blindrecalc - An R Package for Blinded Sample Size Recalculation

by Lukas Baumann, Maximilian Pilz, and Meinhard Kieser

Abstract Besides the type 1 and type 2 error rate and the clinically relevant effect size, the sample size of a clinical trial depends on so-called nuisance parameters for which the concrete values are usually unknown when a clinical trial is planned. When the uncertainty about the magnitude of these parameters is high, an internal pilot study design with a blinded sample size recalculation can be used to achieve the target power even when the initially assumed value for the nuisance parameter is wrong. In this paper, we present the R-package **blindrecalc** that helps with planning a clinical trial with such a design by computing the operating characteristics and the distribution of the total sample size under different true values of the nuisance parameter. We implemented methods for continuous and binary outcomes in the superiority and the non-inferiority setting.

1 Introduction

Determining the sample size that is necessary to achieve a certain target power is a fundamental step in the planning phase of every clinical trial. The sample size depends on the type 1 error rate α, the type 2 error rate β, the effect size Δ and so-called nuisance parameters, which are parameters that affect the distribution of the test statistic but are not of interest in the test problem. While the type 1 and type 2 error rates are usually predetermined and the minimal clinically important effect size is known, there is often uncertainty about the magnitude of the nuisance parameters, such as the variance of the data σ^2 for tests with continuous outcomes or the overall response rate p for tests with binary outcomes.

Consider, as an example, the meta analysis by Nakata et al. (2018) that compares minimally invasive preservation with splenectomy during distal pancreatectomy. Among others, the overall morbidity of the two groups is compared and data from 13 studies are reported (cf. Figure 2(c) in Nakata et al. (2018)). Within these 13 studies, overall morbidity rates pooled over both groups between 0.10 and 0.62 are reported. This illustrates the high uncertainty about the "true" overall morbidity rate, which is the nuisance parameter in this setting.

In these cases, an internal pilot study design with blinded sample size recalculation can be used. In such a design, the nuisance parameter is estimated in a blinded way (i.e., without using information about the group assignment of the patients) after a certain number of outcome data is available, and the sample size is recalculated using this information (Wittes and Brittain, 1990). 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). Using this method to recalculate the sample size is an attractive option because the cost in terms of additional sample size is very small (depending on the outcome) and in most scenarios the type 1 error rate is unaffected by the blinded sample size recalculation. Hence, whenever there is uncertainty about the value of a nuisance parameter and the logistics of the trial allow it, blinded sample size recalculation can be used. Meanwhile, this is even recommended by regulatory authorities. For instance, the Committee for Medical Products for Human Use (CHMP) (2006) states that "(w)henever possible, methods for blinded sample size reassessment (...) should be used".

Methods to reassess the sample size in a blinded manner in an internal pilot study design have been developed for a variety of outcomes. Based on the early work by Stein (1945), Wittes and Brittain (1990) introduced the internal pilot study design for continuous outcomes. Their work was extended in different manners by different authors (cf. among others Birkett and Day (1994), Denne and Jennison (1999), and Kieser and Friede (2000)). In all these papers, the main task is to re-estimate the variance of a continuous outcome in a blinded way. These ideas can be applied to binary outcomes as well where the re-estimated nuisance parameter is the overall response rate over both treatment arms. Associated methods were, for instance, presented by Gould (1992) and Friede and Kieser (2004) for superiority trials and by Friede et al. (2007) for non-inferiority trials.

However, despite the clear benefit of a blinded sample size recalculation and a great number of publications on that topic, **blindrecalc** is to the knowledge of the authors the first R-package on

CRAN, and thus a freely available software, that helps with the planning of a clinical trial with such a design by computing the operating characteristics and the distribution of the total sample size of the study. The package can be used for pre-planned and midcourse implemented blinded sample size reassessments in order to evaluate the potential scenarios the blinded sample size re-estimation may imply. For continuous outcomes, we implemented the t-test for superiority trials and the shifted t-test for non-inferiority trials. For binary outcomes, we implemented the chi-squared test for superiority trials and the Farrington-Manning test for non-inferiority trials.

The structure of the paper is as follows: In the Statistical methods section, we explain the general way of proceeding when conducting a trial with an internal pilot study and how to obtain a blinded estimate of the nuisance parameter for continuous and binary outcomes. The structure of the package is introduced in Package structure. We demonstrate how **blindrecalc** can be utilized to plan a trial with an internal pilot study design and blinded sample size recalculation in Usage and example. In Development principles, we outline the principles of the development process and how we ensure the quality of our code. Finally, a brief Conclusion complements this paper.

