Dear Dr. Urbanek,

We would like to thank you and the two reviewers for the great reviews that enormously help to improve our work. We are particularly thankful for the thorough investigation of our package. In the revised version of our manuscript, we address all the points that were raised by the reviewers. In the following, we reply point-by-point to the reviewers' comments in blue color.

Review 1

Page 1 : The rationale for using blinded sample size recalculation could be spelled out more clearly. For example, authors suggest that when there is high uncertainty about a "true" effect, blinded sample size recalculation can be used. Are the authors suggesting that apriori power analysis should not be used, but instead, blinded sample size recalculation can used to re-estimate the variance during an ongoing RCT and modify the initially planned sample size if necessary?

Thank you for this important comment. While in principle blinded sample size recalculation could be done without any a priori sample size calculation, it is still advisable to calculate an initial sample size based on the best guess for the nuisance parameter available in the planning phase and to determine when to recalculate the sample size based on this initial calculation. This is done to avoid conducting the recalculation too early (so that there is still a great uncertainty about the magnitude of the nuisance parameter when recalculation is performed), or too late (so that there may be no room for adjusting the sample size any longer as the recalculated sample size is already exceeded). We clarified this aspect in the Introduction.

In the first paragraph they also discuss this package as a tool for planning clinical trials, but this technique is generally used once an RCT is already underway, which means it was approved by the funder (e.g., NIH) long before (e.g., months or even years). Additionally, these funding agencies require apriori power analysis before PIs can start recruiting, randomizing, and enrolling participants into arms of an RCT. Are the authors suggesting this can be used as a planning tool before the trial starts or in addition during the internal pilot design?

If there is high uncertainty on nuisance parameters in the planning stage of a clinical trial, blinded sample size recalculation can be pre-planned after a certain fraction of the fixed design's sample size. From our own experience, this is how blinded sample size recalculation is commonly used. The methodology was also primarily developed with this use case in mind. Our package facilitates the analysis of different scenarios that may appear during blinded sample size recalculation and thereby enables the trial statistician to choose an appropriate sample size recalculation strategy.

In addition, our package could also be used when a blinded sample size recalculation is conducted midcourse without having been pre-planned. We clarified this in the Introduction.

Page 1: The authors write that this is the first R package on CRAN. But is that other R packages that are not on CRAN? Are there other software options outside of R? If so, what are the benefits of this R package?

Thank you for this important comment. We were choosing the expression “on CRAN” to demonstrate that the software is publicly available on a frequently used platform. To avoid any misunderstandings, we extended the statement accordingly.

I think it may be helpful to provide some guidance on how to think about values of the nuisance parameter in this context. Authors show that multiple values can be used in the n_fix function, but is there some concern about data dredging by allowing using to enter as many values as they wish?

Allowing the user to pass several nuisance parameters to the n_fix function is to allow investigating how different assumptions result in different (fixed) sample sizes, which is a standard practice in planning clinical trials. Please note that the users could also use the function several times (or use ‘sapply’) if there wasn’t the option to insert as many values as desired.

Data dredging is not an issue in blinded sample size recalculation, since in most scenarios the type 1 error rate is unaffected by the blinded recalculation and therefore control of the nominal significance level also doesn’t depend on the sample size for the internal pilot study.

Eq. 2 seems to reflect a two-tailed test, whereas the preceding narrative text refers to a one-sided test. Is this correct?

Unfortunately, we do not understand this comment. The text as well as the formulated hypotheses describe a one-sided test, as it is natural for non-inferiority trials.

Review 2

Major Points

1. The authors note that they have 100% code coverage with their unit tests, and this fact is to be commended. The nature of their unit tests is not made clear in the present manuscript, however. The online repository shows that the authors did indeed include many types of test cases – relative and absolute value expectations, error triggering, argument specifications, and such. This is a strength, and I believe the authors should include it in the manuscript (even if just as a footnote).

Thanks for the comment and the thorough review of our unit tests! We added a more detailed description of the nature of our unit tests in the manuscript.

