Author’s response to reviews

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

Authors:

Francesca Bovis (francesca.bovis@gmail.com)
Luca Carmisciano (lucacarmisciano@gmail.com)
Alessio Signori (alessio.signori@medicina.unige.it)
Matteo Pardini (matteo.pardini@gmail.com)
Joshua Steinerman (jrsteinerman@gmail.com)
Thomas Li (Thomas.Li@tevapharm.com)
Aaron Tansy (Aaron.Tansy@tevapharm.com)
Maria Pia Sormani (mariapia.sormani@unige.it)

Version: 2 Date: 30 Apr 2019

Author’s response to reviews:

Dear Editor,

We thank Dr Barlow and Dr Roes for their useful suggestions that we tried to follow in this revised version of the paper entitled: “Defining responders to therapies by a statistical modelling approach applied to randomized clinical trial data”. We think their comments helped us to improve the manuscript and to correct some mistakes.

We do hope that the paper could be now be considered suitable to be published on BMC Medicine. Here below a point by point response to the Reviewers’ comments, and we have uploaded a revised version of the paper in track changes and a clean version of it.

I look forward to hearing from you.

Sincerely,

Maria Pia Sormani
Reviewer #1: This manuscript uses statistical methods to get a continuous marker of likelihood to benefit from laquinimod versus placebo for multiple sclerosis. A cutpoint is chosen in order to designate who is likely to benefit and who may not. The study takes advantage of three large randomized trials with the same study design and reasonably comparable populations. While the overall effect of treatment was not significant in all three trials there may be subpopulations that do benefit. The manuscript follows a very planned course in deriving candidate markers from one trial, using the second trial to determine the preferred algorithm, and finally applying the dichotomized marker to the third trial. The marker performed well in identifying a population which benefits (HR=0.44) and one which does not (HR=1.24) with a statistically significant interaction.

Overall the manuscript is very clear and provides an outstanding example of how the statistical methods may be used.

There are a few minor issues but there are two more major issues to bring forward:

1. The abstract seemed to suggest that the statistical methods were developed by the authors. Only in the end of the introduction does the method of Zhao et al. get discussed. It was unclear how much of the Zhao paper was used versus what was novel here. It is perfectly acceptable if this is an illustration of a published technique, but it should be clear to the reader.

Authors’ reply. Thanks for pointing this out. We added in the Abstract a clarification about the fact that we applied here a method previously published by others, and we better explained in the Methods what we added to the proposed methodology (Abstract, line 54, page 3 and Methods section, lines 191 to 193, page 9 of the clean version of the revised Manuscript)

2. There were some concerns based on Figure 2 about true separation of the distributions. It appears that the sensitivity would be just over 50% and specificity maybe 70%. One might want a stronger marker than this before deciding on a therapy.

Authors’ reply. The shape of the score distribution should not be interpreted as the standard graph representing the performance of a score to distinguish responder vs non-responders (that are not a priori defined). The main aim of presenting the distribution of the expected treatment effect is just to have a visual inspection of its shape. The shape of the score distribution gives us a qualitative indication on how to split the cohort in subgroups and it supports the evaluation of the AD(q) curve presented in revised Figure 3. We now better explained that in the Methods section (Methods section, lines 212 to 224, page 10 of the clean version of the revised Manuscript).
More minor issues:

3. The Introduction is one long paragraph. It may need smaller paragraphs but with more discussion of the Zhao method.

Authors’ reply. Thanks for this suggestion. We broke the Introduction in smaller sections, adding some more discussion about the Zhao method (Methods section, line 91, page 5 and Methods section, lines 105 to 110, page 6 of the clean version of the revised Manuscript).

4. The three steps of model creation, evaluation, and validation were well explained and clear.

Authors’ reply. We thank the Reviewer for his appreciation.

5. The one place that was less clear was how the cutpoint was chosen. It seemed to be based on qualitative judgment rather a fixed procedure.

Authors’ reply. We better explained in the Methods section how the cutpoint was selected. A visual inspection was used to identify the range of possible cutpoints, and then a systematic procedure was run testing each cutpoint in that range. We choose the cutpoint dividing the cohort in two subgroups with the lowest p value for treatment by subgroup interaction on the training set. This is now better detailed in the Methods (Methods section, lines 225 to 236, pages 10 to 11 of the clean version of the revised Manuscript).

6. AUC's in Table 2 are mostly around 0.35 or so. In my experience most AUC's are reported as greater than 0.50 (chance) with higher numbers. These seem to be the complements.

Authors’ reply. The reviewer is right, this is not the standard area under a ROC curve, but rather an area under the AD(q) curve, that has a different interpretation, as detailed in the Zhao et al paper: the lower the area the higher the discriminant ability of the score. In this context, the area under the AD(q) curve is not used as an absolute assessment of the goodness of the score, but it is rather used to find the optimal cutpoint and compare different models. We explained in higher details the meaning of the AD(q) in the Methods section (Methods section, lines 212 to 224, page 10 of the clean version of the revised Manuscript). Also, we changed the “AUC” definition that can be misleading, substituting it with Area under the AD(q) curve in the text and in the Table 2 (page 23 of the clean version of the revised Manuscript).

7. The AUC of the interaction model for Table 3 would be useful.
Authors’ reply. As previously stated, the AUC is not a standard Area under a ROC curve, but it is an area under the AD(q) curve. This area was already reported in Table 2 and displayed in the revised Figure 3.

8. The hazard rates on page 8 suggest they are computed separately rather than part of a large single interaction model:

$$\log \text{hazard} = \log(h_0) + \delta \text{trt} + \theta \text{z} + \beta (\text{trt} \times \text{z})$$

with $\beta$ being the deviation in treatment effects due to the covariates. This model only estimates one baseline hazard rather than two and then assuming they are the same.