2 Statistical methods

The general procedure for planning and conducting a trial with a blinded sample size recalculation is as follows: At first, an initial sample size n_{init} is calculated by using a best guess for the value of the nuisance parameter. The sample size for the first stage of the trial, n_1, is then calculated as a fraction of n_{init}, e.g., 0.25, 0.5 or 0.75. After n_1 observations are available, the total sample size n_{rec} is recalculated in a blinded way based on the available data. The final total sample size n is then determined as:

$$n = \min\left(\max\left(n_1, n_{rec}\right), n_{max}\right),$$

where n_{max} is a prespecified maximal sample size. This is called the unrestricted design with upper boundary. A restricted design would use n_{init} as a lower boundary (Wittes and Brittain, 1990). Often n_{max} is set to a multiple of n_{init}. The special case of $n_{max} = \infty$ results in the unrestricted design (Birkett and Day, 1994). After the final total sample size is calculated, the $n_2 = n - n_1$ observations for the second stage are gathered. Finally, the specified statistical test can be conducted with the data of all n patients.

In the following, we shortly introduce the implemented tests and how to obtain a blinded estimate of the nuisance parameter in each case.

Continuous outcomes

Assume a clinical two-arm trial with normally distributed outcomes where a higher value is deemed to be favorable, with mean values μ_E (experimental group) and μ_C (control group) and common unknown variance σ^2. The outcome of interest is the mean difference $\Delta := \mu_E - \mu_C$. By introducing a non-inferiority margin $\delta > 0$, the test problem is given by

$$H_0 : \Delta \leq -\delta \text{ vs. } H_1 : \Delta > -\delta.$$

The null hypotheses can be tested by a shifted t-test taking the non-inferiority margin δ into account. Note that the special case $\delta = 0$ corresponds to the standard t-test for superiority. The approximate total sample size for a one-sided t-test to detect a mean difference of $\Delta = \Delta^* > -\delta$ with a power of $1 - \beta$ while controlling the type 1 error rate at level α equals

$$n = n_E + n_C = \frac{(1+r)^2}{r} \frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2 \sigma^2}{\left(\Delta^* + \delta\right)^2}.$$

Here, r refers to the allocation ratio of the sample sizes between the experimental and the control group, i.e., $r = n_E/n_C$, and z_{1-q} denotes the $1-q$ quantile of the standard normal distribution.

In this framework, the nuisance parameter is the unknown variance σ^2. Due to potential uncertainty on the value of σ^2, it seems appropriate to re-estimate it in a blinded interim analysis to ensure that the desired power level is met. Inserting $\widehat{\sigma}^2$ observed mid-course into the above sample size formula may lead to a more reasonable sample size for the respective trial than sticking to the value assumed in the planning stage. There exist different methods for estimating σ^2 in a blinded manner

(Zucker et al., 1999). In **blindrecalc**, the one-sample variance estimator is implemented. It is defined as

$$\widehat{\sigma}^2 := \frac{1}{n_1 - 1} \sum_{j \in \{E,C\}} \sum_{k=1}^{n_{1,j}} \left(x_{j,k} - \bar{x} \right)^2,$$

where $x_{j,k}$ is the outcome of patient k in group j, $n_{1,j}$ denotes the first-stage sample size in group j, i.e., $n_1 = n_{1,E} + n_{1,C}$, and $\bar{x}$ equals the mean over all n_1 observations. Since the patient's group allocation is not considered when computing $\widehat{\sigma}^2$, blinding is maintained when using this variance estimator.

In the superiority case, i.e., if $\delta = 0$, blinded sample size reassessment can be performed without relevant type 1 error rate inflation (Kieser and Friede, 2003). In the non-inferiority case, however, the type 1 error rate may be inflated by the internal pilot study design and a correction of the applied significance level may become necessary to protect the nominal type 1 error rate at level α (Friede and Kieser, 2003). In particular, this inflation arises if the sample size recalculation is performed too early, i.e., if n_1 is chosen too small.

Interestingly, the cost of this procedure in terms of sample size is quite low. Since the one-sample variance estimate slightly overestimates the variance, an increase in sample size arises. However, this increase amounts to only 8 patients with $\alpha = 0.025$ and $\beta = 0.2$ or 12 patients with the same significance level and $\beta = 0.1$ (Friede and Kieser, 2001). In return, the sample size recalculation procedure implies that the trial's power meets the target value $1 - \beta$ for a wide range of values of σ^2.

