Peer Review File

Article information: https://dx.doi.org/10.21037/atm-21-6027

Reviewer A:

In “Challenges and Opportunities in Biomarker-Driven Trials: Adaptive Randomization”, the author compares different trial arm allocation schemes in clinical trials. The author selected four different covariate-adjusted response adaptive randomisation approaches (RAR) incorporating biomarker information and compared them to the common randomised controlled trial (RCT) design with fixed randomisation ratio and to a RAR approach without adjusting for covariates. The aim is well motivated, but there are some issues to resolve.

Overall, it sounds strange if the author uses “we” in the manuscript. However, I would leave it to the editor whether to keep “we” or replace it adequately, e.g. by avoiding “I” and “we”.

Major issues:

1) The author investigated whether the inclusion of biomarker knowledge could be beneficial in trial conduct. Therefore, the author selected covariate-adjusted RAR (CARA) in a group-sequential design. As comparator, the RAR without covariate adjustment in a group-sequential design and a traditional RCT design with fixed randomisation ratio and without the option of early stopping were chosen. During analysis, the author also compared the number of patients allocated and those who exhibited a treatment failure. As the investigated RAR designs (including CARA) allowed for early stopping but not the selected RCT approach, I recommend including an additional RCT design with a comparable interim monitoring and a comparable option for early stopping.

Reply: Sorry for the confusion. We considered group sequential trials for all designs including Trad, RAR, CARA1 - CARA4 to make fair comparisons. In other words, our results showed the RCT design (Trad) with the same early stopping rule used in RAR and CARA1 - CARA4. We have clarified this.

2) The author performed a simulation study. Is it possible to use additionally real data and to receive a (semi-) synthetic data set (mimicking these trial designs) for method comparison? For example, by sampling from a real data set and relying on variables included in this data set?

Reply: Thank you for the comment. Unfortunately, we don’t have real data for the trials to evaluate the performance of the randomization methods. However, our simulation study supports examples of real trials. We have added a section (i.e., “an example using biomarker-driven randomization”) to provide some information about the application in real trials. Method comparisons were already well addressed in the simulation study. We also suggested practical considerations for biomarker-driven randomization, and our methods and suggestions can be applied to any real data or synthetic data.

3) Please provide the documented source code to reproduce the analysis.

Reply: Thank you for this suggestion. We have added the source code in Appendix.

4) Please separate results and discussion and avoid redundant/repeated issues in the presentation/discussion. Please combine (and condense), e.g., the discussion about type I error.

Reply: Thank you for this suggestion. We have moved “preservation of type I error rate” to “Discussion” to avoid redundant issue in the presentation.

5) Format and definitions

a. In the section “Biomarker-driven randomization”, the introduction of the different approaches for biomarker-driven randomisation would benefit from further information and from the introduction of subsections to guide the reader. Furthermore, this section should contain the description of all trial designs, that are investigated in this manuscript, in general and the specific design settings used in the analyses. Additionally, advantages and limitations/challenges of each design should be mentioned briefly, e.g. as keywords (with corresponding references). Please provide a figure for illustration of the trial designs that also include the essential time points, tasks at these points and notations. The following suggestion of subsections could be a possible structure for this section:

(-) The author could first briefly introduce/name the three types of randomisation discussed in this manuscript (traditional RCT, RAR, CARA).

(subsection i) RCTs with fixed allocation ratios for randomisation (design, randomisation scheme, advantages and limitations/challenges)

(subsection ii) Group-sequential designs with RAR (design, definition of RAR scheme, advantages and limitations/challenges)

(subsection iii) Group sequential designs with CARA (design, definition of CARA scheme, advantages and limitations/challenges)

(subsection iv) CARA1-CARA4 (definitions, motivation for the selection)

(subsection v) Specification of designs (e.g., number of trial arms), that are compared in this manuscript

Reply: To avoid misleading the reader, we have changed the name of the section to “Biomarker-driven randomization in group sequential trials.” We have followed the suggestions to provide further description and information of the approaches. To describe CARA1 - CARA4 with the formula, the specification of designs should be mentioned earlier than the formulas.

