Author’s response to reviews

Title: Adaptive propensity score procedure improves matching in prospective observational trials

Authors:

Dorothea Weber (weber@imbi.uni-heidelberg.de)
Lorenz Uhlmann (uhlmann@imbi.uni-heidelberg.de)
Silvia Schönenberger (silvia.schoenenberger@med.uni-heidelberg.de)
Meinhard Kieser (meinhard.kieser@imbi.uni-heidelberg.de)

Version: 1 Date: 07 May 2019

Author’s response to reviews:

Response to comments made by the Editor and the Reviewers:

We would like to thank the Editor and the two reviewers very much for their careful and detailed reviews as well as the helpful comments which gave us the possibility to make considerable improvements to the manuscript. We thank the reviewers in the Acknowledgement. In the following we explain point by point how we addressed the comments in our careful revision of the manuscript providing section, page and line information of changes. To facilitate the re-reviewing process, we marked changes in the manuscript in red.

Editor:

- Please include the email addresses of all authors on the title page

- Please include an "Abbreviations" section before the declaration section

- Please include a “Declarations” heading

- Please rename “Ethics approval” to "Ethics approval and consent to participate"

- Please include a "Consent for publication" section, if not applicable then include it anyways and state "Not applicable"

- Please include a “Funding” section

- Please remove funding info in ACK
- Rename "Figures" to "Figure Legends"

- Please remove duplicate copy of MS (Reference PDF.pdf)

Thank you for the editorial points, we changed the manuscript accordingly.

Kit Cb Roes (Reviewer 1):

This is a well written paper on a comprehensive idea to improve matching (and thus power) in a (admittedly) specific design. It is well presented, and the adaptive CI approach can be implemented in practice without too much difficulty. It is based on an actual case study, which also inspired/directed the modeling and simulation model.

Before going to more general comments, I provide my understanding of the modeling underlying the simulation study - where to me it seems there is a discrepancy between text and tables. If I understand the modeling correctly from Table 1, baseline variables X1 and X3 determine the propensity for group membership Z, and X4 (together with Z) determines the outcome. I do think the Table 1 differs from the text (lines 40-50, p3, second column): there it is suggested that the outcome is modeled (for the simulation) with logistic regression based on X5. Subsequently, it is a strong point that the variables that are included in the model to estimate the propensity score (X2, X3 and X5?) differ partially from the simulation (avoiding over-optimism). But it would help to clarify this point of text versus Table 1 (or my possible mis-reading). For the remainder I am assuming that simulations are correct and in line with Table 1.

Thank you for this very useful comment. We corrected the text to be in line with the table (and the simulations). We also added a clarification of the propensity score model and its misspecification according to the (here known) true model. (Methods section, line 59ff, page 4/page 5)

The key point is to improve matching by taking random fluctuation at interim analysis of the matching rate conservatively into account. To account for this fluctuation by a (essentially 1:1) sampling mechanism is an elegant idea. I agree with implementing the conservatism with an approach (CI) that depends on sample size. The choice for the 99% level for the CI is not clearly motivated (why not 95%?), and I do think that is needed. The choice of 99% CI leads to an increase in power, but in my view at the expense of a non-trivial increase in mean recruited sample size (order of magnitude of 1/3 for the new group). I would expect more reflection on this sample size increase, as well as on results of alternative choices for the confidence level.

Thank you very much for this comment. We changed the formulation in the Methods section to a more generalized definition of the lower limit of the confidence interval for the matching rate which is used for recalculation. The changed definition includes a variable confidence level (Methods section, line 59/60, page 2 second column). Additionally, we added the results of the mean matching rate of the naïve method and the mean lower CI limits at interim and final analysis for the matching rate using the 99%, 95%, 90%-CI limits (Table 2), as well as the mean
total recruited patients in the treated group in Table 3. The results are described and discussed in
the revised manuscript. (Results section, line 23-35, page 5, first column; Discussion and
Conclusion section, line 32-34, page 6, first column)

The simulation study is geared towards the practical example, but a broader range of sample
sizes is explored. These are driven by effect size achieving 80% power and expressed in Beta4,
outcome. However, these Betas I could not connect to the model stated in Table 1 (which I
assumed was based on a sample size of about 150 and the outcome model has a parameter value
of 0.2 for X4).