Lu (2016) gives closed formulas for the exact distribution of the test statistic of the two-sample t-test in this setting. This allows the simulation of error probabilities and of the sample size distribution in an acceptable amount of time. The proposals made by Lu (2016) are implemented in **blindrecalc**. Thus, the design characteristics for continuous outcomes presented in **blindrecalc** are obtained by simulation and not by exact computation. This is the case for for binary outcomes that are presented in the following.

Binary outcomes

In a superiority trial with binary outcomes where a higher response probability is assumed to be favorable, the one-sided null and alternative hypothesis are

$$H_0 : p_E \leq p_C \text{ vs. } H_1 : p_E > p_C,$$

where p_E and p_C denote the event probabilities in the experimental and the control group, respectively. While several tests exist for this test problem, the widely used chi-squared test is implemented in **blindrecalc**. The sample size for this test can be approximated with the formula (Kieser, 2020):

$$n = \frac{1+r}{r} \frac{\left(z_{1-\alpha/2} \sqrt{(1+r) \cdot p_0 \cdot (1-p_0)} + z_{1-\beta} \sqrt{r \cdot p_{C,A} \cdot (1-p_{C,A}) + p_{E,A} \cdot (1-p_{E,A})} \right)^2}{\Delta^2}.$$

Again, r denotes to the allocation ratio of the sample sizes, and $z_{1-\alpha/2}$ and $z_{1-\beta}$ are the $1-\alpha/2$ and the $1-\beta$ quantiles of the standard normal distribution. Furthermore, $p_{C,A}$ and $p_{E,A}$ are the response probabilities in the control and the experimental group under the assumed alternative, p_0 is the overall response probability, i.e., $p_0 = (p_{C,A} + r \cdot p_{E,A}) / (1+r)$, and Δ is the effect under the alternative, i.e. $\Delta = p_{E,A} - p_{C,A}$. The nuisance parameter here is p_0, which can be estimated in a blinded way after n_1 observations with

$$\hat{p}_0 = \frac{X_{1,E} + X_{1,C}}{n_{1,E} + n_{1,C}},$$

where $X_{1,E}$ and $X_{2,C}$ denote the number of observed events in the experimental and the control group and $n_{1,E}$ and $n_{1,C}$ represent the first-stage sample sizes in the two groups. Blinded estimates of the event rates in each group can then be obtained by $\hat{p}_{C,A} = \hat{p}_0 - \Delta \cdot r / (1+r)$ and $\hat{p}_{E,A} = \hat{p}_0 + \Delta / (1+r)$. These estimates are used to recalculate the sample size. The benefit of the blinded recalculation is that the desired power can be maintained, even if the initially assumed value for p_0 was wrong.

It is well known that the chi-squared test in a fixed design does not maintain the nominal significance level, hence the same can be expected for a chi-squared test with a blinded sample size recalculation. In fact, Friede and Kieser (2004) showed that the actual levels of the test with and without recalculating the sample size are very close.

In a non-inferiority trial, the null and alternative hypothesis are

$$H_0 : p_E - p_C \leq -\delta, H_1 : p_E - p_C > -\delta,$$

where $\delta > 0$ is the fixed non-inferiority margin. The most commonly used test for this problem was

proposed by Farrington and Manning (1990). An approximate sample size formula for this test is

$$n = \frac{1+r}{r} \cdot \frac{\left(z_{1-\alpha/2}\sqrt{r \cdot \tilde{p}_C \cdot (1-\tilde{p}_C) + \tilde{p}_E \cdot (1-\tilde{p}_E)} + z_{1-\beta}\sqrt{r \cdot p_{C,A} \cdot (1-p_{C,A}) + p_{E,A} \cdot (1-p_{E,A})}\right)^2}{(\Delta+\delta)^2},$$

where $\tilde{p}_E$ and $\tilde{p}_C$ are large sample approximations of the restricted maximum likelihood estimators under the null hypothesis restriction $p_E - p_c = -\delta$ (see the Appendix of Farrington and Manning (1990) for the computation). The same formulas as for the chi-squared test can be used to estimate $\hat{p}_E$ and $\hat{p}_C$ in a blinded way, and these estimates have to be used to obtain blinded estimates $\hat{\tilde{p}}_E$ and $\hat{\tilde{p}}_C$ for the restricted maximum likelihood estimates. Plugging these estimates into the sample size formula gives the re-estimated sample size.

Like the chi-squared test, the Farrington-Manning test is also no exact test and can exceed the nominal significance level. Friede et al. (2007) showed that in general no further inflation of the type 1 error rate is caused by blinded re-estimation of the sample size. Nevertheless, it is possible for the chi-squared test as well as for the Farrington-Manning test to choose the nominal significance level smaller than α in order to protect the type 1 error rate at level α. Such an adjustment of α is implemented in **blindrecalc** for the binary and the continuous case.