b. Similarly, in the section about the simulation study, including additional subsections would facilitate reading. Furthermore, please provide more information on the simulation study approach (see below). Please provide a figure illustrating the simulation approach. Furthermore, is a graphical summary for lines 123-151 possible? Suggestion for the section structure and the content of each subsection:

(subsection i) general information (e.g., number of repetitions, sample size, sample size calculation)

(subsection ii) simulation of the population (e.g., biomarker values, covariates, outcome, treatment)

(subsection iii) trial design (e.g., allocation schemes (including the information on what happens, if one arm is already closed, i.e. comprises the targeted number of patients), early stopping rules, motivation for the selected time points for interim analyses, analysis methods for final analyses (including aspects of

biomarker/randomisation incorporation in the analysis))
(subsection iv) analysis methods of simulation study (e.g., performance measures definition and calculation)
(subsection v) results

Reply: Thank you for the suggestions. We have revised the section for simulation study and results. One clarification is that we allow early stopping of the trial due to the superiority or futility of treatment A against B, but we do not drop one of arms during the trial. This is for a comparative two-arm randomized clinical trial. O'Brien-Fleming alpha spending function provides the stopping boundaries for sequential tests (at interims and at final analysis), and a chi-square test was performed. We have provided Figure 2 describing the graphical summary of scenarios 1-20, but the best way to illustrate the simulation approach is to provide the mathematical equations and the explanations.

6) Abstract:
a. Please include more details on methods and results.

Reply: Thank you for this comment. We have added more sentences on methods and results in the abstract (It has 224 words).

7) Introduction:
a. Application and investigation of these designs increased over the last years and probably still increases. Is there additional, more recent literature (lines 42, 45, 48 and 53)?
b. Please clarify “operating characteristics” (line 57). Which characteristics were assessed?

Reply: We have added more recent literature and clarified what “operating characteristics” mean.

8) Biomarker driven randomization:
a. Please add references for the reasons RCTs are criticised for (line 64) and provide more details on these ethical issues. Are references available for the reason clinical investigators would like to avoid fixed randomization schemes? Additionally, please clarify “the”. “This”?
b. RAR cannot completely resolve the issues related to randomisation mentioned by the author (line 67). Please rephrase. Furthermore, RAR favours the treatment that performs best based on the data available prior to randomisation. Please clarify in “the best performing treatment” (line 69) as this is a quite general statement. Additionally, please name the utilities that are maximised by RAR application (line 70).
c. Please include few words (keywords) about some already known limitations of RAR and points to consider in the application of RAR. For example, time points for updates, time-dependent outcome, need of a short-term outcome probably as proxy for a long-term outcome, available biomarkers, inclusion of dependent/interacting biomarkers (with regard to response to therapy) or biomarkers with an impact on the outcome itself; e.g.
i. McShane et al. L, Hunsberger S. 5. In: Matsui S, Buyse M, Simon R, editors. An Overview of Phase II Clinical Trial Design with Biomarkers. Chapman & Hall / CRC Biostatistics Series. Boca Raton, USA: CRC Press; 2015. p. 71–87
ii. Talisa et al. (2018, https://doi.org/10.3389/fimmu.2018.01502), Kesselmeier and Scherag (2018, https://doi.org/10.3389/fimmu.2018.02507) and related references

iii. Strzebonska and Waligora, 2019, https://doi.org/10.1186/s12910-019-0395-5
iv. Janiaud et al., 2019, https://doi.org/10.1016/j.ctrv.2018.12.003
v. Kesselmeier et al., 2020, https://doi.org/10.1371/journal.pone.0237441
d. Is there additional, more recent literature (line 74)?
e. Please provide an explanation of the probabilities and formulae in words (lines 93-111).
f. Please clarify, whether the biomarker values are included in x.

Reply: Thank you for the suggestions. We have revised the words as suggested and have added references. Note that we focused on the short-term endpoint which is a binary response, and the long-term endpoint is not an interest in this manuscript. We believe that considering the long-term primary endpoint and using the surrogate endpoint is a very interesting topic to be a sole paper.

