Design and analysis of clinical non-inferiority trials with active and placebo control for count data

Masterarbeit in Mathematik
eingereicht an der Fakultät für Mathematik und Informatik
der Georg-August-Universität Göttingen
am 22. Oktober 2013

von

Tobias Mütze

Erstgutachter:

Prof. Dr. Axel Munk

Zweitgutachter:

Prof. Dr. Tim Friede

Contents

1 Introduction

In the clinical development of a new treatment, its safety and efficacy has to be assessed. Thereto, a clinical trial, which compares the new treatment with an already established treatment or a placebo, has to be performed. In this thesis, we denote the new treatment as the experimental treatment and an already established treatment as the reference treatment. Concerning these clinical trials, it is in general recommended to compare the experimental treatment, if possible, with a reference treatment and not with a placebo due to ethical concerns, confer point 32 in the declaration of Helsinki from the World Medical Association WMA (2008) as well as D'Agostino et al. (2003). The aim of such a trial is to prove that the experimental treatment is either superior or non-inferior to the reference treatment. Superiority of the experimental treatment over the reference treatment means that the experimental treatment is more effective than the reference treatment. The experimental treatment is non-inferior to the reference treatment if the difference between the treatments is negligible from a medical point of view, i.e. clinically not significant. For the principles of superiority and non-inferiority trials confer ICH (2010). To illustrate the statistical hypotheses for superiority and non-inferiority, let $\lambda_E > 0$ and $\lambda_R > 0$ be parameters of a specific distribution which are associated with the efficacy of the experimental and the reference treatment. If we assume that smaller values correspond to a more efficient treatment, the statistical hypothesis for testing superiority of the experimental versus the reference treatment is given by

$$H_0 : \lambda_E \geq \lambda_R \qquad \text{versus} \qquad H_1 : \lambda_E < \lambda_R.$$

With δ being a prespecified, positive real number denoting the non-inferiority margin, the statistical hypothesis testing whether the experimental treatment is non-inferior to the reference treatment is given by

$$H_0 : \lambda_E \geq \lambda_R + \delta \qquad \text{versus} \qquad H_1 : \lambda_E < \lambda_R + \delta.$$

Superiority and non-inferiority are shown if the corresponding hypothesis is rejected. The hypotheses reveal the important difference between superiority and non-inferiority that superiority does not include the possibility of experimental and reference treatment being equally effective, i.e. $\lambda_E = \lambda_R$. Thus, superiority trials cannot determine whether the experimental treatment is at least as effect as the reference treatment. Moreover, superiority trials have the disadvantage that if the efficacy difference of the treatments is small, many patients are needed to prove superiority. Thus, such trials can get very cost-intensive and take many years. Therefore, non-inferiority trials becoming increasingly popular, confer Figure 1.1 in Mielke (2010) which shows the increasing number of publications and citations for this topic up to 2009. However, non-inferiority trials have also several weaknesses, for instance the difficulty of determining the non-inferiority margin δ. For further discussions about non-inferiority trials we refer to Snapinn et al. (2000), Rothmann et al. (2003), and Fleming (2008).

So far, we have only focused on trials with just an active control. However, just actively controlled trials have several disadvantages which are addressed by various publications including, among others, Hill (1994), Temple and Ellenberg (2000), D'Agostino et al. (2003), and Koch and Röhmel (2004). In particular, Lewis et al. (2002) discusses why and when trials with a placebo control are consistent with the declaration of Helsinki WMA (2008). In the following, we discuss these disadvantages of trials and reason why a placebo should be included. A disadvantage of trials with just an active control is that they do not prove whether the experimental treatment is superior to placebo, which is crucial to guarantee the efficacy of the treatment. The superiority of the reference treatment versus the placebo has in general already been proved but for instance due to a different study design or a general change in medical practice, this superiority has to be proved again. For instance, a different study design results from a modification of the study duration, patient population, or doses. In general, the property that historical evidence also holds in a new trial is called *constancy assumption*, confer Section 3.2.1 in D'Agostino et al. (2003). Moreover, if the constancy assumption holds for the superiority of the reference treatment over the placebo, including placebo can still be necessary to prove that the experimental treatment is more effective than the placebo which does not follow necessarily from the non-inferiority $\lambda_E < \lambda_R + \delta$ and the superiority $\lambda_R < \lambda_P$. Therefore, including a placebo in a clinical study can be necessary. Nevertheless, if a placebo is included, the reference treatment should still be part of the study, since it might be well-established and the experimental treatment has still to be compared with it. A study design including an experimental and a reference treatment as well as a placebo is called gold standard design. Studies with this design will hereinafter just be denoted as three-arm trials.

Due to their increasing importance, in this thesis we focus on trials testing non-inferiority instead of superiority of the experimental versus the reference treatment. We define non-inferiority through the so-called retention of effect hypothesis. The particularity of a retention of effect hypothesis is that non-inferiority of the experimental versus the reference treatment is defined with respect to the placebo response, i.e. we study the hypothesis

$$H_0^{RET}: (\lambda_P - \lambda_E) \leq \Delta(\lambda_P - \lambda_R) \qquad \text{versus} \qquad H_1^{RET}: (\lambda_P - \lambda_E) > \Delta(\lambda_P - \lambda_R)$$

with $\Delta \in (0,1)$ the prespecified clinical relevance, also called non-inferiority margin. Analogously to Pigeot et al. (2003), we motivate the retention of effect hypothesis H_0^{RET} by comparing the efficacies λ_E and λ_R with the non-inferiority hypothesis $H_0 : \lambda_E \geq \lambda_R + \delta$. Now, if the prespecified clinical relevance δ is defined by a fraction $f \in (0,1)$ of how much more effective the reference is compared to placebo, i.e. $\delta = f(\lambda_P - \lambda_R)$, we obtain the testing problem

$$H_0 : \lambda_E \geq f\lambda_P + (1-f)\lambda_R \qquad \text{versus} \qquad H_1 : \lambda_E < f\lambda_P + (1-f)\lambda_R.$$

Substituting $\Delta := 1 - f$ and rearranging the hypothesis yield the retention of effect hypothesis H_0^{RET}. Thus, the retention of effect hypothesis is basically a non-inferiority hypothesis with clinical relevance defined by the efficacy difference of the reference treatment and the placebo.