2. Testing of the package’s functionality and internal consistency was cleverly and efficiently accomplished; however, in my view, even more crucial are test cases which validate the

mathematical/statistical accuracy of a tool (especially simulation-based ones), for example, by comparing final output and/or internally-created objects against those obtained via other methods (e.g., closed-form solutions, other tools). I found some test cases that validate power/sample size values against external criteria; however, unless I'm mistaken, they all seem to be for the same limited range (around 0.8 power). Test cases should cover a range of values. A few more tests per analytic method (i.e., chi-squared, Farrington-Manning, t-test, and shifted t-test) would suffice (e.g., for expected power around 0, 0.3, 0.5, or some other reasonable points, and the corresponding sample sizes).

This is indeed a good point to further increase the package's quality. Unfortunately, we are not aware of any publications about blinded sample size recalculation that examine power values below 0.8, because those are usually not of relevance for clinical trials. However, we decided to use those power values for consistency checks, i.e., we investigated whether smaller power values imply that smaller sample sizes are needed. Furthermore, we used smaller power values and tested if they can be met by our blinded recalculation procedure and added some tests that compare the power values for the fixed design with the results of the validated sample size software PASS. We hope that these extensions sufficiently fulfill your request.

Minor Points

1. The authors mention extending the functionality of the tool as a future direction, and they provide the example of implementing further endpoints. (I assume that by this they mean accommodating other distributions/types of outcomes, such as non-binary discrete variables; however, this was also not totally clear to me.) It may be helpful also to list additional potential extensions in the manuscript more explicitly (e.g., analytic methods/tests).

Thank you for this comment. In the revised version of the manuscript, we explain this aspect in more detail and give reference-based examples for possible extensions.

2. The word "endpoints" may not be clear to all readers. Briefly defining this term and/or linking it to terminology that may be more familiar across research areas would be helpful.

We replaced the word "endpoints" by "outcomes" throughout the manuscript.

3. The code appears clean and well-organized; however, it is minimally commented. While there is documentation on the function/class-level (e.g., usage examples, argument and method descriptions), and the code at its present length is fairly easy to follow, the use of in-line comments would contribute to the authors' stated goals of transparency and community participation in further developing the tool.

Thank you for mentioning this. We added some in-line comments where it seemed appropriate.

4. It is possible the code runs as quickly as is currently feasible; however, it did take me nearly 14 minutes to run the code in the manuscript (system: 128 GB RAM; 64 CPUs, AMD Ryzen Threadripper 2990WX; Ubuntu 18.04.5). I suspect the authors have already profiled their code (given their use of Rcpp/C++ to improve speed), but if they haven't, it might be worth checking for any inefficiencies that could be ironed out.

Thank you for mentioning this important point. Indeed, when developing the functions for the binary outcomes our main concern was speed, which is why we used Rcpp to calculate the exact rejection probabilities. However, the computations in the recalculation design require a four-fold for-loop, which takes quite some time for larger sample sizes, even using C++. This is unfortunately the price to pay for computing exact rejection probabilities.

Extra Notes:

- The authors provide an effective, concise description of the steps for using the tool (with the corresponding code) in the manuscript. Including these examples in the README would ensure an extremely quick start-up process for users. I see that the authors have an open issue regarding the creation of a vignette, which I believe could also be helpful.

Thank you for pointing this out. Indeed, we are planning to use a reduced version of the R Journal paper as vignette and want to implement this as soon as we can refer to the R Journal paper.

Following your suggestion, we already extended the README file to improve the usability of our package.

- The authors may consider including some of the equations provided in the manuscript in the R documentation for functions as well.

We followed your suggestion to further explain the methodology in the package documentation.

- The aesthetic quality of the figures produced by the tool could be improved; however, as I imagine plots will typically be used for researcher convenience rather than published, and recognizing the potential benefits of minimizing dependencies on external packages, I don't necessarily see this as a sticking point.

We tried to reduce dependencies whenever possible. Producing publication-ready figures would have required the inclusion of ggplot2 or some other graphics package, which seemed excessive for a relatively small benefit, since as you mentioned we expect that the figures will mainly be used for planning purposes. However, since all functions are vectorized it is very easy for users to save the output of the function and to create their own figures with external packages when desired.

- Two very minor typographical errors: Page 1, paragraph 4, line 1 seems to have an extra “how” (“[m]ethods how to reassess”), and Page 2 under the first equation, “boundary” is missing a y.

The typos are fixed now, thank you.