3 Package structure

When a clinical trial with an internal pilot study is planned, it is essential to know the characteristics of the applied design. To this end, the performance in terms of achieved power levels, type I error rates, and sample size distribution has to be known for different values of the nuisance parameter and the first-stage sample size n_1. The package **blindrecalc** provides all necessary tools that are needed to plan a trial with a blinded sample size recalculation with only a small number of functions, which makes using the package very accessible.

blindrecalc makes use of R's S4 class system. This allows the application of the same methods for different design classes and facilitates the usage of the package. Furthermore, this approach makes the package easily extendable without any changes in the current source code.

The usage of **blindrecalc** is intended to be as intuitive as possible. To obtain characteristics of a blinded sample size recalculation procedure, two steps have to be made. At first, the user has to define a design object to indicate which test and which characteristics such as the desired type 1 and type 2 error rates are to be applied. To this end, the three functions `setupChiSquare`, `setupFarringtonManning`, and `setupStudent` exist to define a design object of the class corresponding to the respective test.

Secondly, the trial characteristic of interest can be calculated. Currently, the following methods are implemented: The method `toer` allows the computation of the actual type 1 error rate for different values of the nuisance parameter and the sample size of the internal pilot study. By means of `adjusted_alpha`, the adjusted significance level can be calculated that can be applied as nominal significance level when strict type 1 error rate control is desired. The method `pow` computes the achieved power of the design under a given set of nuisance parameters or internal pilot sample sizes. With `n_fix`, the sample size of the corresponding fixed design can be computed. Finally, the method `n_dist` provides plots and summaries of the distribution of the sample size. For all these methods (except for `n_dist`), the logical parameter `recalculation` allows to define whether a fixed design or a design with blinded sample size recalculation is analyzed.

4 Usage and example

For each test, there is a `setup` function (e.g., `setupChiSquare` for the chi-squared test) that creates an object of the class of the test. Each `setup` function takes the same arguments:

- `alpha`: The one-sided type 1 error rate.
- `beta`: The type 2 error rate.
- `r`: The allocation ratio between experimental and control group, with a default of 1.
- `delta`: The difference in effect size between alternative and null hypothesis.
- `alternative`: Whether the alternative hypothesis contains greater (default) or smaller values than the null.
- `n_max`: The maximal total sample size, with a default value of `Inf`.

In this example, the nuisance parameter is the overall response rate p. A difference in response rates between the two treatment groups of $\Delta = p_E - p_C = 0.2$ is to be detected. Using **blindrecalc**, a

chi-squared test that achieves a power of $1 - \beta = 0.8$ to detect this effect of $\Delta = 0.2$ and that uses a nominal type 1 error rate of $\alpha = 0.025$ can be set up by

```
design <- setupChiSquare(alpha = 0.025, beta = 0.2, delta = 0.2)
```

The sample size for a fixed design given one or multiple values of the nuisance parameter (argument `nuisance`) can then be calculated with the function `n_fix`:

```
n_fix(design, nuisance = c(0.2, 0.3, 0.4, 0.5))
#> [1] 124 164 186 194
```

The function `toer` calculates the actual level of a design with blinded sample size recalculation or of a fixed design (logical argument `recalculation`) given either one or more values of the total sample size in a fixed or the sample size for the first stage in a recalculation design (argument `n1`) or one or more values of the nuisance parameter. Note that all functions are only vectorized in one of the two arguments `n1` and `nuisance`. In this example, it is assumed that the internal pilot study contains half of the fixed sample size that would be needed if the overall response rate p equals 0.2. In this setting, **blindrecalc** can be used to compare the actual levels of a fixed design and a recalculation design with the same parameters.

```
n <- n_fix(design, nuisance = 0.2)
p <- seq(0.1, 0.9, by = 0.01)
toer_fix <- toer(design, n1 = n, nuisance = p, recalculation = FALSE)
toer_ips <- toer(design, n1 = n/2, nuisance = p, recalculation = TRUE)
```

In Figure 1, the type 1 error rate in dependence of the nuisance parameter is depicted for the designs with and without sample size recalculation. Note that, as mentioned in Section Binary outcomes, the level of significance exceeds the pre-defined level of $\alpha = 0.025$ in both cases. If strict control of the type 1 error rate is desired, the function `adjusted_alpha` can be used to to calculate an adjusted significance level, such that the nominal significance level is preserved.

```
adj_sig <- adjusted_alpha(design, n1 = n/2, nuisance = p, precision = 0.0001,
                          recalculation = TRUE)
design@alpha <- adj_sig
toer_adj <- toer(design, n1 = n/2, nuisance = p, recalculation = TRUE)
```

In this example, the adjusted significance level equals 0.0232 for the trial with internal pilot study, i.e., using this value as nominal level ensures that the actual significance level does not exceed $\alpha = 0.025$. Figure 1 demonstrates that the type 1 error rate is protected at level $\alpha = 0.025$ if the adjusted significance level is applied.

