Sample Size Planning for Nonparametric Non-Inferiority Analyses in the 'Gold Standard' Design

20 weeks Master Thesis
In partial fulfillment of the requirements for the degree
Master of Science (M.Sc.) in Applied Statistics
at the University of Göttingen

Maxi Schulz

Supervisors:
Prof. Dr. Tim Friede
Dr. Thomas Asendorf

University of Göttingen
Department of Medical Statistics
Submitted on June 1, 2023

Acknowledgements

First of all, I would like to express my gratitude to Prof. Dr. Tim Friede for his guidance and input in the process of writing this thesis. I would also like to thank Dr. Tobias Mütze for his role in fostering constructive discussions and providing valuable feedback, as well as Prof. Dr. Frank Konietschke for his insights. A note of thanks also goes to Dr. Thomas Asendorf, who serves as the second supervisor for this thesis. Lastly, I would like to acknowledge my study colleagues and friends who have been a constant source of mental support throughout the process of writing this thesis.

Abstract

This study investigates sample size planning methods for nonparametric analyses of three-arm non-inferiority trials using the retention-of-effect approach. Through extensive simulation studies, a methodology for sample size estimation using the studentized permutation test, as proposed by Mütze et al. (2017), is developed and considerations are provided for the nonparametric approach using classical mid-ranks by Munzel (2009).

The proposed sample size planning method for analysing with the studentized permutation test methodology incorporates the use of the parametric sample size formula by Hasler et al. (2008), which effectively ensures the desired power level of the studentized permutation test, even when data deviates from normality. This method offers a convenient approach where only the expectation and variance parameters need to be specified in advance.

To validate the assumptions on the nuisance parameters as specified in the planning stage, a sample size re-estimation procedure based on data from an internal pilot study is proposed. Two variance estimators, the unblinded group-variance estimator and the blinded adjusted one-sample variance estimator, are suggested, with the former demonstrating broader applicability. Additionally, an inflation factor is introduced to enhance the reliability of achieving the target power level in scenarios with small pilot study sizes and non-normal data. However, its effectiveness is limited as it may result in overestimated sample sizes or fail to guarantee the desired power level, particularly in non-normal data settings.

Regarding the nonparametric approach using classical mid-ranks by Munzel (2009), this study provides insights into the type I error and empirical power of the test when assuming parametric distributions for the data. However, due to the finding that assuming parametric distributions does not directly translate into relative effects, this study presents considerations for sample size planning when analysing with the Munzel test without providing a definitive result.

Contents

List of Figures

List of Tables

1 Introduction

In drug development, new treatments often present potential advantages such as reduced toxicity, improved administration, or lower cost compared to standard treatments. In such cases, the primary objective is often not to establish the superiority of the new drug over the control but rather to demonstrate that the new treatment maintains the well-established effectiveness of the active control. In other words, the goal is to establish that the new treatment is not clinically worse than the active control by more than an irrelevant amount. These types of clinical trials are commonly known as non-inferiority trials. The efficacy of a new treatment is proven by demonstrating it as non-inferior to the standard reference treatment which has been demonstrated to be efficacious in previous trials.

While a two-arm trial consisting of the experimental and reference arm is typically used for this purpose, it has two major drawbacks. Firstly, the choice of the non-inferiority margin must be justified through previous studies. Secondly, two-arm non-inferiority trials lack a direct comparison with placebo. As a result, assay sensitivity, which refers to the "ability to distinguish an effective treatment from a less effective or ineffective treatment" (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2001, p. 11), may not be guaranteed since the effectiveness of the active control cannot be directly demonstrated within the trial. Hence, including a placebo arm in non-inferiority trials becomes recommendable whenever ethically justifiable (European Medicines Agency, 2005). This design is referred to as 'gold standard' design and has gained popularity across various areas of clinical research (see for instance Daniels et al., 2009; Pratley et al., 2019; Vanden Bossche and Vanderstraeten, 2015).

In a three-arm gold standard design, the non-inferiority hypothesis can be formulated in two ways: either by defining the non-inferiority margin as an absolute margin or by using the retention-of-effect approach, which considers the relative difference between the reference and placebo group. Over the past two decades, the retention-of-effect approach has been extensively studied in the analysis of three-arm non-inferiority trials for various clinical endpoints. However, when assumptions about the underlying distributional properties of the clinical endpoint become untenable, it is reasonable to consider nonparametric approaches for testing non-inferiority instead.

To our knowledge, Munzel (2009) introduced the first nonparametric approach for three-arm non-inferiority trials. This approach is based on Kruskal-Wallis-type functionals, commonly known as relative effects. These relative effects quantify the influence of each treatment by measuring the deviation of its distribution relative to where probability mass is concentrated in the experiment. This characteristic makes the test suitable for analysing both continuous and ordinal data, without requiring any parametric assumptions about the endpoint. The estimation of these relative effects leads to mid-ranks,

which is why the test is referred to as nonparametric test based on classical mid-ranks. In 2017, Mütze et al. developed a second nonparametric test based on a studentized permutation test using a Wald-type test statistic. This method proved particularly useful for small sample sizes. Unlike the Munzel test, however, the studentized permutation test compares expected values under the hypothesis rather than relative effects and remains mean-based in its approach. Consequently, this test requires the data to be metric. However, since it derives the rejection area of the test through permutations, it does not rely on parametric assumptions and is nonparametric in nature. Other nonparametric approaches have been proposed as well. S. Ghosh et al. (2017) introduced a testing strategy that relies on transformations of the data, however, the method still requires the data to be continuous. More recently, Li et al. (2023) investigated a nonparametric approach for non-inferiority trials that accounts for non-ignorable missing data.

A critical aspect of designing a clinical trial involves the determination of the sample size. The goal in planning the number of subjects is to ensure a population sufficiently large to yield reliable answers to the research questions posed (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 1998, p. 16). Yet, it is equally important to avoid excessive sample sizes that might unnecessarily subject participants to interventions and result in a waste of resources. Therefore, determining the sample size requires a sound justification, with the assumed treatment effect being one key aspect of this determination (for more details refer to International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 1998, p. 17). Failing to adequately specify parameters during sample size planning can potentially result in underestimated or overestimated sample sizes. To validate the assumptions made in the planning stage, trial designs may allow for the initially planned sample size to be adjusted while the trial is in progress. This procedure is referred to as sample size re-estimation.

Owing to the fact that nonparametric approaches, however, do not rely on parametric assumptions of the data, defining the anticipated treatment effect for the purpose of sample size estimation can pose an additional challenge. This thesis aims to investigate possible strategies to estimate sample sizes when using a nonparametric procedure for the retention-of-effect approach, with a primary focus on the studentized permutation test by Mütze et al. (2017). Extensive simulation studies have led to the development of a sample size planning method for this test. Additionally, considerations regarding the test by Munzel (2009) have been explored.

The first part of this thesis provides a brief overview of the existing literature on the analysis of three-arm clinical trials. The subsequent thesis is structured around the two testing procedures. Chapter 3 focuses on the studentized permutation test, covering the statistical model and hypothesis test. The operating characteristics of the test under both the null and alternative hypothesis are explored and the results are then utilised

to propose a sample size planning method. Chapter 4 examines the nonparametric test proposed by Munzel (2009), introducing the statistical model and hypothesis and investigating the test's operating characteristics under both the null and alternative hypothesis. Considerations for sample size planning methods are provided. The findings for each testing procedure are summarised and discussed in the corresponding section of the chapter. Finally, the thesis concludes with a review of the results and an outlook on how further research can build upon these findings.

2 Review of three-arm gold standard design analyses in existing literature

Two-arm non-inferiority trials are widely used to demonstrate the non-inferiority of an experimental treatment versus a reference treatment. D'Agostino et al. (2003) and Röhmel (1998) provide a detailed discussion on the design of such trials. Despite this, research over the past two decades has increasingly focused on three-arm trials. This period witnessed ongoing analysis of three-arm non-inferiority trials across various clinical endpoints. Pigeot et al. (2003) proposed the first approach for normally distributed endpoints, assuming a common variance in the three treatment groups. The method was then extended to normally distributed endpoints with heterogeneous variances by Hasler et al. (2008), and a novel approach based on the Fieller-Hinkley distribution for normal endpoints was presented by Koti (2007). The first approach to testing non-inferiority for binary data was introduced by Kieser and Friede (2007). Since then, a variety of additional methods for binary data have been presented (M.-L. Tang and Tang, 2004; N.-S. Tang et al., 2014; Chowdhury et al., 2019b; N. Tang and Yu, 2020; and Paul et al., 2021). Two non-inferiority tests for survival data have been proposed; Mielke et al. (2008) proposed one assuming exponentially distributed endpoints and Kombrink et al. (2013) another assuming Weibull distributed endpoints. A testing strategy for count data following a Poisson distribution was established by Mielke and Munk (2009). S. Ghosh et al. (2022) introduced a novel approach for Poisson-count data. Mütze et al. (2016) suggested the non-inferiority testing approach for count data following a negative binomial distribution. The analysis of ordinal data has been discussed in the context of two-arm non-inferiority trials (see Lui and Chang, 2013). When it comes to three-arm designs, to the best of our knowledge, the nonparametric test proposed by Munzel (2009) is the only available method for testing non-inferiority with ordinal data.

In addition to the previously mentioned frequentist approaches, several testing strategies rooted in Bayesian analysis have been developed for non-inferiority trials. Simon (1999) introduced the first Bayesian approach for two-arm non-inferiority trials, and a Bayesian approach in the three-arm design was later proposed by P. Ghosh et al. (2011). Several studies, including P. Ghosh et al. (2011) and S. Ghosh et al. (2016), have discussed the advantages of Bayesian methods in analysing non-inferiority trials. These benefits encompass the incorporation of prior information through specific prior distributions and the grounding of inference in the posterior distribution, eliminating the need for reliance on asymptotics – a feature especially valuable for sparse data. Another advantage of Bayesian approaches is their inherent flexibility, which enables the accommodation of a wider range of models and diverse types of data. The Bayesian method has seen specific refinements for certain endpoints, particularly for binary data (see Chowdhury et al., 2019a; S. Ghosh et al., 2018) and normally distributed data (Gamalo et al., 2016).

Another approach that offers greater flexibility in testing non-inferiority is to avoid making assumptions about the underlying distributional properties of the clinical endpoint by using nonparametric methods. As mentioned in the introduction (Section 1), several nonparametric tests for non-inferiority trials have been introduced, including those by Munzel (2009), Mütze et al. (2017), S. Ghosh et al. (2017), and Li et al. (2023). This thesis will investigate methods for sample size planning in three-arm non-inferiority trials using the proposed studentized permutation test by Mütze et al. (2017) and the nonparametric test based on classical mid-ranks by Munzel (2009).

3 The studentized permutation test

3.1 Statistical model and hypothesis testing

Non-inferiority hypothesis Denote X_{ik} with $k = 1, ..., n_i$ and $i = \text{EXP}, \text{REF}, \text{PLA}$ the outcomes of independent real-valued random variables under the experimental treatment (EXP), reference treatment (REF) and placebo (PLA) of a three-arm clinical trial. It is assumed that the respective random variable X_{ik} follows a distribution F_i with finite mean $\mathbb{E}[X_{ik}] = \mu_i$, finite positive variance $Var[X_{ik}] = \sigma_i^2 > 0$ and finite fourth moment $\mathbb{E}[X_{ik}^4]$.

Let μ_{EXP}, μ_{REF} and μ_{PLA} denote the parameters of interest for the experimental, reference and placebo group respectively where higher values are associated with a higher treatment effect. To demonstrate non-inferiority of the experimental treatment compared to the existing reference treatment, it is necessary to show that the difference between the experimental and reference treatment exceeds a pre-specified, clinically irrelevant amount, that is the margin δ with $\delta < 0$. The statistical testing problem of non-inferiority can be formulated as

$$H_0 : \mu_{\text{EXP}} - \mu_{\text{REF}} \leq \delta \text{ vs. } H_1 : \mu_{\text{EXP}} - \mu_{\text{REF}} > \delta. \tag{1}$$

Simultaneously, the trial needs to ensure assay sensitivity, which refers to its ability to distinguish an effective treatment from an ineffective one (International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2001). This is typically done by showing that the reference treatment is superior to placebo. In three-arm non-inferiority trials, this is achieved by using the information from the placebo arm to define the margin δ as a fraction f of the difference between the reference treatment and placebo, that is

$$\delta = f(\mu_{\text{REF}} - \mu_{\text{PLA}}) \tag{2}$$

with $f \in (-1, 0)$. The fraction f thereby quantifies by how much the reference treatment is superior to placebo. By plugging f into (1) one obtains

$$H_0 : \mu_{\text{EXP}} - \mu_{\text{REF}} \leq f(\mu_{\text{REF}} - \mu_{\text{PLA}}) \text{ vs. } H_1 : \mu_{\text{EXP}} - \mu_{\text{REF}} > f(\mu_{\text{REF}} - \mu_{\text{PLA}}). \tag{3}$$

Assuming that $\mu_{\text{REF}} - \mu_{\text{PLA}} > 0$, the definition of Δ becomes $\Delta = 1 + f$. The above testing problem can then be rewritten in terms of the ratio of the differences in the group-specific expectation parameters μ_i, that is

$$H_0 : \frac{\mu_{\text{EXP}} - \mu_{\text{PLA}}}{\mu_{\text{REF}} - \mu_{\text{PLA}}} \leq \Delta \text{ vs. } H_1 : \frac{\mu_{\text{EXP}} - \mu_{\text{PLA}}}{\mu_{\text{REF}} - \mu_{\text{PLA}}} > \Delta. \tag{4}$$

This way of rearranging the hypotheses allows a straightforward interpretation of the effect under the alternative hypothesis. It implies that, under the alternative hypothesis, the experimental treatment achieves more than $(\Delta \cdot 100)\%$ of the efficacy of the reference treatment where the experimental and reference treatment are each compared to placebo. Δ therefore represents the minimum fraction of the reference treatment effect relative to placebo that the experimental treatment effect relative to placebo needs to preserve in order to demonstrate non-inferiority. The margin Δ is typically referred to as the non-inferiority margin. This kind of hypothesis where the non-inferiority margin is defined by a fraction f is also known as retention-of-effect hypothesis and was first described by Koch and Tangen (1999). A separate approach in defining the hypotheses was proposed by Hida and Tango in 2011. The absolute margin approach defines the non-inferiority margin Δ by a pre-specified constant. The following work, however, will focus on the retention-of-effect approach.

The studentized permutation test This paragraph introduces the studentized permutation test published by Mütze et al. (2017), which evaluates the non-inferiority hypothesis for the retention-of-effect approach. A permutation approach for testing the absolute margin hypothesis in the three-arm design is outlined in Appendix A.

Rather than relying on parametric assumptions, the studentized permutation test approximates the distribution of the test statistic under the null hypothesis by permutation of the data and therefore represents a nonparametric approach for testing the non-inferiority hypothesis. The construction of the Wald-type test statistic involves rearranging (4) as follows:

$$H_0 : \mu_{\text{EXP}} - \Delta\mu_{\text{REF}} + (\Delta - 1)\mu_{\text{PLA}} \leq 0 \text{ vs. } H_1 : \mu_{\text{EXP}} - \Delta\mu_{\text{REF}} + (\Delta - 1)\mu_{\text{PLA}} > 0. \quad (5)$$

Let $\boldsymbol{X}_n$ denote the random vector that contains all observations of the trial, that is

$$\boldsymbol{X}_n = (X_{\text{EXP},1}, \ldots, X_{\text{EXP},n_{\text{EXP}}}, X_{REF,1}, \ldots, X_{\text{REF},n_{\text{REF}}}, X_{\text{PLA},1}, \ldots, X_{\text{PLA},n_{\text{PLA}}})$$

with $\mathbb{P}$ denoting its probability measure. Denote n as the total sample size and n_i the respective group sample size. Further, suppose that none of the three treatment groups vanishes asymptotically $w_i = \lim_{n_i, n\to\infty} \frac{n_i}{n} \in (0, 1)$ with $w_i = \frac{n_i}{n}$. The Wald-type statistic T_n can then be derived by replacing the μ_i's in (5) by the group-specific sample means $\bar{X}_{i\cdot}$ and dividing the term by an estimate of its standard deviation, that is

$$T_n = T_n(\boldsymbol{X_n}) = \sqrt{n}\frac{\bar{X}_{\text{EXP}\cdot} - \Delta\bar{X}_{\text{REF}\cdot} + (\Delta - 1)\bar{X}_{\text{PLA}\cdot}}{\hat{\sigma}} \quad (6)$$

where the variance estimator $\hat{\sigma}^2$ is defined as

$$\hat{\sigma}^2 = \frac{\hat{\sigma}^2_{\text{EXP}}}{w_{\text{EXP}}} + \Delta^2 \frac{\hat{\sigma}^2_{\text{REF}}}{w_{\text{REF}}} + (1-\Delta)^2 \frac{\hat{\sigma}^2_{\text{PLA}}}{w_{\text{PLA}}} \tag{7}$$

with the group-specific sample variances

$$\hat{\sigma}_i^2 = \frac{1}{n_i - 1} \sum_{k=1}^{n_i} (X_{ik} - \bar{X}_{i\cdot})^2.$$

In a first step, the test statistic T_n is calculated based on the observed data $\boldsymbol{X_n}$. However, rather than using parametric assumptions on the distribution of the test statistic, the data is permutated to approximate the distribution under the null hypothesis. Let $(\tau(i))_{i \leq n}$ denote a random variable that is uniformly distributed on the group of all permutations of the first n natural numbers with probability measure $\tilde{\mathbb{P}}$ and denote $\tau_n(\boldsymbol{X}_n) = (X_{n,\tau(1)}, \ldots, X_{n,\tau(n)})$ as a randomly permutated vector of $\boldsymbol{X}_n$. The test statistic is then again calculated based on the permutated data. For a given vector $\boldsymbol{X}_n$, the test statistic calculated with the permutated vector $\tau_n(\boldsymbol{X}_n)$ is referred to as permutation statistic. Hence, the permutation statistic is a result of the mapping

$$(\tau(i))_{i \leq n} \to T_n\big(X_{n,\tau(1)}, \ldots, X_{n,\tau(n)}\big) | \boldsymbol{X_n}.$$

For a given significance level $\alpha \in (0,1)$, the function ϕ_n^{Perm} then refers to the studentized permutation test with

$$\phi_n^{\text{Perm}}(\boldsymbol{X}_n) = \begin{cases} 1 & T_n(\boldsymbol{X}_n) > c_n(\alpha) \\ 0 & T_n(\boldsymbol{X}_n) \leq c_n(\alpha) \end{cases}$$

where $c_n(\alpha)$ denotes the α-quantile of the permutation distribution, that is the distribution of the test statistic based on the permutated data, which is the largest number such that

$$\tilde{\mathbb{P}}\big(T_n(\tau_n(\boldsymbol{x}_n)) > c_n(\alpha)\big) \leq \alpha$$

holds. It can be shown that the expected value of the test function ϕ_n^{Perm} converges to α with respect to the probability measure $\mathbb{P}$ at the boundary of the null hypothesis, that is

$$\lim_{n \to \infty} \mathbb{E}_{\mathbb{P}}\big[\phi_n^{\text{Perm}}(\boldsymbol{X}_n)\big] = \alpha.$$

Refer to Mütze et al. (2017) for a detailed derivation of the asymptotic behaviour of the studentized permutation test.

In practice, the procedure to derive the p-value for the non-inferiority hypothesis based on the studentized permutation test can be broken down into the following steps:

1. Computation of the test statistic $T_n(\boldsymbol{X}_n)$ for observed data $\boldsymbol{X}_n$

2. Permutation of the data $\tau_n(\boldsymbol{X}_n)$

3. Computation of the test statistic based on permutated data $T_n(\tau_n(\boldsymbol{X}_n))$

4. Repetition of steps 2 and 3 for J times, e.g. 10,000 times (number of permutation replications)

5. Computation of the p-value as the number of times that the permutated test statistic $T_n(\tau_n(\boldsymbol{X}_n))$ is as or more extreme than the test statistic on the observed data $T_n(\boldsymbol{X}_n)$ divided by J, that is

$$\frac{1}{J}\sum_{j=1}^{J} I(T_n(\boldsymbol{X}_n) \leq T_n(\tau_n(\boldsymbol{X}_n))$$

   where I denotes the indicator function.

3.2 Operating characteristics of the studentized permutation test

To develop a sample size planning method, the first step involves examining the operating characteristics of the studentized permutation test by means of simulation studies. As a primary step, the operating characteristics of the studentized permutation test are assessed under the null hypothesis. This step is undertaken to verify that the studentized permutation test maintains the nominal significance level across the simulation scenarios being investigated. The subsequent step involves investigating the empirical power of the test under the alternative hypothesis. The findings from the power simulation then serve as a basis to derive an approach for sample size planning.

Table 1 displays the scenarios used in the simulation study. The parameters are categorised based on the situation under the hypothesis, that is under the null hypothesis (column 2) for the investigation of the type I error rate and the alternative hypothesis (column 3) for the investigation of the power of the test. Parameters that remain the same under both scenarios are represented by merged cells.

Table 1: Scenarios for the simulation study investigating the operating characteristics of the studentized permutation test under the null and alternative hypothesis.

Parameter	Values under H_0	Values under H_1
Distributions	Normal, t, Lognormal, Chi-squared	
Non-inferiority margin Δ	0.8	
Ratio in the mean differences $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$	0.8	0.9, 1, 1.1, 1.2
Group standard deviations $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}})$	(1;1;1); (1;2;3); (3;2;1)	$(1; 1; 1)$; $(3; 2; 1)$
Sample size allocations $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	$(1:1:1)$; $(2:2:1)$; $(3:2:1)$; $(3:3:1)$; (1:Δ:1-Δ)	
Total sample size n	30, 60, 120, 210, 300	420
One-sided nominal level α	0.025	
Permutation replications	10,000	
Simulation replications	5,000	

The simulation includes continuous data generated from various distributions, including normal, t, lognormal, and chi-squared distributions. This broad range of distributions is considered to capture scenarios where the data deviates from normality, exhibiting characteristics such as asymmetry, skewness, and heavier tails. Hereby, data is generated from a lognormal distribution with location parameter $\mu = 0$ and scale parameter $\sigma = 1$, from a t-distribution with 4 degrees of freedom and a chi-square-distribution with 2 degrees of freedom, denoted as χ^2. In the following, the observation X_{ik} will be referred to as t, lognormal or χ^2-distributed when it is calculated from a standardized t, lognormal or χ^2-distributed random variable, that is

$$X_{ik} = \left(\frac{X_{ik} - \mathrm{E}[X_{ik}]}{\sqrt{\mathrm{Var}[X_{ik}]}} \right) \cdot \sigma_i + \mu_i \qquad i = \mathrm{EXP}, \mathrm{REF}, \mathrm{PLA} \text{ and } k = 1, ..., n_i. \tag{8}$$

This notation was adapted from Mütze et al. (2017). It guarantees that the data conforms to the appropriate expectations and standard deviations, regardless of their underlying distribution.

In the simulation, the non-inferiority margin Δ is fixed at 0.8. The choice of the non-inferiority margin is motivated following the assumption of Pigeot et al. (2003) that a difference of 20% between the population means of the experimental and reference treat-

ment is considered as clinically unimportant, that is $f(\mu_{\text{REF}} - \mu_{\text{PLA}}) = -1/5$. Expressed in the ratio of the group means (5) this means that in order to be accepted as non-inferior, the experimental treatment must demonstrate an effect size of more than 80% of the mean effect size of the reference treatment, each in comparison to placebo. The effect under the hypothesis $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ is then varied accordingly to the hypothesis. For both scenarios, μ_{REF} and μ_{PLA} are fixed with $\mu_{\text{REF}} = 1$ and $\mu_{\text{PLA}} = 0$. Under the null hypothesis, μ_{EXP} is fixed at 0.8 while μ_{EXP} is varied under the alternative hypothesis to generate the different ratios $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$. It is expected that greater ratios result in more evidence for the alternative hypothesis, rendering higher power levels obtained with the studentized permutation test.

Homogeneous as well as heterogeneous standard deviation scenarios across the three treatment groups are investigated. The homogeneous scenario is defined as $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 1; 1)$. In case of heterogeneity, the standard deviations are chosen as $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ and $(3; 2; 1)$. In part, these scenarios represent very extreme scenarios of variation. This is done to show possible limitations of the applied methods. For a depiction and brief description of the generated data within the simulation study for the considered expectation and standard deviation parameters under both the null and alternative hypothesis, refer to Appendix A.

Typical sample size allocations for three-arm clinical trials $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$, namely (1:1:1), (2:2:1), (3:2:1), (3:3:1) and $(1{:}\Delta{:}1-\Delta)$ (here: 1:0.8:0.2), are included. Specifically, the latter is supposed to be the optimal allocation for any parametric family in the sense that a maximum power is achieved while minimising the total required sample size n (refer to Mielke and Munk, 2009). Under the null hypothesis, the total sample size n is varied from 30 to 300 to investigate whether the one-sided nominal level can be ensured for increasing n. Under the alternative hypothesis, the total sample size n is fixed at $n = 420$ to demonstrate the impact of an increasing effect size on the power of the test instead. A one-sided significance level of $\alpha = 0.025$ is assumed. To ensure that the one-sided significance level can be attained asymptotically, the rejection area of the studentized permutation test is obtained from 10,000 permutations replications. Each scenario is then replicated 5,000 times.

For comparison, the parametric equivalent Wald-type test derived by Hasler et al. (2008) is also conducted for the presented scenarios and its results are included in the following analysis. The test by Hasler et al. (2008), referred to as Hasler test in the following, assumes that the endpoints follow a normal distribution with means μ_{EXP}, μ_{REF} and μ_{PLA} and allows for heterogeneous variances across the three groups. The Hasler test is based on the same test statistic as the studentized permutation test, that is T_n as in (6). Based on a Welch approximation, the distribution of T_n under the null hypothesis is thereby approximated by a t-distribution with ν^{het} degrees of freedom, which are given

by

$$\nu^{\text{het}} = \frac{\left(\frac{1}{n_{\text{EXP}}}\sigma^2_{\text{EXP}} + \frac{\Delta^2}{n_{\text{REF}}}\sigma^2_{\text{REF}} + \frac{(1-\Delta)^2}{n_{\text{PLA}}}\sigma^2_{\text{PLA}}\right)^2}{\frac{1}{n^2_{\text{EXP}}(n_{\text{EXP}}-1)}\sigma^4_{\text{EXP}} + \frac{\Delta^4}{n^2_{\text{REF}}(n_{\text{REF}}-1)}\sigma^4_{\text{REF}} + \frac{(1-\Delta)^4}{n^2_{\text{PLA}}(n_{\text{PLA}}-1)}\sigma^4_{\text{PLA}}}$$

where the unknown parameters σ^2_{EXP}, σ^2_{REF} and σ^2_{PLA} can be estimated by the sample variances S^2_{EXP}, S^2_{REF} and S^2_{PLA}. This yields the estimated degrees of freedom $\hat{\nu}^{\text{het}}$. The hypothesis of inferiority is then rejected if the test statistic is greater than the $(1-\alpha)$-quantile of the central t-distribution with $\hat{\nu}^{\text{het}}$ degrees of freedom. This test can be expressed as the function

$$\phi_n^{\text{Hasler}}(\boldsymbol{X}_n) = \begin{cases} 1 & T_n(\boldsymbol{X}_n) > t_{1-\alpha}(\hat{\nu}^{\text{het}}) \\ 0 & T_n(\boldsymbol{X}_n) \leq t_{1-\alpha}(\hat{\nu}^{\text{het}}). \end{cases}$$

The simulation was conducted using the program R (R Core Team, 2022). Both tests were carried out using the functionality of the R-package `ThreeArmedTrials` which is available on CRAN (Mütze, 2023). In the process of this thesis, a minor coding error was discovered in the functionality of the studentized permutation test, leading to erroneous calculations specifically for unbalanced group designs (refer to Appendix A for detailed information). The error was reported to the maintainer of the package, who promptly addressed and resolved the issue (see the commit of April 18, 2023). As a result, a corrected version of the studentized permutation test is now implemented in the `ThreeArmedTrials` package. The scripts for conducting the simulation study are provided in Appendix A.

3.2.1 Type I error rate

The simulation under the null hypothesis investigates whether the studentized permutation test holds the nominal level of $\alpha = 0.025$ under the various simulation scenarios. For this purpose, the effect under the hypothesis $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$, that is $\mu_{\text{EXP}} - \Delta\mu_{\text{REF}} + (\Delta - 1)\mu_{\text{PLA}}$ is set to $\Delta = 0.8$. The results of the simulation under the null hypothesis are shown in Figure 1. The Figure is divided into columns and rows. The columns represent the distributions that generated the data, namely the normal, t, log-normal and χ^2-distribution. The rows represent the different standard deviation scenarios for the three groups, namely the homogeneous scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 1; 1)$ in row 1 and the two heterogeneous scenarios of $(1; 2; 3)$ and $(3; 2; 1)$ in rows 2 and 3. The different sample size allocation schemes are represented by different line types. The type I error rate on the y-axis is then displayed by the varying total sample size n on the x-axis. The two grey lines indicate the area of the nominal level $\alpha = 0.025 \pm$ two times the Monte Carlo error. The Monte Carlo error for a nominal level $\alpha = 0.025$ by 5,000

replications is given by

$$\sqrt{\frac{1}{5{,}000} \cdot (0.025 \cdot (1 - 0.025))} \approx 0.002.$$

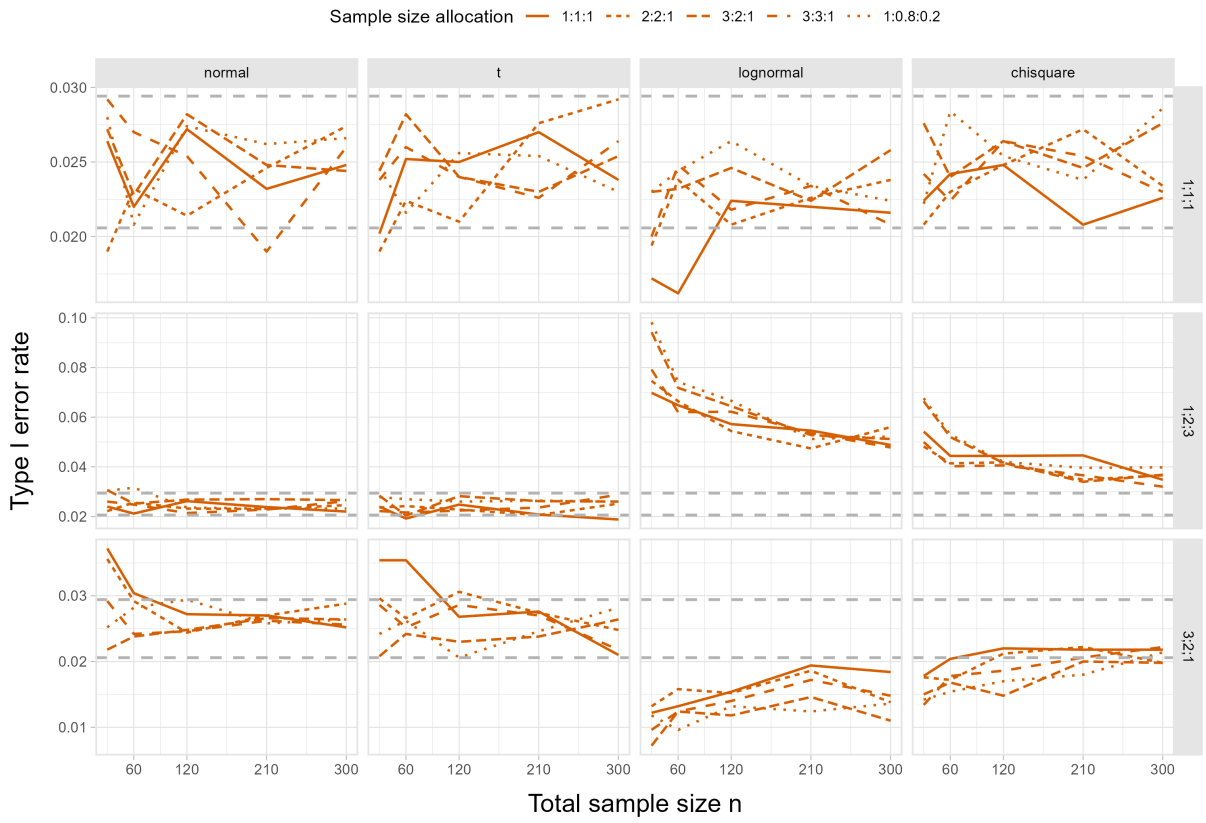


Figure 1: Actual significance level $\hat{\alpha}$ of the studentized permutation test against the total sample size n. The dashed grey lines depict the area of $\alpha = 0.025 \pm$ two times the Monte Carlo error.

