Author’s response to reviews

Title: The trans-DATA study: Aims and Design of a Translational Breast Cancer Prognostic Marker Identification Study

Authors:

Tim de Ruijter (t.deruijter@maastrichtuniversity.nl)
Kim Smits (kim.smits@maastrichtuniversity.nl)
Maureen Aarts (mjb.essers.aarts@mumc.nl)
Irene van Hellemond (irene.van.hellemond@mumc.nl)
Leander Van Neste (leandervanneste@gmail.com)
Bart de Vries (ba.devries@zuyderland.nl)
Petronella Peer (Petronella.peer@radboudumc.nl)
Jürgen Veeck (kim.smits@mumc.nl)
Manon van Engeland (m.vanengeland@maastrichtuniversity.nl)
Vivianne Tjan-Heijnen (vcg.tjan.heijnen@mumc.nl)

Version: 1 Date: 21 Jun 2019

Author’s response to reviews:

Response to Reviewers:

We thank the reviewers for their helpful comments and suggestions. Below, we have endeavoured to address the comments as fully as possible.

Reviewer 1

1. Question 1A: In the 'marker discovery procedure', roughly, the first step is to identify differentially methylated promoter regions between cases vs. non-cases, and second step is to use cox proportional hazards model approach to identify markers that are associated with DRFI (are the markers evaluated in the cox models restricted to those probes that are differentially methylated?).
Answer: All promoter regions are assessed in the Cox proportional Hazards model, in order to apply the FDR correction for multi-hypothesis testing as strictly as possible. Selected potential Candidate markers are those markers that were both differentially methylated and related to 6-year DRFI, we have added a clarification of this to the manuscript.

Question 1B: are the 60 non-cases included in the cox model analysis, and if so, what role they are playing in the analysis since they are not really contributing (all are free from DR for 6-yrs, ie., no events within 6-yrs)?

Answer: The 60 non-cases were included in the cox model analysis, these patients contribute to the analysis as they completed the full 6-year survival.

2. Question 2A: will any non-cases included at all (if yes, how are those are defined)?

Answer: There were non-cases included in the early and late subset analyses. Each patient with distant metastasis was paired with a non-case patient based on propensity score during the formation of the discovery and validation cohorts. The included non-case patients are those that have been paired to the early and late recurrent case patients. Per your request the number of patient pairs included in these analyses have been added to the paper.

Question 2B: Please provide n for each subset. How will findings from these subset analysis be used?

Answer: The markers identified by the discovery method from all three subsets were all further verified on the validation subset as described in the validation method. We have added a clarification of this to the manuscript.

3. Question 3A: The validation part described in the paper is not really a 'validation', since although it is based on fixed set of DNA methylation markers already identified in the discovery phase, it also involves model selection/building, but now incorporating clinical factors.

Answer: We agree with the reviewer that the validation part of our proposed study is not a full validation, and consider this analysis a first step in assessing the prognostic value of our final model. We have addressed this issue in the Discussion.

Question 3B: Also, this section needs much more elaboration/clarification regarding how exactly the procedure is being planned/done. For example, it is not clear how the 'most powerful markers' are selected, what model the backward elimination procedure is applied to. Also, how Harrell’s c-statistic and AIC will both be used to compare and select models, and the (bootstrap validated) c-statistic be used to evaluate the model performance?
Answer: We have clarified the section on model building in our validation procedure. Harrell’s C-Statistic and AIC will be used by selecting the model with the lowest AIC and the highest c-statistic.

Question 3C: Also, in ‘…the initial prediction model will be internally validated using bootstrapping…’, please clarify which model does the ‘initial prediction model’ refers to?

Answer: We have clarified which model will be internally validated in the text (end of page 11), this refers to the preferred model with the lowest AIC and the highest c-statistic.

4. Question 4: The marker discovery phase includes 60 of the 93 cases, and the validation step includes the rest 33 cases. Will the procedure have sufficient power? Also, since # cases are under-represented in the validation cohort compared to the whole cohort/population, how does this impact the procedure result here? Has weighted analysis been considered?

Answer: We recognise the low number of cases in our validation cohort, and thank the reviewer for the suggestion of a weighted analysis, as this approach may result in better P values. However, we fear that due to the low number of cases, a weighted analysis may overamplify chance based differences between case and non-case patients. We would rather accept larger confidence intervals, that signal further validation is required, than amplify our results using a method that may promote chance findings. We consider this validation analysis a first step in assessing the prognostic value of the selected marker model and therefore accept a larger confidence interval.

5. Question 5: Although the term 'prognostic or predictive' have been used at several places, e.g., the Abstract, the method/procedure described in the paper is only relevant to identifying 'prognostic' markers.

Answer: We agree with the reviewer that our proposed method cannot currently prove predictive value of the identified markers. As addressed in the discussion section, due to the setup of the DATA study in which patient treatment in the two study arms only differed 3 years after randomisation, it is currently not possible to distinguish between prognostic or predictive markers in our study. To avoid misleading information we have therefore removed the mention of predictive markers from the abstract. However, as the relation between the identified markers and the provided therapy can currently not be assessed, we choose to retain the section on possible predictive value in the discussion of the paper.
6. Question 6: Clinical utility of the final model is not discussed. Is the purpose of the model to predict risk of distant recurrence?

Answer: The purpose of this study is indeed to devise a model to identify patients with increased risk of distant recurrence. We have further clarified this purpose in the discussion of our paper.

7. Question 7: p7, last paragraph first sentence, should 'revised' be 'reviewed'?

Answer: "Revised" has been changed to "reviewed" according to the suggestions of the reviewer.

8. Question 8: Figure 1, in the box '93 patients had a 6-year distant disease-free interval event', please change 'distant disease-free interval event' to 'distant recurrence event'.

Answer: The proposed change to the figure was made as requested.

Reviewer 2.

1. Question 1: Please consider matching the case to a control whose follow-up time is at least the same as the case which could be less than six years. This is described as a time-matched case control study.

Answer: We thank the reviewer for the suggestion of time-matched control study, however we have consciously chosen to limit the control patients to patients that completed the full 6 years follow-up without a recurrence. We made this choice because patients that were censored during the follow-up period may have had a distant recurrence event after censoring, effectively misclassifying these patients as non-recurrent. As the differences in methylation we aim to detect is small misclassified samples could lead to loss of potential markers early in the first step of our discovery process would negatively impact the differential methylation analysis which is the first step of our discovery analysis.

2. Question 2: Secondly only 60 controls are used in the discovery cohort, but 661 are used in the validation cohort. It would be more efficient to increase the number of controls in the discovery cohort (e.g. 1 case to 3 controls) as controls are overrepresented in the validation cohort.

Answer: We agree with the reviewer that ideally we would increase the number of non-recurrent cases in our discovery analysis to better match the representation of controls in the discovery
cohort. Unfortunately, this is not possible because of financial reasons. The genome wide technique used for our discovery analysis, Methylation EPIC Bead Chip platform, costs about 400 Euro per sample, our budget does not allow for more analyses to be performed.

3. Question 3: Lastly consider a case-cohort approach instead of case control. The analytic work is more difficult, but the case-cohort design can accommodate different endpoints (such as DFS) and additional outcomes that occur with continued follow-up

Answer: We thank the reviewer for the suggestion of a case cohort approach. The manner of selecting our discovery cohort has been debated extensively in our group. We finally settled on the case control setup as discussed in this paper because this approach allows us to correct for differences in base line characteristics of known prognostic factors via propensity matching. This way, we increase the chance of newly discovered markers of prognostic potential independently from these known prognostic factors. Using a case cohort based approach this would not be possible.