9) Simulation study and results
a. Please clarify whether the biomarkers are independent. If they are independent, please discuss in the discussion possible implications of dependent biomarkers, especially if their interaction and consequently the biomarker distribution (and the individual status of a patient, i.e. positive for one or for multiple biomarkers) in the trial also have an impact on the outcome. Furthermore, please discuss the impact of the number of integrated biomarkers on the final trial result.
b. Please clarify the type of covariates. Furthermore, which influence has the type of covariate (binary, categorical, ordinal or numeric)?
c. Performance measures:
i. Please clearly define all measures to assess the design performance, e.g. overall type I error rate or the statistical power.
ii. Please provide the estimated treatment effects (as boxplot with indicated true effect) of the final analyses, as a shift in the estimated effect might also induce a change in type I error rate or statistical power.
iii. If several investigated biomarkers has an impact of the outcome, the differing biomarker status distributions (combination of the biomarker positive indicator as tuple) in the trial arms might also have an impact on the estimated treatment effect. Hence, please provide the biomarker status distributions.
iv. Please provide the number of screened and included participants as boxplot.
v. How often was a trial stopped early? Which impact had it on the statistical power and type I error rate?
d. Covariates and definition of the response probability:
vi. Please clarify whether the biomarker are part of x.
vii. Currently, the response probability is only defined based on two covariates. Please check.
e. Lines 220-245: Please link the complete results.

Reply: Thank you for the useful suggestions. For (a) and (b), we have clarified that two binary biomarkers are independent. We have provided the discussions of dependent biomarkers, the number of biomarkers, and the type of covariates in the Discussion section. For (c), we have clearly defined all measures we used to assess the design performance. We have added Figures 3-4 for (ii) and (iii) and have updated Figure 5 including the early stopping probability. Specifically, for (c4), there are no screened or additionally included participants in the trial. All N subjects are evaluated and used for testing. For (d), Yes, you are right. We have clarified the biomarker profiles x with two biomarkers x1 and x2. For (e), we have added tables showing the results when the prevalence rate of x1 is either 0.7 or 0.25 (which links the previous results in Line 220-245).

10) Discussion:
a. Please discuss further limitations/challenges of the investigated designs (see also comments above).
b. Please discuss the impact of varying update time point, trial duration (related to, e.g., medical or environmental changes) and quality of the selected biomarker.
c. How can these results be translated to similar designs and to the planning/analysis of a clinical trial? Please add.
d. Please include a section on strengths and limitations of the current investigation.

Reply: For (a), the objective of the paper is to find the challenges and opportunities in biomarker-driven trials, and we focused on biomarker-driven randomization. We found the problems of the type I error rate inflation using the biomarkers for adaptive randomization in group sequential trials through the simulation study. We also discussed the limitations and challenges of biomarker-driven randomization (i.e., we have moved them to the Discussion section as suggested) and provided useful suggestions and considerations. We didn't intend to propose the design. However, it is interesting to mention briefly as suggested in (b) for readers who are interested in implementing it under the group sequential trials. We have discussed them in the Discussion section. For (c), we have added more practical guidelines for the cutoff calibration and stratified test in the discussion. For (d), we have created a "conclusion" section to briefly mention the strengths and limitations of our investigation.

11) Appendix A:
a. Line 405: Please provide i.
b. Please add some information about the application in a real trial.
c. Line 516: What about CARA2?

Reply: Thank you for your comments. For (a), we have provided "i". For (b), as noted above, we have created a section "an example using biomarker-driven randomization" to provide some information about the real applications. For (c), we mentioned that CARA2 preserved the type I error rate when the informative prior is used. Therefore, we mainly investigated the approaches to preserve the type I error rate under CARA1, CARA3, and CARA4 designs. We have clarified this.

Minor issues:

1) Please introduce all abbreviations on first use, e.g. CARA is missing.

Response: Thank you for catching this. We have added "covariate-adjusted RAR (CARA)" in the Introduction.