Figure 1 shows that the studentized permutation test holds the one-sided nominal level for the scenario of homogeneous variances for all four data-generating mechanisms (first row). For normal and t-distributed data (first and second column), the significance level is also controlled under almost all remaining scenarios of group standard deviations (rows 2 and 3). An exception is the case of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$ for small sample sizes under a balanced group design or design of $(3 : 2 : 1)$ where the test tends to be liberal. For data following a lognormal and χ^2-distribution (third and fourth column), however, the test becomes too liberal in case of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ and too conservative in case of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$. The former is especially the case for lognormal data. When a test is too liberal, i.e. exceeds the nominal significance level, it implies that it has a higher tendency to reject the null hypothesis although it is true. This behaviour is more concerning than a conservative behaviour where the test is less likely to produce false positive results. In Figure 1, higher sample sizes reduce the elevated type

I error rate in that scenario but cannot control it to the desired level with the considered sample sizes. The lines indicating the different allocation schemes suggest that there is variation in the observed type I error rate depending on the chosen allocation ratio of the groups. However, no pattern is visible across all scenarios.

Additionally, it is also possible to compare the operating characteristics of the Hasler test with those of the studentized permutation test under the null hypothesis. Figure 2 shows the type I error rate for the studentized permutation test as well as the Hasler test. Note that the different colours in the figure now indicate the considered test statistic, that is the studentized permutation test and the test by Hasler et al. (2008). Figure 2 specifically illustrates the type I error rate solely for the unbalanced $(1 : \Delta : 1-\Delta)$ design. This choice is due to the Hasler test demonstrating comparable behaviour across different allocation schemes. Moreover, this focus enhances clarity in the visual representation.

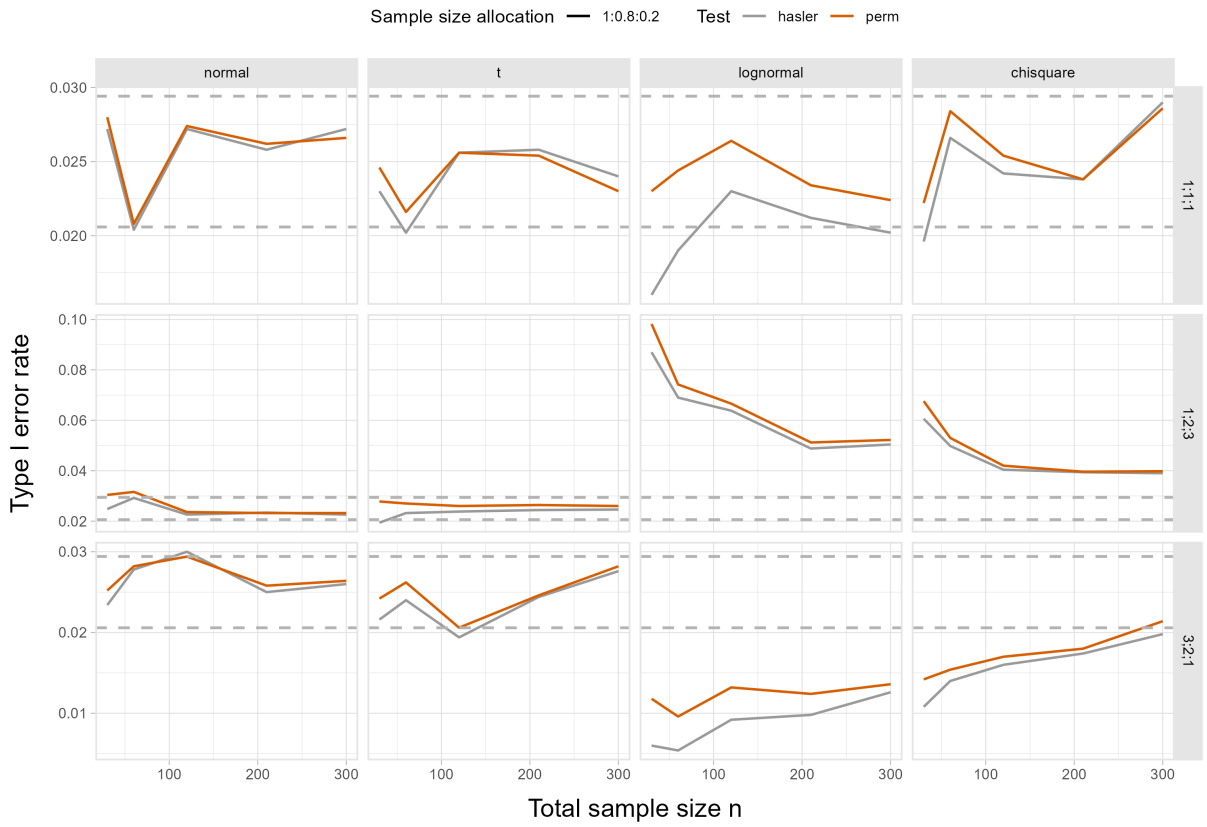


Figure 2: Actual significance level $\hat{\alpha}$ of the studentized permutation test and the Hasler test against the total sample size n in the group design $(1 : \Delta : 1 - \Delta)$. The dashed grey lines depict the area of $\alpha = 0.025 \pm$ two times the Monte Carlo error.

Overall, Figure 2 shows that both tests behave very similarly in terms of the type I error rate. The nominal significance level is maintained by both tests for almost all scenarios under normal and t-distributed data. In case of skewed data with group standard deviations of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ the Hasler test also shows a liberal behaviour as well as a conservative behaviour under $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$. However, it even

seems that the studentized permutation test reports a greater type I error rate than the Hasler test in both of these scenarios.

Also, it should be noted that both tests report slight differences in the type I error rate according to the chosen allocation scheme. However, there does not seem to be a pattern across all group standard deviation scenarios and/or data-generating mechanisms, explaining the variation by allocation scheme.

The liberal behaviour observed in both the studentized permutation test and the Hasler test can be attributed to the variability in the variance estimation of data that follows lognormal and χ^2-distributions, particularly for smaller sample sizes. For a detailed explanation of the liberal behaviour of these tests in the scenario where $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$, please refer to Appendix A.

The type I error rate of the studentized permutation test was also investigated by Mütze et al. (2017). In their study, the operating characteristics of the test were examined under the variance scenario of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 2; 3)$. They observed a conservative behaviour of the test for skewed data in that case, which contrasts with the liberal behaviour observed in the previous results for increasing standard deviation scenarios. The difference in findings can be attributed to the fact that Mütze et al. (2017) focused on the reversed effect under the hypotheses, where smaller values of the outcome are associated with a higher treatment effect. As evidenced in the Appendix A, it can be demonstrated that the results presented here align with those reported by Mütze et al. (2017).

In summary, the simulation study conducted under the null hypothesis demonstrates that the studentized permutation test effectively maintains the nominal significance level in various cases. Particularly for normal and t-distributed data, the test exhibits excellent performance. However, for skewed data, the test tends to be too liberal when the standard deviations increase, as observed in the scenario $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$. Conversely, it becomes too conservative when the standard deviations decrease, as seen in the scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$. It is important to note that this characteristic is also observed in the parametric equivalent Hasler test. Therefore, caution should be exercised when handling scenarios with increasing group standard deviations, particularly when the data is skewed. The type I error rate of the studentized permutation test exhibits minor variations among the different allocation schemes but does not display a consistent pattern across all scenarios. Moreover, its behaviour closely resembles that of the Hasler test in terms of the type I error rate.

3.2.2 Power

The simulation under the alternative hypothesis aims at investigating the power behaviour of the studentized permutation test to ultimately derive possible strategies to plan sample sizes. The effect under the alternative hypothesis $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ is now

varied from 0.9 to 1.2 by steps of 0.1 to demonstrate the impact of an increasing effect on the power of the test. The total sample size is fixed with $n = 420$. In light of the detected liberal behaviour of the studentized permutation test and the Hasler test under the null hypothesis for skewed data under $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$, this scenario is excluded from the power simulation. Refer to column 3 of Table 1 for the parameters of the simulation. The results for the studentized permutation test are shown in Figure 3 where the empirical power curve is displayed on the y-axis against the varying ratio under the hypothesis on the x-axis. The Figure is divided into columns and rows where the columns represent the four data-generating mechanisms and the rows represent the two group standard deviation scenarios. The different sample size allocations $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$ are distinguished by different line types.

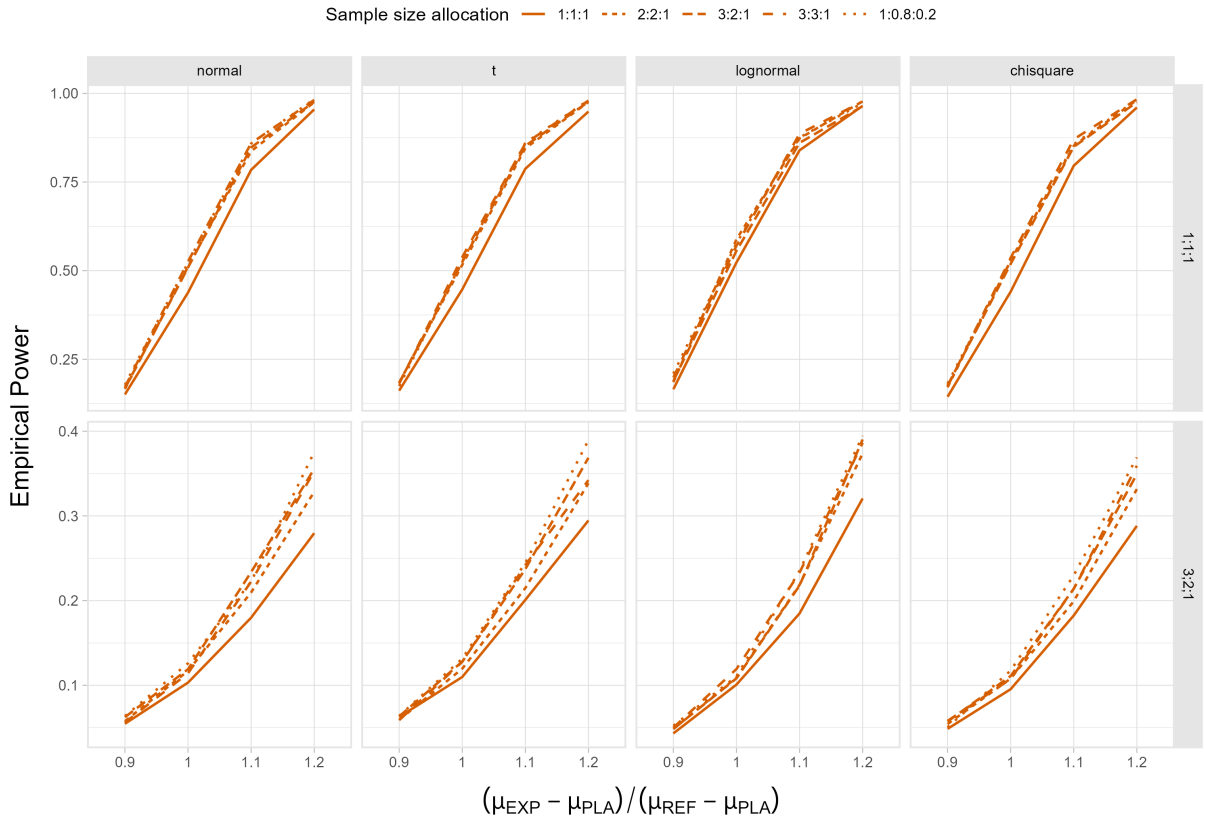


Figure 3: Observed power of the studentized permutation test against $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ for a total sample size $n = 420$.

As expected, the power of the test increases for an increasing ratio in the mean differences. That is because an increasing ratio means more evidence for the alternative of non-inferiority. If one aims at a power level of 80%, it is reached for a ratio $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ of approximately 1.1 for all four underlying distributions in case of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 1; 1)$ for a total sample size of $n = 420$. For the scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 1; 1)$, the power curve is considerably higher compared to the scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$. In the heterogeneous scenario, a power

level of 80% is not reached. Not surprisingly, a higher total sample size n is needed for heterogeneous standard deviations to obtain an adequate power level.

The power of the studentized permutation test also varies by the chosen allocation scheme. Table 2 displays the mean power of the studentized permutation test by sample size allocation and underlying distribution of the data. The last column shows the overall mean power of the respective allocation scheme and the last row shows the mean power by the respective underlying distribution of the data.

Table 2: Mean observed power of the studentized permutation test in percentage by group design and underlying distribution of the data for a total sample size $n = 420$ across all $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$.

$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	normal	$t(4)$	lognormal	$\chi^2(2)$	Mean power
$(1 : 1 : 1)$	36.81	37.66	39.27	36.95	37.67
$(2 : 2 : 1)$	40.03	40.67	42.18	40.33	40.80
$(3 : 3 : 1)$	41.05	41.61	42.76	41.27	41.67
$(3 : 2 : 1)$	40.89	41.66	41.66	40.64	41.21
$(1 : \Delta : 1 - \Delta)$	41.63	42.14	42.84	41.41	42.01
Mean power	40.08	40.75	41.74	40.12	

Table 2 demonstrates that the studentized permutation test is able to obtain a slightly greater power level under lognormality of the data with a mean power of about 41.74% across all scenarios. For the other three underlying distributions, the mean power is constant by about 40%. The highest mean power levels are reached under the allocation schemes of $(1 : \Delta : 1 - \Delta)$ and $(3 : 3 : 1)$. The lowest mean power is reported for the balanced design with about 38%. Hence, unbalanced designs are able to achieve the desired power level more rapidly and therefore require fewer participants. Therefore, they should be preferred over balanced group designs. In particular, assuming normal data, it was demonstrated by Pigeot et al. (2003) that the allocation of $(1 : \Delta : 1 - \Delta)$ is optimal in the sense that it maximises power while minimising the total required sample size n. This holds even true for any parametric family when variances are estimated unrestrictedly, as shown by Mielke and Munk (2009). The simulation results mentioned above are consistent with this finding, even in the nonparametric scenario when using the studentized permutation test.

For comparison, the power of the parametric test by Hasler et al. (2008) can be compared to the power of the studentized permutation test. In Figure 4 the power behaviour of the Hasler test is shown additionally to the power of the studentized permutation test. Since the power curves show an analogous behaviour with respect to the chosen allocation scheme, the results for the allocation scheme $(1 : \Delta : 1 - \Delta)$ are shown solely. Note that

the colours now distinguish the two tests, namely the studentized permutation test and the Hasler test.

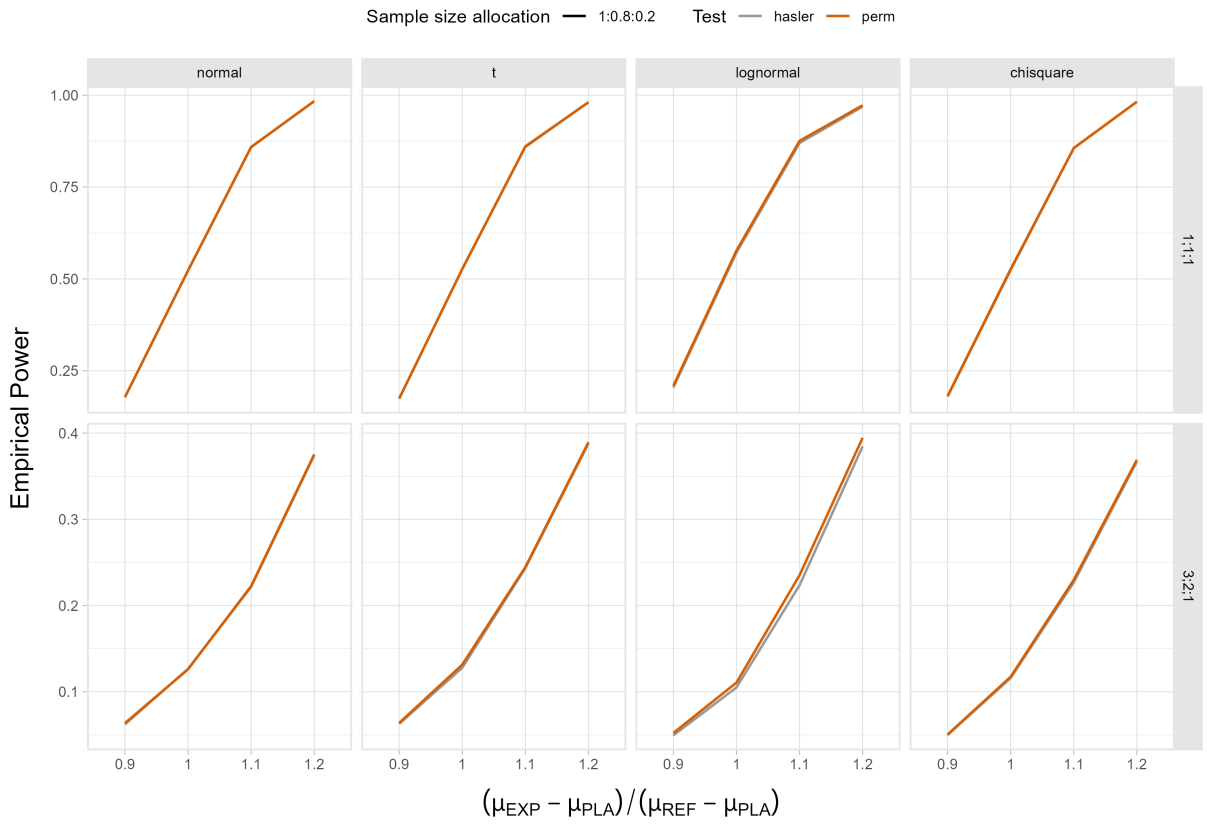


Figure 4: Observed power of the studentized permutation test and the Hasler test against $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ for a total sample size $n = 420$ in the group design $(1 : \Delta : 1 - \Delta)$.

Figure 4 shows that, overall, the studentized permutation test behaves very similarly to the Hasler test with respect to the empirical power. The power curves seem to overlap for most of the considered scenarios. For lognormal data, the studentized permutation test is able to obtain a slightly higher power level than the Hasler test, especially for the heterogeneous standard deviation scenario. This also seems to be the case for data following a χ^2-distribution whereby the difference is not so strong as for lognormal data. In Table 3, the mean power differences between the studentized permutation test and the Hasler test are displayed by the standard deviation scenario and underlying distribution of the data. The last column and the last row show the respective overall mean power difference. The difference is computed by substracting the power of the Hasler test from that of the studentized permutation test.

Table 3: Mean difference between the observed power of the studentized permutation test and the observed power of the Hasler test in percentage points by standard deviation scenario and underlying distribution of the data for a total sample size $n = 420$ across all $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ and group designs.

$(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}})$	normal	$t(4)$	lognormal	$\chi^2(2)$	Mean Difference
$1; 1; 1$	-0.02	0.11	1.28	0.55	0.48
$3; 2; 1$	0.19	0.36	2.13	0.91	0.90
Mean Difference	0.09	0.23	1.70	0.73	0.007

Except for the case of homogeneous variances under normality, the studentized permutation test performs better than the Hasler test in terms of power across all scenarios. On average, the power of the studentized permutation test is greater by 0.007 percentage points, indicating, however, that this difference is rather marginal. The highest power difference is obtained under lognormality where the studentized permutation test reports 1.70 percentage points greater power than the Hasler test. This is followed by 0.73 percentage points greater power for data following a χ^2-distribution. The smallest positive difference occurs for normal data with a mean power difference of 0.09. Hence, the simulation could show a greater power of the test compared to the Hasler test for non-normal data.

Also, the studentized permutation test performs especially well under heterogeneous standard deviations compared to the homogeneous case across all data-generating mechanisms. On average, the studentized permutation test reports 0.9 percentage points greater power than the Hasler test under the heterogeneous setting. Overall, however, the magnitude of the difference is relatively small.

The simulation demonstrated that the power curve of the studentized permutation test closely resembles the power curve of the Hasler test. This suggests that the power of the studentized permutation test can be approximated by the power of the Hasler test. However, it is also noteworthy that the studentized permutation test can be the better choice in terms of power. This is especially the case for skewed data and scenarios of heterogeneous standard deviations.

3.3 Using the Hasler sample size formula for sample size planning for the studentized permutation test

The simulation results under the alternative hypothesis for continuous data showed that the power of the studentized permutation test behaves very similarly to the power of the Hasler test in three-arm non-inferiority designs. Differences in the power between those two tests occur mostly due to the underlying data-generating mechanism and the group

standard deviations. That is, for skewed data and heterogeneous standard deviations the studentized permutation test performs slightly better. However, these differences were found to be rather small. The observation that the power curve of the studentized permutation test resembles the curve of the Hasler test consequently suggests the use of the Hasler formula for sample size planning when using the studentized permutation test for analysis.

The upcoming paragraph will introduce the sample size estimation approach as derived by Hasler et al. (2008), and examine the potential challenges that arise during the process of estimating sample sizes. To obtain a power of at least $1-\beta$ for the Hasler test, it must hold that

$$P\left\{T_n > t_{1-\alpha}(\nu^{\text{het}}) \middle| \frac{\mu_{\text{EXP}} - \mu_{\text{PLA}}}{\mu_{\text{REF}} - \mu_{\text{PLA}}} > \Delta, \sigma^2_{\text{EXP}}, \sigma^2_{\text{REF}}, \sigma^2_{\text{PLA}}\right\} \geq 1-\beta$$

where ν^{het} denote the degrees of freedom of the central t-distribution which is given by

$$\frac{\left(\frac{1}{n_{\text{EXP}}}\sigma^2_{\text{EXP}} + \frac{\Delta^2}{c_{\text{REF}} n_{\text{EXP}}}\sigma^2_{\text{REF}} + \frac{(1-\Delta)^2}{c_{\text{PLA}} n_{\text{EXP}}}\sigma^2_{\text{PLA}}\right)^2}{\frac{\sigma^4_{\text{EXP}}}{n^2_{\text{EXP}}(n_{\text{EXP}}-1)} + \frac{\Delta^4\sigma^4_{\text{REF}}}{(c_{\text{REF}} n_{\text{EXP}})^2(c_{\text{REF}} n_{\text{EXP}}-1)} + \frac{(1-\Delta)^4\sigma^4_{\text{PLA}}}{(c_{\text{PLA}} n_{\text{EXP}})^2(c_{\text{PLA}} n_{\text{EXP}}-1)}} \tag{9}$$

Therein, the sample sizes of reference and placebo group are denoted as proportions c_{REF} and c_{PLA} of the experimental groups sample size, that is $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = 1 : c_{\text{REF}} : c_{\text{PLA}}$. Then, the sample size for the experimental group n_{EXP} for an unbalanced design with fixed c_{REF} and c_{PLA} is given by the smallest n_{EXP} for which

$$n_{\text{EXP}} \geq \left(t_{1-\alpha}(\nu^{\text{het}}) - t_{\beta}(\nu^{\text{het}})\right)^2 \quad \frac{\sigma^2_{\text{EXP}} + \frac{\Delta^2}{c_{\text{REF}}}\sigma^2_{\text{REF}} + \frac{(1-\Delta)^2}{c_{\text{PLA}}}\sigma^2_{\text{PLA}}}{\left(\mu_{\text{EXP}} - \Delta\mu_{\text{REF}} - (1-\Delta)\mu_{\text{PLA}}\right)^2} \tag{10}$$

holds. The solution is derived iteratively. Refer to Hasler et al. (2008) for a detailed derivation of the sample size formula.

The sample size formula (10) requires the specification of the expected values μ_i and variance parameters σ^2_i for all three groups $i =$ EXP, REF, PLA.

The specification of the expected treatment effects is based on determining the clinically relevant effect between the experimental and reference treatments, each compared to placebo, in order for the experimental treatment to be considered non-inferior to the reference treatment. Thereby, it is common to plan the trial under the assumption that the experimental treatment and the reference treatment have the same treatment effect, that is $\mu_{\text{EXP}} = \mu_{\text{REF}}$. The treatment effect of the reference treatment compared to placebo is often determined based on information from previous studies and/or pre-clinical data. Consequently, specifying the expected treatment effects usually does not pose a challenge

during the planning stage.

In contrast, the specification of variances for the three groups is often subject to greater uncertainty. The magnitudes and interrelationships of the variances depend more on the specific characteristics and circumstances of the trial. Ideas about the variances in the reference and placebo groups, as well as their relationship to each other, may be informed by findings from previous studies. Determining the variance of the experimental group and its relationship to the variances of the reference and placebo groups, however, is typically more challenging and uncertain.

Failing to adequately specify nuisance parameters during sample size planning could potentially result in underpowered or overpowered trials. Underpowering occurs when the estimated sample size is insufficient to achieve the desired power level, resulting in a higher probability of failing to detect non-inferiority when it truly exists. On the other hand, overpowering occurs when the power level exceeds the desired one. In such cases, the estimated sample size is larger than necessary, raising concerns about exposing too many individuals to treatment and the potential wastage of resources. The subsequent analysis will therefore explore scenarios wherein the nuisance parameters, specifically the variance parameters for the three treatment groups, are either correctly specified or misspecified during the planning of a trial using the Hasler sample size formula.

3.4 Sample size re-estimation based on nuisance parameter estimates

To account for potential misspecifications of nuisance parameters when planning a trial, trial designs might allow that the initially planned sample size is adjusted while the trial is in progress. In order to estimate the nuisance parameter, one might use the data of an internal pilot study and adjust the final sample size according to the nuisance parameter estimates. This procedure is referred to as sample size re-estimation based on nuisance parameter estimates and distinguishes itself from re-estimation procedures that rely on treatment effect estimates.

The idea of re-estimating sample sizes based on data from internal pilot studies in the setting of clinical trials was introduced by Wittes and Brittain (1990). Nuisance parameters are estimated based on the internal pilot study data whereby the internal pilot study data is also used for the final statistical analysis. Various sample size re-estimating strategies were studied for the two-arm design for normally distributed endpoints (see Gould and Shih, 1992; Kieser and Friede, 2003; Friede and Kieser, 2013; Friede and Kieser, 2011a; Xing and Ganju, 2005; Glimm and Läuter, 2013).

Building upon these ideas, Mütze and Friede (2017) extended the methodology to three-arm trials and investigated the performance of the proposed methods for normally distributed data. However, their focus was primarily on methods under the absolute margin hypothesis for non-inferiority trials. Furthermore, their analysis predominantly

involved variance estimators that preserved the blinding of the pilot study data, limiting the examination to scenarios with homogeneous variances across the three groups. In this study, one of the methods proposed by Mütze and Friede (2017), specifically the blinded adjusted one-sample variance estimator, will be applied to the retention-of-effect hypothesis. Additionally, an unblinded estimator capable of considering heterogeneous variances across the three groups will be compared. These two procedures will be introduced and subsequently their performance will be assessed in simulation studies by means of power and type I error rate. The main questions to address in the following are:

1. Does the studentized permutation test reach the desired power level through planning with the Hasler formula in a fixed sample size design when nuisance parameters are correctly specified? How does this compare to the case when nuisance parameters are misspecified?

2. Is re-estimation better than a fixed sample size design in terms of attaining the desired power when nuisance parameters are either correctly specified or misspecified during the planning phase?

3. Among the considered variance estimators, which one is more effective in re-estimating sample sizes and achieving the desired power level? What explains their differing behaviour? How can the scenarios, where re-estimation fails to reach the desired power level, be explained?

4. In situations where the desired power level cannot be achieved, can the power performance of the re-estimation procedure be improved by inflating the re-estimated sample size using an inflation factor?

5. Do the sample size re-estimation procedures maintain the pre-specified type I error rate, which is a prerequisite from a regulatory point of view?

To our knowledge, a sample size re-estimation procedure for the retention-of-effect hypothesis based on the Hasler sample size formula has not yet been investigated. The ultimate goal of this work is to find a possible sample size re-estimating strategy when analysing with the studentized permutation test. The presented method, however, also represents a valid approach when analysing with the Hasler test.

Estimating the nuisance parameters Two variance estimators are considered for the estimation of the nuisance parameters based on internal pilot study data: the unblinded group-variance estimator and the blinded adjusted one-sample variance estimator. Denote n_1 as the size of the internal pilot study.

1. Unblinded group-variance estimator: The within-group variances $\sigma^2_{n_{1,i},i}$ can be estimated unbiased by the group-specific sample variances of the internal pilot data, that is

$$\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i} = \frac{1}{n_{1,i}-1}\sum_{k_i=1}^{n_{1,i}}(Y_{ik_i} - \bar{Y}_i)^2 \qquad i = \text{EXP, REF, PLA} \tag{11}$$

where $n_{1,i}$ are the respective sample sizes of the three groups $i = \mathrm{EXP}, \mathrm{REF}, \mathrm{PLA}$ in the internal pilot study and Y_{ik_i} the unblinded oberservations of the internal pilot study. The group estimates $\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}$ are then plugged in the Hasler sample size formula and the sample size is re-adjusted accordingly. The re-estimation procedure based on the unblinded group-variance estimates $\hat{\sigma}^2_{1,i,\mathrm{UG}_i}$ is abbreviated with UG in the following.

2. Blinded adjusted one-sample variance estimator: The blinded one-sample variance estimator (OS) by Kieser and Friede (2003) estimates a single nuisance parameter for all three groups by the sample variance of the blinded observations $Y_1, \ldots, Y_{n_1}$ of the internal pilot study, that is

$$\hat{\sigma}^2_{n_1,\mathrm{OS}} = \frac{1}{n_1-1}\sum_{k=1}^{n_1}(Y_k - \bar{Y})^2 \qquad k = 1, \ldots, n_1. \tag{12}$$

By ignoring the fact that the observations come from three groups, this estimator yields an estimate of the overall variance, that is the sum of the between-group variance and the within-group variances. If the group means differ from each other, this estimator, however, is biased. Mütze and Friede (2017) derived the bias of the one-sample variance estimator $\hat{\sigma}^2_{n_1,\mathrm{OS}}$ for three-arm designs under the absolute margin hypothesis. Denote μ_i as the group-specific means with $i = \text{EXP, REF, PLA}$ and $\bar{\mu}$ the weighted mean of the group means, that is, $\bar{\mu} = w_{1,\mathrm{EXP}} \cdot \mu_{\mathrm{EXP}} + w_{1,\mathrm{REF}} \cdot \mu_{\mathrm{REF}} + w_{1,\mathrm{PLA}} \cdot \mu_{\mathrm{PLA}}$ with $w_{1,i} = n_{1,i}/n_1$. The bias is then given by

$$\mathrm{Bias}(\hat{\sigma}^2_{n_1,\mathrm{OS}}, \sigma^2) = \frac{n_1}{n_1-1}\sum_i w_{1,i}(\mu_i - \bar{\mu})^2. \tag{13}$$

Refer to the Appendix of Mütze and Friede (2017) for a derivation of the bias. The bias remains the same under the retention-of-effect hypothesis as the considered models are equivalent. Gould and Shih (1992) then showed that an unbiased version of the one-sample variance estimator, denoted as $\hat{\sigma}^2_{n_1,\mathrm{OSU}}$, can be derived when one substracts the respective bias $\mathrm{Bias}(\hat{\sigma}^2_{n_1,\mathrm{OS}}, \sigma^2)$ from the one-sample variance estimator $\hat{\sigma}^2_{n_1,\mathrm{OS}}$, that is

$$\hat{\sigma}^2_{n_1,\mathrm{OSU}} = \hat{\sigma}^2_{n_1,\mathrm{OS}} - \mathrm{Bias}(\hat{\sigma}^2_{n_1,\mathrm{OS}}, \sigma^2).$$

In the following simulation, the adjusted version of the one-sample variance estima-

tor $\hat{\sigma}^2_{n_1,\text{OSU}}$ is employed and will be abbreviated by OSU. For this, the OS estimator is calculated based on the blinded observations in a first step. The bias is subsequently calculated based on the group-specific parameters μ_i^* that were assumed in the planning stage. The sample size is then re-estimated by plugging the estimate $\hat{\sigma}^2_{n_1,\text{OSU}}$ in the Hasler formula (10) for each of the three groups. However, when subtracting the bias from the one-sample variance estimator $\hat{\sigma}^2_{n_1,\text{OS}}$, it is possible that the resulting adjusted estimator becomes negative. In such instances, sample size re-estimation is not feasible. Instead, the one-sample variance estimator is employed without the adjustment. It is worth noting, however, that in the upcoming simulation study, this situation occurred in only 0.001% of all cases. It should also be noted that the adjusted version is only unbiased, if and only if, the assumptions on the group means μ_i^* are correct.

In two-arm trials, the blinded one-sample variance estimator proved to be the best nuisance parameter estimator for sample size re-estimation in meeting the target power (Friede and Kieser, 2013) while generally overpowering a trial in a three-arm trial design (Mütze and Friede, 2017).

It is important to note that the OSU estimator is limited to estimating a single nuisance parameter only in order to preserve the blinding of the internal pilot data. As a result, it is expected to perform well in scenarios with homogeneous variances and may be able to accommodate minor deviations from the homogeneity assumption. In contrast, the UG estimator provides variance estimates for each of the three treatment groups, making it a suitable choice for both homogeneous and heterogeneous variance settings. However, unblinding the data possibly introduces a bias to the study (Schulz and Grimes, 2002) and is generally not recommended by regulators (Committee for medicinal products for human use, 2007). Methods that maintain the blinding of the study should generally be preferred. For comparison, both methods will be applied to homogeneous and heterogeneous variance scenarios.