In the setting of binary outcomes, adjusting the level such that the nominal type 1 error rate is protected for any realization of the nuisance parameter in its domain $[0, 1]$ is feasible. However, when the nuisance parameter has an infinite domain, such as the variance in the case of continuous outcomes, this is not possible. The solution in these cases is to compute a $(1 - \gamma)$ confidence interval of the nuisance parameter in the blinded interim analysis and adjust the significance level such that the actual level is below $\alpha - \gamma$ for all values in this confidence interval (Friede and Kieser, 2011). If this approach is applied, the user can set the parameter `gamma`.

To calculate the power of either the internal pilot study design or the fixed design, the function `pow` can be used. Again, the function is vectorized in either `n1` or `nuisance`. This function can be used to compare the power values of the two designs under different actual values of the nuisance parameter.

```
pow_fix <- pow(design, n1 = n, nuisance = p, recalculation = FALSE)
pow_ips <- pow(design, n1 = n/2, nuisance = p, recalculation = TRUE)
```

As we can see in Figure 2, the power achieved by the internal pilot study design is very close to the target power of 0.8 in most cases. Only when the overall response rate is very close to 0 or 1, the power is exceeded. On the other hand, the fixed design is much more sensitive to the actual value of the nuisance parameter and the actual power can either be way too large or way too small if the sample size was calculated under wrong assumptions.

Finally, the distribution of the total sample size can be computed under different assumptions on the nuisance parameter with the function `n_dist`. This is particularly useful for the planning of internal pilot study designs since it allows the investigation of what could happen in a certain clinical trial and helps the applicant to prepare for different scenarios.

```
p <- seq(0.2, 0.8, by = 0.1)
n_dist(design, n1 = n/2, nuisance = p, plot = TRUE)
```

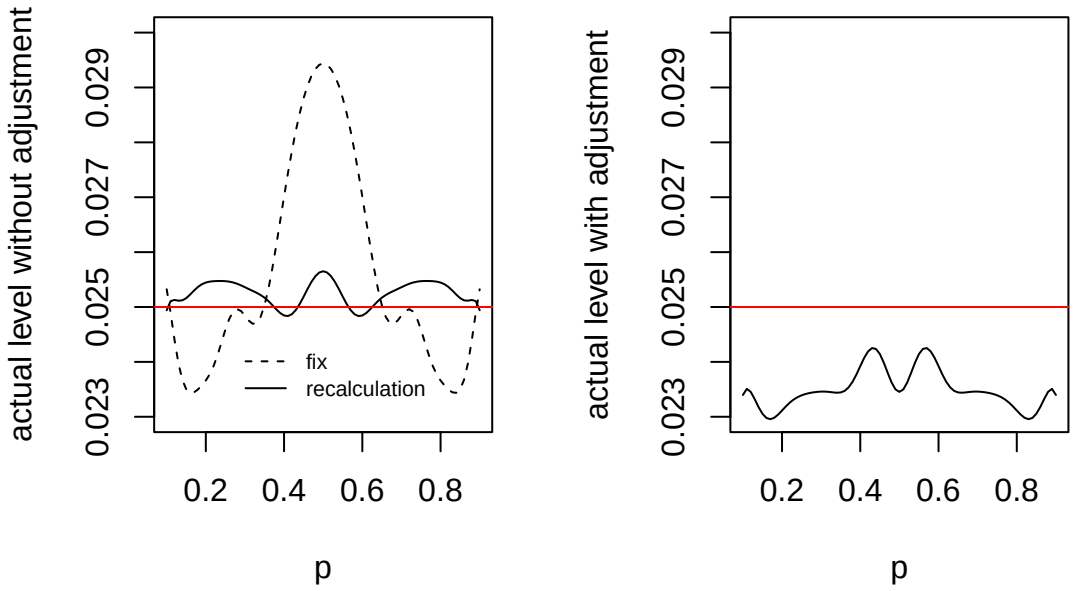

Figure 1: Actual significance level for different nuisance parameters with (right panel) and without (left panel) adjustment of the nominal significance level. In the adjusted case, the actual level is below the desired level (red line) for all nuisance parameters in contrast to the un-adjusted case.