Authors’ reply. The reviewer is right. However, estimating two separate models rather than one with interaction terms has advantages from the computational point of view. A single model with many interaction terms can have convergence problems. Moreover, the assumption of the same baseline hazard is tenable in a randomized context. Both Zhao et al and Li et al in a subsequent technical paper (Biometrics 2015) suggests better properties of the use of two separate models rather than a single one with interactions.

9. Figures 1 and 3 could be moved to a supplement or deleted.

Authors’ reply. We enriched Figure 3 with an explanation of the meaning of the AD(q) curve that was requested also by Reviewer 2 (Figure Legend, lines 460 to 464, page 21 of the clean version of the revised Manuscript). We kept the flow diagram in Figure 1, but we have no problems in removing it if needed.

Overall the manuscript is well written with an important technique illustrated well.

Reviewer #2: 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 overstatement 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.

Authors’ reply: we thank the Reviewer for his overall appreciation and we smoothed a bit the conclusions in the Abstract, as suggested (Abstract, line 77, page 4 of the clean version of the revised Manuscript). Also we expanded the Discussion in this direction (Discussion section, lines 347 to 350, page 15 of the clean version of the revised Manuscript).

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.

Authors’ reply. We thanks the reviewer for these comments, since it is very important to us to make the goals of this analysis clear. We presented this methodology on a drug that was not approved on the market, so the implications of this study are not related to this specific drug. The main aim of this study is to show the potential of the method, especially in the field of Multiple
Sclerosis (MS) and to stimulate similar exercises (with practical implications) on approved drugs. In fact, we have now many drugs available for MS, and one of the major challenges for clinicians is how to choose one among a set of drugs with similar efficacy profiles. In the absence of criteria to identify patients who will respond to each drug (being equal the safety), the standard approach is to choose one drug, closely monitor the patient and switch in case of efficacy failure. The presented methodology could give clinicians additional criteria helping in this initial treatment choice.

We must point out that we choose to apply the method to a non-approved drug mainly because Pharma companies were very reluctant to allow us to apply it using data from clinical trials of approved treatments. They did not like to show subgroups of non-responders to their drugs. We do hope that this paper can stimulate the discussion and encourage data analyses on all the other approved drugs in MS, but also in other fields of medicine where many therapies are available and no specific markers predicting response are known.

We added some comments to the Discussion better explaining these points, as suggested (Conclusion section, lines 356 to 359, page 16 of the clean version of the revised Manuscript).

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.

Authors’ reply. We fully agree with this comment. Presenting the p value for the treatment effect in the subgroup of responders was a mistake and we removed it from the text and from Figure 4 (Results section, lines 305, 306, 308, 316 and 317, page 14 of the track-changes version of the revised Manuscript). Our analysis is focused on the interaction analysis testing whether there is any evidence of a heterogeneous response to the drug according to different levels of the score.

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)?

Authors’ reply. This is a very good point. We deeply discussed this issue when drafting the protocol for this analysis. We then decided to base our choice on the p value for interaction as the criterion to choose the model rather than on biological knowledge for two reasons: first of all,
as previously noted, the focus is on the methodology and not on the specific drug, so we preferred to choose an objective criterion. Second, among the variables included in the analysis, there is no pre-defined factors known to have any specific biological meaning in relation to this specific drug (we were using mainly demographic and standard clinical factors). Another criterion to be used could be the one of simplicity and parsimony, especially if the score has to be used in clinical practice. We added to the Discussion a paragraph discussing these issues to be considered in future similar analyses (Discussion section, lines 347 to 350, page 15 and Conclusion session, lines 356 to 359, page 16 of the clean version of the revised Manuscript).

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.

Authors’ reply. Sex alone was not selected in the first round as a predictor of treatment effect (the p value for sex by treatment interaction was>0.05), but it has a strong role when combined with the other variables. We deeply discussed the biological meaning of the results that came out from this analysis: we do think that it will be interesting to try to understand post-hoc whether the subgroup defined by an unsupervised analysis has any biological meaning. The advantage of this approach is that we can test the characteristics of the responders and non-responders subgroups on other more dedicated studies (maybe including genetic and immunological data) to give a biological interpretation to these findings.

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)?

Authors’ reply. We used the disability progression as defined and estimated in the pivotal studies. Drop outs and early terminations were censored at the time of termination.

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.

Authors’ reply. We fully agree, this was a mistake. We removed all the subgroup-specific p values from the text and from Figure 4 (Results section, lines 305, 306, 308, 316 and 317, page 14 of the track-changes version of the revised Manuscript).

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?).

Authors’ reply. This was a completely arbitrary choice. We added to the Methods section a sentence pointing out that a p value<0.05 was selected as a priori cutoff by an arbitrary choice (Methods section, line 200, page 9 of the clean version of the revised Manuscript).

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?).

Authors’ reply. We fully agree with this comment and we amended the text accordingly (Results section, lines 270 to 273, page 12 of the track-changes version of the revised Manuscript).

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).

Authors’ reply. This was a comment also from the first reviewer. This is not the standard area under a ROC curve, but rather an area under the AD(q) curve, that has a different interpretation, as detailed in the Zhao et al paper: the lower the area the higher the discriminant ability of the score. We explained in higher details the meaning of the AD(q) in the Methods section (Methods section, lines 212 to 224, page 10 of the clean version of the revised Manuscript). Also, we changed the “AUC” definition that can be misleading, substituting it with Area under the AD(q) curve in the text and in the Table 2 (page 23 of the clean version of the revised Manuscript). The model selection was based on p-values alone and we amended the text accordingly (Methods section, lines 208 to 209, page 10 of the clean version of the revised Manuscript).