Sample size re-estimation procedure In the simulation, sample sizes are estimated according to the Hasler formula (10) for a desired power level of 80%. The re-estimation procedure then proceeds as follows: The variance terms in (10) are substituted by their estimates after reaching n_1, the internal pilot study sample size. The sample size is recalculated as in (10) and the newly estimated sample size is denoted as $\tilde{n}(x)$. For this simulation, it is assumed that the re-estimated sample size must be at least as large as the size of the internal pilot study n_1. Following the suggestion by Gould (1992), an upper limit for the final sample size is defined at 10,000 subjects to prevent excessively large sample sizes and reduce computational effort. Therefore, the re-estimated final sample size, denoted as $\tilde{n}(x)$, must satisfy

$$n_1 \leq \tilde{n}(x) \leq 10,000. \tag{14}$$

It should be noted that the variance estimation based on the pilot study data is not used in the final analysis of the data. Instead, the variance of the test statistic is estimated unblinded.

In order to adress the proposed Questions 1. to 5., the performance of the sample size re-estimation (SSR) based on the UG and OSU estimator will be assessed in simulation studies by means of power and type I error rate. The abbreviation SSR will be used to denote the employed sample size re-estimation procedure in the ensuing simulation. The scenarios for the simulation study are listed in Table 4. Column 2 displays the scenarios considered for examining the power performance, while column 3 presents the parameters utilised to investigate the type I error rate.

Table 4: Scenarios for the simulation study investigating the behaviour of the proposed sample size re-estimation procedure based on the UG and OSU estimator under the null and alternative hypothesis.

Parameter	Values under H_1	Values under H_0
Distributions	Normal, t, Lognormal, Chi-squared	
Non-inferiority margin Δ	0.8	
Ratio in the mean differences $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$	1	0.8
Group variances at planning stage $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}})$	$(1; 1; 1)$	
True group variances $(\sigma^{2}_{\text{EXP}}; \sigma^{2}_{\text{REF}}; \sigma^{2}_{\text{PLA}})$	$(1; 1; 1)$; $(3; 2; 1)$	
Sample size allocations $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	$(1 : 1 : 1)$; $(1 : \Delta : 1 - \Delta)$	
Internal pilot study size n_1	30,40, ... 120	
Target power $1 - \beta$	0.8	
One-sided nominal level α	0.025	
Permutation replications	10,000	
Simulation replications	5,000	

The Hasler sample size formula is derived based on the assumption of normality of the data. The studentized permutation test, however, also proved feasible in scenarios where data does not conform to normality (refer to Sections 3.2.1 and 3.2.2). Accounting for situations where the normality assumption may be violated, the simulation also covers

lognormal, t, and χ^2-distributed data, similar to the previous simulations. The non-inferiority margin Δ, again, is fixed at 0.8. Under the alternative hypothesis, the typical planning assumption of equal treatment effects between the experimental and reference treatment is employed, represented by $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}}) = 1$. Under the null hypothesis, the true effect under the hypothesis $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ is set to 0.8 while the planning is, again, made under the assumption of equal treatment effects, that is $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}}) = 1$. In the planning stage, homogeneous variances, that is $\sigma_i^{2*} = 1$, are assumed. Note that in previous simulations, the heterogeneous scenario was defined in terms of standard deviations (see Table 1). In order to reduce computational effort and simulate a more realistic situation, the heterogeneous scenario is now expressed in terms of variances. Two cases for the true variances are considered, the first one being that true variances coincide with the ones specified in the planning stage and a second one where the true variances deviate from the planning assumptions with $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$. Due to the liberal behaviour under the null hypothesis for skewed data and an increasing variance structure, only a decreasing variance scenario is considered. Our hypothesis is that the UG estimator would perform well under both variance settings, whereas the OSU estimator is expected to be less effective in a heterogeneous variance setting. Both estimators will be applied to both variance settings for comparison. Two sample size allocations are included, a balanced group design and the optimal allocation based on the assumption of $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$, that is $(1 : \Delta : 1 - \Delta)$. The required total sample sizes n based on the Hasler sample size formula (10) for a target power of 80% in a fixed design are 993 for the balanced group design and 787 for the unbalanced design, respectively. These calculations assume $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$.

The size of the internal pilot study is varied from 30 to 120 by steps of 10. It is expected that the power of the studentized permutation test will increase towards the desired power level with increasing internal pilot study data when sample sizes are re-estimated, since the precision of the variance estimation will improve. The data for the pilot study is generated with the same allocation as the final trial. Again, 10,000 permutation replications are used to obtain the rejection area of the studentized permutation test. Each scenario is replicated 5,000 times which results in a Monte Carlo error of roughly 0.006 for the targeted power level of 0.8. The scripts for conducting the simulation study in R are provided in Appendix A.

3.4.1 Question 1. Power in a fixed sample size design

The first question investigates how correct versus incorrect specification of variance parameters during the planning phase impacts the power of the studentized permutation test. Thus far, no procedure for re-estimating the sample size is considered. This means that the total sample size n remains fixed at 993 for the balanced group design and 787 for the unbalanced design. Data from the internal pilot study is not used for re-estimation.

Nevertheless, in order to facilitate comparison with the re-estimation procedures at a later point, the empirical power of the studentized permutation test is depicted in Figure 5 against the size of the internal pilot study n_1 on the x-axis. It should be noted, once again, that the pilot study data is not utilised, and therefore, the power is not anticipated to vary based on the pilot study size n_1. Again, the columns indicate the four underlying data-generating mechanisms and the rows display the two variance scenarios. The first row corresponds to the scenario where the variances are appropriately specified during the planning stage, with values of $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$. The second row, on the other hand, reflects the situation where the variances deviate from the planning assumptions, with values of $(\sigma^{2}_{\text{EXP}}; \sigma^{2}_{\text{REF}}; \sigma^{2}_{\text{PLA}}) = (3; 2; 1)$. The two allocation schemes are indicated by the two line types, that is the balanced design (dot-dashed line) and the optimal allocation of $(1 : \Delta : 1 - \Delta)$ (solid line). The boundaries of the area representing the targeted power level of 80% $\pm$ two times the Monte Carlo error, which is approximately 0.006, are demarcated by the two grey lines. Note that achieving the desired power level is understood in relation to the Monte Carlo error of this simulation study. Specifically, if the attained power level falls within the boundaries of $1 - \beta = 0.8 \pm$ two times the Monte Carlo error, it is considered that the desired power level of 80% is met. This terminology will be used consistently throughout the subsequent analysis.

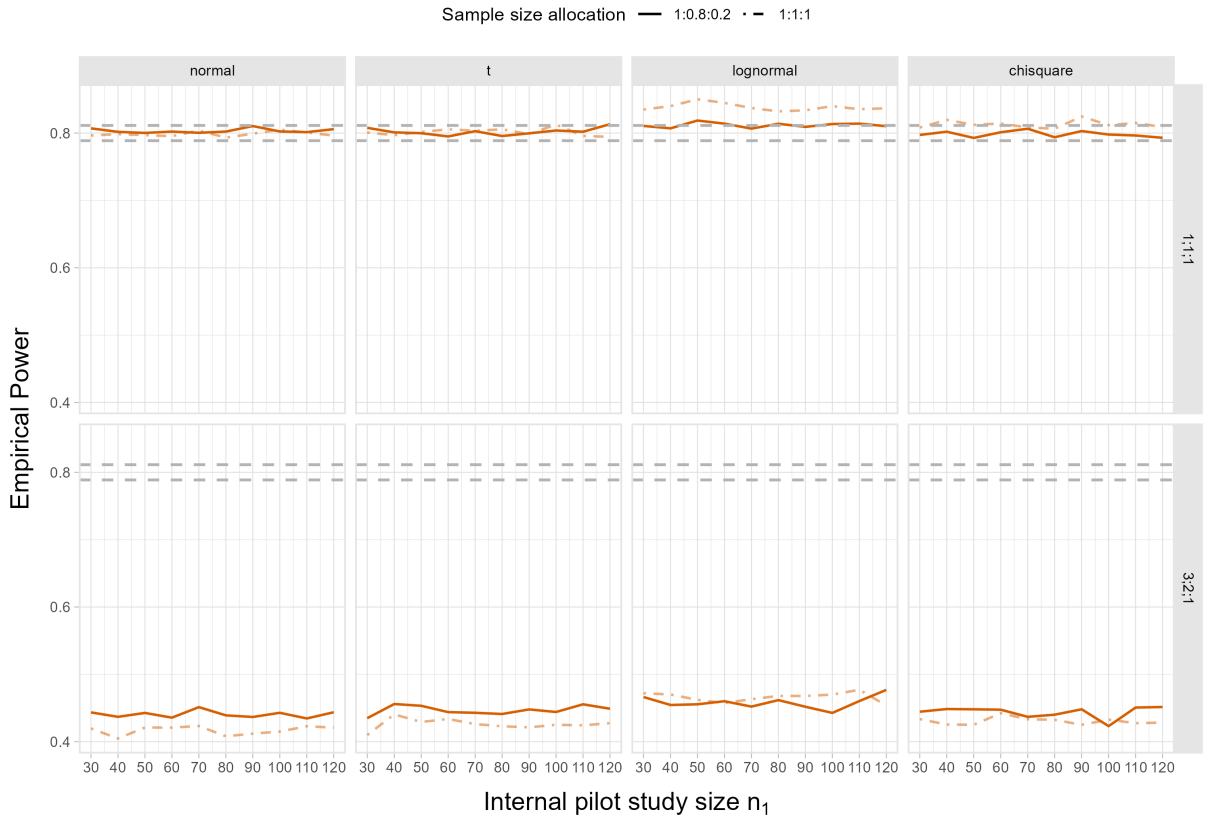


Figure 5: Observed power of the studentized permutation test in the fixed sample size design without sample size re-estimation. The dashed grey lines depict the area of $1 - \beta = 0.8 \pm$ two times the Monte Carlo error.

When variances are correctly specified in the planning stage, as shown in the first row, the power curves meet or even exceed the target area, as defined by the two grey lines, for all four examined data-generating mechanisms. For normal and t-distributed data the power curves for both allocation schemes hit the targeted region precisely, showing a similar performance between the two allocation ratios. In the case of lognormal data, both power curves exceed the targeted region. The χ^2-distributed data's power curve exceeds the targeted region when using a balanced design. Notably, the balanced design demonstrates higher power levels for both skewed data scenarios (The achieved power levels are displayed separately in Table S4 in the Appendix).

Since the power simulation results in Section 3.2.2 showed the studentized permutation test having a higher power level compared to the Hasler test for these data types (see Table 3), it was anticipated that the power exceeds the target level under lognormal and χ^2-distributed data. Still, the level of the difference is higher than expected. Specifically, this is the case for lognormal data. For a brief analysis of this phenomenon, please refer to the Appendix A.

Overall, when applying the studentized permutation test for the analysis of normal and t-distributed data, the sample sizes estimated using the Hasler formula in the fixed design result in achieving the desired power level. For lognormal and χ^2-distributed data, it is even possible to anticipate a power level higher than the targeted one, especially in the balanced design. However, it is important to note that these observations hold only true when the parameters, specifically the variance parameters, are correctly specified in the planning stage.

The question remains what happens to the power of the test when parameters are misspecified in the planning stage, specifically if there are misspecifications in the variance parameters. Consider row two of Figure 5 that shows the impact on the power of the studentized permutation test when the true underlying variances deviate from the planning stage by $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$. The power levels for all four data-generating mechanisms lie within the range of 0.4 to 0.5, indicating considerable underpowering (refer to Table S4 in the Appendix for the achieved power levels). For lognormal data, again, the power level is slightly elevated compared to the other data types, while χ^2-distributed data does not appear to differ significantly from normal and t-distributed data. Additionally, both normal and t-distributed data report a slightly greater power level in the unbalanced design while for both skewed data, no pattern between the allocation schemes is visible. Despite these variations, it is clear that the power curves fall far below the target area outlined by the two grey lines, emphasising a notable loss in power when variances are misspecified in the planning stage.

In summary, when variances are accurately specified in the planning stage, Figure 5 demonstrates that the studentized permutation test can meet or even exceed the target power level. For lognormal and χ^2-distributed data, using the Hasler sample size formula

in planning may even lead to higher power levels. Therefore, it can be said once again that the sample size estimation using the Hasler formula is highly effective when analysing with the studentized permutation test in terms of successfully reaching the desired power level. However, this holds only true when parameters are correctly specified during the planning phase. Under misspecification of the variance parameters Figure 5 showed that the desired power level cannot be reached. Particularly in this scenario, the trial is hugely underpowered with a power level of roughly 0.4 to 0.5. In such cases, the estimated sample size is not sufficient to adequately reach the desired power and a fixed design is not recommended. Rather, this motivates a procedure that estimates the sample variance while the trial is ongoing and re-adjusts sample sizes accordingly.

3.4.2 Question 2. Power with sample size re-estimation

Utilising the two proposed variance estimators, the second question of interest addresses a comparison between the power performance of the fixed sample size design and the design with sample size re-estimation. For the sample size re-estimation the unblinded group-variance estimator (UG) and the blinded adjusted one-sample variance estimator (OSU) are employed (refer to Section 3.4 for a derivation of the estimators). The goal is to determine if sample size re-estimation equals or surpasses the fixed sample size design in achieving the desired power level. After addressing this question, a discussion on the differences between the two estimators will follow in the next paragraph.

At this point, a distinct evaluation of the two variance scenarios is provided. In the first scenario, variances are rightly specified in the planning stage, with values of $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$. Figure 6 shows the empirical power on the y-axis against the varying internal pilot study size n_1. Again, the columns represent the four underlying data-generating mechanisms and the two allocation schemes are represented by two different line types.

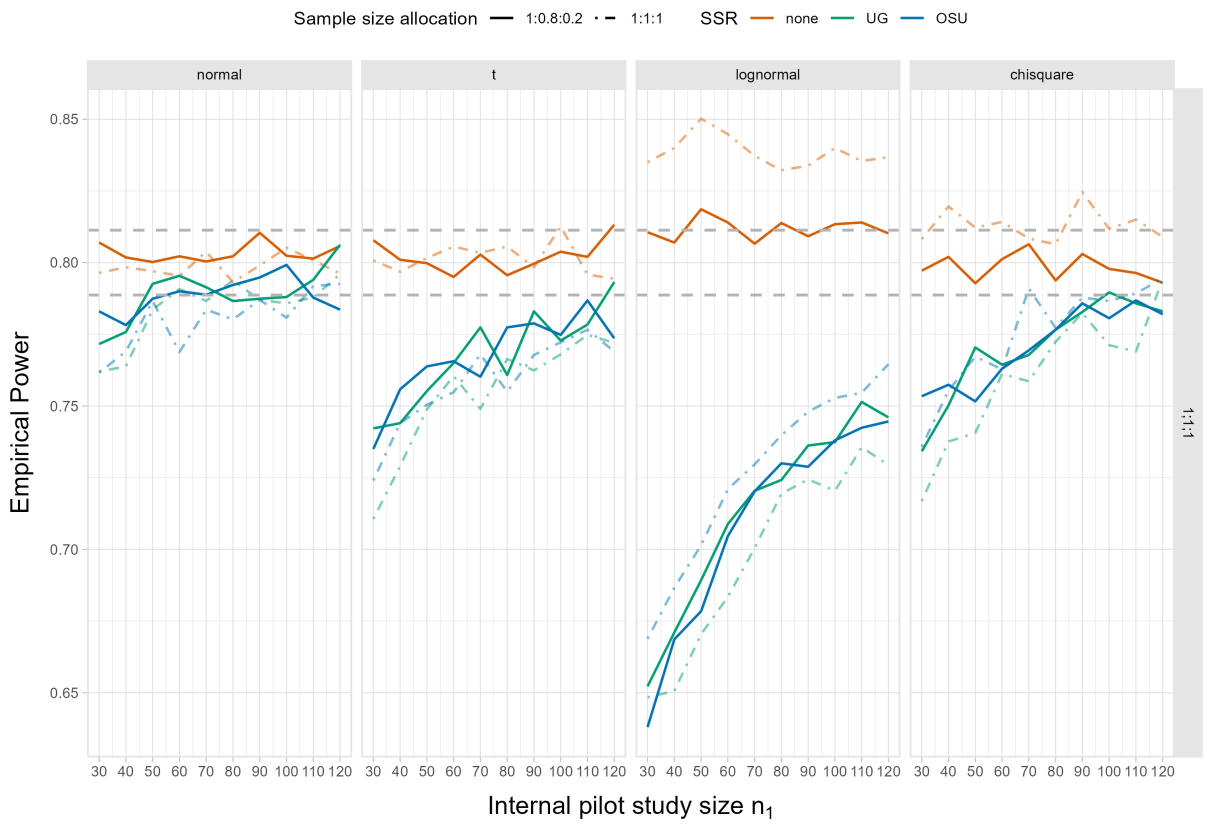


Figure 6: Observed power of the studentized permutation test without sample size re-estimation compared to the observed power with sample size re-estimation when variances are correctly specified in the planning stage with $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$ against the internal pilot study size n_1. The dashed grey lines depict the area of $1 - \beta = 0.8 \pm$ two times the Monte Carlo error.

The power curves represented by the orange lines demonstrate a fixed sample size design that does not employ any of the proposed variance estimators for sample size re-estimation. These two curves are identical to the ones in row 1 of Figure 5. The power curves represented by the two other colours show the power behaviour of the two re-estimation procedures. The green line illustrates the re-estimation procedure that uses the UG estimator, while the blue line illustrates the re-estimation procedure that uses the OSU estimator.

As mentioned above, the fixed sample size design has a constant power curve that meets or exceeds the desired power level of 80% across the internal pilot study size n_1. In contrast, the power curves of the two re-estimation procedures now increase as the internal pilot study size n_1 increases, but only reach the desired power level in a few cases. Specifically, the desired power level is achieved for normal data when n_1 is at least 60 using the UG estimator in the balanced group design and at least 50 in the unbalanced group design. When using the OSU estimator, re-estimation achieves the target level in the case of normal data for $n_1 \geq 110$ in the balanced design and $n_1 \geq 60$ in the unbalanced design. Both procedures achieve the desired power level in a few instances of non-normal

data (refer to Table S5 in the Appendix for the achieved power levels).

Thereby, both re-estimation procedures seem to attain similar power levels in the unbalanced design. For the UG estimator, the power curve of the unbalanced design is slightly elevated to the power curve in the balanced design across all scenarios, indicating a better performance in the unbalanced design. For the OSU estimator, a slightly improved power performance can be observed in the unbalanced design for normal and t-distributed outcomes, whereas the balanced design appears to perform better for lognormal and χ^2-distributed data. The difference between the allocation schemes, though, seems to be subordinate.

Nevertheless, both re-estimation procedures perform best for normal data where the highest power curves are achieved. However, for non-normal data (t, lognormal, and χ^2-distributed data), the attained power levels are noticeably lower than for normal data. In fact, the desired power level can only be achieved in a few instances (refer to Section 3.4.3 for a discussion on this phenomenon). This contrasts with the fixed sample size design, in which the desired power level is met across all considered data types (as depicted by the orange lines in Figure 6). For lognormal data, the disparity is particularly noticeable. The power curves of the re-estimation procedures are remarkably lower than for other data types and it fails to approach the desired level. For t- and χ^2-distributed data, the power curves behave similar, remaining slightly below the targeted area even for greater pilot study sizes.

At this point, it is crucial to note that the power curves of both re-estimation procedures are consistently below the power curve in the fixed sample size design. This indicates that, when variance parameters are correctly specified in the planning stage, the fixed sample size design is able to achieve the desired power level, whereas designs with sample size re-estimation do not. In fact, neither of the proposed re-estimation procedures can achieve the desired power level across all n_1. It is only met for greater n_1.

The second variance scenario, wherein $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$, presents a deviation of the variances from the planning assumption $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$. Figure 7 illustrates the power performance of the re-estimation procedures for this case.

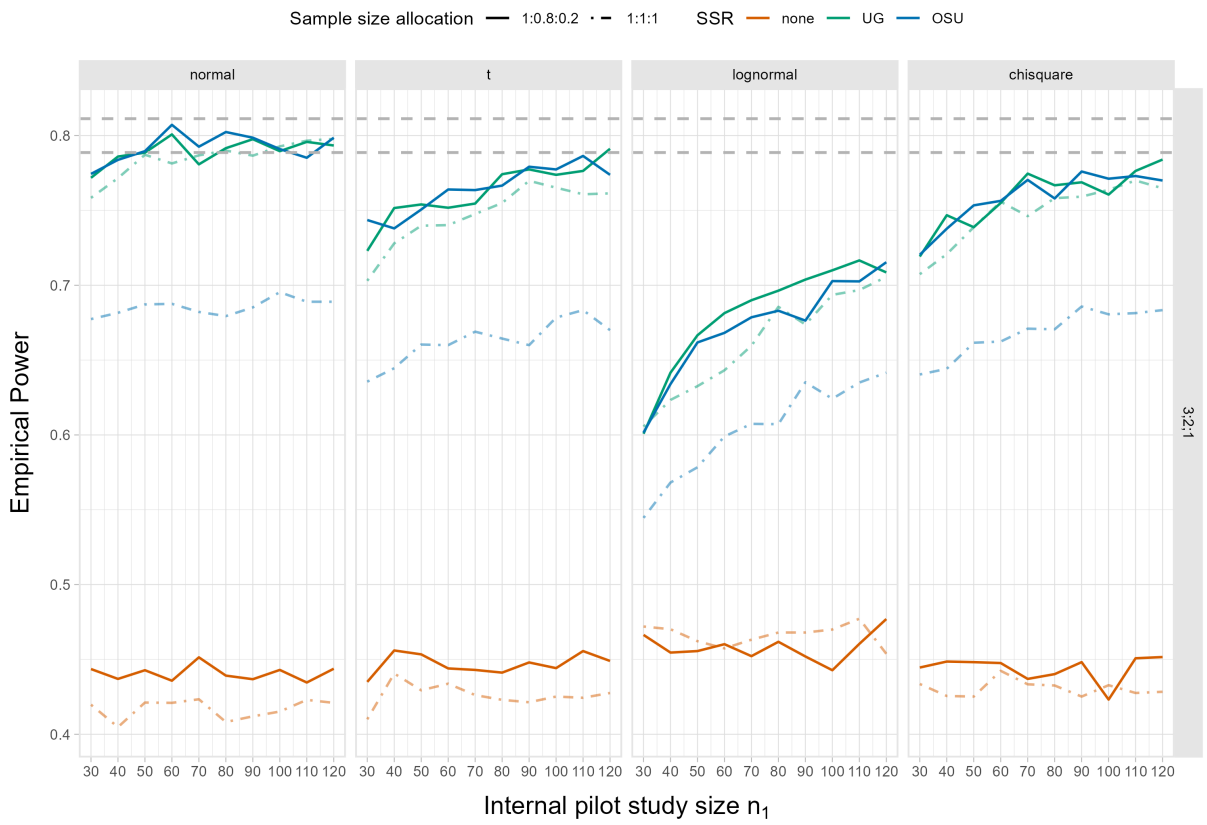


Figure 7: Observed power of the studentized permutation test without sample size re-estimation compared to the observed power with sample size re-estimation when variances deviate by $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$ from the planning assumption of $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$ against the internal pilot study size n_1. The dashed grey lines depict the area of $1 - \beta = 0.8 \pm$ two times the Monte Carlo error.

As illustrated by the orange lines, a power range between 0.4 and 0.5 is exhibited by the studentized permutation test in the fixed sample size design. As previously highlighted, the fixed sample size design shows considerable underpowering of the trial. The impact of re-estimation is demonstrated by the green and blue lines, which represent the power after re-estimation based on the UG and the OSU estimator, respectively.

In contrast to the previous variance scenario, the re-estimation procedures now exhibit higher power curves than the fixed sample size design for all four underlying data-generating mechanisms. This indicates a better power performance compared to the fixed sample size design. Specifically, both re-estimation procedures demonstrate similar power behaviours in the unbalanced design. However, in the balanced design, the power curve of the UG estimator is slightly lower compared to the unbalanced design, while the re-estimation procedure based on the OSU estimator shows a significant decline in performance. This suggests that the behaviour of the OSU estimator differs substantially between the balanced and unbalanced designs. An analysis of the OSU estimator's behaviour will be provided in the next paragraph.

The power curves of the re-estimation procedures increase again with increasing pilot

study size n_1. However, even with the maximum pilot study size considered, the desired power level cannot be achieved for t-distributed, lognormal, and χ^2-distributed data, except for one instance. In contrast, the target power is met in the case of normal data when $n_1 \geq 60$ in the unbalanced design and $n_1 \geq 80$ in the balanced design using the UG estimator, as well as when $n_1 \geq 50$ in the unbalanced design using the OSU estimator. Both t-distributed and χ^2-distributed data display power curves similar to those of normal data, albeit slightly lower. On the other hand, lognormal data shows a considerable deviation. As also illustrated in Figure 6, the attained power level for lognormal data falls far short of the targeted area. Refer to Tables S6 in the Appendix for the achieved power levels.

In short, two variance scenarios were examined. In the first scenario, variances were correctly specified in the planning stage. Figure 6 showed that designs with fixed sample sizes are able to reach the desired power level while the sample size re-estimation procedures only managed to reach it in a limited number of instances. This was the case for higher internal pilot study sizes n_1 under normal data. Therefore, re-estimation performs worse than the fixed sample size design in terms of attaining the target power level when variances are correctly specified in the planning stage. That means the re-estimation procedure cannot guarantee that the studentized permutation test will achieve the desired power level, unlike the fixed design. For that reason, a fixed sample size design should be preferred over re-estimation procedures when variances are specified under certainty.

In the second scenario, the true variances differed from the planning assumptions with heterogeneous variances. This resulted in a considerable underpowering issue with the fixed sample size design. The re-estimation procedures, using both the UG and OSU estimator, on the other hand, were able to effectively increase the power level. Although the OSU estimator falls short of reaching the target level in the balanced group design, the re-estimation procedures are both more effective than the fixed sample size design in terms of achieving greater power levels that are closer to the targeted one. Therefore, a design with sample size re-estimation is recommended over a fixed sample size design in case of uncertain variance specification. It should be noted though that neither of the proposed re-estimation procedures achieved the desired power of 80% for all pilot study sizes n_1 and all distributions of the data.

To better understand the varying power behaviour for both re-estimation procedures, as well as the instances in which the re-estimation procedures fail to achieve the desired power level, the following discussion will offer an analysis of the behaviour of both variance estimators.

3.4.3 Question 3. Comparison between the two sample size re-estimation procedures

In the next step, the effectiveness of the re-estimation procedures using the OSU estimator and the UG estimator will be compared. The simulation results showed that in the

first scenario with homogeneous variances, the re-estimation based on the OSU estimator showed comparable power to the UG estimator, with only minor differences observed in the balanced design. In the second scenario with heterogeneous variances, both procedures performed equally well in the unbalanced group design. However, the power performance of the OSU estimator fell significantly short of the UG estimator's performance in the balanced design.

Although the power behaviour of both re-estimation procedures might appear to be similar at first sight, the estimators exhibit distinct characteristics that are inherent to each of them. Therefore, the following paragraph will provide separate investigations of the behaviour of the OSU estimator and the UG estimator. Following that, a brief discussion comparing both re-estimation procedures will be provided, along with an exploration of the scenarios in which the estimators fail to achieve the desired power level.

Sample size re-estimation based on the OSU estimator We hypothesised earlier in Section 3.4 that the re-estimation based on the OSU estimator may not be as effective as the UG estimator in a heterogeneous variance scenario. Interestingly, in the simulation, the power performance of the OSU estimator was satisfactory in the unbalanced group design but showed inadequate power levels in the balanced design. To explain this discrepancy, the behaviour of the OSU estimator when used as a procedure for sample size re-estimation will be investigated.

To gain a comprehensive understanding of the OSU estimator's performance in the context of variance estimation, its estimates will be compared to the actual underlying pooled variance, which is given by

$$\sigma^2_{\text{pool}} = \sigma^2_{\text{EXP}} \cdot \frac{n_{\text{EXP}}}{n} + \sigma^2_{\text{REF}} \cdot \frac{n_{\text{REF}}}{n} + \sigma^2_{\text{PLA}} \cdot \frac{n_{\text{PLA}}}{n}.$$

In the homogeneous variance scenario, $\sigma^2_{\text{pool}} = 1$ for both allocation schemes. In the heterogeneous variance scenario, $\sigma^2_{\text{pool}} = 2.4$ in the unbalanced design and $\sigma^2_{\text{pool}} = 2$ in the balanced design. For illustration purposes, the variance estimation will be considered based on the OSU estimator for normal data with $n_1 = 120$. Figure 8 displays the density of the variance estimates obtained from the OSU estimator for both variance scenarios, as depicted by the rows. The two different line types represent the group designs, and the true pooled variance is shown as a grey line for each variance scenario and group design.

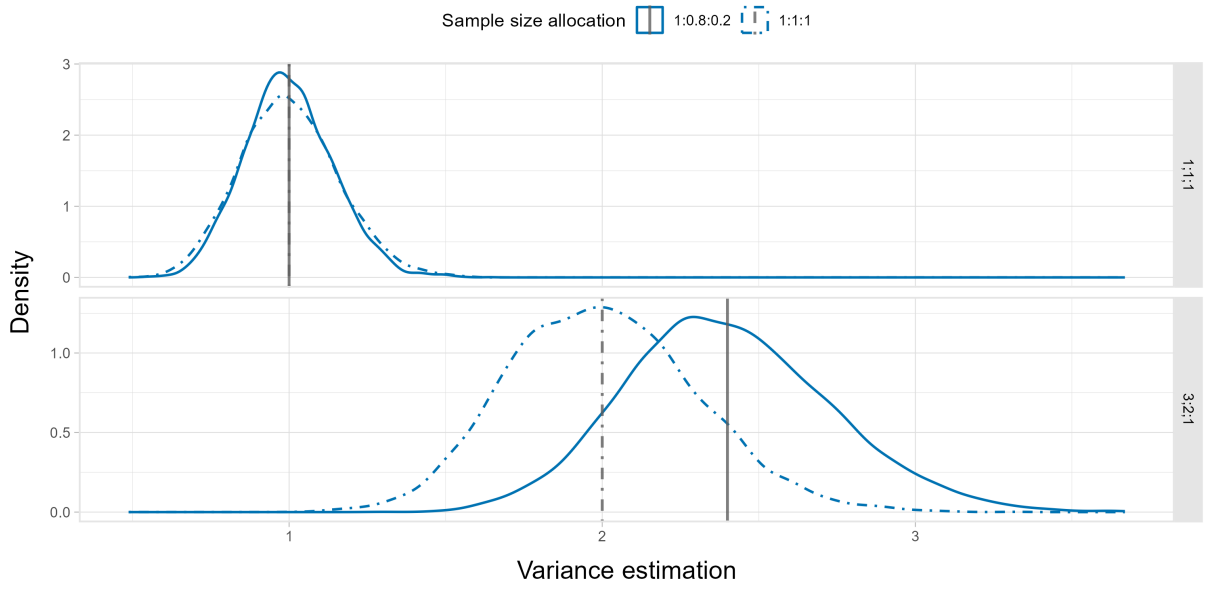


Figure 8: Density plot of the variance estimation based on the OSU estimator for data following a normal distribution with $n_1 = 120$. The grey lines depict the true underlying pooled variance.

The density plot clearly shows that for both variance scenarios and allocation schemes, the peak of the variance estimate aligns, though approximately, with the true underlying pooled variance. This indicates that, on average, the OSU estimator accurately estimates the true underlying pooled variance of the data, demonstrating its unbiasedness.

In the homogeneous variance scenario, the variance estimation in the balanced design shows greater tails compared to the estimation in the unbalanced design. This suggests that the variance estimation exhibits greater variation in the balanced design. The greater variability in the variance estimation, in turn, leads to more variability in the re-estimated sample sizes, explaining why the procedure based on the OSU estimator performed slightly worse in the balanced design compared to the unbalanced design (see Figure 6).

In contrast to the homogeneous variance scenario, the variance estimate in the heterogeneous variance scenario shows a broader range, indicating higher variability in the variance estimation. However, the broadness appears to be similar for both allocation schemes. Furthermore, in both allocation schemes of the heterogeneous setting, the OSU estimator, on average, provides an unbiased estimation of the true underlying pooled variance.

One might assume that being an unbiased estimator would result in the re-estimation of the required total sample size needed to achieve the desired power level. While this was true in the homogeneous variance setting, it was not the case in the heterogeneous setting with the balanced design (see Figure 7). Therefore, these findings do not explain the differing power behaviours observed between the two allocation schemes in the heterogeneous variance scenario. For a deeper understanding, it becomes essential to examine the re-estimated sample sizes that are derived from the OSU estimator.

By employing the Hasler sample size formula (10) as an approximation for the power of the studentized permutation test, it becomes possible to derive the required total sample sizes n needed to achieve a power of at least $1-\beta$ for both variance scenarios under both group designs as follows:

Table 5: Required total sample sizes n based on the Hasler sample size formula for the considered variance scenarios and group designs.