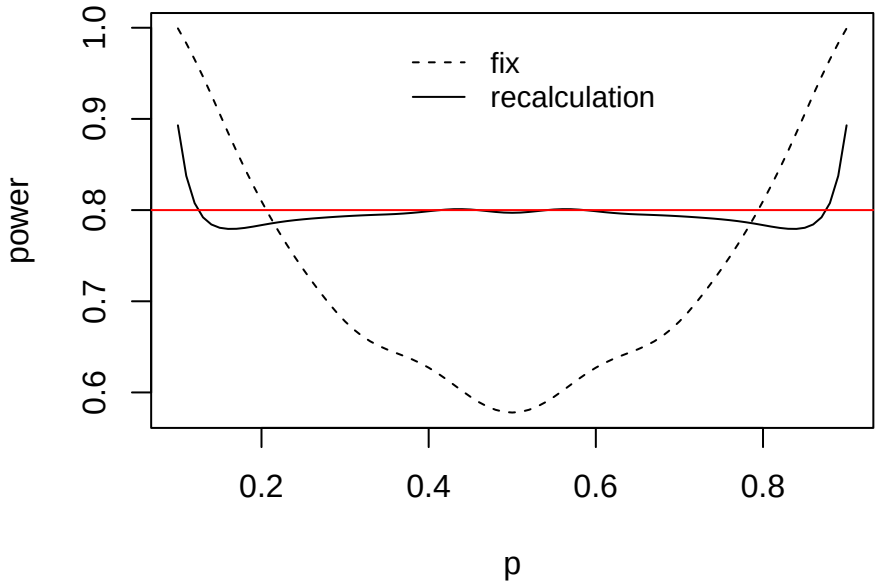

Figure 2: Power values for different nuisance parameters for a fixed design and a design with blinded sample size recalculation. The design with recalculation meets the target power (red line) for a wider range of nuisance parameters.

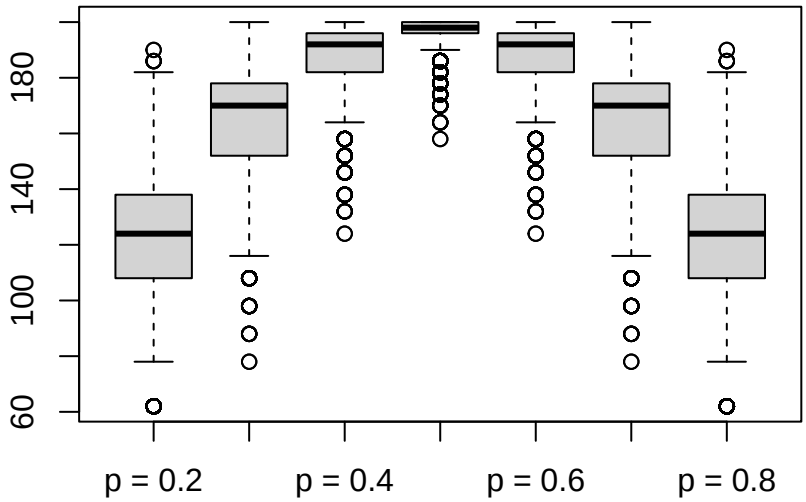

Figure 3: Distribution of the sample size for a design with blinded sample size recalculation in dependence of the nuisance parameter. For more extreme values of the nuisance parameter the variance of the sample size distribution becomes larger.

```
#>         p = 0.2  p = 0.3  p = 0.4  p = 0.5  p = 0.6  p = 0.7 p = 0.8
#> Min.     62.000  78.0000 124.0000 158.0000 124.0000  78.0000  62.000
#> 1st Qu. 108.000 152.0000 182.0000 196.0000 182.0000 152.0000 108.000
#> Median  124.000 170.0000 192.0000 198.0000 192.0000 170.0000 124.000
#> Mean    125.021 164.6057 188.4801 196.5075 188.4801 164.6057 125.021
#> 3rd Qu. 138.000 178.0000 196.0000 200.0000 196.0000 178.0000 138.000
#> Max.    190.000 200.0000 200.0000 200.0000 200.0000 200.0000 190.000
```

By default, `n_dist` prints a summary of the sample size distribution for each nuisance parameter. With `plot = TRUE`, a series of boxplots is drawn (cf. Figure 3). Since the maximum sample size is obtained if the overall response rate is estimated to be 0.5 in the sample size recalculation, this maximum can occur under any true value of the nuisance parameter (except for 0 and 1), albeit with very small probability. For this reason, sample sizes that occur with a probability of less than 0.01% are ignored. This is not the case for continuous outcomes since there, the sample size distributions are determined by simulation.

