same sub-population)? And what would an actual personalised treatment entail? Patients with a score above \(-0.31\) would not receive the treatment, and those below would? Could this decision also depend on the level of side effects? And a step before: this drug was not allowed on the market for the intended indication. Should it be granted approval based on analyses as these, but then for this score defined population? And what would then happen in practice (i.e. real life, where patients and doctors make decisions on treatments)? It would be helpful if authors discuss briefly the intended implications of such analyses, including directions for additional trials needed before treatment reaches patients.

The verification of the models in the CONCERTO trial leads to a treatment effect in the score defined responder group, associated with a p-value of 0.03 for the treatment effect in the subgroup, and of 0.033 for the interaction. However, the overall result of CONCERTO on the same (primary) endpoint was negative. Hence, the type 1 error of the trial as such (declaring a positive result overall or in a pre-defined fixed subgroup) is inflated well above 0.05 (see (2)), and presenting a p-value of 0.03 actually overstates the strength of evidence for a treatment effect in the subgroup. Thus, although the analyses are important and insightful, it is probably not strong enough to claim a differential treatment effect.

The shortlist of five models can hardly be distinguished in their "p-value" or AUC performance (Table 2); differences are not likely beyond random fluctuation. They do differ in composition: the effect of Sex and REL is quite consistent across models, but coefficients (or presence) of the other variables vary strongly (e.g. Age) between models. Would it not be relevant (rather than "picking the best") to include biological/medical knowledge in the final selection step? And how much worse would a sparser model do (e.g., including only Sex, REL and NBV)? Combined with Table 4, other questions come up: In CONCERTO there are no Males among Responders: Did a more classical (unsophisticated) male/female subgroup analysis in this study show a relevant effect? And if the Table 4 results hold (no male responders) in the confirmatory (verification) study, would it still be okay to conclude the treatment works in a subgroup of male patients as well, or would we first need a stronger biological foundation? The paper could thus be strengthened by including both the biological credibility as well as the more classical view in the results, to appreciate the modeling exercise.

More detailed points

The primary endpoint used is disability progression. It is not clear from the paper how drop-out / early termination / missing data were dealt with. I assume these were all classified as non-responder or as progressed (treating them as censored may actually bias the results)?

In the abstract and full paper p-values are reported for the HRs in BRAVO and ALLEGRO, for each subgroup. I do not think these have any useful meaning, as for these studies the subgroups
cannot be treated as fixed, but are maximized on contrast in these studies. If anything is reported, it should reflect the uncertainty of the entire model building process under an H0. Not sure if this is useful, and the p-values presented are likely heavily biased.

For model selection, in the first selection step \( p < 0.05 \) for interaction is chosen as cut-off. In essence this arbitrary (why not smaller or larger?). In the Results section (lines 289 and beyond) this selection is referred to as "significant" interaction, but this is wrong use of the term "significant" - no relevant hypothesis is tested, it is just an arbitrary model selection criterion. I would suggest to present it as such and also motivate the choice for 0.05 as cut off (why does that work well in model building?).

In lines 203/204 it is indicated that model selection is based on lowest p-value and lowest AUC. It may be my shortcoming, but which AUC and in prediction would you not usually aim for a large AUC? Finally, it makes no difference and in Results it is indicated that selection was based on the smallest p-value (alone).


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

Yes

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

Yes

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

Yes

**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:

Acceptable

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 manuscript?

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

Statement on potential review bias
Please complete a statement on potential review bias, considering the following questions:

1. Did you co-author any publication with an author of this manuscript in the last 5 years?

2. Are you currently or recently affiliated at the same institution as an author of this manuscript?

If you can answer no to all of the above, write 'I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria’. If your reply is yes to any, please give details below.

I declare that I did not publish with these authors in the last 5 years and also meet the affiliation criteria
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