$(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}})$	$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	Required total sample size n
$1; 1; 1$	$(1:1:1)$	993
	$(1:\Delta:1-\Delta)$	787
$3; 2; 1$	$(1:1:1)$	2547
	$(1:\Delta:1-\Delta)$	1886

Figure 9 displays the density of the re-estimated final sample sizes for both variance scenarios and both allocation schemes for data following a normal distribution with $n_1 = 120$. The line types indicate the two group designs.

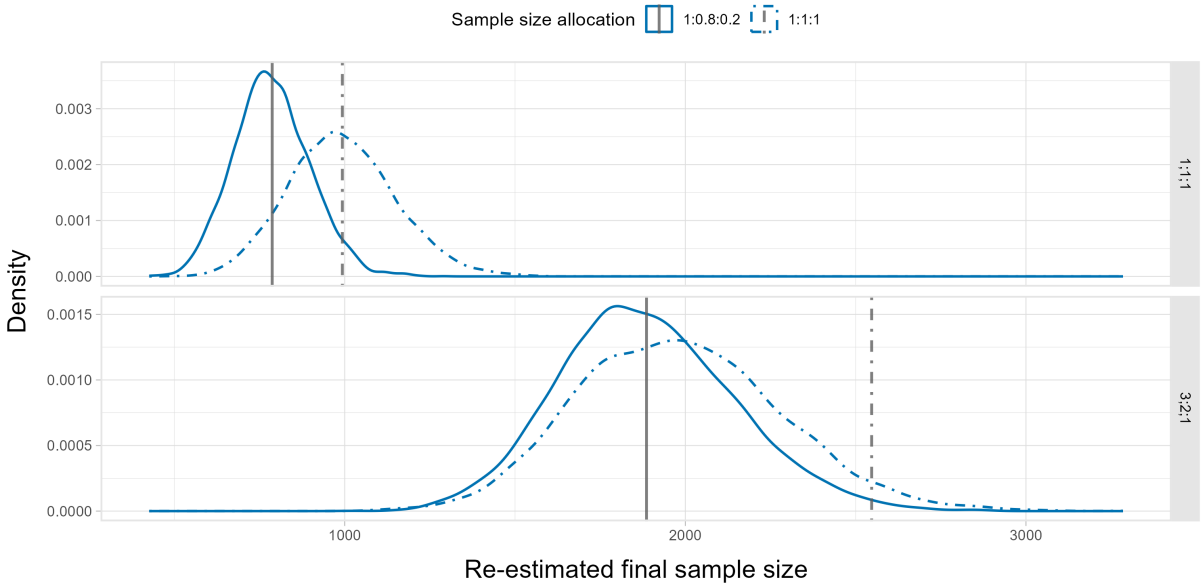


Figure 9: Density plot of the re-estimated final sample sizes based on the re-estimation using the OSU estimator for data following a normal distribution with $n_1 = 120$. The grey lines depict the required total sample size to attain a power of at least $1-\beta$ based on the Hasler sample size formula.

In the first scenario with homogeneous variances, the peak of the re-estimated final sample sizes under both group designs aligns with the required total sample size calculated using the Hasler sample size formula (see Table 5). It is also apparent that the re-estimated sample sizes exhibit greater variability in the balanced design, as evidenced by the wider

range depicted in the density plot. As briefly mentioned, the variance estimation of the balanced design demonstrated slightly heavier tails compared to the unbalanced design (see Figure 8). As a result, the higher power level in the unbalanced design, as shown in Figure 6, can be attributed to the fact that the variance estimation is denser around the true underlying pooled variance, indicating less variability and thus more precise sample size estimation. The slightly worse power performance in the balanced design can then be explained by the greater variability in the variance estimation.

In the second scenario of heterogeneous variances, the peak of the re-estimated sample sizes aligns closely with the required total sample size in the case of the unbalanced design. However, in the balanced design, the average re-estimated sample size falls considerably below the required total sample size. On average, the re-estimated sample size amounts to 1981.715, whereas the required total sample size is 2547 subjects. This renders the re-estimation procedure insufficient to achieve the desired power level (as seen in Figure 7).

As previously discussed, the OSU estimator is unbiased in estimating the true underlying pooled variance. In the heterogeneous variance scenario where the variances are $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$, the true underlying pooled variance is 2, i.e. $\sigma^2_{\text{pool}} = 2$. According to the Hasler sample size formula, the required total sample size for this variance scenario in the balanced group design is 2547 subjects (refer to Table 5). However, if one assumes homogeneous variances in the planning stage and provides the pooled variance estimate of 2, i.e. $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (2; 2; 2)$, the Hasler sample size formula now estimates a required total sample size of 1980.

Unsurprisingly, the average re-estimated sample size based on the OSU estimator coincides with 1980, which is significantly lower than the actual required total sample size of 2547. The underestimated sample size re-estimate is therefore insufficient to achieve the desired power level, as observed in Figure 7. This suggests that assuming homogeneous variances generally leads to inadequate re-estimated sample sizes to achieve the desired power. That is because the power of the studentized permutation test, as well as the Hasler test, is not solely determined by the pooled variance across the three groups, but rather by the variances specific to each group. A more detailed discussion on the impact of assuming homogeneous variances, when in reality the variances are heterogeneous, can be found in Hasler et al. (2008). Thus, the inability to achieve the target power in the balanced design with heterogeneous variances can be attributed to the fact that the OSU estimator fails to account for differing variances among the three groups, since it provides a single estimate only.

However, in the unbalanced group design, the re-estimation was able to achieve the desired power level. This can be explained as follows: In the unbalanced group design, the pooled variance of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$ is 2.4, denoted as $\sigma^2_{\text{pool}} = 2.4$. By providing this pooled variance estimate in the Hasler sample size formula, the required

total sample size is calculated as 1886. Surprisingly, this matches the required total sample size for $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$ (compare Table 5). However, this occurrence can be considered a coincidence. Calculating the required total sample size for a scenario of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 2; 3)$ results in a figure of 1258. Although the pooled variance remains the same, the required total sample size is significantly lower than the 1980 estimated by the OSU estimator. Therefore, in that scenario, a design with sample size re-estimation is expected to be overpowered.

Nevertheless, these results suggest that estimating a single nuisance parameter in unbalanced group designs and providing this estimate in the sample size formula might balance out the fact that planning with homogeneous variances does not lead to the actual required total sample size. Further research is needed to investigate the power behaviour when assuming homogeneous variances in the sample size re-estimation for unbalanced group designs. Specifically, the relationship between the group design and the pooled variance estimate should be further examined. It is of interest to determine whether blinded re-estimation based on the OSU estimator in unbalanced designs can generally achieve the desired power level when variances are heterogeneous.

In summary, it can be concluded that the OSU estimator performs well in homogeneous variance scenarios by providing an unbiased estimate for the variance, leading to the attainment of the desired power level particularly in the case of normal data. It should be noted that the re-estimation procedure still fails to attain the desired power level for non-normal data and smaller pilot study sizes n_1. The reader may refer to Mütze and Friede (2017) for a more detailed investigation of the re-estimation performance using the OSU estimator in homogeneous variance settings. However, in heterogeneous variance scenarios, the OSU estimator is unable to accurately re-estimate the sample size as it provides a single estimate only. As a result, it fails to achieve the desired power level. However, there is evidence suggesting that the estimates in unbalanced group designs result in re-estimated sample sizes that are closer to the actual required total sample sizes. Further research is needed to investigate the relationship between group design and pooled variance estimation in heterogeneous variance settings.

Sample size re-estimation based on the UG estimator Observations from the power simulation suggest that the re-estimation based on the UG estimator is able to achieve an adequate power level in both variance scenarios. However, a slight difference in power levels was observed between the unbalanced and balanced group designs, with higher power in the unbalanced design. To gain insights into the behaviour of the UG estimator when used for sample size re-estimation, an examination on the variance estimates obtained with the UG estimator will be conducted. As the UG estimator provides group-specific sample variances, it is now possible to examine the density of the variance estimate for each treatment group. For the purpose of illustration, the focus will be on the estimate

obtained with normal data for $n_1 = 120$. Figure 10 displays the density of the variance estimation by treatment group in the columns and the variance scenario in the rows. The two line types indicate the two allocation schemes.

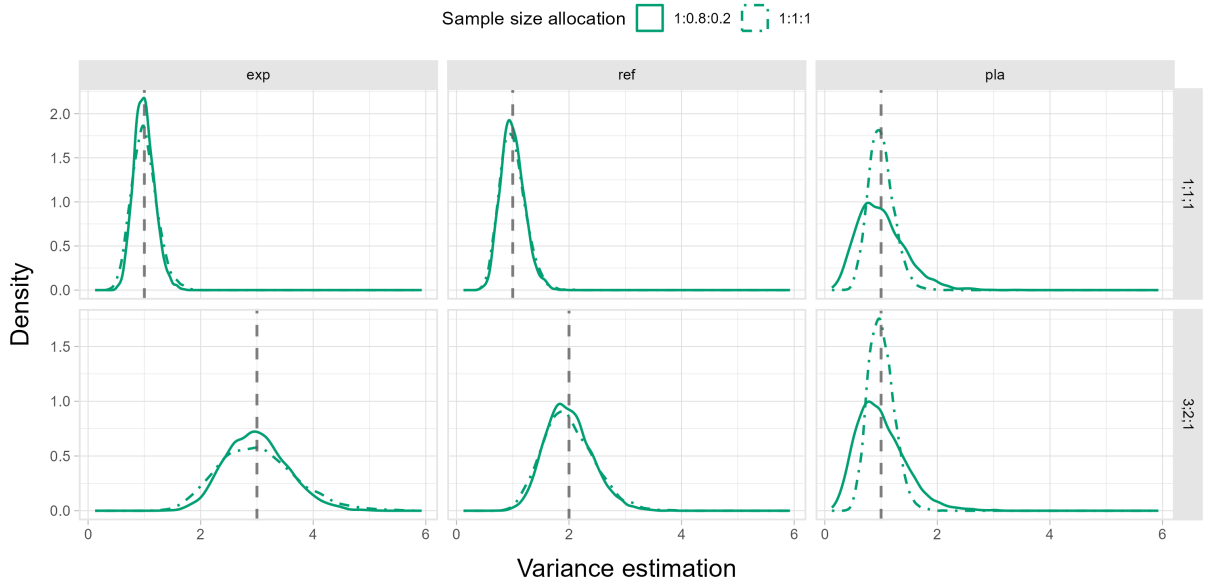


Figure 10: Density plot of the group-specific variance estimation based on the UG estimator for data following a normal distribution with $n_1 = 120$. The grey lines depict the true underlying group-variances.

Figure 10 reveals that the peak of the variance estimation coincides with the true underlying group variances, indicating that the UG estimator is an unbiased estimator for the group-specific variances. Consequently, it is also an unbiased estimator for the pooled variance. However, depending on the allocation scheme, the density of the variance estimate can be either wider or denser, indicating more or less variability in the estimate. For example, in the heterogeneous variance scenario, the variance estimate for the placebo group (row 2, column 3) is much more peaked compared to the experimental group (row 2, column 1). This suggests that with a smaller variance and lower proportion of the total sample size, a denser variance estimate is obtained.

In comparison to the OSU estimator, it can be concluded that both the UG estimator and the OSU estimator provide unbiased estimates of the pooled variance. However, the UG estimator has the additional advantage of also providing unbiased estimates of the group-specific variances. The re-estimated sample sizes based on the UG estimator are depicted in Figure 11 for normal data with a pilot study size of $n_1 = 120$. The required total sample sizes based on the Hasler sample size formula, as calculated in Table 5, to achieve a power of at least $1 - \beta$, are indicated by the grey lines.

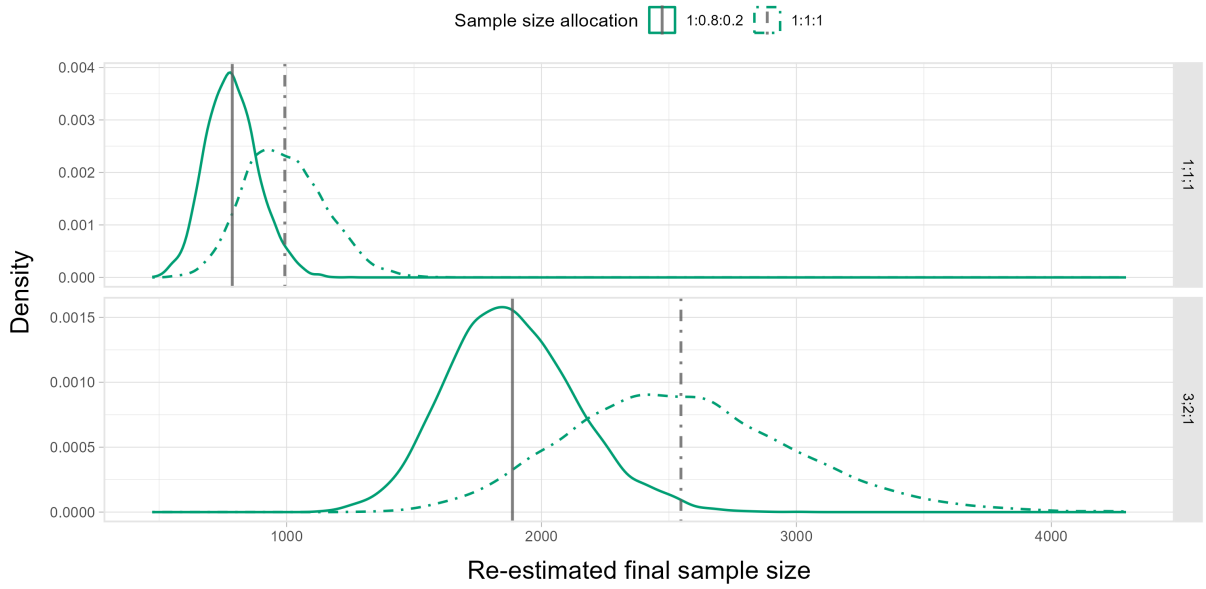


Figure 11: Density plot of the re-estimated final sample sizes based on the re-estimation using the UG estimator for data following a normal distribution with $n_1 = 120$. The grey lines depict the required total sample size to attain a power of at least $1 - \beta$ based on the Hasler sample size formula.

Other than for the OSU estimator, the peak of the re-estimated final sample sizes now aligns with the required total sample sizes across all scenarios. This suggests that, on average, the re-estimated sample size is accurately determined when the variances are specified based on the UG estimator.

Additionally, Figure 11 also helps explain why the observed power with the UG estimator in the unbalanced design was slightly higher than in the balanced design. The density curve of the unbalanced design is denser than that of the balanced design in both variance scenarios, indicating more variation in the re-estimated sample sizes for the balanced design. However, since the UG estimator provides an adequate power level in both group designs it can be considered a reasonable estimator for both allocation schemes, despite a slightly worse performance in the balanced design. Furthermore, it can be observed that the variation in the re-estimated sample sizes is greater in the heterogeneous variance scenario. However, this increased variation does not result in a worse power behaviour in the heterogeneous variance scenario.

Comparison between the two sample size re-estimation procedures To explain the behaviour of both re-estimation procedures, an investigation was conducted on the variance estimation and the resulting re-estimated sample sizes for both procedures. It was shown that both variance estimators provide unbiased estimates of the pooled variance, regardless of the variance scenario being homogeneous or heterogeneous. In the homogeneous scenario, the re-estimated sample sizes were accurately determined, resulting in both re-estimation procedures achieving adequate power levels that were approximately the same.

Compared to the fixed design, however, both re-estimation procedures are ineffective in attaining the desired power level consistently.

In the heterogeneous variance scenario, the OSU estimator failed to adequately re-estimate sample sizes, as it provided a single estimate for the variances only, resulting in a much lower power level in the balanced group design. In contrast, the UG estimator ensures an unbiased estimate of the group-specific variances. By providing group-specific variance estimates, the UG estimator is able to accurately re-estimate the required total sample size to achieve the desired power level in both variance settings, homogeneous and heterogeneous. This was not the case for the OSU estimator. Surprisingly, the average re-estimated sample size based on the OSU estimator in the unbalanced design, however, coincided with the required total sample size, thereby achieving adequate power levels. Still, the simulation results demonstrate that re-estimation based on both estimators can be recommended over a fixed sample size design when variances are misspecified in the planning stage

In summary, the OSU estimator and the UG estimator are unbiased estimators for the variance in the homogeneous variance setting, enabling them to achieve adequate power levels. However, in heterogeneous variance settings, the OSU estimator cannot guarantee an accurate re-estimation of sample sizes, while the UG estimator can. Therefore, for the considered scenarios, the UG estimator can be declared more effective in re-estimating sample sizes and reaching the target power level. Still, there remain instances where the desired power level cannot be achieved, particularly for non-normal data and smaller pilot study sizes n_1. In the subsequent analysis, an explanation for these cases will be provided by examining the behaviour of the UG estimator.

Scenarios where sample size re-estimation based on the UG estimator fails to reach the desired power level As discussed earlier, both re-estimation procedures fail to reach the target power level across all n_1 and underlying distributions of the data, particularly in case of non-normal data.

The behaviour for non-normal data may not come as a surprise, since planning with the Hasler formula is based on the assumption of normal data, and the presence of heavier tails and skewness in the data, as represented by the t, lognormal and χ^2-distributions, deviate from this assumption. However, in a fixed sample size design the studentized permutation test was able to attain the desired power level for these data types when parameters were correctly specified (refer to Figure 6). Hence, it can be concluded that the disparity in power performance for the re-estimation procedures across different data types cannot be attributed to the studentized permutation test performing inadequately for non-normal data. Instead, it appears that the reason for this difference lies within the re-estimation procedures themselves. To investigate this further, the focus will be on the variance estimation using the UG estimator. Specifically, the pooled variance estimate

obtained from the sample variances of the three groups will be considered, that is

$$\hat{\sigma}^2_{\text{pool}} = \frac{(n_{1,\text{EXP}} - 1)\hat{\sigma}^2_{n_{1,\text{EXP}},\text{UG}_{\text{EXP}}} + (n_{1,\text{REF}} - 1)\hat{\sigma}^2_{n_{1,\text{REF}},\text{UG}_{\text{REF}}} + (n_{1,\text{PLA}} - 1)\hat{\sigma}^2_{n_{1,\text{PLA}},\text{UG}_{\text{PLA}}}}{n_1 - 3}.$$

It holds true that the pooled variance estimator is unbiased for any underlying distribution. However, in order to comprehend the disparity in performance across the four data types, it becomes necessary to examine the distribution of the variance estimation. Figure 12 displays the density of the pooled variance estimation for the four different underlying data-generating mechanisms in the unbalanced design, with a fixed internal pilot study size of 120. The rows represent the underlying variance scenarios, while the four colours indicate the data-generating mechanisms. The dashed grey line indicates the true pooled variance, that is

$$\sigma^2_{\text{pool}} = \sigma^2_{\text{EXP}} \cdot \frac{n_{\text{EXP}}}{n} + \sigma^2_{\text{REF}} \cdot \frac{n_{\text{REF}}}{n} + \sigma^2_{\text{PLA}} \cdot \frac{n_{\text{PLA}}}{n} = 1$$

in the homogeneous scenario (Figure 12, row 1) and $\sigma^2_{\text{pool}} = 2.4$ in the heterogeneous scenario (Figure 12, row 2). Please note that the x-axis limits in the figure are manually set from 0 to 6 to enhance the visibility of the distribution of the variance estimation.

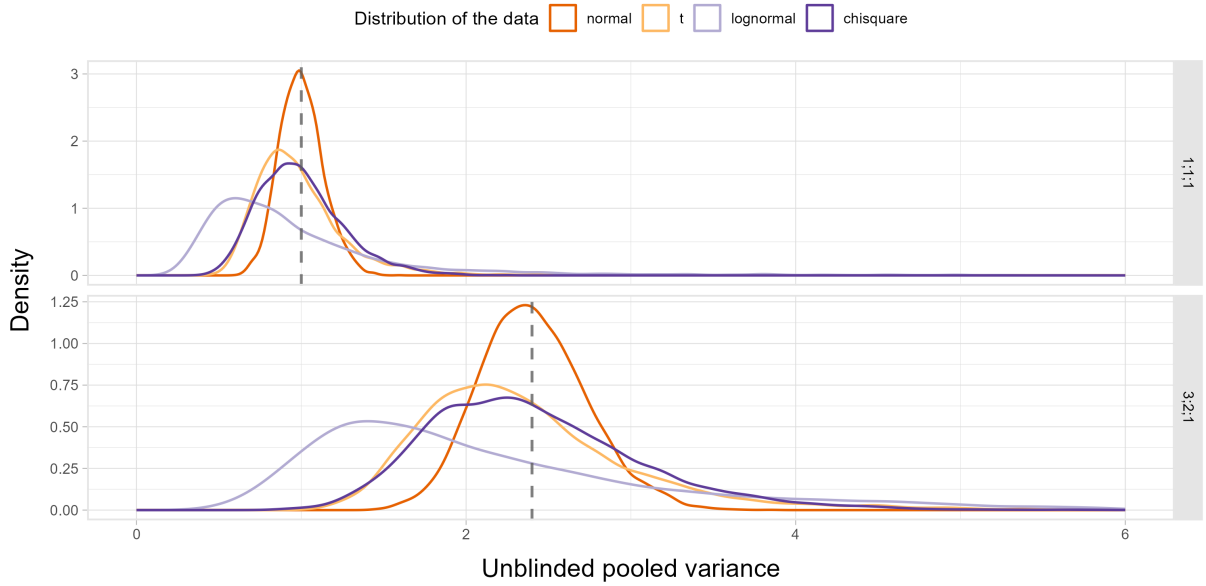


Figure 12: Density plot of the pooled variance estimation based on the UG estimator for data following a normal, $t(4)$, lognormal and $\chi^2(2)$-distribution with $n_1 = 120$ in the group design $(1 : \Delta : 1 - \Delta)$.

The distribution of the variance estimation for normal data is approximately symmetrical around the true value and displays a prominent peak (dark orange line). The distribution of the variance estimation for non-normal data, however, clearly differs from that, with a less pronounced peak, higher skewness, and more variability in the variance

estimation.

Recall that the UG variance estimate for each of the three groups is simply the sample variance of each group based on the internal pilot data. Therefore, $\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}$ can be represented as

$$\big((n_{1,i}-1)\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}\big)/\sigma_i^2$$

Assuming the data follows a normal distribution, it can be shown that this random variable follows a χ^2-distribution with $(n_{1,i}-1)$ degrees of freedom. As the value of n_1 increases, the χ^2-distribution starts to resemble the curve of a normal distribution more closely. Specifically, for $n_{1,i} > 100$, the χ^2-distribution approximates a normal distribution with mean $n_{1,i}$ and variance $2n_{1,i}$. This explains why the density of the variance estimation for normal data exhibits characteristics similar to a normal density curve in Figure 12.

For non-normal data, however, the distribution of $\big((n_{1,i}-1)\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}\big)/\sigma_i^2$ can deviate from the χ^2-distribution with $(n_{1,i}-1)$ degrees of freedom. The shape of the distribution then depends on the underlying distribution of the data, and it can exhibit heavier tails and/or more skewness. This is also evident in Figure 12. Specifically, the variance estimation for lognormal data (light purple line) differs significantly from that for normal data, exhibiting considerable skewness and greater variability. Consequently, the variance of the variance estimation is higher for non-normal data, resulting in greater variability of the re-estimated sample sizes and a higher probability of misestimating sample sizes to achieve the desired power level. This explains the poor performance of the re-estimation procedure for lognormal data. On the other hand, the density curves for t-distributed data (light orange) and χ^2-distributed data (dark purple line) behave similarly, explaining their comparable power behaviour when sample sizes are re-estimated. The symmetry observed in the variance estimation for χ^2-distributed data, however, is unexpected, given that it is a skewed distribution.

Thus, it can be inferred that re-estimation procedures based on variance estimation can result in greater deviation from the targeted power level due to the higher variability in the variance estimation for non-normal data. Nonetheless, the degree of this effect relies on the extent to which the non-normal data deviates from the characteristics of normal data. In this simulation, the re-estimation procedures fail to perform well for lognormal data (see Figures 6 and 7). Consequently, lognormal data will no longer be considered for the sample size re-estimation procedures in the following. For t- and χ^2-distributed data, the power levels come at least close to the target level, although not reaching it entirely.

Next, the power simulation also revealed that the desired power level is often met only for larger pilot study sizes. For illustrative purposes, the behaviour of the UG estimator for normal data will be examined once again. In this case, the effect of increasing pilot study sizes on the re-estimated final sample sizes will be explored. Figure 13 shows the variation in the re-estimated final sample sizes based on the UG estimator, represented by

the interquartile range (IQR) in the unbalanced group design. The required total sample size is indicated by the dashed grey lines.

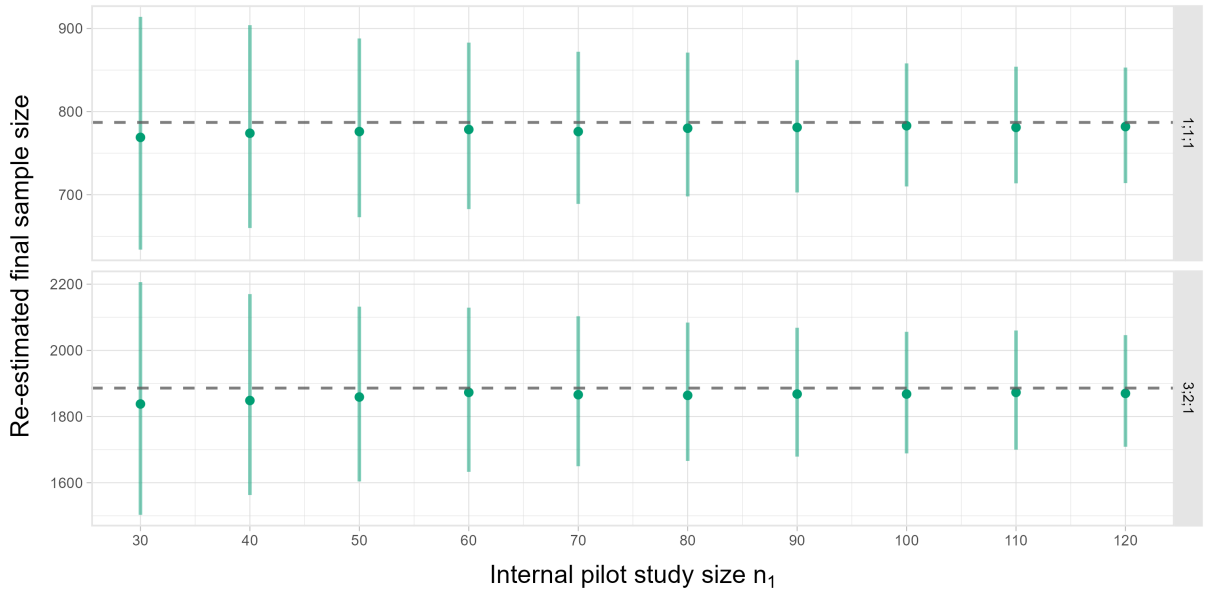


Figure 13: Median and interquartile range of the distribution of the re-estimated final sample sizes based on the UG estimator against the internal pilot study size n_1 for data following a normal distribution in the group design $(1 : \Delta : 1 - \Delta)$. The dashed grey lines depict the required total sample size based on the Hasler sample size formula.

Figure 13 demonstrates that as the size of the pilot study n_1 increases, the variability in the re-estimated sample sizes $\tilde{n}(x)$ decreases. The median of the re-estimated sample sizes becomes closer to the required total sample size and the IQR decreases. The density of the re-estimated sample sizes therefore becomes denser around the required total sample size, indicating more accurate estimation.

The variation in the re-estimated sample sizes again can be explained by the underlying distribution of the variance estimation. As mentioned earlier, it can be shown that $\big((n_{1,i}-1)\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}\big)/\sigma_i^2$ follows a χ^2-distribution with $(n_{1,i}-1)$ degrees of freedom under normality of the data. From this, it is easy to see that the accuracy of the variance prediction depends on the pilot study size $n_{1,i}$. That is because for a normal population, it holds that

$$\mathrm{Var}(\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}) = \frac{2\sigma^4}{n_{1,i}-1}.$$

Consequently, as $n_{1,i}$ increases, the variance of the sample variances decreases. This is because the variance of $\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}$ is inversely proportional to $n_{1,i}-1$, so a larger pilot study size will lead to a smaller variance of the sample variance.

However, as the pilot study size increases, the variance estimate becomes more precise, resulting in more accurate re-estimated final sample sizes, as depicted in Figure 13. This improved accuracy enables the attainment of the desired power level for larger pilot study sizes (refer to Figures 6 and 7).

To sum up, it can be concluded that for both scenarios in which the sample size re-estimation procedure fails to attain the desired power level, the underlying cause can be attributed to the distribution of the variance estimation. In the case of non-normal data, the variance estimation exhibits greater variability, leading to increased variability when recalculating sample sizes. Furthermore, for smaller pilot study sizes, the variance estimation is less precise compared to larger pilot study sizes, owing to its properties as a scaled χ^2-distributed random variable. This highlights the need for approaches that can enhance the power performance. In the following, an approach will be explored to adjust the sample size re-estimation procedure by incorporating the uncertainty associated with variance estimation.

3.4.4 Question 4. Improvement of power performance by inflating the re-estimated sample sizes

The simulation revealed that the re-estimation procedures consistently fell short of the desired power level of 80% for small pilot study sizes and non-normal data. This failure can be attributed to the greater variability in the prediction of the true variance by the variance estimate (see Section 3.4.3), which in turn leads to inaccurate sample size estimation. Essentially, the sample size re-estimation procedure does not effectively account for the uncertainty associated with the variance estimation, resulting in a failure to reach the desired power level across all pilot study sizes and underlying distributions of the data.

In order to address the limitations observed in the power performance, the focus will be on improving the re-estimation procedure based on the UG estimator. This choice is based on our previous findings, which showed that the UG estimator was the most suitable option for the scenarios considered in this study.

One such approach, suggested by Zucker et al. (1999), involves adjusting the sample size formula to account for the uncertainty associated with the variance estimation. In this approach, the re-estimated sample size $\tilde{n}$ is increased by an inflation factor ζ to ensure that the desired power level is achieved. To determine this inflation factor, the assumption is made that the power of the studentized permutation test can be approximated by that of the Hasler test. Therefore, the power function of the Hasler test will be used as a basis to derive the inflation factor.

It is worth noting that another possible suggestion to ensure that the desired power level is met is to simply increase the pilot study size n_1, as the simulation study results indicated that increasing n_1 brings the power closer to the target level. However, this approach has two potential drawbacks. Firstly, it remains unclear how large the pilot study size must be in order to achieve the target power level with certainty. Secondly, in scenarios where the internal pilot study size exceeds the actually required total sample size to meet the target power, it would result in a waste of resources.

In the following, the proposed method by Zucker et al. (1999) will be introduced by

deriving the inflation factor for the two-arm design, noting that the derivation assumes homogeneous variances. The concept will then be extended to a three-arm design with heterogeneous variances for the UG estimator, and the resulting power behaviour of the re-estimation will be evaluated through a simulation study.

Sample size inflation factor for the two-arm design with homogeneous variances as proposed by Zucker et al. (1999) Let $B(n)$ denote the power function of the test under consideration, where the power depends on the total sample size n. In order to achieve the desired power level, the total sample size is chosen such that the power is equal to the desired level. In a homogeneous variance setting, the total sample size n is determined as a function of the nuisance parameter σ^2, that is $n(\sigma^2)$. This implies that the sample size n varies depending on the value of σ^2, and hence the power is dependent on σ^2.

Consider a two-arm non-inferiority trial, consisting of an experimental and reference arm, that is tested using the two-sample t-test, assuming homogeneous variances. To obtain a power of at least $1-\beta$, the total sample size n can be approximated by the smallest n that satisfies

$$n = \frac{1}{r(1-r)} \frac{\left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)\right)}{\Delta^* - \delta} \sigma^2 \tag{15}$$

where r denotes the allocation ratio between the two groups, $\Phi^{-1}(1-\alpha)$ and $\Phi^{-1}(\beta)$ the $(1-\alpha)$th and βth percentile of the standard normal distribution, σ^2 the variance and $\Delta^* - \delta$ the difference between the assumed treatment group difference Δ^* and the non-inferiority margin δ with $\delta < 0$.

In a fixed sample size design, the required total sample size n to achieve the desired power level is determined by using a pre-specified value of σ^{2*} in the sample size formula (15). To validate the assumption regarding this nuisance parameter, one approach, as proposed in this work, is to estimate the variance using data from an internal pilot study, obtaining an estimate denoted as $\hat{\sigma}^2_{n_1}$. At this stage, it is important to note that $\hat{\sigma}^2_{n_1}$ is considered as an arbitrary variance estimator of the variance of the internal pilot study.

Subsequently, the sample size is re-estimated based on this estimate, denoted as $\tilde{n}(x)$ with $n_1 \leq \tilde{n}(x) \leq 10,000$. However, the sample size formula does not account for the uncertainty associated with the variance estimate $\hat{\sigma}^2_{n_1}$. Hence, the re-estimation cannot generally guarantee that the desired power is met.