2) Please check for "CARA 1" instead of "CARA1", e.g. line 263.

Response: We have fixed it.

3) Biomarker driven randomization:
a. Lines 61-62: Please add "randomly" to "assigned and, e.g., "allocation" to "ratio".

b. Lines 78-80:
i. Please check for missing comma and “and”.
ii. Suggestion: (CARA1) … (CARA4) instead of (1) … (4) and corresponding rephrasing.
c. Line 80: Please include the reference for CARA4.
d. Line 93: “or” should be “and”.
e. Line 97: One “the” must probably be deleted, i.e. “… define the following, biomarker-driven …”.
f. Lines 108-111: Please provide the dimension of x (e.g. in lines 94/95) to assess the transpose. Please check for all vectors and matrices.

Response: We have modified our texts as advised.

4) Simulation study and results
a. Line 115: “… sample size of 210”. “of” is missing.
b. Line 126 (and others): Please clarify “no any effect”.
c. Please provide a reference for the O’Brien-Fleming alpha spending function (line 152).
d. Please provide the R version as well as the package version.
e. Line 187: Please clarify “rounding issue”.
f. Lines 271/272: Please rephrase. “…, (1) Except” does not fit.
g. Please provide result description in the main text and not in the supplement.

Response: We have modified our texts as advised. For (e), we have updated the decimal digits in Tables to represent the probability, and we have deleted some sentences saying “…was a rounding issue to report...” For (f), the sentences we described in “preservation of type I error rate” saying “When we looked at the results from using informative priors, (1) except the ... with an average of 0.10” are redundant, because we mentioned type I error rate inflation in “simulation study and results”. Thus, we have deleted the sentences.

5) Tables
a. Please always introduce all parameters and abbreviations.
b. Please always provide the identical number of digits, e.g. in Table 1 always one and for probabilities (results) always three.
c. Please clearly link the tables (in the description/title) to the analysis underlying the respective table.
d. Table 1: Please clarify “all x_1”, “x_1=1” and so on.
e. Table 2:
i. Suggestion: “estimated rejection probability”.
ii. Suggestion: separate columns for the three kinds of priors below CARA1-CARA4; similarly for tables 5 and 6 for the multiplicity correction (or check for space in front of the round bracket, e.g “0.705 (0.565)” instead of “0.705(0.565)”)
iii. Please add “informative prior” and “uninformative prior”, respectively, to the diagonal elements.
iv. Please add the information which scenario are the null and which the alternative scenarios, i.e. type I error rate versus statistical power.

Response: We have modified our tables as advised.

6) Figure 1

a. Please clarify “difference”. Might the proportions be easier to assess for the reader?
b. For better comparability, please provide equal y-axis scaling per row.
c. Please provide per panel the model (null or alternative) as title.
d. Left, lower panel: Please consider jittered point, so that the results can be completely assessed.

Response: For (a), it describes the difference of the number of patients between A and B. We have clarified this. For (b) and (c), we have followed the suggestions. For (d), we have changed the style of the lines. Note that the actual values we made the plot are very close to each other except scenarios 13-15 (See the below tables).

Designs	# of failures								
Trad	104.5	92.2	69.0	54.7	164.7	104.6	105.3	153.3	130.6	95.4
RAR	104.6	92.2	68.9	54.5	164.6	104.6	104.8	153.4	130.1	95.1
CARA1	104.6	92.2	68.7	54.6	164.0	104.3	104.5	145.1	113.1	69.5
CARA2	104.7	92.5	68.8	54.7	165.0	104.8	104.7	153.4	130.5	95.2
CARA3	104.8	92.7	68.6	54.3	164.2	104.1	104.8	153.1	129.9	95.0
CARA4	104.6	92.0	68.4	54.2	164.5	104.4	104.6	153.5	130.5	95.1

7) Appendix B:
a. Line 430: Please clarify “but ethic”. Missing verb?

Response: We have moved appendices to the main results as you suggested. Sentences in the previous Appendix B were updated.