Thanks for this very important comment. The notation was misleading. The Beta4,outcome is
renamed in BetaZ,outcome to clarify that it is the regression coefficient for the group variable in
the outcome model. The regression coefficient for sample size 150 can now be connected to the
model stated in Table 1. (Methods section, line 14-17, page 4, second column)

Secondly, I would expect more reflection on the level of generalization that can be done based
on the results (of essentially only one model). What are consequences of more complex models,
and higher levels of misspecification of the propensity score model? I do expect some
robustness, but further substantiation will increase the practical value.

Thanks for this useful comment. We simplified the simulations by restricting to a single model
which is clearly a limitation of the generalizability of the results. Nevertheless, we assume that
model complexity should not strongly influence the performance of the method if convergence
of the propensity score model is guaranteed. If the level of misspecification in the propensity
score model increases, we would expect a decrease in performance (lower matching rate and
therefore lower power), but this would be the case for both methods. We added a paragraph to
discuss the points in the manuscript. (Discussion and Conclusion section, line 41-50, page 6, first
column)

Reviewer 2 (Reviewer 2):

GENERAL COMMENTS: The authors proposed a new method to compute the total required
sample size after the interim analysis. The idea is great. But, this approach seems to be a "Ad
Hoc" approach by using the lower 99% confidence limit as the weight, instead of the mean rate.
The motivation example is great. I like that very much.

The simulation studies assume the independence between Xs, which may not be the case in
practice. I would expect them to be correlated in the simulation studies.

Thanks for this comment. It is correct, we assume Xs to be uncorrelated and simulated them
independently. For sure, in practice correlations between baseline variables occur, but one should
evaluate those correlations and take them into account for the decision on inclusion in the
propensity score model. In general, propensity score matching behaves like regression models
and therefore highly correlated variables should not both be used as matching variables. We
added a comment on correlations in the Methods section: (Methods section, line 51-55, page 3, second column)

REQUESTED REVISIONS:

The conclusion of overestimation of the matching rate (page 2, line 28) was based on the simulation studies, which may not be true in other simulation studies, unless they could theoretically prove it.

Thanks for this comment. We are aware that the conclusion that there may occur an overestimation of the matching rate is not theoretically proven. We changed the sentence now stating that there might be practical situations in which the matching rate is overestimated. (Introduction section, line 33/34, page 2, first column)

The lower 99% limit was used. Reviewers may have the concern of the underestimation of matching rate if the 99% limit is used. How about 95% 90% limits?

Thank you very much for this useful comment. We changed the formulation in the Methods section to a more generalized definition of the lower limit of the confidence interval for the matching rate which is used for recalculation. The changed definition includes a variable confidence level (Methods section, line 59/60, page 2 second column). Additionally, we added the results of the mean matching rate of the naïve method and the mean lower CI limits at interim and final analysis for the matching rate using the 99%, 95%, 90%-CI limits (Table 2), as well as the mean total recruited patients in the treated group in Table 3. The results are described and discussed in the revised manuscript. (Results section, line 23-35, page 5, first column; Discussion and Conclusion section, line 32-34, page 6, first column)

ADDITIONAL REQUESTS/SUGGESTIONS:

page 3 line 27, the authors do not explain why n_treat.interim patients are sampled from the control, not n_treat+5?

We added an explanation in the Methods section. The idea is to calculate the matching rate on equally sized groups at interim analysis. The treated patients are “newly” recruited and at interim analysis only n_treat.interim patients are already recruited for the treated group. This number (n_treat.interim) is fixed at the beginning by the timepoint of interim analysis. (Methods section, line 49-52, page 2, second column)

Page 3 line 36 second column, before “By using a logistic regression”, I would suggest adding the logit model from page 2 here.

Thanks for this useful comment. It will help the reader to understand which model we address in this sentence. To avoid duplicate formulas, we referenced Table 1 which includes the logit model equation we address. (Methods section, line 47/48, page 3, second column)
page 4 line 62, you may provide the detailed functions from the R packages?

Thank you for this comment; we added the used functions in brackets in this sentence. (Methods section, line 26/27, page 4, second column)