In fact, the actual power of the re-estimation is no longer solely determined by $B(n)$, but instead depends on the properties of the variance estimator $\hat{\sigma}^2_{n_1}$ as well. Denote $f_{\hat{\sigma}^2_{n_1}}(\cdot)$ as the density of the nuisance parameter estimator $\hat{\sigma}^2_{n_1}$. Taking the uncertainty of $\hat{\sigma}^2_{n_1}$ into account, the power of the design with sample size re-estimation can now be approximated

by

$$\text{Power} \approx \int_0^\infty B\big(\tilde{n}(x)\big) f_{\hat{\sigma}^2_{n_1}}(x)\mathrm{d}x.$$

The power based on the re-estimated sample size $B\big(\tilde{n}(x)\big)$ is therefore weighted by the density of the nuisance parameter estimator $f_{\hat{\sigma}^2_{n_1}}(x)$.

We now wish to determine an inflation factor ζ that guarantees that the desired power level is met. For that reason, the re-estimated sample size $\tilde{n}(x)$ is inflated by an inflation factor ζ with $\zeta \in (0, \infty)$. The inflated re-estimated sample size based on the variance estimate $\hat{\sigma}^2_{n_1}$ is given by $\tilde{n}_\zeta(x) = \zeta \cdot \tilde{n}(x)$. Again, a lower and upper boundary is defined for the inflated re-estimated sample size with

$$n_1 \leq \tilde{n}_\zeta(x) \leq 10,000.$$

The power of the design with inflated re-estimated sample size $\tilde{n}_\zeta(x)$ can then be approximated by

$$\text{Power} \approx \int_0^\infty B\big(\tilde{n}_\zeta(x)\big) f_{\hat{\sigma}^2_{n_1}}(x)\mathrm{d}x. \tag{16}$$

Setting (16) equal to $1 - \beta$, that is

$$\text{Power} \approx \int_0^\infty B\big(\tilde{n}_\zeta(x)\big) f_{\hat{\sigma}^2_{n_1}}(x)\mathrm{d}x = 1 - \beta \tag{17}$$

and solving for ζ yields the solution for the inflation factor. For this, the density of the variance estimator $f_{\hat{\sigma}^2_{n_1}}(x)$ needs to be derived. For the unblinded pooled variance estimate in the two-arm design, that is

$$\hat{\sigma}^2_{\text{pool}} = \frac{(n_{1,\text{EXP}} - 1)\hat{\sigma}_{\text{EXP}} + (n_{1,\text{REF}} - 1)\hat{\sigma}_{\text{REF}}}{n_1 - 2},$$

with the group-specific sample variances $\hat{\sigma}_{\text{EXP}}$ and $\hat{\sigma}_{\text{REF}}$, the approximated power function (16) can be rearranged such that the density $f_{\hat{\sigma}^2_{\text{pool}}}$ can be substituted by the density $f_{\hat{\sigma}^2_{\text{pool}}/\sigma^2}$, which is given by a scaled χ^2-density with $2(n_1 - 1)$ degrees of freedom (Zucker et al., 1999). The known formula for $f_{\hat{\sigma}^2_{\text{pool}}/\sigma^2}$ can then be substituted in (17) and the inflation factor ζ can be obtained with

$$\zeta = \left[\frac{t_{1-\alpha}(n_1 - 2) + t_\beta(n_1 - 2)}{\Phi^{-1}(1-\alpha) + \Phi^{-1}(\beta)}\right]^2 \tag{18}$$

where $t_{1-\alpha}$ and t_β denote the $(1-\alpha)$th and βth percentile of the distribution function of the central t-distribution with $(n_1 - 2)$ degrees of freedom, and $\Phi^{-1}(1-\alpha)$ the $(1-\alpha)$th

and $\Phi^{-1}(\beta)$ the βth percentile of the standard normal distribution. The inflation factor is a ratio between the quantiles of the central t-distribution with degrees of freedom that are dependent on the size of the internal pilot study size n_1 and the quantiles of the standard normal distribution which represent the quantiles if one had planned the trial in a fixed sample size design with (15).

In equation (18), the denominator remains constant for all pilot study sizes n_1, while the numerator changes. Specifically, for smaller pilot study sizes n_1, the quantiles of the corresponding t-distribution are larger than those for larger pilot study sizes. As a result, the numerator is larger for small pilot study sizes and smaller for larger pilot study sizes. This implies that the inflation factor ζ is greater for smaller pilot study sizes and smaller for larger pilot study sizes. Consequently, the re-estimated sample size is more inflated for smaller internal pilot study sizes than for larger pilot study sizes. Expectedly, this leads to an increase in power levels, particularly for smaller pilot study sizes.

A sample size inflation factor for the three-arm design with heterogeneous variances based on the UG estimator In a heterogeneous variance setting, the sample size n is now a function of the three group variances σ_i^2 with $i = \text{EXP}, \text{REF}, \text{PLA}$, that is $n(\sigma^2_{\text{EXP}}, \sigma^2_{\text{REF}}, \sigma^2_{\text{PLA}})$. To obtain a power of at least $1 - \beta$, the sample size of the experimental group is given by the smallest n_{EXP} that satisfies (10) for fixed group ratios of the reference and placebo group. The total required sample size is then given by $n = n_{\text{EXP}} + n_{\text{REF}} + n_{\text{PLA}}$.

In a fixed sample size design, the total sample size n that is needed to attain the desired power level is now determined by specifying values σ^{2*}_{EXP}, σ^{2*}_{REF} and σ^{2*}_{PLA} in the Hasler sample size formula (10). The estimation of the variances of the three groups $i = \text{EXP}, \text{REF}, \text{PLA}$ was proposed using the UG estimator $\hat{\sigma}^2_{n_{1,i},\text{UG}_i}$, which corresponds to the sample variance of each treatment group. Based on these estimates, the sample size is re-estimated with (10), denoted as $\tilde{n}(x)$ with $n_1 \leq \tilde{n}(x) \leq 10,000$. Again, however, the uncertainty associated with the group-specific variance estimates $\hat{\sigma}^2_{n_{1,i},\text{UG}_i}$ is not taken into account in the re-estimation of the sample size. Consequently, it was observed in the simulation study that the re-estimation does not guarantee the attainment of the desired power across all scenarios (see Figure 6 and 7).

Similar to the two-arm design, the objective is to inflate the re-estimated sample size $\tilde{n}(x)$ by a factor ζ with $\tilde{n}_\zeta(x) = \zeta \cdot \tilde{n}(x)$ that ensures that the power of the re-estimation equals at least $1 - \beta$. Based on the UG estimates, the approximated power with the inflated re-estimated sample size is now given by

$$\text{Power} \approx \int_0^\infty B(\tilde{n}_\zeta(x) f_{\hat{\sigma}^2_{n_{1,i},\text{UG}_i}}(x) \text{d}x = 1 - \beta. \tag{19}$$

In order to solve this equation for ζ, the density of the UG estimator $f_{\hat{\sigma}^2_{n_{1,i},\text{UG}_i}}(x)$ needs

to be derived. Considering that the UG estimator $\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}$ is equivalent to the sample variance, it can be represented as $\big((n_{1,i}-1)\hat{\sigma}^2_{n_{1,i},\mathrm{UG}_i}\big)/\sigma_i^2$, which follows a χ^2-distribution with $(n_{1,i}-1)$ degrees of freedom. Therefore, $\big(\hat{\sigma}^2_{n_1,\mathrm{UG}_i}\big)/\sigma_i^2$ follows a scaled χ^2-distribution with $2(n_{1,i}-1)$ degrees of freedom. Given the similarity to the pooled variance estimator as discussed by Zucker et al. (1999), the inflation factor can be directly applied as indicated in equation (18) to the three-arm design, accounting for the quantiles given in the Hasler sample size formula (10). It is important to note that a direct solution for ζ in equation (17) was not obtained. Instead, the underlying structure of equation (18) was transferred to the three-arm design.

Following the structure of (18), the inflation factor ζ for the three-arm design using the Hasler sample size formula may be given by

$$\zeta = \left[\frac{(t_{1-\alpha}(\nu^{het}_{n_1}) + t_{\beta}(\nu^{het}_{n_1}))}{(t_{1-\alpha}(\nu^{het}_{\tilde{n}(x)}) + t_{\beta}(\nu^{het}_{\tilde{n}(x)}))}\right]^2. \tag{20}$$

Here, the quantiles of the central t-distribution with ν^{het} degrees of freedom are used on both the numerator and denominator of the term. The degrees of freedom ν^{het} are defined as in equation (9). The degrees of freedom ν^{het} in the numerator of the fraction are dependent on the pilot study size n_1, while those in the denominator are dependent on the re-estimated sample size $\tilde{n}(x)$. In contrast to the two-sample design, the denominator of the inflation factor for the three-arm design is not based on the quantiles of the standard normal distribution, but rather on the quantiles of the central t-distribution. The reason for this is that the Hasler sample size formula (10) requires the use of the t-distribution. Therefore, the denominator in (20) varies with the re-estimated sample size $\tilde{n}(x)$ instead of being constant. This was done to ensure that the denominator in the inflation factor for the three-arm design corresponds to the quantiles used for planning, as if the planning was carried out in a fixed design with the re-estimated sample size $\tilde{n}(x)$, similar to (18) for the two-sample design. It is possible to argue that approximating the denominator using the quantiles of the standard normal distribution would be sufficient, given that $\tilde{n}(x)$ is expected to be relatively large. However, to ensure precision and accuracy even for smaller re-estimated sample sizes, this approach is not adopted. Despite the now varying denominator, the anticipated effect of the inflation factor remains the same as in the two-sample design. Specifically, smaller pilot study sizes will result in a higher ratio, leading to a greater inflation factor and an increased power level, enhancing the power level particularly for smaller pilot study sizes. As before, the adjusted final sample size using the inflation factor is given by $n_1 \leq \tilde{n}_{\zeta}(x) \leq 10,000$.

The impact of the sample size inflation factor on the power of the studentized permutation test In a power simulation, the impact of the inflation factor on the attained power of the studentized permutation test when sample sizes are re-estimated is inves-

tigated. The simulation follows a similar structure to the power simulation conducted under both re-estimation procedures, as described in column 1 of Table 4. However, this simulation focuses solely on the re-estimation based on the UG estimator and adjusts the re-estimated sample sizes by the corresponding inflation factor.

Figure 14 displays the power curves for the re-estimation based on the UG estimator in both variance scenarios, as indicated by the rows. The transparent curves represent the power curves without the use of the inflation factor, while the solid curves represent the observed power curves with the inflation factor. The two line types differentiate between the group designs.

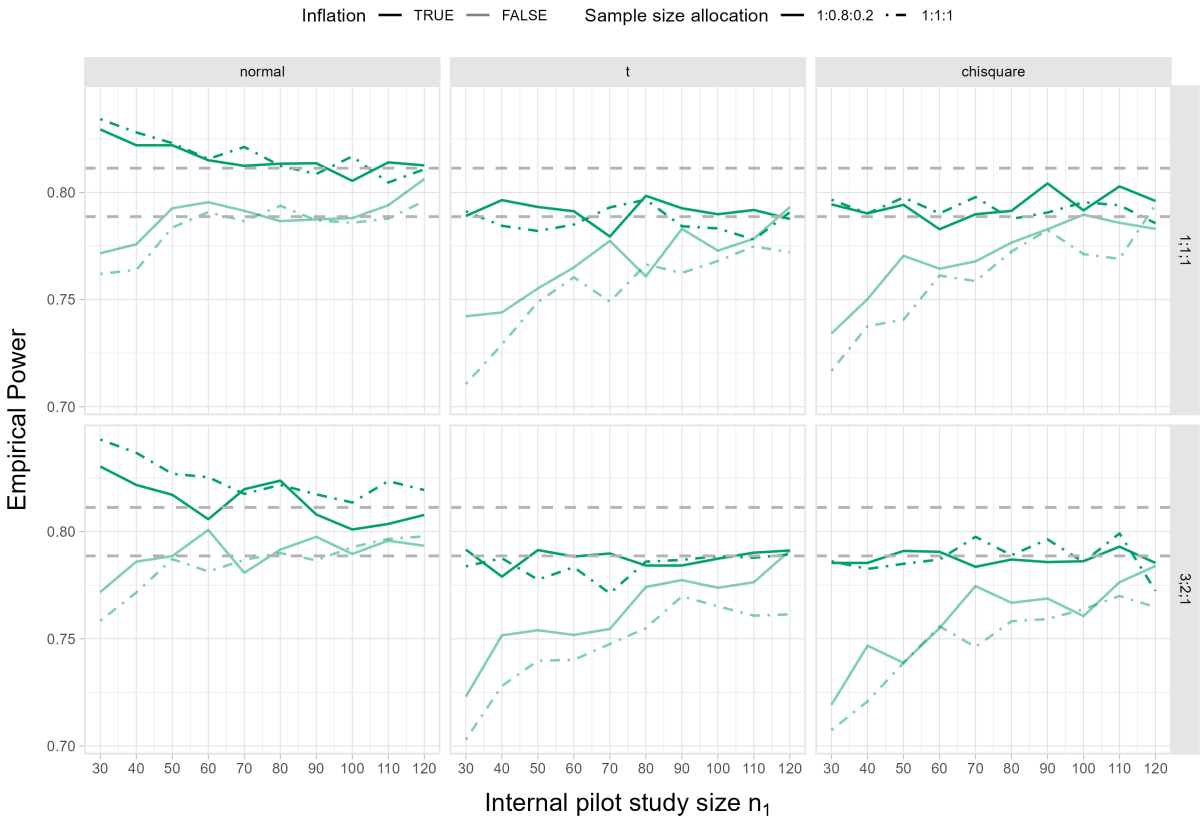


Figure 14: Observed power of the studentized permutation test with inflated sample size re-estimation based on the UG estimator compared to the observed power without inflated sample size re-estimation against the internal pilot study size n_1. The dashed grey lines depict the area of $1 - \beta = 0.8 \pm$ two times the Monte Carlo error.

It is evident that the inflation factor considerably increases the power curves compared to when it is not used. Specifically, the power curves now meet the targeted area more often and the power curves are more stable across the internal pilot study sizes.

The power curves no longer increase with larger pilot study sizes, but instead remain relatively constant. As a result, the power level is improved for smaller pilot study sizes, achieving the desired level of power even with reduced pilot study sizes more often (refer to Table S7 and S8 in the Appendix for the achieved power levels when the inflation factor

is applied).

That is specifically the case for normal data. While the power curves without the inflation factor were already able to approximate the targeted power level and reached the targeted level for larger pilot study sizes, the power curves were unable to reach the target level for smaller pilot study sizes. When the inflation factor is applied, the resulting power curves now meet or even exceed the targeted area for all the pilot study sizes considered. Notably, the power curves for smaller pilot study sizes under the UG estimator now surpass the targeted area. With increasing pilot study size, however, the power curves slightly decrease and come closer to the targeted area.

Still, it generally holds that desired effect of stabilising the power curve and increasing the power level for smaller pilot study sizes is achieved by the inflation factor.

Also for non-normal data, the power curves of the re-estimation procedures show a significant improvement when the inflation factor is applied. They are noticeably higher compared to the power curves without the inflation factor, and they exhibit greater stability across all pilot study sizes. Figure 14 specifically shows that, without the inflation factor, the power curves for t-distributed and χ^2-distributed data reach the desired power level range in only a few instances, and the introduction of the inflation factor brings the power curves closer to the targeted area indicated by the two grey lines. However, even with the inflation factor, it seems that the power curves still only touch the lower bound of the desired power level range. Table S7 and S8 in the Appendix provide a detailed report on the power levels for all underlying data-generating mechanisms. While the desired power level is achieved in most instances in the homogeneous variance scenario for t-distributed and χ^2-distributed data, the attained power levels fall short of the desired level in the heterogeneous variance scenario, particularly in the balanced group design. Thus, it can be concluded that the inflation factor can enhance the power level when data deviates from normality. However, this increase is not always sufficient to attain the desired power level. Hence, it should be noted that the inflation factor alone does not guarantee the attainment of the desired power level for all data types.

Interestingly, the introduction of the inflation factor also leads to a more comparable power performance between the unbalanced and balanced group designs. Without the inflation factor, the power curve of the unbalanced design was higher compared to the balanced design, primarily due to the greater variability in the re-estimation of the required total sample size (refer to Figure 11). However, with the inflation factor, both power curves converge to a similar level.

The resulting power performance, as described, can be understood by examining the distribution of the re-estimated final sample sizes with the application of the inflation factor. This distribution is depicted in Figure 15 which illustrates the effect of the inflation factor on the re-estimated sample sizes by means of the interquartile range (IQR). The transparent lines represent the re-estimated sample sizes without the inflation factor, while

the solid line represents the inflated re-estimated sample sizes. For clarity, Figure 15 focuses on the scenario of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$, with the two rows distinguishing between the balanced and unbalanced group designs.

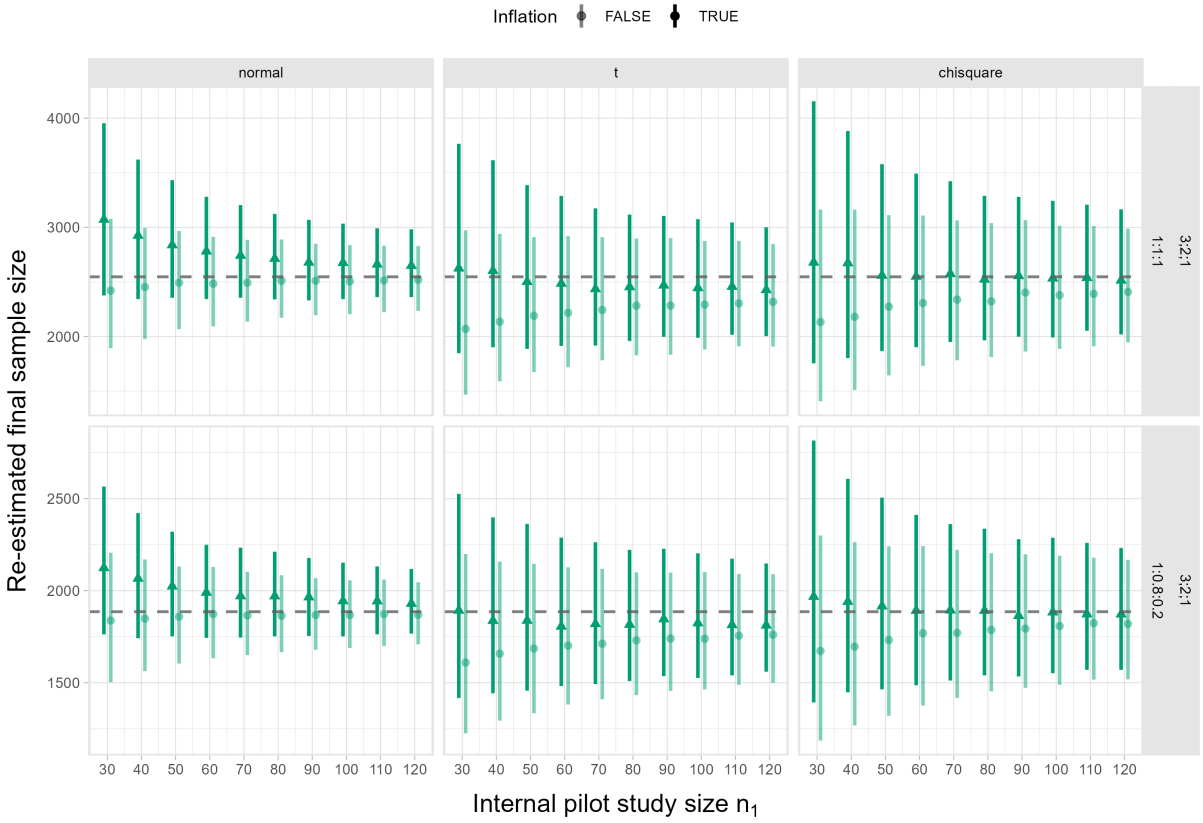


Figure 15: Median and interquartile range of the distribution of the inflated re-estimated final sample sizes based on the UG estimator against the internal pilot study size n_1 in the scenario of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$. The dashed grey lines depict the required total sample size based on the Hasler sample size formula.

Figure 15 illustrates that, for all data-generating mechanisms, the IQR is generally higher when the inflation factor is applied compared to without it. However, it is worth noting that the length of the IQR is generally not reduced by applying the inflation factor.

In the case of normal data, the mean and median of the re-estimated sample sizes are higher than the required total sample size. Without the inflation factor, the median of the re-estimated sample size was already very close to the actual required sample size, resulting in the attainment of the desired power level for larger pilot study sizes. With the inflation factor, the total sample size is now overestimated. This explains why the target power level is exceeded across all pilot study sizes, as seen in Figure 14.

While achieving higher power levels may seem attractive, it is important to note that re-estimation procedures that result in overpowering also lead to larger re-estimated sample sizes and overestimation of sample sizes can raise ethical and resource-related concerns. It could therefore be argued that applying the inflation factor may not always

be a reasonable choice when applying to the sample size re-estimation based on the UG estimator. Using the case of $n_1 = \{30, 40, 50\}$, Table 6 shows the impact on the re-estimated sample sizes when overpowering with the UG estimator in the case of normal data, after adjusting with the inflation factor.

Table 6: Mean and Median (in brackets) inflated re-estimated final sample sizes based on the UG estimator compared to the mean and median (in brackets) re-estimated sample sizes without the inflation factor for data following a normal distribution with $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$.

	$(1:1:1)$		$(1:0.8:0.2)$	
	w/o IF	w/ IF	w/o IF	w/ IF
$n_1 = 30$	2542.13	3242.26	1878.33	2196.17
	(2421)	(3072)	(1838)	(2124)
$n_1 = 40$	2546.44	3044.31	1885.76	2103.13
	(2454)	(2925)	(1848.5)	(2066.5)
$n_1 = 50$	2554.19	2927.10	1884.19	2056.46
	(2493)	(2838)	(1859)	(2024)

In the balanced design, the re-estimated sample sizes without the inflation factor already approximate the required total sample size quite accurately, with an average and median close to the required size of 2547 (refer to Table 5). However, when the inflation factor is applied, the average and median re-estimated sample size increase significantly, exceeding the required total sample size by a considerable margin. For example, with a pilot study size of $n_1 = 50$, the median re-estimated sample size increases by 345, exceeding the required size of 2547 by 291 subjects. This corresponds to an increase of 115 subjects per treatment group, which is substantial relative to the total sample size.

In the unbalanced design, the difference between the median re-estimated sample size and the required total sample size is smaller compared to the balanced design. For example, with $n_1 = 50$, the difference between the medians is 165 subjects, corresponding to an overestimation of 55 subjects per treatment group. However, this still represents a considerable overestimation relative to the required total sample size of 1886, with an inflation factor leading to an overestimation of 138 subjects. Overall, the application of the inflation factor results in a notable increase in sample size, particularly in the balanced design.

In the case of non-normal data, an elevated interquartile range (IQR) can be observed in Figure 14, with the median approaching the required total sample size. However, compared to normal data, the IQR is still larger, indicating greater variation in the re-estimated sample sizes. This is a result of the increased variability in the variance

estimation for non-normal data (refer to Section 3.4.3 for a discussion on the this phenomenon). Therefore, while applying the inflation factor improves the average estimation of the required total sample size, it does not address the inherent variability in the variance estimation for non-normal data. This explains why the power curves for non-normal data are more elevated and stable but still do not reliably reach the targeted range (as shown in Figure 14).

Moreover, the numbers show that the use of the inflation factor enables re-estimation to achieve the desired power level, even in the case where variances are rightly specified in the planning stage. This makes re-estimation comparable to the fixed sample size design in terms of achieving the desired power level. While the power curves in the fixed sample size design still demonstrate more stable power curves that meet the targeted area precisely (compare Figure 6), re-estimation can now be equally recommended as the fixed sample size design for scenarios where variances are correctly specified. Thus, re-estimation with the inflation factor emerges as a feasible approach for both variance scenarios. However, as mentioned, it should be noted that even with the inflation factor, the re-estimation procedure may still fall short of achieving the desired power level for non-normal data. In cases where the data exhibits higher skewness and variability, the effectiveness of re-estimation with the inflation factor may still be compromised compared to the fixed sample size design.

In summary, Figure 14 clearly demonstrates that applying the inflation factor increases the power levels associated with the re-estimation using the UG estimator. As a result, the inflation factor achieves the desired effect of stabilising the power curve and increasing the power level to the desired power level, particularly for smaller pilot study sizes. However, it is important to note that the inflation factor may also result in overestimating the sample size and consequently overpowering the trial. Particularly, this is observed in the simulation study for small pilot study sizes when data follows a normal distribution. The simulation results also indicate that both re-estimation procedures with inflation can be recommended under misspecification and correctly specifying the variance parameters, as the target power can be met in both variance scenarios. While the re-estimation procedure using the inflation factor is generally successful in achieving the desired power level (and even surpassing it) across all n_1 values for normal data, there are still instances where the desired power level cannot be attained, that is when data is non-normal. Therefore, the inflation factor can effectively increase the power level for small pilot study sizes and account for the uncertainty associated with variance estimation. However, it may not always fully compensate for the variability in variance estimation for non-normal data.

It is important to note that the inflation factor utilised in this study was derived solely from the proposed inflation factor by Zucker et al. (1999). Further research could solve equation (17) for the group-specific sample variance estimates and obtain the correct inflation factor that meets the desired power level. Therefore, the findings presented

here should only be considered as a preliminary indication of the potential usefulness of the inflation factor. However, it can be concluded that, in general, there are methods available to enhance the power behaviour of the re-estimation procedures by incorporating the uncertainty that comes along with the variance estimation.

3.4.5 Question 5. Type I error rate

Besides the attainment of the desired power level, it is important that a sample size re-estimation procedure controls for the type I error rate from a regulatory point of view (refer to Committee for medicinal products for human use, 2007). While research has shown that blinded sample size re-estimation typically does not inflate the type I error rate in superiority trials (refer to Friede and Kieser, 2013), violations in the type I error rate were observed for two-arm non-inferiority trials (see for example Friede and Kieser, 2003; Friede and Stammer, 2010; Friede and Kieser, 2011b). Specifically, Glimm and Läuter (2013) showed cases in which the one-sample estimator proposed by Kieser and Friede (2003) violates the type I error rate. The type I error rate violation has been observed not only for blinded methods of sample size re-estimation but also for unblinded methods (Wittes et al., 1999; Zucker et al., 1999; Friede and Kieser, 2013). In the case of three-arm designs, Mütze and Friede (2017) also observed a slight increase in the type I error rate when re-estimating sample sizes using the OSU and UG estimators under the absolute margin approach. Therefore, the subsequent simulation investigates how the two proposed methods of sample size re-estimation impact the type I error rate in the specific scenarios under consideration, as outlined in Question 5. An inflation of the type I error rate is anticipated in this simulation study.

To simulate a situation under the null hypothesis, the effect $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ is set to 0.8 for a non-inferiority margin Δ of 0.8. In the planning stage, homogeneous variances across the three groups are again assumed, that is $(\sigma^{2*}_{\text{EXP}}; \sigma^{2*}_{\text{REF}}; \sigma^{2*}_{\text{PLA}}) = (1; 1; 1)$. After reaching n_1, the variances are re-estimated based on the UG and OSU estimator. Based on these estimates, the sample size is re-calculated with the Hasler sample size formula (10) assuming $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}}) = 1$. The type I error rate of each re-estimation procedure is then calculated as the number of times that the studentized permutation test rejects the null hypothesis divided by the number of replications. Refer to column 3 of Table 4 for the setup of the simulation.

The simulation results are depicted in Figure 16, which presents the variability in the observed type I error rate for both re-estimation procedures through boxplots. Similar to previously, the two colours differentiate between the two re-estimation procedures, and two boxes are shown for each re-estimation procedure to represent the considered allocation scheme. Note that the observed significance level $\hat{\alpha}$ is only shown for the homogeneous variance scenario for the OSU estimator, as it has been demonstrated that the OSU estimator is not applicable for heterogeneous variance scenarios. The rows correspond to

the underlying data-generating mechanism, while the columns represent the true variance scenarios. The two dashed grey lines indicate the one-sided nominal level α, with a margin of 0.025 $\pm$ two times the corresponding Monte Carlo error.

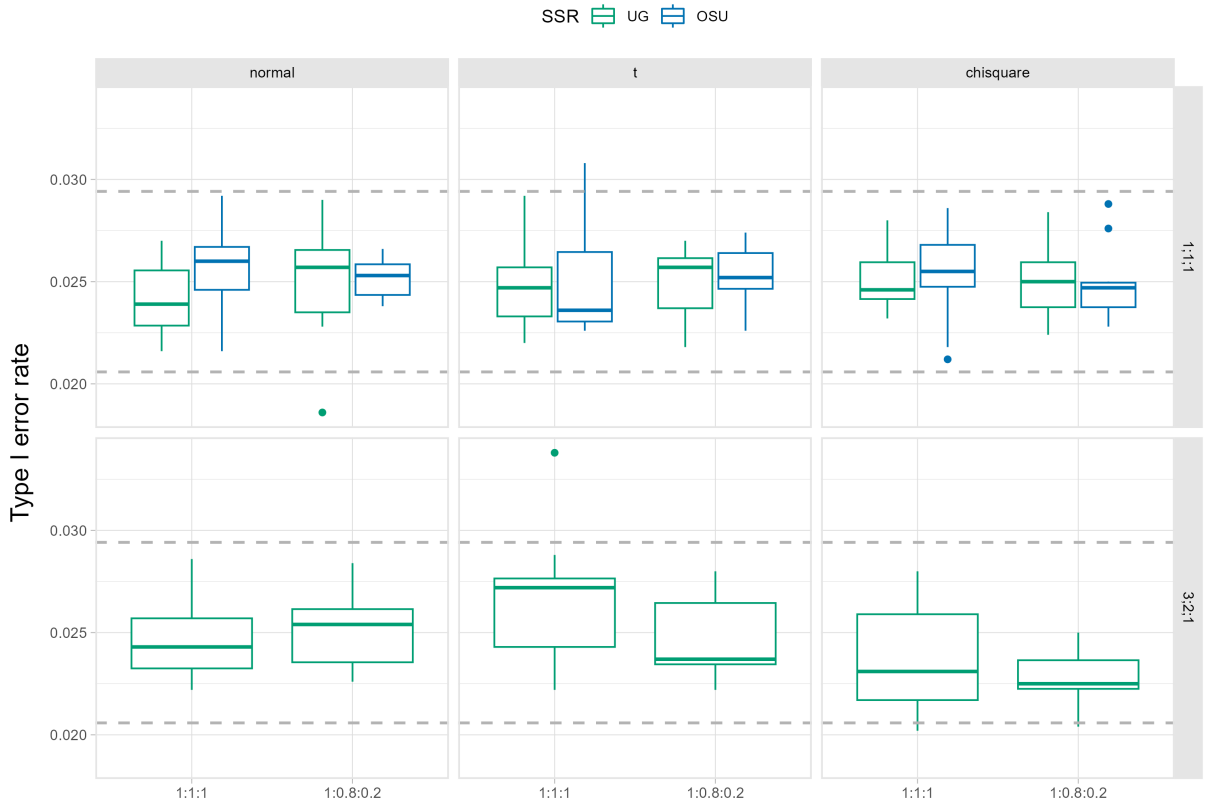


Figure 16: Actual significance level $\hat{\alpha}$ of the studentized permutation test with sample size re-estimation based on the UG and OSU estimator. The dashed grey lines depict the area of $\alpha = 0.025 \pm$ two times the Monte Carlo error.

Figure 16 illustrates that the variation in the type I error rate for both re-estimation procedures remains within the range indicated by the two grey lines, despite showing occasional outliers. This indicates that the re-estimation procedures do not result in any inflation of the type I error rate that can be distinguished from the Monte Carlo error. Thus, the observed type I error rate remains consistent with the predetermined one-sided α-level in a fixed design. Therefore, from a regulatory point of view, the proposed re-estimation procedures can be regarded as a valid method for re-estimating sample sizes in a three-arm design while maintaining the one-sided nominal significance level specified in advance. This, however, contrasts with the findings of a slight inflation in the type I error rate for both re-estimation procedures by Mütze and Friede (2017). It should be noted also that the impact of the inflation factor on the type I error rate was not investigated and should be addressed by further research.

3.5 Summary and Discussion

The presented section of this thesis aimed at deriving a procedure to estimate sample sizes when utilising the nonparametric studentized permutation test in a three-arm clinical trial.

The studentized permutation test The first step involved introducing the non-inferiority hypothesis within the context of the retention-of-effect approach, followed by the derivation of the studentized permutation test. To gain insights into the test's operating characteristics, its performance under the null and alternative hypothesis was investigated. Through this investigation, two key contributions were made. Firstly, the study observed that the studentized permutation test exhibits a liberal behaviour under the null hypothesis in scenarios involving skewed data with increasing heterogeneous variance structures, extending the findings of Mütze et al. (2017). Subsequently, these scenarios were excluded from the analysis.