For continuous outcomes, i.e., the (shifted) *t*-test, the functional content of **blindrecalc** is the same as in the binary case that was presented in this example. The only difference is that in the continuous case, the numbers are computed by simulation. Thus, the user can set the parameters `iters`, defining the number of simulation iterations, and `seed`, the random seed for the simulation.

5 Development principles

The utilization of R's object-oriented programming capabilities implies that the example that was presented for the chi-squared test could very similarly be applied to the Farrington-Manning test or the *t*-test. Besides using S4 classes, the following development principles of **blindrecalc** should be briefly described.

All calculations for binary outcomes are exact and require nested for-loops. Since for-loops are known to be very slow in R, all computation-intensive functions for the chi-squared test and the Farrington-Manning test are implemented in C++ via the **Rcpp** package (Eddelbuettel and François, 2011) to speed up the calculations significantly.

blindrecalc is developed open-source on GitHub.com. The entire source code can be found at `https://github.com/imbi-heidelberg/blindrecalc`. This allows anyone to contribute to **blindrecalc** and, furthermore, provides maximal transparency. To ensure a certain quality of the provided code, **blindrecalc** is checked by unit tests using the package **covr** (Hester, 2020). The unit tests compare numbers for the sample size, type 1 error rate, and power calculated with **blindrecalc** with numbers from peer-reviewed publications and, furthermore, check the technical functionality of the package such as vectorization and display of error messages. Thus, the unit tests do not only monitor the technical accuracy of the package's results but also their content-related correctness. The current version **blindrecalc** 0.1.3 achieves a code coverage of 100%, i.e., each line of the source code is checked by at least one unit test.

6 Conclusion

In this paper, we introduced the R-package **blindrecalc** that can be used to plan clinical trials with a blinded sample size recalculation in an internal pilot study design when either continuous or binary outcomes in a superiority or non-inferiority setting are of interest. We introduced the basic methodology of internal pilot studies and explained how the package can be used to calculate the operating characteristics of a trial with such a design.

The scope of **blindrecalc** can simply be extended due to its modular character. Blinded sample size recalculation can be applied to many different types of clinical trials. For instance, there exists research on further kinds of outcomes (e.g., see Friede and Schmidli (2010) for count data) or on different study designs (e.g., see Golkowski et al. (2014) for bioequivalence trials). The implementation of internal pilot studies for such cases in **blindrecalc** is an exciting area of future work.

Bibliography

M. A. Birkett and S. J. Day. Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13 (23-24):2455–2463, 1994. URL https://doi.org/10.1002/sim.4780132309. [p142, 143]

Committee for Medical Products for Human Use (CHMP). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-methodological-issues-confirmatory-clinical-trials-planned-adaptive-design_en.pdf, 2006. Accessed: March 10, 2022. [p142]

J. S. Denne and C. Jennison. Estimating the sample size for a t-test using an internal pilot. *Statistics in Medicine*, 18(13):1575–1585, 1999. URL https://doi.org/10.1002/(SICI)1097-0258(19990715)18:13<1575::AID-SIM153>3.0.CO;2-Z. [p142]

D. Eddelbuettel and R. François. Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, 40(8):1–18, 2011. URL https://doi.org/10.18637/jss.v040.i08. [p148]

C. P. Farrington and G. Manning. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9(12):1447–1454, 1990. URL https://doi.org/10.1002/sim.4780091208. [p145]

T. Friede and M. Kieser. A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, 20(24):3861–3873, 2001. URL https://doi.org/10.1002/sim.972. [p144]

T. Friede and M. Kieser. Blinded sample size reassessment in non-inferiority and equivalence trials. *Statistics in Medicine*, 22(6):995–1007, 2003. URL https://doi.org/10.1002/sim.1456. [p144]

T. Friede and M. Kieser. Sample size recalculation for binary data in internal pilot study designs. *Pharmaceutical Statistics*, 3(4):269–279, 2004. URL https://doi.org/10.1002/pst.140. [p142, 144]

T. Friede and M. Kieser. Sample size reassessment in non-inferiority trials. *Methods of Information in Medicine*, 50(3):237–243, 2011. URL https://doi.org/10.3414/ME09-01-0063. [p146]