Secondly, this study conducted the first investigation of the test's operating characteristics under the alternative hypothesis. To our knowledge, the power behaviour of this test has not been previously studied, nor have there been any considerations on sample size planning. It was observed that the power curve of the studentized permutation test resembles the one of the parametric equivalent, known as the Hasler test, with minor variations depending on the underlying data-generating mechanism and variance scenario. Based on this observation, the proposal was made to use the sample size formula by Hasler et al. (2008) as a planning method for the studentized permutation test. This approach offers a straightforward way to determine sample sizes. It requires specifying only the expectation and variance parameters in advance, making the estimation of sample sizes convenient without the need for extensive simulation studies.

Sample size planning Simulation studies demonstrated that the studentized permutation test can effectively achieve the desired power level when sample sizes are planned using the Hasler sample size formula, even when data deviates from normality. However, this only holds when expectation and variance parameters are correctly specified in the planning stage. While the accurate specification of parameters is important for any sample size formula, variances in the context of three-arm designs, in particular, are often subject to greater uncertainty when determining sample sizes. Simulation studies have shown that when variances are misspecified in the planning stage, the initially estimated sample sizes are inadequate to achieve the desired power level. To address this issue, a sample size re-estimation procedure based on nuisance parameter estimates has been proposed. Sample size re-estimation is a widely recognised and feasible approach for maintaining the desired power level in a clinical trial, which involves estimating the sample variance using data from an internal pilot study and adjusting sample sizes accordingly while the trial is ongoing.

Sample size re-estimation based on nuisance parameter estimates Within the sample size re-estimation procedure, two variance estimators have been suggested: the unblinded group (UG) estimator, which requires unblinding of the internal pilot data, and the blinded adjusted one-sample variance estimator (OSU), which provides a blinded approach. The findings of this study highlight that when variances are accurately specified in the planning stage, the re-estimation of sample sizes may lead to a loss of power compared to the fixed sample size design. However, in the studied cases where variances are misspecified, the re-estimation approach considerably improves the power level.

Both re-estimation procedures are straightforward to implement and perform equally well in the homogeneous variance scenario. While each estimator provides an unbiased estimate of the pooled variance, the OSU estimator fails to account for varying variances between treatment groups. This limitation makes the OSU estimator impractical in scenarios involving heterogeneous variances. In the considered scenarios, the procedure tended to underestimate sample sizes, which consequently led to lower power levels. To the best of our knowledge, there has been no investigation of blinded variance estimators in the context of heterogeneous variances yet. By demonstrating the inability of the OSU estimator to accurately re-estimate sample sizes, this study contributes to our understanding of the feasibility of using blinded estimators in three-arm clinical trials.

Interestingly, in the considered heterogeneous variance scenario within the unbalanced design, the OSU estimator still managed to provide reasonably accurate estimates, despite estimating a single nuisance parameter only. This outcome raises the question about the potential advantages of unbalanced designs in re-estimation scenarios when heterogeneous variances are present. Further research could explore the behaviour of the OSU estimator across different group design scenarios and in situations with varying degrees of heterogeneous variances.

Moreover, it could be argued that the OSU estimator still offers a valid re-estimation of sample sizes in heterogeneous scenarios when the variances between the treatment groups are not drastically different. The sensitivity of the OSU estimator to the extent of heterogeneity and its performance under varying levels of heterogeneity could be the subject of further investigation.

Furthermore, there is a need to explore additional blinded variance estimators for the retention-of-effect approach in heterogeneous variance scenarios. This would allow for a more comprehensive understanding of the performance of different estimators in such settings and potentially provide alternative approaches for sample size re-estimation.

In contrast to the OSU estimator, the UG estimator demonstrated itself to be a reliable and valid tool for sample size re-estimation in both variance scenarios. Hence, for the scenarios considered, it demonstrated broader applicability, making it a recommended choice for re-estimation purposes. In terms of performance, the UG estimator exhibited the best results with normal data, where the desired power level was achieved for larger

pilot study sizes. However, with non-normal data and smaller pilot study sizes, the targeted power level was often not achieved, highlighting that sample size re-estimation alone may not always be sufficient to attain the desired power level.

Inflation factor for sample size re-estimation procedures In such cases, employing an inflation factor can be a practical solution. This approach involves inflating the re-estimated sample size based on the UG estimator using an inflation factor ζ, as previously proposed by Zucker et al. (1999). The aim of this approach is to account for the uncertainty of the variance estimation in the sample size re-estimation, thereby improving the reliability of reaching the target power level. The inflation factor was derived for the two-arm design and then extended to the three-arm design. Simulation results showed that the inflation factor helped stabilise the power curve, enabling the attainment of the desired power level across all pilot study sizes in the case of normal data. However, it should be noted that the use of the inflation factor resulted in overestimation of the sample sizes, leading to an overpowered trial in case of normal data. On the other hand, for non-normal data, although the inflation factor considerably improved the power curve, it was not always sufficient to achieve the desired power level.

Regarding the inflation factor used in this study, it is important to acknowledge that its derivation for the three-arm design was not based on a mathematical derivation specific to this design. Instead, the inflation factor was derived by adapting ideas from the two-arm design to the three-arm design. Future research could focus on deriving the inflation factor mathematically and refine the approach to balance the trade-off between attained power level and sample size. Additionally, there may be other approaches to enhance the power performance of sample size re-estimation procedures, especially when using the studentized permutation test. One possible approach could be the use of quantiles based on the permutation distribution instead of t-quantiles, as in the Hasler sample size formula, for the inflation factor.

Limitations of sample size re-estimation based on nuisance parameter estimates It should also be highlighted that this study provides first insights into the performance of re-estimation procedures based on nuisance parameter estimates when dealing with non-normal data. So far, re-estimation procedures have not been studied for non-normally distributed continuous data in the context of three-arm non-inferiority trials. This study revealed that the limitations of re-estimation for non-normal data primarily arises from the variance estimation itself rather than being a result of the proposed sample size planning method or the performance of the studentized permutation test. The estimation of variances for non-normal data tends to exhibit greater variability, leading to less accurate estimation of re-estimated sample sizes. Consequently, lower power levels are observed in the re-estimation procedure. The extent of this impact depends on the degree to which the

data deviates from normality. Therefore, caution should be exercised when using sample size re-estimation procedures based on nuisance parameter estimates for data that does not conform to the characteristics of normal data.

The observed limitations of the re-estimation procedures in the presence of non-normal data also emphasise the need for additional strategies to enhance the power level when data deviates from normality. One potential approach could involve adjusting the Hasler sample size formula to account for non-normality when re-estimating sample sizes. This could be achieved by incorporating estimates of higher moments of the internal pilot study data into the re-estimation within the Hasler sample size formula. However, it should be noted that the adjustment of a sample size formula for higher moments has not yet been explored in the context of the equivalent two-sample t-test situation, making it even more challenging to apply in the three-arm setting.

Type I error rate of the proposed sample size re-estimation procedures Moreover, this work found no evidence of inflation in the type I error rate when using the proposed methods for sample size re-estimation relative to the Monto Carlo error. Therefore, this study shows that both methods are feasible from a regulatory point of view in terms of maintaining the nominal significance level as predefined during the planning stage. This contrasts to previous literature that suggested an inflation in the type I error rate in three-arm non-inferiority trials.

Further contributions of this work Although not explicitly tested in this study, it is worth noting that the proposed re-estimation procedure and inflation factor can also be applied when analysing data with the Hasler test. This work therefore provides an initial approach for sample size re-estimation based on nuisance parameter estimates when analysing with the Hasler test under the retention-of-effect hypothesis, allowing for heterogeneous variance.

It is also worth highlighting that a minor coding error was discovered in the functionality of the studentized permutation test within the package `ThreeArmedTrials` during this study. The error was reported to the package owner and a corrected version of the studentized permutation test has been implemented. As a result, the `ThreeArmedTrials` package now correctly supports the analysis of data using the studentized permutation test in R.

Short summary In summary, the proposed strategy for sample size planning involves using the Hasler sample size formula and employing the UG variance estimator in cases of uncertain variance specification. Additionally, the use of an inflation factor can help enhance the power level, but it may not guarantee the attainment of the desired power level in all scenarios, particularly when data deviates significantly from normality. The

presented work therefore provides an initial approach to determining sample sizes in the analysis of three-arm non-inferiority trials using the studentized permutation test. It should, however, be noted that the presented work focused solely on continuous data. The analysis using the studentized permutation test for discrete data remains an area for further research.

4 Nonparametric test based on classical mid-ranks

4.1 Statistical model and hypothesis testing

Nonparametric relative effects in the three-arm design Denote X_{ik} with $k = 1, ..., n_i$ and $i = \text{EXP}, \text{REF}, \text{PLA}$ the outcomes of independent random variables under the experimental treatment (EXP), reference treatment (REF) and placebo (PLA) of a three-arm clinical trial. It is assumed that the respective random variable X_{ik} follows a distribution F_i that is described by the so-called normalized version of the cumulative distribution function (CDF), that is, the mean of the right-continuous and the left-continuous version

$$F_i(x) = 0.5 \cdot [P(X_{ik} < x) + P(X_{ik} \leq x)].$$

The relative treatment effects describing the influence of the treatment i to the observation are derived by

$$p_i = \int H(x) \mathrm{d}F_i(x) \tag{21}$$

where

$$H(x) = n^{-1} \sum_{i=1}^{3} n_i F_i(x) \tag{22}$$

denotes the mean of the CDFs for all $n = \sum^i n_i$ observations in the experiment. The relative treatment effects p_i can therefore be interpreted as a relative deviation from $H(x)$, the weighted average.

An unbiased estimate for the relative effect p_i can be obtained by replacing the CDF by its empirical counterpart $\hat{F}_i(x)$ with $\hat{H}(x)$, that is

$$\hat{p}_i = \int \hat{H}(x) \mathrm{d}\hat{F}_i(x). \tag{23}$$

This leads to means $\bar{R}_i$ of the mid-ranks R_{ik} among all n observations. The estimation (23) can then be expressed as

$$\hat{p}_i = \int \hat{H}(x) \mathrm{d}\hat{F}_i(x) = n^{-1}\big(\bar{R}_i - \tfrac{1}{2}\big). \tag{24}$$

The relative effects based on $H(x)$ are typically called weighted relative effects as they are dependent on the sample sizes n_i. To avoid this dependence, $H(x)$ can be replaced

by the unweighted mean

$$G(x) = \frac{1}{3}\sum_{i=1}^{3} F_i(x). \tag{25}$$

The relative treatment effects can then be derived as

$$p_i = \int G(x)\mathrm{d}F_i(x). \tag{26}$$

Replacing $G(x)$ and $F_i(x)$ by their empirical counterparts $\hat{G}(x)$ and $\hat{F}_i(x)$ leads to so-called pseudo-ranks. In this case, the estimation of the relative treatment effects p_i does not rely on the group-specific sample sizes n_i. The relative effects based on $G(x)$ are typically referred to as unweighted relative effects. To our knowledge, the concept of unweighted relative effects was first mentioned by Brunner and Puri (2001), with a more detailed discussion provided by Brunner et al. (2017), who also established the term pseudo-ranks to describe the estimation process. When the sample sizes in the groups are equal, both procedures yield the same estimated effects. However, if there is imbalance in the group sample sizes, the weighted relative effects may produce distorted results, as they assign more weight to observations in the larger groups. For a more detailed discussion on the results obtained with weighted relative effects in contrast to unweighted relative effects, refer to Brunner et al. (2020).

Munzel (2009) proposed using weighted relative effects for the nonparametric analysis of three-arm non-inferiority trials. It is worth noting that Munzel did not specifically consider the effect of group imbalance on the estimated effects. However, the impact of typical unbalanced group designs in three-arm trials on the estimated effects is generally minimal. To adhere to the proposed framework and ensure consistency, the analysis will therefore focus on the weighted relative effects as described in (21).

Non-inferiority hypothesis In a setting where higher values of the response are associated with a higher treatment effect, non-inferiority of the experimental treatment to the reference treatment is demonstrated by the following hypothesis by means of nonparametric relative effects, as defined in (21)

$$H_0 : p_{\mathrm{EXP}} - p_{\mathrm{REF}} \leq \delta \text{ vs. } H_1 : p_{\mathrm{EXP}} - p_{\mathrm{REF}} > \delta \tag{27}$$

where δ refers to a pre-specified, clinically irrelevant amount that the difference between the experimental and reference treatment needs to exceed in order to demonstrate non-inferiority, with $\delta < 0$.

Following the retention-of-effect approach, non-inferiority in the three-arm design can now be formulated as a fraction of the trials sensitivity using the information of the

placebo arm by defining the margin δ as a fraction f of the difference between the reference treatment and placebo

$$\delta = f(p_{\text{REF}} - p_{\text{PLA}})$$

with $f \in (-1, 0)$. This leads to the hypotheses

$$H_0 : p_{\text{EXP}} - p_{\text{REF}} \leq f(p_{\text{REF}} - p_{\text{PLA}}) \text{ vs. } H_1 : p_{\text{EXP}} - p_{\text{REF}} > f(p_{\text{REF}} - p_{\text{PLA}}). \tag{28}$$

Let $\Delta = 1 + f$, then the hypothesis pair can be rewritten as:

$$H_0 : \frac{p_{\text{EXP}} - p_{\text{PLA}}}{p_{\text{REF}} - p_{\text{PLA}}} \leq \Delta \text{ vs. } H_1 : \frac{p_{\text{EXP}} - p_{\text{PLA}}}{p_{\text{REF}} - p_{\text{PLA}}} > \Delta. \tag{29}$$

The nonparametric test based on classical mid-ranks This paragraph introduces the nonparametric test based on classical mid-ranks published by Munzel (2009), which evaluates the non-inferiority hypothesis for the retention-of-effect approach. To construct the test statistic, (29) can be rearranged as

$$H_0 : p_{\text{EXP}} - \Delta p_{\text{REF}} + (\Delta - 1)p_{\text{PLA}} \leq 0 \text{ vs. } H_1 : p_{\text{EXP}} - \Delta p_{\text{REF}} + (\Delta - 1)p_{\text{PLA}} > 0. \tag{30}$$

Replacing the relative treatment effects p_i by their estimates $\hat{p}_i$ and dividing the term by estimates of the variance and covariance components yields the test statistic T_n with

$$T_n = \sqrt{n} \frac{\hat{p}_{\text{EXP}} - \Delta \hat{p}_{\text{REF}} + (\Delta - 1)\hat{p}_{\text{PLA}}}{\sqrt{\hat{s}_{11} + \Delta^2 \hat{s}_{22} + (1 - \Delta)^2 \hat{s}_{33} - 2\Delta \hat{s}_{12} - 2(1 - \Delta)\hat{s}_{13} + 2(1 - \Delta)\Delta \hat{s}_{23}}} \tag{31}$$

where $\hat{s}_{ij}$ denote the estimates of the variance and covariance components (see Appendix A for their definition).

The distribution of T_n under the null hypothesis can be approximated by a t-distribution with $(n - 3)$ degrees of freedom (for a derivation of the approximation refer to Munzel, 2009). The hypothesis of inferiority is then rejected if the test statistic is greater than the $(1 - \alpha)$-quantile of the central t-distribution with $(n - 3)$ degrees of freedom. For the vector $\boldsymbol{X}_n = (X_{\text{EXP},1}, \ldots, X_{\text{EXP},n_{\text{EXP}}}, X_{\text{REF},1}, \ldots, X_{\text{REF},n_{\text{REF}}}, X_{\text{PLA},1}, \ldots, X_{\text{PLA},n_{\text{PLA}}})$ that contains all observations, this test can be expressed as the function

$$\phi_n^{\text{Munzel}}(\boldsymbol{X}_n) = \begin{cases} 1 & T_n(\boldsymbol{X}_n) > t_{1-\alpha}(n-3) \\ 0 & T_n(\boldsymbol{X}_n) \leq t_{1-\alpha}(n-3). \end{cases}$$

The test based on classical mid-ranks will be referred to as the Munzel test in the subsequent analysis. The Munzel test was implemented in R using the terminology and coding structure of the non-inferiority tests implemented in the `ThreeArmedTrials` pack-

age to facilitate its adoption into the package. The estimation of relative effects within the Munzel test is performed using the functionality of the `rankFD` package (Konietschke et al., 2022). The code is provided in Appendix A.

4.2 Operating characteristics of the nonparametric test

In order to develop a sample size planning method for the nonparametric Munzel test, a similar approach is taken as for the studentized permutation test. First, the one-sided nominal level is tested under the null hypothesis for the power simulation scenarios considered. Then, a situation under the alternative hypothesis is studied to derive a sample size planning method based on the observed power behaviour. Table 7 displays the simulation scenarios for the Munzel test.

Table 7: Scenarios for the simulation study investigating the operating characteristics of the nonparametric Munzel test under the null and alternative hypothesis.

Parameter	Values under H_0	Values under H_1
Distributions	Normal, Lognormal, Chi-squared	
Non-inferiority margin Δ	0.8	
Ratio in the mean differences $(\mu_{EXP} - \mu_{PLA})/(\mu_{REF} - \mu_{PLA})$	0.8	0.9, 1, 1.1, 1.2
Group standard deviations $(\sigma_{EXP}; \sigma_{REF}; \sigma_{PLA})$	For all distributions of the data: (1;1;1); For normal data: (1;2;3); (3;2;1)	
Sample size allocations $(n_{EXP} : n_{REF} : n_{PLA})$	$(1:1:1)$; $(2:2:1)$; $(3:2:1)$;	$(1{:}\Delta{:}1\text{-}\Delta)$
Total sample size n	30, 60, 120, 210, 300	420
One-sided nominal level α	0.025	
Simulation replications	1,000	

The simulation scenarios for the Munzel test are similar to those used for the studentized permutation test, as described in Table 1, and in Figure S1 of the Appendix A. The simulations focus on continuous data, specifically normal, lognormal, and chi-square-distributions. Data underlying a lognormal or chi-square distribution, denoted as χ^2, are simulated using the formula (8) to ensure that data conforms to the appropriate expectations and standard deviations. t-distributed data are not simulated. The non-inferiority margin Δ is again fixed at 0.8. As for the studentized permutation test, the simulation involves varying the effect of the ratio in the expectation parameters, that is

$(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$, to represent and simulate the relative effects under the hypothesis. However, in the process of this thesis it became evident that assuming parametric distributions with specific parameters does not directly translate into nonparametric relative effects defined under the hypothesis. The relative effects vary with different distributions, expectations and standard deviations, making it challenging to establish a direct relationship between the expectation parameters and the relative effects. For data following a normal distribution with $F_i(x) \sim N(\mu_i, \sigma)$ the weighted relative effects in a three-arm design correspond to

$$p_i = \int H(x) \mathrm{d}F_i(x) = \frac{1}{n} \sum_{j=1}^{3} n_j \cdot \Phi\Big(\frac{\mu_i - \mu_j}{\sigma\sqrt{2}}\Big).$$

However, while a relationship between relative effects and parameters can be derived for normal data, to our knowledge, no closed-form solution exists for deriving relative effects under non-normal data, requiring simulations instead. Consequently, simulating the situation under the respective hypothesis is not straightforward.

It can be shown that heteroskedastic and skewed data lead to a shift in the relative effects and consequently the hypothesis. Therefore, the group standard deviation scenarios are differentiated based on normal and skewed data in this simulation study, representing symmetric and asymmetric distributions respectively. That is, for data following a log-normal or χ^2-distribution, the heteroskedastic standard deviation scenarios are excluded. Additionally, fewer group designs are included, specifically $(1:1:1)$, $(2:2:1)$, $(3:2:1)$, and $(1:\Delta:1-\Delta)$. Under the null hypothesis, the total sample size n is varied while it is fixed at $n = 420$ under the alternative hypothesis. The simulation scenarios are replicated 1,000 times. The scripts for conducting the simulation study in R are provided in Appendix A.

Given the inherent challenges in accurately simulating the type I error rate and power of the Munzel test, it was decided during the course of this thesis to shift the focus towards the studentized permutation test as the primary approach for deriving a sample size planning method. Consequently, the results for the Munzel test may lack completeness, as they were not further explored. Nevertheless, the following section will briefly present the results for the type I error rate and power of the Munzel test.

4.2.1 Type I error rate

Figure 17 illustrates the simulation results under the null hypothesis, that is the type I error rate of the Munzel test as a function of the total sample size n. The different line types correspond to the group designs, while the columns represent the underlying distribution of the data and the rows depict the group standard deviation scenarios. It is important to note that only the results for normal data are shown for all standard deviation

scenarios, as the relative effects under the hypothesis shift in skewed data, indicating a situation under the alternative hypothesis. Consequently, the results for skewed data and heteroskedastic standard deviations are not displayed.

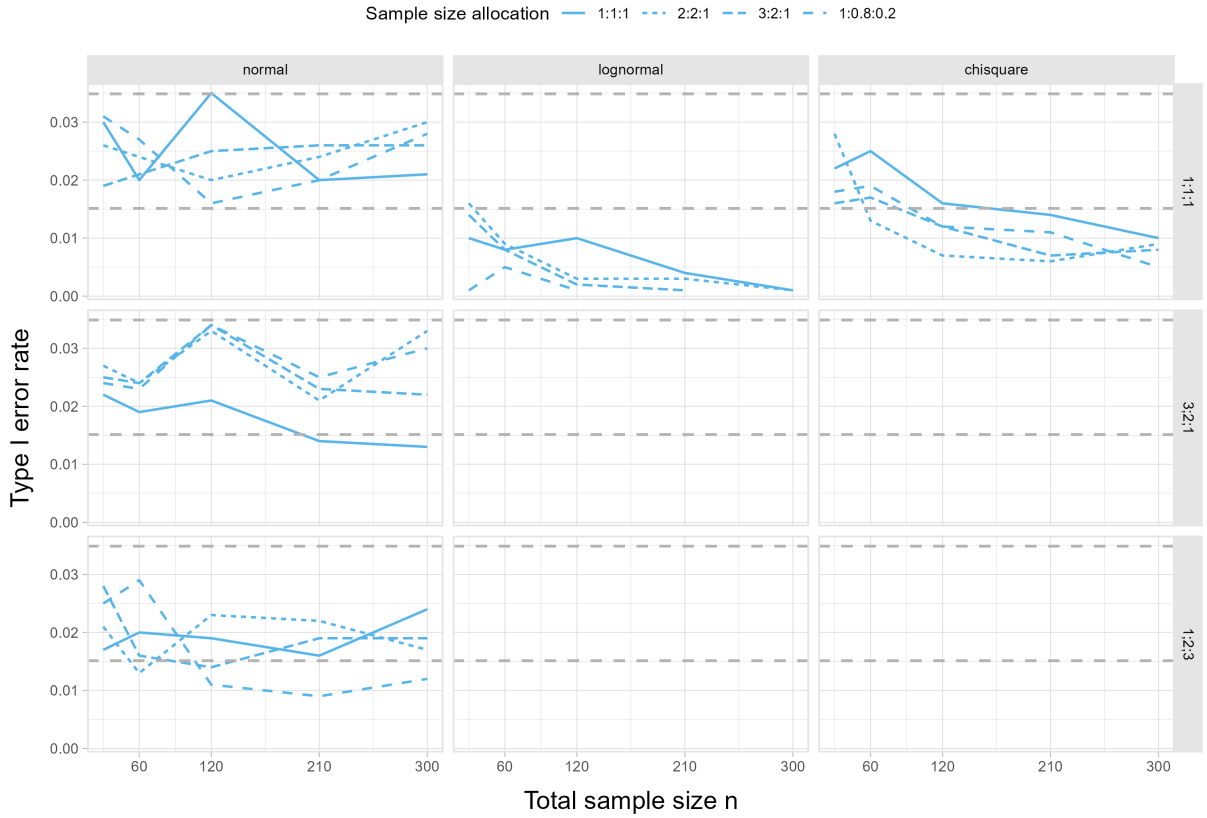


Figure 17: Actual significance level $\hat{\alpha}$ of the Munzel test against the total sample size n. The dashed grey lines depict the area of $\alpha = 0.025 \pm$ two times the Monte Carlo error.

Figure 17 shows that in the case of normal data, the one-sided nominal significance level can be maintained across all standard deviation scenarios. This is indicated by the lines that remain within the boundaries defined by the grey dashed lines. For data following a lognormal and χ^2-distribution, there is a conservative behaviour observed as the sample size increases. This suggests a shift in the effect under the hypotheses, with more evidence supporting the null hypothesis as the sample size grows. It is also observable that there exist differences in the type I error rate by group design, however no pattern across all scenarios becomes apparent.

4.2.2 Power

Figure 18 depicts the simulation results under the alternative hypothesis, that is the observed power of the Munzel test against a varying ratio in the expectation parameters $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$. The different line types correspond to the group designs, while the columns represent the underlying distribution of the data and the rows depict

the group standard deviation scenarios. Similar to the previous Figure 17, the results for normal data and skewed homoskedastic data are presented, while the results for other distributions and standard deviation scenarios are not displayed.

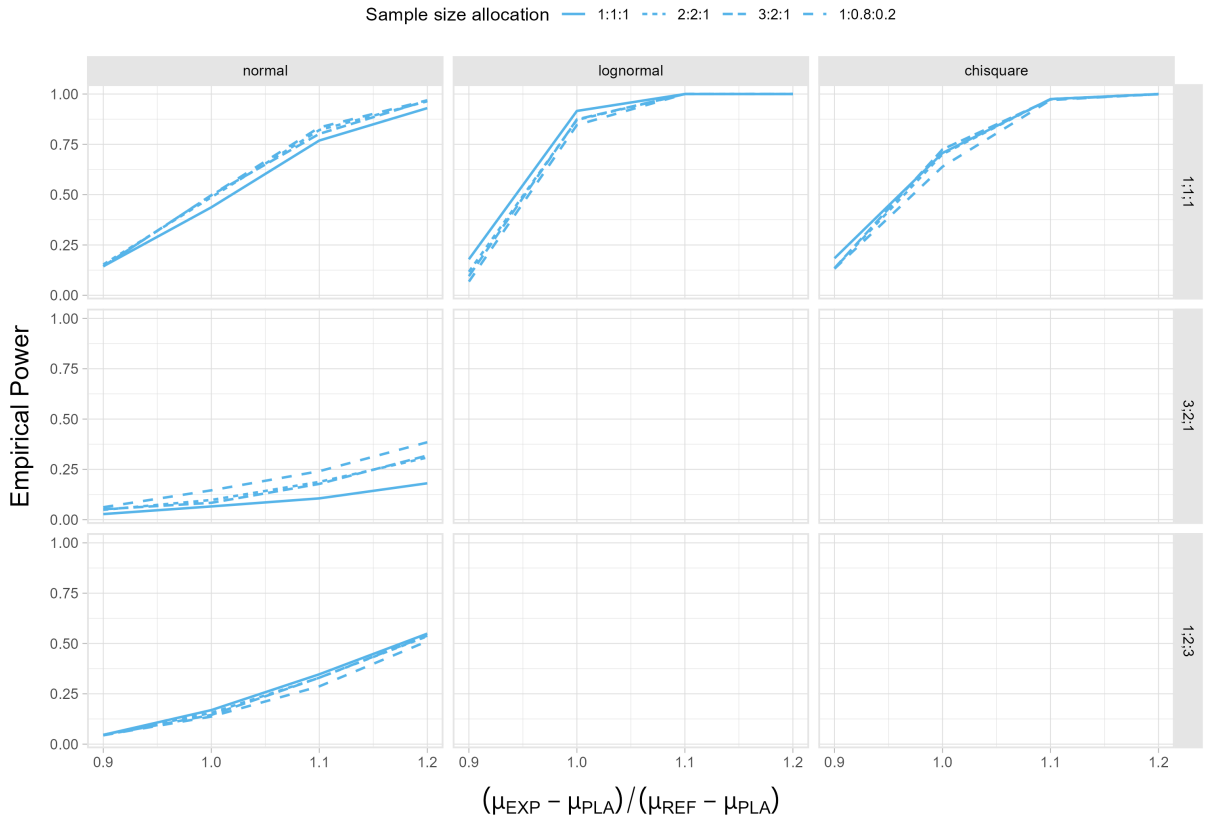


Figure 18: Observed power of the Munzel test against $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ for a total sample size $n = 420$.

In the case of normal data, the power of the test increases as the effect under the hypothesis increases. The highest power is observed in scenarios with homogeneous variances, while the lowest power is observed in the $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$ structure. For lognormal and skewed data, the power curves are higher compared to normal data. This observation suggests a potential shift in the effect under the hypothesis, providing more evidence in support of the alternative hypothesis. Some minor differences can be observed due to the chosen group design.

4.3 Sample Size Planning

The simulation results for the Munzel test demonstrate that, similar to the studentized permutation test, the one-sided nominal level can be maintained across the considered scenarios. In the power simulation it is shown that the power is influenced by the expectations parameters and the underlying group standard deviations. However, the underlying distribution of the data and the standard deviation significantly affects the relative effects

of the treatment groups, making it challenging to derive a sample size planning method based on parametric assumptions.

One possible approach to derive a sample size planning method is to examine the power function of the test instead. The power function is dependent on various factors, including the non-inferiority margin Δ, the relative effects p_i, the total sample size n, and the variance and covariance parameters of the test statistic s_{ij}. However, deriving the variance and covariance parameters is difficult since they, like the relative effects, depend on mid-ranks (see Appendix A). These, in turn, depend on the data. Without information on the data or the ranks of the data, it is not possible to obtain a power function.

Nevertheless, one alternative option is to specify relative effects and simulate them rather than defining parametric distributions that yield unknown relative effects. This method allows for a more flexible and data-driven approach to sample size determination for the Munzel test. By employing this method, a more comprehensive investigation into the power behaviour of the test can be conducted, providing valuable insights for potential sample size planning strategies.

As this thesis primarily focused on the studentized permutation test, the considerations for sample size planning based on the Munzel test are limited. The Munzel test was investigated in terms of its power and performance, but its application for sample size determination was not extensively explored.

4.4 Summary and Discussion

The study mainly focused on the studentized permutation test, thus limiting insights into the Munzel test's use for sample size planning. While the Munzel test's power and performance under the null hypothesis were examined, the application for sample size determination wasn't extensively addressed. Simulation results indicate that the power of the Munzel test is impacted by the expectation parameters and group standard deviations. However, the relative effects of the treatment groups are greatly influenced by the data's distribution and standard deviation. The complex interplay of these elements signifies that a straightforward sample size determination based on parametric assumptions for the Munzel test may not be feasible. Assuming and simulating relative effects rather than simulating parameters of a distributions can potentially provide a more practical and flexible approach for the Munzel test's sample size determination. However, deriving a generalised power function for sample size planning is challenging due to the variance and covariance components of the test statistic.

Nevertheless, it is recommended to explore and refine the sample size planning method using the Munzel test in future research, given the limitations of the current study. It is necessary to understand how different factors, such as the non-inferiority margin, relative effects, total sample size, and the variance and covariance parameters of the test statistic,

interact and impact the power of the test.

While this study focused primarily on continuous data, the true strength of the non-parametric approach by Munzel (2009) is best utilised in the case of ordinal data. Thus, future research could benefit from a concentrated exploration of this test's application to ordinal data.

5 Conclusion

Through extensive simulation studies, this thesis examined how sample sizes can be determined for different distributional properties of data in three-arm gold standard designs, focusing primarily on continuous data. The nonparametric analysis approaches employed in this investigation included the studentized permutation test by Mütze et al. (2017), and the nonparametric test based on classic mid-ranks by Munzel (2009).

When analysing with the studentized permutation test, this thesis presents a convenient method for sample size planning utilising the sample size formula of its parametric counterpart, the test by Hasler et al. (2008). While the Hasler sample size formula effectively achieves the target power level of the studentized permutation test, it depends on precise parameter specification during the planning stage. To address parameter uncertainties, the proposed method includes a sample size re-estimation procedure that uses data from an internal pilot study to validate assumptions on the nuisance parameters, thus ensuring that the desired power is achieved. The comparison of two sample size re-estimation procedures has shown the unblinded sample variance estimator to be more versatile, demonstrating effectiveness in both homogeneous and heterogeneous variance scenarios. However, achieving the target power level remains challenging for non-normal data, even with the application of an inflation factor, due largely to increased variability in variance estimation for such data.

In addition to pioneering a method for determining sample sizes in three-arm non-inferiority trials using the nonparametric studentized permutation test, this research also provides valuable insights into the test's operating characteristics under both hypotheses. It identifies scenarios in which the test exhibits a liberal behaviour and provides insights into the power characteristics of the test. Furthermore, this study explores sample size re-estimation procedures based on nuisance parameter estimates for scenarios with heterogeneous variances, thereby offering a valid approach for sample size re-estimation when analysing data using the Hasler test as well. By including an estimator that maintains the blinding of the data, this study also contributes to our understanding of the feasibility of using blinded estimators in three-arm clinical trials. Moreover, the study demonstrates the limitations of sample size re-estimation based on nuisance parameter estimates in cases of non-normal data due to the increased variability of the variance estimation.

This study highlights several potential areas for further research. Firstly, it suggests exploring additional blinded variance estimators in heterogeneous variance scenarios. Specifically, investigating the performance of these estimators in unbalanced designs and their sensitivity to the extent of heterogeneity would provide a better understanding of their applicability and limitations in such scenarios. Secondly, the study proposes further investigation of the applied inflation factor to enhance the reliability of achieving the desired power level in sample size re-estimation procedures. This involves conducting

a mathematical derivation of the inflation factor, refining it to prevent overestimation of sample sizes, and optimising its performance for non-normal data. Furthermore, future research could examine sample size re-estimation based on nuisance parameter estimates for non-normal data in more detail, with the goal of uncovering its limitations and exploring strategies to adjust the recalculation of sample sizes. This could involve considering observed higher moments of the internal pilot data as a basis for adjustment. Lastly, extending the presented method for the studentized permutation test to discrete data could be a valuable avenue for further exploration.

In analysing with the nonparametric test based on classic mid-ranks by Munzel (2009), this thesis identified the simulation of relative effects under the hypothesis as a challenge for sample size planning. The relative effects of the treatment groups are influenced by the distribution of the data, as well as the standard deviations and expectation parameters. This interplay suggests that simulating the relative effects based on parametric assumptions of the data is not straightforward, making it difficult to investigate the operating characteristics of the test under the respective hypothesis. This study provided brief results for the operating characteristics of the test under both the null and alternative hypothesis and suggests that presuming and simulating relative effects, rather than simulating parameters of a distribution, could offer better insights into the test's power behaviour. These findings could then contribute to the development of a sample size planning method.

Furthermore, future research aimed at developing a sample size planning method for the Munzel test could focus on deriving a generalised power function for sample size planning. However, deriving such a function remains challenging due to the variance and covariance components of the test statistic. In addition to providing insights into the operating characteristics of the test, this work also contributes by implementing the test for analysis in R. The inclusion of the test in the `ThreeArmedTrials` package is planned. Moreover, the potential of the Munzel test for analysing ordinal data, which was not the primary focus of this study, warrants further exploration.

References

Brunner, E., Konietschke, F., Bathke, A. C., & Pauly, M. (2020). Ranks and pseudo–ranks: Surprising results of certain rank tests in unbalanced designs. *International Statistical Review*, *89*(2), 349–366. https://doi.org/10.1111/insr.12418

Brunner, E., Konietschke, F., Pauly, M., & Puri, M. L. (2017). Rank-based procedures in factorial designs: Hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *79*(5), 1463–1485. https://doi.org/10.1111/rssb.12222

Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, *42*, 1–52. https://doi.org/10.1007/s003620000039

Chowdhury, S., Tiwari, R. C., & Ghosh, S. (2019a). Bayesian approach for assessing non-inferiority in three-arm trials for risk ratio and odds ratio. *Statistics in Biopharmaceutical Research*, *11*(1), 34–43. https://doi.org/10.1080/19466315.2018.1554504

Chowdhury, S., Tiwari, R. C., & Ghosh, S. (2019b). Non-inferiority testing for risk ratio, odds ratio and number needed to treat in three-arm trial. *Computational Statistics & Data Analysis*, *132*, 70–83. https://doi.org/10.1016/j.csda.2018.08.018

Committee for medicinal products for human use. (2007). 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

D'Agostino, R. B., Massaro, J. M., & Sullivan, L. M. (2003). Non-inferiority trials: Design concepts and issues: The encounters of academic consultants in statistics. *Statistics in Medicine*, *22*(2), 169–186. https://doi.org/10.1002/sim.1425

Daniels, S., Casson, E., Stegmann, J.-U., Oh, C., Okamoto, A., Rauschkolb, C., & Upmalis, D. (2009). A randomized, double-blind, placebo-controlled phase 3 study of the relative efficacy and tolerability of tapentadol ir and oxycodone ir for acute pain. *Current Medical Research and Opinion*, *25*(6), 1551–1561. https://doi.org/10.1185/03007990902952825

European Medicines Agency. (2005). Guideline on the choice of the non-inferiority margin. https://www.ema.europa.eu/en/choice-non-inferiority-margin-scientific-guideline

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

Friede, T., & Kieser, M. (2011a). Blinded sample size recalculation for clinical trials with normal data and baseline adjusted analysis. *Pharmaceutical Statistics*, *10*(1), 8–13. https://doi.org/10.1002/pst.398

Friede, T., & Kieser, M. (2011b). Sample size reassessment in non-inferiority trials. Internal pilot study designs with ANCOVA. *Methods of Information in Medicine*, *50*(3), 237–243. https://doi.org/10.3414/ME09-01-0063

Friede, T., & Kieser, M. (2013). Blinded sample size re-estimation in superiority and noninferiority trials: Bias versus variance in variance estimation. *Pharmaceutical Statistics*, *12*(3), 141–146. https://doi.org/10.1002/pst.1564

Friede, T., & Stammer, H. (2010). Blinded sample size recalculation in noninferiority trials: A case study in dermatology. *Drug Information Journal*, *44*(5), 599–607. https://doi.org/10.1177/009286151004400507

Gamalo, M. A., Wu, R., & Tiwari, R. C. (2016). Bayesian approach to non-inferiority trials for normal means. *Statistical Methods in Medical Research*, *25*(1), 221–240. https://doi.org/10.1177/0962280212448723

Ghosh, P., Nathoo, F., Gönen, M., & Tiwari, R. C. (2011). Assessing noninferiority in a three-arm trial using the bayesian approach. *Statistics in Medicine*, *30*(15), 1795–1808. https://doi.org/10.1002/sim.4244

Ghosh, S., Ghosh, S., & Tiwari, R. C. (2016). Bayesian approach for assessing non-inferiority in a three-arm trial with pre-specified margin. *Statistics in Medicine*, *35*, 695–708. https://doi.org/10.1002/sim.6746

Ghosh, S., Paul, E., Chowdhury, S., & Tiwari, R. C. (2022). New approaches for testing non-inferiority for three-arm trials with poisson distributed outcomes. *Biostatistics*, *23*(1), 136–156. https://doi.org/10.1093/biostatistics/kxaa014

Ghosh, S., Chatterjee, A., & Ghosh, S. (2017). Non-inferiority test based on transformations for non-normal distributions. *Computational Statistics & Data Analysis*, *113*, 73–87. https://doi.org/10.1016/j.csda.2016.10.004

Ghosh, S., Tiwari, R. C., & Ghosh, S. (2018). Bayesian approach for assessing noninferiority in a three-arm trial with binary endpoint. *Pharmaceutical Statistics*, *17*(4), 342–357. https://doi.org/10.1002/pst.1851

Glimm, E., & Läuter, J. (2013). Some notes on blinded sample size re-estimation. *arXiv*, *1301.4167*. https://doi.org/10.48550/arXiv.1301.4167

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

Gould, A. L., & Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics - Theory and Methods*, *21*(10), 2833–2853. https://doi.org/10.1080/03610929208830947

Hasler, M., Vonk, R., & Hothorn, L. A. (2008). Assessing non-inferiority of a new treatment in a three-arm trial in the presence of heteroscedasticity. *Statistics in Medicine*, *27*(4), 490–503. https://doi.org/10.1002/sim.3052

Hida, E., & Tango, T. (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Statistics in Medicine*, *30*(3), 224–231. https://doi.org/10.1002/sim.4099

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (1998). Statistical principles for clinical trials E9. https://database.ich.org/sites/default/files/E9_Guideline.pdf

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use. (2001). Choice of control group in clinical trials E10. https://database.ich.org/sites/default/files/E10_Guideline.pdf

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

Kieser, M., & Friede, T. (2007). Planning and analysis of three-arm non-inferiority trials with binary endpoints. *Statistics in Medicine*, *26*(2), 253–273. https://doi.org/10.1002/sim.4113

Koch, G. G., & Tangen, C. M. (1999). Nonparametric analysis of covariance and its role in noninferiority clinical trials. *Drug Information Journal*, *33*(4), 1145–1159. https://doi.org/10.1177/009286159903300419

Kombrink, K., Munk, A., & Friede, T. (2013). Design and semiparametric analysis of non-inferiority trials with active and placebo control for censored time-to-event data. *Statistics in Medicine*, *32*, 3055–3066. https://doi.org/10.1002/sim.5769

Konietschke, F., Friedrich, S., Brunner, E., & Pauly, M. (2022). *rankFD: Rank-based tests for general factorial designs* [R package version 0.1.1]. https://CRAN.R-project.org/package=rankFD

Koti, K. M. (2007). Use of the fieller-hinkley distribution of the ratio of random variables in testing for noninferiority. *Journal of Biopharmaceutical Statistics*, *17*(2), 215–228. https://doi.org/10.1080/10543400601177335

Li, W., Zhang, Y., & Tang, N. (2023). Non-parametric non-inferiority assessment in a three-arm trial with non-ignorable missing data. *Mathematics*, *11*(1), 246. https://doi.org/10.3390/math11010246

Lui, K.-J., & Chang, K.-C. (2013). Notes on testing noninferiority in ordinal data under the parallel groups design. *Journal of Biopharmaceutical Statistics*, *23*(6), 1294–1307. https://doi.org/10.1080/10543406.2013.834923

Mielke, M., & Munk, A. (2009). The assessment and planning of non-inferiority trials for retention of effect hypotheses - towards a general approach. *arXiv*. https://doi.org/10.48550/arXiv.0912.4169

Mielke, M., Munk, A., & Schacht, A. (2008). The assessment of non-inferiority in a gold standard design with censored, exponentially distributed endpoints. *Statistics in Medicine*, *27*(25), 5093–5110. https://doi.org/10.1002/sim.3348

Munzel, U. (2009). Nonparametric non-inferiority analyses in the three-arm design with active control and placebo. *Statistics in Medicine*, *28*, 3643–3656. https://doi.org/10.1002/sim.3727

Mütze, T. (2023). *ThreeArmedTrials: Design and analysis of clinical non-inferiority or superiority trials with active and placebo control* [R package version 1.0-5]. https://CRAN.R-project.org/package=ThreeArmedTrials

Mütze, T., & Friede, T. (2017). Blinded sample size re-estimation in three-arm trials with 'gold standard' design. *Statistics in Medicine*, *36*(23), 3636–3653. https://doi.org/10.1002/sim.7356

Mütze, T., Konietschke, F., Munk, A., & Friede, T. (2017). A studentized permutation test for three-arm trials in the 'gold standard' design. *Statistics in Medicine*, *36*, 883–898. https://doi.org/10.1002/sim.7176

Mütze, T., Munk, A., & Friede, T. (2016). Design and analysis of three-arm trials with negative binomially distributed endpoints. *Statistics in Medicine*, *35*, 505–521. https://doi.org/10.1002/sim.6738

Paul, E., Tiwari, R. C., Chowdhury, S., & Ghosh, S. (2021). A more powerful test for three-arm non-inferiority via risk difference: Frequentist and bayesian approaches. *Journal of Applied Statistics*, *12*, 1–23. https://doi.org/10.1080/02664763.2021.1998391

Pigeot, I., Schäfer, J., Röhmel, J., & Hauschke, D. (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, *22*(6), 883–899. https://doi.org/10.1002/sim.1450

Pratley, R., Amod, A., Hoff, S. T., Kadowaki, T., Lingvay, I., Nauck, M., Pedersen, K. B., Saugstrup, T., & Meier, J. J. (2019). Oral semaglutide versus subcutaneous liraglutide and placebo in type 2 diabetes (pioneer 4): A randomised, double-blind, phase 3a trial. *Lancet*, *394*(10192), 39–50. https://doi.org/10.1016/S0140-6736(19)31271-1

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Röhmel, J. (1998). Therapeutic equivalence investigations: Statistical considerations. *Statistics in Medicine*, *17*(15-16), 1703–1714. https://doi.org/10.1002/(sici)1097-0258(19980815/30)17:15/16<1703::aid-sim972>3.0.co;2-g

Schulz, K. F., & Grimes, D. A. (2002). Blinding in randomised trials: Hiding who got what. *Lancet*, *359*(9307), 696–700. https://doi.org/10.1016/S0140-6736(02)07816-9

Simon, R. (1999). Bayesian design and analysis of active control clinical trials. *Biometrics*, *55*(2), 484–487. https://doi.org/10.1111/j.0006-341x.1999.00484.x

Tang, M.-L., & Tang, N.-S. (2004). Tests of noninferiority via rate difference for three-arm clinical trials with placebo. *Journal of Biopharmaceutical Statistics*, *14*(2), 337–347. https://doi.org/10.1081/BIP-120037184

Tang, N., & Yu, B. (2020). Simultaneous confidence interval for assessing non-inferiority with assay sensitivity in a three-arm trial with binary endpoints. *Pharmaceutical Statistics*, *19*(5), 518–531. https://doi.org/10.1002/pst.2010

Tang, N.-S., Yu, B., & Tang, M.-L. (2014). Testing non-inferiority of a new treatment in three-arm clinical trials with binary endpoints. *BMC Medical Research Methodology*, *14*, 134. https://doi.org/10.1186/1471-2288-14-134

Vanden Bossche, L., & Vanderstraeten, G. (2015). A multi-center, double-blind, randomized, placebo-controlled trial protocol to assess traumeel injection vs dexamethasone injection in rotator cuff syndrome: The traumeel in rotator cuff syndrome (traro) study protocol. *BMC Musculoskeletal Disorders*, *16*(1), 8. https://doi.org/10.1186/s12891-015-0471-z

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*(1-2), 65–71, discussion 71–2. https://doi.org/10.1002/sim.4780090113

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., & Proschan, M. (1999). Internal pilot studies i: Type i error rate of the naive t-test. *Statistics in Medicine*, *18*(24), 3481–3491. https://doi.org/10.1002/(SICI)1097-0258(19991230)18:24<3481::AID-SIM301>3.0.CO;2-C

Xing, B., & Ganju, J. (2005). A method to estimate the variance of an endpoint from an on-going blinded trial. *Statistics in Medicine*, *24*(12), 1807–1814. https://doi.org/10.1002/sim.2070

Zucker, D. M., Wittes, J. T., Schabenberger, O., & Brittain, E. H. (1999). Internal pilot studies ii: Comparison of various procedures. *Statistics in Medicine*, *18*(24), 3493–3509. https://doi.org/10.1002/(sici)1097-0258(19991230)18:24<3493::aid-sim302>3.0.co;2-2

A Appendix

Absolute margin approach

Rather than establishing non-inferiority of an experimental treatment to a standard reference treatment based on the observed effect of the reference treatment over a placebo, the absolute margin approach following the idea by Hida and Tango (2011) defines non-inferiority by means of a pre-defined fixed margin. That is, the experimental treatment must show that it is not worse than the reference treatment by more than this fixed margin. This concept mirrors the approach used in a two-arm design where the superiority of the reference treatment over placebo is assumed to be consistent across trials. However, this might not hold true resulting in a lack of of assay sensitivity in the two-arm design. Therefore, a placebo arm is included to ensure assay sensitivity. As a consequence, the trial must also prove the superiority of the reference treatment over the placebo.

That means, in order to establish non-inferiority in the three-arm design, two key elements need to be shown, namely that

1. the reference treatment is superior to placebo by more than Δ_{SUP}
2. the experimental treatment is non-inferior to the reference treatment with non-inferiority margin Δ_{NI}

Both margins Δ_{SUP} and Δ_{NI} must be non-negative. Assume that the clinical endpoints X_{ik} with $k = 1, ..., n_i$ and $i = \text{EXP}, \text{REF}, \text{PLA}$ are mutually independent and follow a distribution F_i with finite mean $\mathbb{E}[X_{ik}] = \mu_i$, common finite positive variance $Var[X_{ik}] = \sigma^2 > 0$ and finite fourth moment $\mathbb{E}[X_{ik}^4]$. In a setting where higher values of the outcome are associated with higher efficacy of the treatment, the following inequality must hold for the relationship between the respective μ_i:

$$\mu_{\text{PLA}} + \Delta_{\text{SUP}} < \mu_{\text{REF}} < \mu_{\text{EXP}} + \Delta_{\text{NI}}.$$

This results in the following two sets of hypotheses:

$$K_0 : \mu_{\text{REF}} \leq \mu_{\text{PLA}} + \Delta_{\text{SUP}} \quad \text{vs} \quad K_1 : \mu_{\text{REF}} > \mu_{\text{PLA}} + \Delta_{\text{SUP}} \tag{32}$$

$$H_0 : \mu_{\text{EXP}} \leq \mu_{\text{REF}} - \Delta_{\text{NI}} \quad \text{vs} \quad H_1 : \mu_{\text{EXP}} > \mu_{\text{REF}} - \Delta_{\text{NI}} \tag{33}$$

where K_0 refers to the hypothesis of assay sensitivity and H_0 to the non-inferiority hypothesis. Rearranging the hypotheses K_0 and H_0, replacing the respective μ_i's by the group-specific sample means $\bar{X}_{i\cdot}$ and dividing by an estimate of the respective pooled

standard deviations yields the following two test statistics

$$T_1 = \frac{\bar{X}_{\text{REF}\cdot} - \bar{X}_{\text{PLA}\cdot} - \Delta_{\text{SUP}}}{\hat{\sigma}_1\sqrt{1/n_{\text{REF}} + 1/n_{\text{PLA}}}} \quad \text{with} \quad \hat{\sigma}_1 = \frac{(n_{\text{REF}} - 1)S^2_{\text{REF}} + (n_{\text{PLA}} - 1)S^2_{\text{PLA}}}{n_{\text{REF}} + n_{\text{PLA}} - 2} \tag{34}$$

$$T_2 = \frac{\bar{X}_{\text{EXP}\cdot} - \bar{X}_{\text{REF}\cdot} + \Delta_{\text{NI}}}{\hat{\sigma}_2\sqrt{1/n_{\text{EXP}} + 1/n_{\text{REF}}}} \quad \text{with} \quad \hat{\sigma}_2 = \frac{(n_{\text{EXP}} - 1)S^2_{\text{EXP}} + (n_{\text{REF}} - 1)S^2_{\text{REF}}}{n_{\text{EXP}} + n_{\text{REF}} - 2}. \tag{35}$$

Here, S_i^2 denote the group-specific sample variances and n_i the respective group sample sizes.

When allowing for differing variances in the three groups, that is $Var[X_{ik}] = \sigma_i^2 > 0$, the pooled sample variances $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are not appropriate any longer for constructing the test statistics as in (34, 35). Replacing the pooled sample variances by the sample variances of the respective treatment groups yields the test statistics allowing for heterogeneous variances between the three groups

$$T_1 = \frac{\bar{X}_{\text{REF}\cdot} - \bar{X}_{\text{PLA}\cdot} - \Delta_{\text{SUP}}}{\sqrt{\frac{S^2_{\text{REF}}}{n_{\text{REF}}} + \frac{S^2_{\text{PLA}}}{n_{\text{PLA}}}}} \qquad T_2 = \frac{\bar{X}_{\text{EXP}\cdot} - \bar{X}_{\text{REF}\cdot} + \Delta_{\text{NI}}}{\sqrt{\frac{S^2_{\text{EXP}}}{n_{\text{EXP}}} + \frac{S^2_{\text{REF}}}{n_{\text{REF}}}}}. \tag{36}$$

Let $\boldsymbol{X}_{n_l}$ denote the random vector that contains the corresponding observations of the trial for the corresponding test statistic T_l with $l \in (1, 2)$, that is

$$\boldsymbol{X}_{n_1} = (X_{\text{REF},1}, \ldots, X_{\text{REF},n_{\text{REF}}}, X_{\text{PLA},1}, \ldots, X_{\text{PLA},n_{\text{PLA}}}) \tag{37}$$

$$\boldsymbol{X}_{n_2} = (X_{\text{EXP},1}, \ldots, X_{\text{EXP},n_{\text{EXP}}}, X_{REF,1}, \ldots, X_{\text{REF},n_{\text{REF}}}) \tag{38}$$

and denote $\mathbb{P}_1$ and $\mathbb{P}_2$ their respective probability measures.

Following the approach by Mütze et al. (2017), a permutation approach can be applied to approximate the distribution of both test statistics under the corresponding null hypothesis. A p-value for the respective test statistic can be derived with the following procedure:

1. Computation of the test statistics $T_1(\boldsymbol{X}_{n_1})$ and $T_2(\boldsymbol{X}_{n_2})$ for observed data $\boldsymbol{X}_{n_1}$ and $\boldsymbol{X}_{n_2}$

2. Permutation of the data with $\tau_{n_1}(\boldsymbol{X}_{n_1}) = (X_{n_1,\tau(1)}, \ldots, X_{n_1,\tau(n_1)})$ and $\tau_{n_2}(\boldsymbol{X}_{n_2}) = (X_{n_2,\tau(1)}, \ldots, X_{n_2,\tau(n_2)})$ denoting the randomly permutated vectors based of $\boldsymbol{X}_{n_1}$ and $\boldsymbol{X}_{n_2}$

3. Computation of the test statistics based on permutated data $T_1(\tau_{n_1}(\boldsymbol{X}_{n_1}))$ and $T_2(\tau_{n_2}(\boldsymbol{X}_{n_2}))$

4. Repetition of steps 2 and 3 for J times, e.g. 10,000 times (number of permutation replications)

5. Computation of the respective p-values as the number of times that the permutated test statistic $T_1(\tau_{n_1}(\boldsymbol{X}_{n_1}))$ or $T_2(\tau_{n_2}(\boldsymbol{X}_{n_2}))$ is as or more extreme than the test statistic based on the observed data $T_1(\boldsymbol{X}_{n_1})$ or $T_2(\boldsymbol{X}_{n_2})$, each divided by J, that is

$$\frac{1}{J}\sum_{j=1}^{J} I\big(T_1(\boldsymbol{X}_{n_1}) \leq T_1(\tau_{n_1}(\boldsymbol{X}_{n_1}))\big) \quad \text{and} \qquad \frac{1}{J}\sum_{j=1}^{J} I\big(T_2(\boldsymbol{X}_{n_2}) \leq T_2(\tau_{n_2}(\boldsymbol{X}_{n_2}))\big)$$

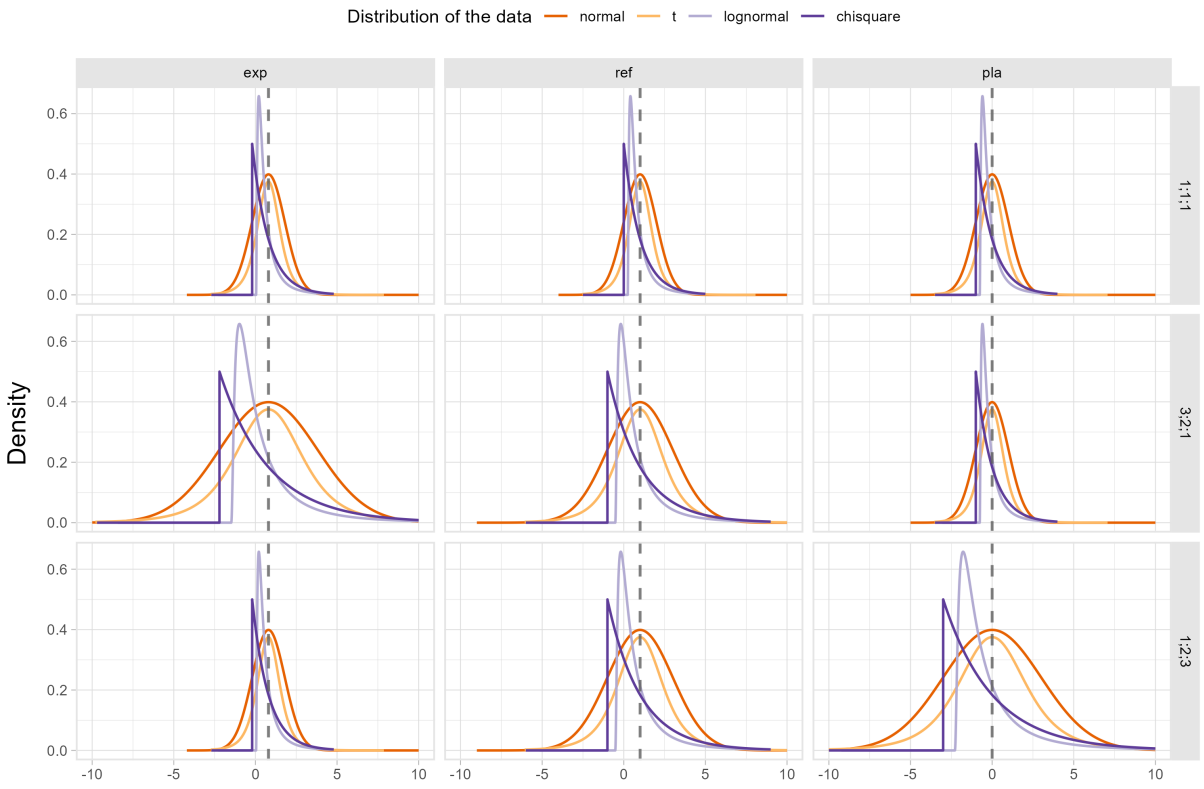
where I denotes the indicator function.

Based on the computed p-value, a test decision can be made. It should be noted that for the demonstration of non-inferiority of the experimental treatment, both test statistics must reject the null hypothesis.

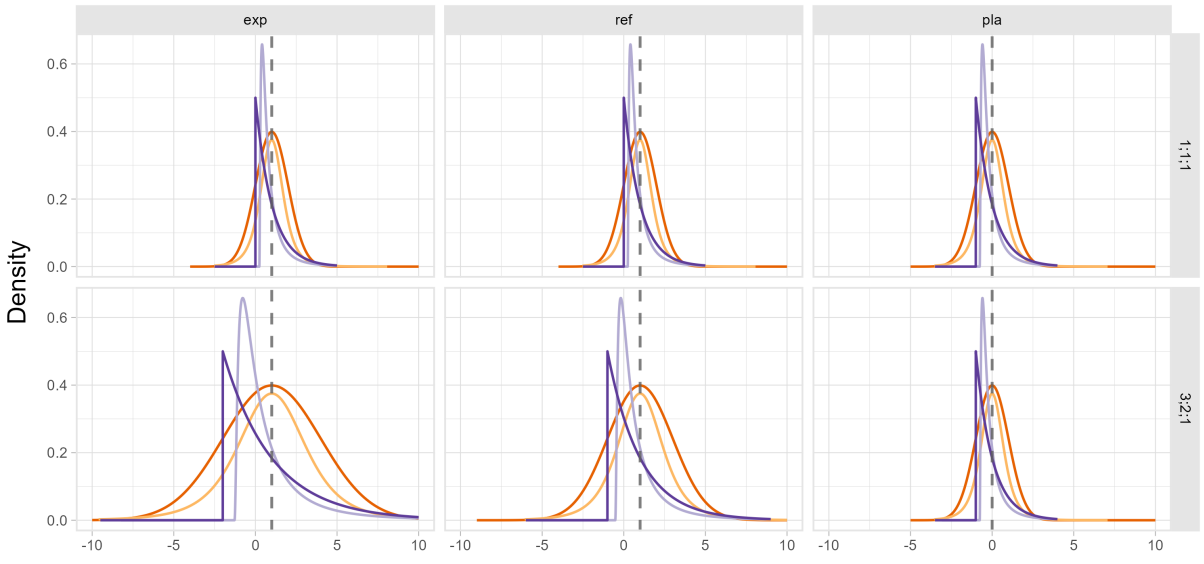
Distribution of the data for the simulation scenarios considered

The following paragraph briefly describes the characteristics of the data in the simulation scenarios considered for investigating the operating characteristics of the studentized permutation test, as outlined in Table 1. These descriptions are also valid for the scenarios considered when investigating the operating characteristics of the Munzel test. However, it should be noted that the heteroskedastic standard deviation scenarios were not considered for the Munzel test with data following a lognormal or χ^2-distribution. Additionally, t-distributed were are not simulated for the Munzel test.

Figure S1 illustrates the density curves of the underlying distributions of the data and standard deviation scenarios under the null hypothesis and alternative hypothesis, respectively. Specifically, Subfigure S1a depicts the scenarios under the null hypothesis, while Subfigure S1b presents the scenarios under the alternative hypothesis. The colours represent the different data-generating mechanisms. The figures are organised into rows and columns, with each column representing one of the three groups (EXP, REF, and PLA), and each row representing a standard deviation scenario. The dashed grey lines represent the true mean of the three groups, with $\mu_{\text{EXP}} = 0.8$, $\mu_{\text{REF}} = 1$, and $\mu_{\text{PLA}} = 0$ under the null hypothesis, and $\mu_{\text{EXP}} = 1$, $\mu_{\text{REF}} = 1$, $\mu_{\text{PLA}} = 0$ under the alternative hypothesis. Note that under the alternative hypothesis, additional values of μ_{EXP} are considered (refer to Table 1, column 3). For simplicity, only the case where $\mu_{\text{EXP}} = \mu_{\text{REF}}$ is depicted. To improve clarity, the x-axis limits are manually set from -10 to 10 in both figures.



(a) Density curves under H_0.



(b) Density curves under H_1.

Figure S1: Density curves for data following a normal, $t(4)$, lognormal and $\chi^2(2)$-distribution with $\mu_{\text{EXP}} = 0.8$ $(H_0)/\mu_{\text{EXP}} = 1$ (H_1), $\mu_{\text{REF}} = 1$, $\mu_{\text{PLA}} = 0$ for the standard deviation scenarios considered in the simulation study. The dashed grey lines depict the respective underlying μ_i.

Figure S1 illustrates that normal and t-distributed data exhibit a symmetrical density with the peak aligning with the true underlying mean of the corresponding group. The t-distributed data shows a higher peak and has heavier tails compared to the normal distribution. In the case of lognormal and χ^2-distributed data, the peak of the data is located to the left of the underlying mean for each group. The tails of both distributions extend to the right, indicating the positively skewed shape of these distributions. It is worth noting that the peak of data following a lognormal distribution is higher compared to the χ^2-distribution with 2 degrees of freedom. Depending on the underlying scenario in the group standard deviation, the density curves exhibit varying widths, with greater standard deviations resulting in wider tails.

Impact of the Coding Error

The following paragraph describes the impact of the coding error that was found in the process of this thesis. As described in section 3.1 the studentized permutation test permutes the original data J times where J defines the number of permutation replications. For each permutated dataset, the test statistic as in (6) is computed. In the package `ThreeArmedTrials` the variance of the permutated test statistic was calculated by

```
sigma2_Tperm <- n * (sigma2ExpEst / nExp +
    Delta^2 * sigma2RefEst / nExp +
    (1-Delta)^2 * sigma2PlaEst / nExp )
```

The formula divides the variance estimation of each group by the sample size of the experimental group `nExp`. While this formula is correct under a balanced group design, it leads to a miscalculation of the teststatistic's variance for unbalanced group designs. The correct formula is given by

```
sigma2_Tperm <- n * (sigma2ExpEst / nExp +
    Delta^2 * sigma2RefEst / nRef +
    (1-Delta)^2 * sigma2PlaEst / nPla )
```

The impact of the coding error became most clearly in the power simulation. In Figure S2 the power of the studentized permutation test and the Hasler test under the erroneous estimation are shown. Note that the rows hereby represent the five different allocation schemes and the two line types represent the group standard deviation scenarios.

Other than in Figure 4, the lines for the two tests, although still running in parallel, differ visibly for the unbalanced designs of $(3 : 2 : 1)$ and $(1 : \Delta : 1 - \Delta)$, that is $(1 : 0.8 : 0.2)$ (rows 3 and 5). It was suggested that more unbalanced designs in general created this difference. In the scenarios $(2 : 2 : 1)$ and $(3 : 3 : 1)$ (rows 2 and 4), however, the power curves aligned again. The mean power difference in percentage points between

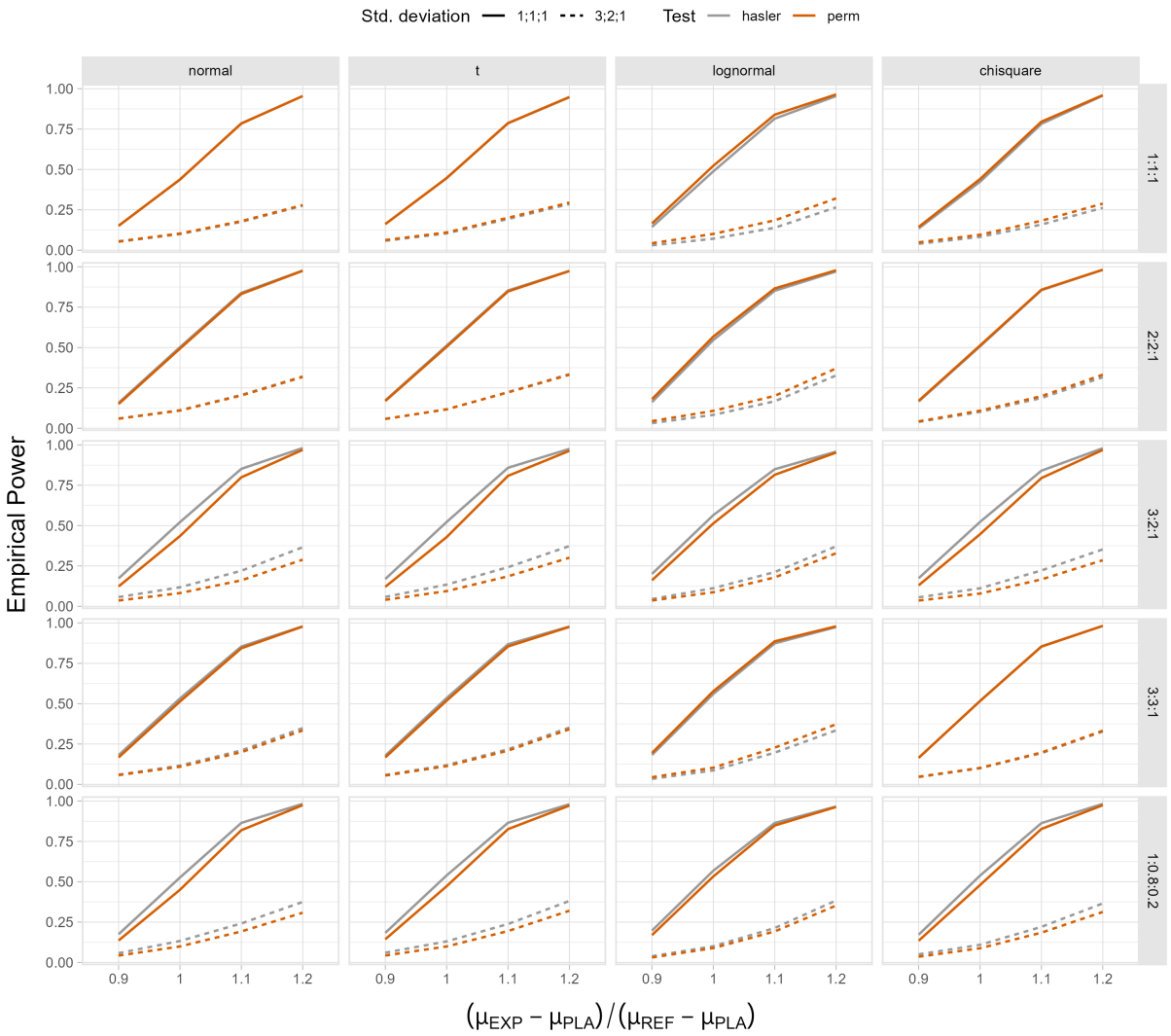


Figure S2: Impact of the coding error on the observed power of the studentized permutation test compared to the observed power of the Hasler test against $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ for a total sample size $n = 420$.

the studentized permutation test and the Hasler test for the different allocation schemes under the erroneous calculation is shown in Table S1.