T. Friede and H. Schmidli. Blinded sample size reestimation with count data: methods and applications in multiple sclerosis. *Statistics in Medicine*, 29(10):1145–1156, 2010. URL https://doi.org/10.1002/sim.3861. [p149]

T. Friede, C. Mitchell, and G. Müller-Velten. Blinded sample size reestimation in non-inferiority trials with binary endpoints. *Biometrical Journal*, 49(6):903–916, 2007. URL https://doi.org/10.1002/bimj.200610373. [p142, 145]

D. Golkowski, T. Friede, and M. Kieser. Blinded sample size re-estimation in crossover bioequivalence trials. *Pharmaceutical Statistics*, 13(3):157–162, 2014. URL https://doi.org/10.1002/pst.1617. [p149]

A. L. Gould. Interim analyses for monitoring clinical trials that do not materially affect the type I error rate. *Statistics in Medicine*, 11(1):55–66, 1992. URL https://doi.org/10.1002/sim.4780110107. [p142]

J. Hester. *covr: Test Coverage for Packages*, 2020. URL https://CRAN.R-project.org/package=covr. R package version 3.5.1. [p148]

M. Kieser. *Methods and Applications of Sample Size Calculation and Recalculation in Clinical Trials*. Springer Series in Pharmaceutical Statistics. Springer Verlag, 2020. URL https://doi.org/10.1007/978-3-030-49528-2. [p144]

M. Kieser and T. Friede. Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19(7):901–911, 2000. URL https://doi.org/10.1002/(SICI)1097-0258(20000415)19:7<901::AID-SIM405>3.0.CO;2-L. [p142]

M. Kieser and T. Friede. Simple procedures for blinded sample size adjustment that do not affect the type I error rate. *Statistics in Medicine*, 22(23):3571–3581, 2003. URL https://doi.org/10.1002/sim.1585. [p144]

K. Lu. Distribution of the two-sample t-test statistic following blinded sample size re-estimation. *Pharmaceutical Statistics*, 15(3):208–215, 2016. URL https://doi.org/10.1002/pst.1737. [p144]

K. Nakata, S. Shikata, T. Ohtsuka, T. Ukai, Y. Miyasaka, Y. Mori, V. V. D. M. Velasquez, Y. Gotoh, D. Ban, Y. Nakamura, Y. Nagakawa, M. Tanabe, Y. Sahara, K. Takaori, G. Honda, T. Misawa, M. Kawai, H. Yamaue, T. Morikawa, T. Kuroki, Y. Mou, W.-J. Lee, S. V. Shrikhande, C. N. Tang, C. Conrad, H.-S. Han, P. Chinnusamy, H. J. Asbun, D. A. Kooby, G. Wakabayashi, T. Takada, M. Yamamoto, and M. Nakamura. Minimally invasive preservation versus splenectomy during distal pancreatectomy: a systematic review and meta-analysis. *Journal of Hepato-Biliary-Pancreatic Sciences*, 25(11):476–488, 2018. URL https://doi.org/10.1002/jhbp.569. [p142]

C. Stein. A two-sample test for a linear hypothesis whose power is independent of the variance. *Annals of Mathematical Statistics*, 16(3):243–258, 1945. URL https://doi.org/10.1214/aoms/1177731088. [p142]

J. Wittes and E. Brittain. The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72, 1990. URL https://doi.org/10.1002/sim.4780090113. [p142, 143]

D. M. Zucker, J. T. Wittes, O. Schabenberger, and E. Brittain. Internal pilot studies II: comparison of various procedures. *Statistics in Medicine*, 18(24):3493–3509, 1999. URL https://doi.org/10.1002/(SICI)1097-0258(19991230)18:24<3493::AID-SIM302>3.0.CO;2-2. [p144]

7 Contribution

The first two authors contributed equally to this manuscript.

Lukas Baumann
Institute of Medical Biometry
University of Heidelberg
Im Neuenheimer Feld 130.3
69120 Heidelberg
Germany
ORCiD 0000-0001-7931-7470
baumann@imbi.uni-heidelberg.de

Maximilian Pilz
Institute of Medical Biometry
University of Heidelberg
Im Neuenheimer Feld 130.3
69120 Heidelberg
Germany
ORCiD 0000-0002-9685-1613
pilz@imbi.uni-heidelberg.de

Meinhard Kieser
Institute of Medical Biometry
University of Heidelberg
Im Neuenheimer Feld 130.3
69120 Heidelberg
Germany
ORCiD 0000-0003-2402-4333
kieser@imbi.uni-heidelberg.de