Table S1: Impact of the coding error on the mean differences between the observed power of the Hasler test and the observed power of the studentized permutation test in percentage points by group design for a total sample size $n = 420$ across all $(\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}})$ and variance scenarios.

$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	Mean power Difference w/ coding error	Mean power Difference w/o coding error
$(1 : 1 : 1)$	1.23	1.23
$(2 : 2 : 1)$	0.53	1.03
$(3 : 2 : 1)$	-4.30	0.00
$(3 : 3 : 1)$	0.00	0.96
$(1 : \Delta : 1 - \Delta)$	-3.31	0.23

On average, the power of the studentized permutation test differed from the power of the Hasler test by about 4.3 percentage points under $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (3 : 2 : 1)$ and by 3.31 percentage points under $(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : \Delta : 1 - \Delta)$. For the other sample size allocations, the mean difference was reported considerably lower. These numbers confirmed the suggestion above. Also in comparison with the Hasler test, the reported type I error rates were considerably lower than the respective ones of the Hasler test for the group designs $(3 : 2 : 1)$ and $(1 : \Delta : 1 - \Delta)$. In contrast, without the coding error, the mean power differences for the unbalanced designs are not so apparent (see Table S1, column 3). Hence, under the respective hypothesis the type I error and the power of the studentized permutation test were estimated too low.

The error was discovered while searching why these differences between the allocation schemes occured and subsequently reported to the maintainer of the package. Based on the code of the package `ThreeArmedTrials` (Mütze, 2023) a corrected version of the studentized permutation test was developed. This version was used to re-conduct the simulations for the unbalanced designs. To account for the loss in simulation time, the replications of each simulation scenario were reduced from 10,000 to 5,000. The maintainer of the package promptly addressed and resolved the issue (see the commit of April 18, 2023). A correct version of the studentized permutation test is now implemented in the package `ThreeArmedTrials`.

Liberal behaviour for $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$

To understand the liberal behaviour of both tests for skewed data, one must keep in mind that the chosen heterogeneous standard deviations represent very extreme scenarios. Consider Figure S3. It shows the density curves of all considered underlying distributions of the data under the null hypothesis for $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$. The colours

represent the data-generating mechanisms and the rows the three groups EXP, REF and PLA. The grey line indicates the true mean of the three groups, that is $\mu_{\text{EXP}} = 0.8$, $\mu_{\text{REF}} = 1$ and $\mu_{\text{PLA}} = 0$. To enhance clarity, the limits on the x-axis were manually set from -10 to 10.

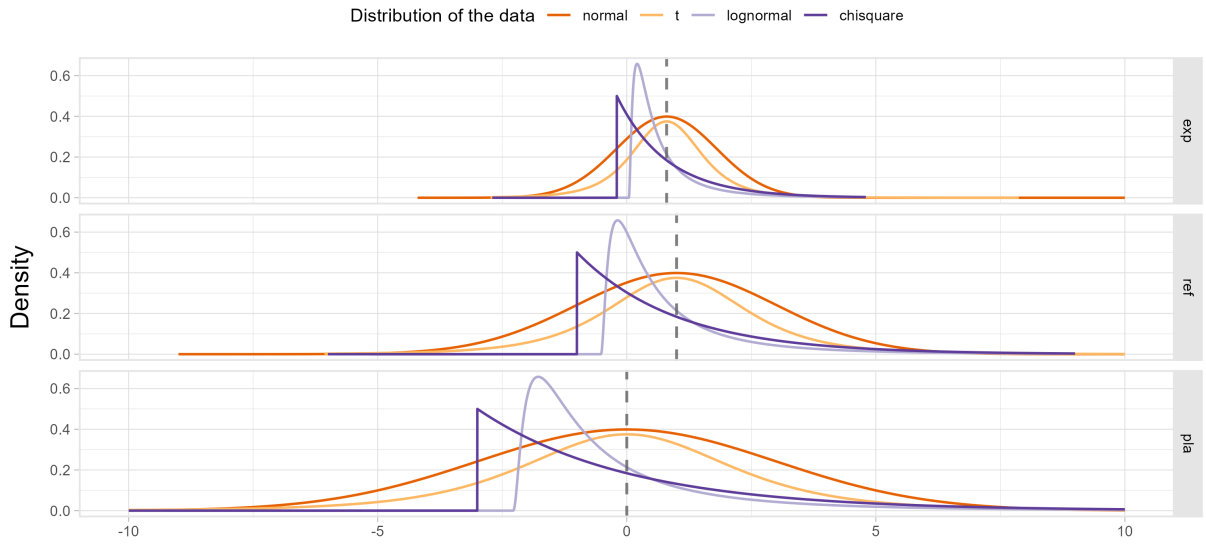


Figure S3: Density curves for data following a normal, $t(4)$, lognormal and $\chi^2(2)$-distribution with $\mu_{\text{EXP}} = 0.8$, $\mu_{\text{REF}} = 1$, $\mu_{\text{PLA}} = 0$ and $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$. The dashed grey lines depict the respective underlying μ_i.

Not surprisingly, the peak of the data coincides with the true underlying mean under normal and t-distributed data. As expected, for both lognormal and χ^2-distributed data, the peak does not coincide with the true mean rather is the mean located to the right of the data peak, representing the skewness of the data. If one simulates data for a balanced group design of $n = 999$ under $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ and evaluates the data, then one can explain the liberal behaviour by the distribution of the variance estimation. Figure S4 shows the variation of the variance estimation by the four underlying data-generating mechanisms for each of the three groups EXP, REF and PLA (on the rows) for a 5,000 replications by means of the density curve. The true variances under $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ are indicated by the grey line. Please note that the limits on the x-axis are manually set from 0 to 20 to provide a clearer view of how the variance estimation behaves across the different data-generating mechanisms. In the case of χ^2- and lognormal distributed data, the estimation ranged up to 100.

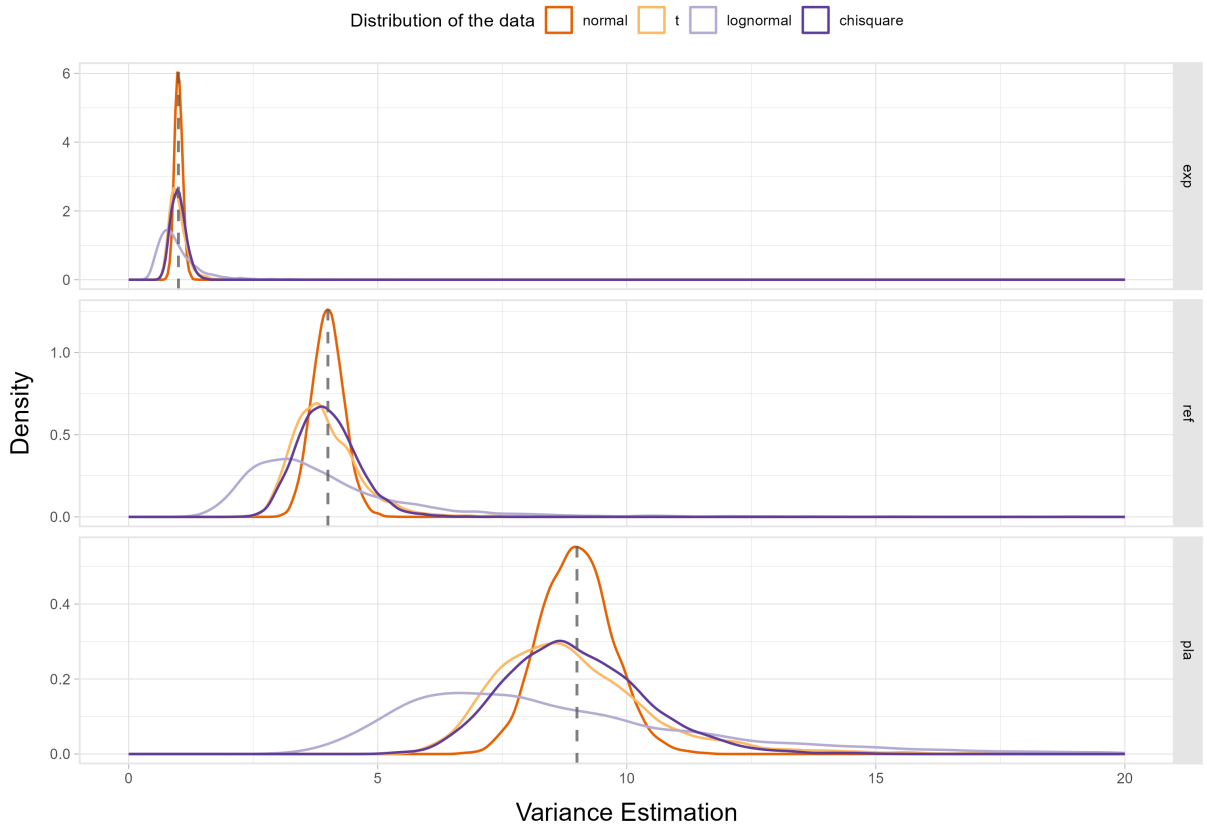


Figure S4: Density plot of the variance estimation of each treatment group for data following a normal, $t(4)$, lognormal and χ^2-distribution with $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ in the group design $(1:1:1)$ with total sample size $n = 999$. The dashed grey lines depict the respective underlying σ_i.

The variance estimation for normal data (dark orange line) appears symmetrical around the true value. However, the variance estimation for the other three data-generating mechanisms deviates considerably from that. The density curves for t-distributed and χ^2-distributed data (light orange and dark purple line) also appear symmetrical around the true value, but with larger tails compared to normal data, indicating greater variability in their variance estimation. Surprisingly, although the χ^2-distribution is also a skewed distribution, it shows a rather symmetrical curve in the variance estimation here. In contrast, the curve for lognormal data (light purple line) is noticeably different from that of normal, t and χ^2-distributed data. It is skewed and exhibits substantial variation across the range of the x-axis. The peak of the curve and the majority of the estimates are located to the left of the true value for all three groups, indicating that the variance estimation tends to be underestimated in this sample. It should be noted that the variance estimation, on average, would estimate the true underlying variance if this experiment were replicated multiple times, as the variance estimator is unbiased for any underlying distribution. Additionally, the variance estimation increases for all four data types as the true variance increases, which can be observed in the placebo group (third row, note the different scales on the y-axis).

Recall that the test statistic for both the studentized permutation test and the Hasler test is given by (6) where the variance estimator on the denominator is calculated by

$$\hat{\sigma}^2 = \frac{\hat{\sigma}^2_{\text{EXP}}}{w_{\text{EXP}}} + \Delta^2 \frac{\hat{\sigma}^2_{\text{REF}}}{w_{\text{REF}}} + (1-\Delta)^2 \frac{\hat{\sigma}^2_{\text{PLA}}}{w_{\text{PLA}}}.$$

Hence, the estimated group variances have a direct impact on the test statistic. If $\hat{\sigma}_i^2$ and thus $\hat{\sigma}^2$ tend to be small, then the denominator of the test statistic is also smaller and ultimately, the test statistic is estimated higher. Table S2 shows the quantiles of the estimated test statistic for a 5,000 replications of the above scenario.

Table S2: Quantiles of the estimated test statistic T_n for data following a normal, $t(4)$, lognormal and $\chi^2(2)$-distribution with $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$.

Data	Q0	Q25	median	mean	Q75	Q100
normal	-3.570	-0.675	-0.008	-0.006	0.666	3.865
$t(4)$	-3.378	-0.682	-0.024	-0.010	0.663	3.397
lognormal	-2.845	-0.656	-0.012	0.035	0.694	3.901
$\chi^2(2)$	-3.886	-0.689	0.029	0.023	0.695	4.148

Under the null hypothesis, the test statistic is expected to equal 0. One can see that, on average, the test statistic for normal and t-distributed data is very close to that, despite showing slight deviations that stem from the limited number of replications. For lognormal data, however, the estimate tends to be higher than for the other three underlying distributions (compare Q0 to Q75). Similarly, the estimated test statistic for χ^2-distributed data is on average more elevated while showing a greater variation. Hence, the test statistics under skewed data are on average higher and the estimation varies more than for normal or t-distributed data. Recall that the null hypothesis of the Hasler test is rejected if

$$T > t_{1-\alpha}(\hat{\nu}^{\text{het}}).$$

Higher values of the test statistic, therefore, yield more evidence for the alternative hypothesis and hence, the test rejects the null hypothesis more often although it is true. Similarly, the studentized permutation test rejects the null hypothesis more often. This leads to an increased type I error rate. The total sample size n acts multiplicatively on the test statistic and therefore reduces the elevated type I error rate. This effect also became evident in Figure 1, although it could not control it the desired level of $\alpha = 0.025$ in the considered scenarios. Therefore, the liberal behaviour of both tests in case of skewed data can be explained by the variation in the variance estimation which yields greater variation

in the test statistic. Ultimately, this leads to a higher rate of false positives. Generally, one might infer from these results that highly skewed data with great heteroskedastic variances tend to shift the data in such a way that tests that investigate the mean structure under the hypotheses are not suitable any longer.

Differing results for the type I error rate compared to the findings by Mütze et al. (2017)

To provide an explanation for the divergent outcomes observed in Section 3.2.1 compared to the findings of Mütze et al. (2017) regarding the type I error rate of the studentized permutation test for skewed data, it is important to consider that Mütze et al. (2017) examined the reverse effect under the hypotheses, which are given by

$$H_0 = \frac{\mu_{\text{EXP}} - \mu_{\text{PLA}}}{\mu_{\text{REF}} - \mu_{\text{PLA}}} \geq \Delta \text{ vs. } H_1 = \frac{\mu_{\text{EXP}} - \mu_{\text{PLA}}}{\mu_{\text{REF}} - \mu_{\text{PLA}}} < \Delta.$$

Under the variance scenario of $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 2; 3)$ Mütze et al. (2017) found a conservative behaviour of the studentized permutation test for skewed data whereas the results in Section 3.2.1 suggested a liberal behaviour for increasing standard deviation scenarios. First, it should be noted that Mütze et al. (2017) investigated variance scenarios rather than standard deviations. This, however, does not explain the differing results, since in the overall tendency, the studentized permutation test should behave similarly depending on the differences in the group variances. Rather, the different results can be explained by the reversed effects under the hypotheses, investigated by Mütze et al. (2017). The variance structure $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 2; 3)$ might be located at the other end of the distribution under the reversed null hypothesis. Therefore, the scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (3; 2; 1)$ under the above simulation corresponds to $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 2; 3)$ in Mütze et al. (2017). Both cases reported a conservative behaviour for lognormal and χ^2-distributed data. In a small simulation, the scenario of $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$ was simulated under the reversed hypotheses for lognormal and χ^2-distributed data to confirm that an increasing variance structure under the reversed hypotheses shows a conservative behaviour, as in Mütze et al. (2017). Figure S5 shows the results of the observed significance level. Note that the simulation setup is based on the setup as in Mütze et al. (2017). Thereby, μ_{PLA} is fixed with 5.5, μ_{REF} is varied from $\{0.5, 1, \ldots, 5\}$ and μ_{EXP} is adjusted accordingly $\mu_{\text{EXP}} = \Delta \cdot \mu_{\text{REF}} + (1 - \Delta) \cdot \mu_{\text{PLA}}$. On the x-axis, the varying μ_{REF} is displayed. The considered sample size allocations $n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}$ are $(1 : 1 : 1)$, $(2 : 2 : 1)$ and $(3 : 2 : 1)$ are indicated by different line types. The simulation is replicated 5,000 times for a total sample size of $n = 30$. The two grey lines depict the area of the nominal significance level $\alpha = 0.025 \pm$ two times the Monte Carlo error.

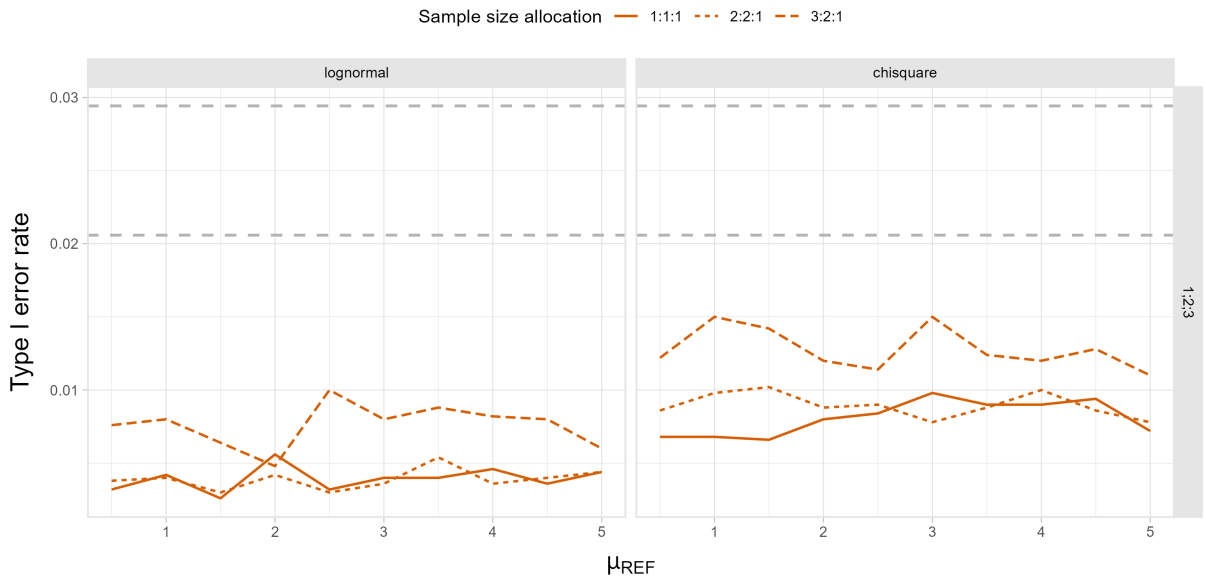


Figure S5: Actual significance level $\hat{\alpha}$ of the studentized permutation test against μ_{REF} for a total sample size $n = 30$ under the hypothesis $H_0 = (\mu_{\text{EXP}} - \mu_{\text{PLA}})/(\mu_{\text{REF}} - \mu_{\text{PLA}}) \geq \Delta$ for data following a lognormal and $\chi^2(2)$-distribution with $(\sigma_{\text{EXP}}; \sigma_{\text{REF}}; \sigma_{\text{PLA}}) = (1; 2; 3)$. The dashed grey lines depict the area of $\alpha = 0.025 \pm$ two times the Monte Carlo error.

Figure S5 demonstrates that, under reversed hypotheses, the increasing standard deviation structure leads to a conservative behaviour under the null hypothesis for skewed data. All lines are considerably lower than the targeted area between the two grey lines. These results confirm the suggestion above. Depending on the direction of the hypotheses, the variance or standard deviation structures lie on either side of the distribution under the null hypothesis rendering the test either conservative or liberal under skewed data. This demonstrates that the results presented here are consistent with the findings reported by Mütze et al. (2017).

Analysis of the power differences for skewed data in the fixed sample size design

Table S3 summarises the power differences of the studentized permutation test relative to the Hasler test as found in Section 3.2.2 and the deviation from the targeted 80% power level as found in row 1 of Figure 5 for data following a lognormal and χ^2-distribution.

Table S3: Differences in the observed power of the studentized permutation test compared to the observed power of the Hasler test and the deviation of the observed power in the fixed sample size design from the target power of 80% for data following a lognormal and $\chi^2(2)$-distribution with $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 1; 1)$.

$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}})$	Mean power difference with Hasler test	Deviation from target 80% power
lognormal-distributed data		
$(1 : 1 : 1)$	2.26	3.85
$(1 : \Delta : 1 - \Delta)$	0.50	1.17
$\chi^2(2)$-distributed data		
$(1 : 1 : 1)$	1.06	1.29

It is reasonable to expect that these two measures align closely. The difference in power relative to the Hasler test provides insight into the power behaviour when the Hasler sample size formula is used for sample size planning. In Section 3.2.2's power simulation, the power of the studentized permutation test surpassed that of the Hasler test by 2.26 and 0.50 percentage points. This was under homogeneous lognormal data for the balanced and the unbalanced design of $(1 : \Delta : 1 - \Delta)$, respectively, as illustrated in Table S3, column 2. However, on average, the power surpasses the targeted 80% power level by 3.85 percentage points in the balanced design and by 1.17 percentage points in the unbalanced design of $(1 : \Delta : 1 - \Delta)$ when planned with the Hasler formula (see Table S3, column 3). This suggests an overestimation of the trial's power, particularly in the balanced design. In other words, when analysing lognormal data using the studentized permutation test, the Hasler formula tends to yield slightly larger sample size estimations than required. Consequently, this leads to a modest overestimation of the trial's power. As seen in Figure 5, data following a χ^2-distribution reveals the same behaviour within a balanced design. The last row of Table S3 summarises the power differences for χ^2-distributed data.

In a balanced design, the power of the studentized permutation test outperformed the Hasler test by an average of 1.06 percentage points, as illustrated in Table S3, column 2. When planned with the Hasler formula, the power surpasses the 80% by 1.29 percentage points (see Table S3, column 3). Although the deviation from the target power level is still higher than the difference observed with the Hasler test, the difference is less apparent compared to lognormal data.

Observed power levels of the studentized permutation test with sample size re-estimation

Observed power levels of the studentized permutation test in the fixed sample size design

Cells colored green denote power levels that fall within the range of 0.8 ± the Monte Carlo error, while red cells mark instances where the targeted range was surpassed.

Table S4: Mean observed power of the studentized permutation test in percentage in the fixed sample size design without sample size re-estimation by variance scenario, underlying distribution of the data and group design across all n_1.

$(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}})$	Normal		$t(4)$		Lognormal		$\chi^2(2)$	
	1:1:1	1:0.8:0.2	1:1:1	1:0.8:0.2	1:1:1	1:0.8:0.2	1:1:1	1:0.8:0.2
1; 1; 1	79.854	80.338	80.15	80.206	83.854	81.174	81.294	79.836
3; 2; 1	41.696	44.08	42.616	44.694	46.62	45.83	43.068	44.4

Observed power levels with sample size re-estimation based on the UG and OSU estimator

Table S5: Observed power of the studentized permutation test in percentage with sample size re-estimation based on the UG and OSU estimator by n_1 and underlying distribution of the data for the scenario $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 1; 1)$.

	Normal		$t(4)$		Lognormal		$\chi^2(2)$	
n_1	UG	OSU	UG	OSU	UG	OSU	UG	OSU
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : 1 : 1)$								
30	76.2	76.16	71.06	72.4	64.84	66.88	71.68	73.58
40	76.38	76.9	72.9	74.4	65.06	68.66	73.76	75.54
50	78.36	78.66	74.88	75.04	67.04	70.14	74.06	76.72
60	79.08	76.88	76.04	75.48	68.34	72.1	76.12	76.28
70	78.66	78.36	74.9	76.8	70.04	72.96	75.86	79.1
80	79.38	78.02	76.64	75.5	71.94	73.98	77.24	77.72
90	78.7	78.7	76.24	76.78	72.44	74.8	78.24	78.78
100	78.58	78.08	76.8	77.22	72.04	75.28	77.12	78.66
110	78.78	79.18	77.48	77.66	73.56	75.46	76.9	78.92
120	79.6	79.26	77.2	76.9	72.94	76.46	79.42	79.46
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : \Delta : 1 - \Delta)$								
30	77.16	78.3	74.22	73.5	65.22	63.8	73.42	75.34
40	77.58	77.82	74.4	75.58	67.1	66.86	75.02	75.74
50	79.26	78.74	75.52	76.38	68.92	67.84	77.04	75.16
60	79.54	79	76.5	76.56	70.88	70.46	76.44	76.3
70	79.14	78.88	77.74	76.02	72.04	72.02	76.78	76.94
80	78.66	79.22	76.08	77.74	72.42	73	77.66	77.66
90	78.74	79.48	78.3	77.88	73.62	72.88	78.28	78.58
100	78.8	79.92	77.28	77.48	73.74	73.8	78.96	78.06
110	79.4	78.78	77.84	78.68	75.14	74.24	78.58	78.68
120	80.62	78.36	79.32	77.36	74.6	74.46	78.3	78.2

Table S6: Observed power of the studentized permutation test in percentage with sample size re-estimation based on the UG and OSU estimator by n_1 and underlying distribution of the data for the scenario $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$.

	Normal		$t(4)$		Lognormal		$\chi^2(2)$	
n_1	UG	OSU	UG	OSU	UG	OSU	UG	OSU
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : 1 : 1)$								
30	75.84	67.74	70.3	63.56	60.56	54.46	70.74	64.04
40	77.16	68.16	72.8	64.46	62.34	56.82	72.08	64.42
50	78.72	68.72	73.98	66.04	63.26	57.84	73.88	66.16
60	78.14	68.76	74.02	66	64.32	59.9	75.58	66.24
70	78.68	68.22	74.76	66.9	65.96	60.74	74.62	67.1
80	79	67.94	75.5	66.44	68.56	60.72	75.82	67.06
90	78.66	68.52	76.98	66	67.38	63.52	75.92	68.58
100	79.28	69.54	76.52	67.84	69.36	62.42	76.38	68.06
110	79.66	68.9	76.08	68.34	69.68	63.5	77	68.14
120	79.78	68.9	76.14	67	70.56	64.16	76.48	68.34
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : \Delta : 1 - \Delta)$								
30	77.18	77.44	72.3	74.36	60.08	60.16	71.92	72.04
40	78.6	78.38	75.16	73.8	64.16	63.4	74.68	73.78
50	78.86	78.96	75.4	75.06	66.66	66.18	73.88	75.34
60	80.08	80.72	75.18	76.4	68.14	66.82	75.5	75.64
70	78.08	79.26	75.46	76.36	69	67.86	77.46	77.04
80	79.16	80.24	77.42	76.66	69.64	68.3	76.68	75.8
90	79.76	79.86	77.74	77.92	70.38	67.64	76.88	77.6
100	78.96	79.1	77.38	77.74	71	70.28	76.06	77.12
110	79.58	78.52	77.64	78.64	71.66	70.26	77.64	77.3
120	79.34	79.86	79.12	77.38	70.86	71.54	78.4	77

Observed power levels with inflated sample size re-estimation based on the UG estimator

Table S7: Observed power of the studentized permutation test in percentage with inflated sample size re-estimation based on the UG estimator by n_1 and underlying distribution of the data for the scenario $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (1; 1; 1)$.

n_1	Normal	$t(4)$	$\chi^2(2)$
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : 1 : 1)$			
30	83.42	79.12	79.66
40	82.8	78.44	79.02
50	82.3	78.2	79.76
60	81.56	78.5	79.02
70	82.12	79.3	79.78
80	81.24	79.66	78.76
90	80.86	78.42	79.06
100	81.68	78.32	79.54
110	80.46	77.8	79.4
120	81.08	79.08	78.56
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : \Delta : 1 - \Delta)$			
30	82.94	78.9	79.44
40	82.2	79.64	79.02
50	82.2	79.32	79.42
60	81.5	79.12	78.28
70	81.24	77.94	78.98
80	81.34	79.84	79.14
90	81.36	79.26	80.42
100	80.54	78.98	79.16
110	81.4	79.18	80.28
120	81.26	78.76	79.6

Table S8: Observed power of the studentized permutation test in percentage with inflated sample size re-estimation based on the UG estimator by n_1 and underlying distribution of the data for the scenario $(\sigma^2_{\text{EXP}}; \sigma^2_{\text{REF}}; \sigma^2_{\text{PLA}}) = (3; 2; 1)$.

n_1	Normal	$t(4)$	$\chi^2(2)$
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : 1 : 1)$			
30	84.3	78.38	78.62
40	83.68	78.76	78.26
50	82.7	77.76	78.5
60	82.54	78.34	78.7
70	81.76	77.12	79.76
80	82.18	78.6	78.9
90	81.74	78.68	79.64
100	81.36	78.86	78.58
110	82.36	78.78	79.92
120	81.94	78.94	77.24
$(n_{\text{EXP}} : n_{\text{REF}} : n_{\text{PLA}}) = (1 : \Delta : 1 - \Delta)$			
30	83.04	79.16	78.54
40	82.18	77.9	78.54
50	81.72	79.14	79.1
60	80.58	78.84	79.06
70	81.98	78.98	78.36
80	82.38	78.42	78.7
90	80.8	78.42	78.58
100	80.1	78.74	78.62
110	80.36	79.02	79.3
120	80.78	79.12	78.54

Variance and covariance components of the nonparametric test statistic

The variance and covariance components of the test statistic T_n, as derived in (31), are given by

$$\hat{s}_{ii} = \frac{1}{n}\Big[\frac{1}{n_i(n_i-1)}\boldsymbol{R}_i^t \cdot \boldsymbol{R}_i + \frac{1}{n_i^2}\sum_{r=1}^{3}\frac{n_r}{n_r-1}\boldsymbol{R}_{ri}^t \cdot \boldsymbol{R}_{ri}\Big] \tag{39}$$

$$\hat{s}_{ij} = \frac{1}{n}\Big[\frac{1}{n_i n_j}\sum_{r=1}^{3}\frac{n_r}{n_r-1}\boldsymbol{R}_{ri}^t \cdot \boldsymbol{R}_{rj} - \frac{1}{n_i(n_j-1)}\boldsymbol{R}_i^t \cdot \boldsymbol{R}_{ij} - \frac{1}{n_j(n_i-1)}\boldsymbol{R}_j^t \cdot \boldsymbol{R}_{ji}\Big] \tag{40}$$

where

$$\boldsymbol{R}_i = \{R_{ik} - R_{ik}^{(i)} - \bar{R}_{i\cdot} + \bar{R}_{i\cdot}^{(i)}\} \tag{41}$$

$$\boldsymbol{R}_{ij} = \{R_{ik} - R_{ik}^{-j} - \bar{R}_{i\cdot} + \bar{R}_{i\cdot}^{-j}\} \text{ for } j \neq i \tag{42}$$

with

- R_{ik} = overall rank of X_{ik} among all n observations
- $R_{ik}^{(i)}$ = internal rank of X_{ik} among all n_i observations in the i-the treatment group
- R_{ik}^{-j} = partial rank of X_{ik} among all $n - n_j$ observations.

R-Code

The scripts used for analysis, along with the results obtained within this thesis, can be accessed from the following GitLab Repository.

Use of ChatGPT

In der hier vorliegenden Arbeit habe ich ChatGPT oder eine andere KI wie folgt genutzt:

- [ ] gar nicht
- [ ] bei der Ideenfindung
- [ ] bei der Erstellung der Gliederung
- [ ] zum Erstellen einzelner Passagen, insgesamt im Umfang von % am gesamten Text
- [ ] zur Entwicklung von Software-Quelltexten
- [X] zur Optimierung oder Umstrukturierung von Software-Quelltexten
- [X] zum Korrekturlesen oder Optimieren
- [X] Weiteres, nämlich: Hilfe mit Latex

Ich versichere, alle Nutzungen vollständig angegeben zu haben. Fehlende oder fehlerhafte Angaben werden als Täuschungsversuch gewertet.