As mentioned above, the retention of effect hypothesis is only meaningful if the reference treatment is more effective than the placebo, i.e. if $\lambda_R < \lambda_P$ holds. Otherwise, we would compare the experimental treatment with a reference treatment which is not even as effective as the placebo. If the superiority of the reference treatment over the placebo has not been established previously or the constancy assumption does not hold, the hypothesis

$$H_0^{RP}: \lambda_R \geq \lambda_P \qquad \text{versus} \qquad H_1^{RP}: \lambda_R < \lambda_P$$

has to be tested additionally to the retention of effect hypothesis. Analogously to the reference treatment, if the superiority of the experimental treatment over the placebo has not been tested before, the hypothesis

$$H_0^{EP}: \lambda_E \geq \lambda_P \qquad \text{versus} \qquad H_1^{EP}: \lambda_E < \lambda_P$$

has to be tested, too. The property of a clinical trial that active treatments are superior to the placebo is called *assay sensitivity*. More precisely the ICH (2000) guideline E10 defines assay sensitivity as “the ability to distinguish an effective treatment from a less effective or ineffective treatment”. Thus, in the setting of a three-arm trial assay sensitivity corresponds

to superiority of the experimental or the reference treatment over placebo.
Overall, the aim of a three-arm non-inferiority trial is to prove both non-inferiority of the experimental versus the reference treatment and assay sensitivity. A trial is successful if all hypotheses can be rejected. Hereafter, we refer to the test which aims to show both assay sensitivity and non-inferiority as the *test procedure*. The level of significance is determined for each test separately to control the rate of a false rejection for each of the hypotheses. However, the level of significance for the test procedure is controlled, i.e. at most α, if each hypothesis is tested with a level of significance α. The exact level of significance of the test procedure depends on the correlation of the different tests. Even if the level of significance is determined for each test, the power is reported for the test procedure.
In clinical trials, an endpoint denotes a specific characteristic which is measured for each patient. We test the hypotheses stated above through these measurements. In this thesis, we assume that the observations can be modelled as overdispersed count data. More precisely, we consider the observations to be negative binomially distributed with the expectation, which is denoted as the rate, indicating the active treatment efficacies λ_E and λ_R as well as the placebo response λ_P. As the name implies, count data describes data where for each patient the measurement of the endpoint is a natural number. Examples for such endpoints are the number of exacerbations in trials with patients suffering from chronic obstructive pulmonary disease (COPD) or the number of lesions in trials in (MS), confer Section 2. Further, a random variable is called overdispersed if the variance exceeds the expectation. For a general consideration of overdispersion models we refer to Hinde and Demétrio (1998) and for modelling overdispersed count data see Chapters 2.3, 2.4, and 2.6 in Winkelmann (2003). One reason for overdispersion is that patients which all receive the same treatment respond to the treatment very differently. Besides, important predictors are not included in the model, for instance because they are not known or cannot be measured.

1.1 State of research

In the following, we focus on the state of research on tests for non-inferiority and assay sensitivity in three-arm trials. To our knowledge, there are no publications about tests for the retention of effect hypothesis for overdispersed count data in general or for the negative binomial distribution in particular. However, the theory of the retention of effect hypothesis for other distributions has been subject of a number of recent publications. Pigeot et al. (2003) studied the retention of effect hypothesis for normally distributed endpoints with homogeneous variance. Regarding this setting, a sample size recalculation procedure has been introduced by Schwartz and Denne (2006). The case of normally distributed endpoints with heterogeneous group variances has been studied by Hasler et al. (2008). The Wald-type test theory for a generalized retention of effect hypothesis has been established by Mielke (2010) for parametric families whose parameter has an asymptotic normally distributed maximum-likelihood estimator and a non-singular covariance matrix. In Mielke (2010), this theory has been applied to binary, Poisson, and censored, exponentially distributed endpoints. The retention of effect hypothesis for binary distributed endpoints has been studied by Kieser and Friede (2007) too. A nonparametric retention of effect hypothesis defined through relative effects has been introduced by Munzel (2009).

Besides the retention of effect hypothesis, Kombrink et al. (2013) established a semiparametric analysis for censored time-to-event data in a three-arm trial. A general approach for calculating the sample size of a three-arm trial where the active treatments and the placebo are compared pairwise has been established by Stucke and Kieser (2012).

Since we assume that the treatment efficacies and the placebo response correspond to rates of negative binomial distributions, testing superiority of the experimental or the reference treatment over placebo corresponds to comparing rates of negative binomial distributions. Thereto, Wald-type tests for the logarithmized rate are commonly used, confer Friede and Schmidli (2010) and Zhu and Lakkis (2013). These publications, however, do not address the actual level of significance of the test. Aban et al. (2009) introduced tests for the equality of two negative binomial rates and compared their actual levels.

1.2 Content and Organization

In this thesis we study three-arm non-inferiority trials with negative binomially distributed endpoints. To assess non-inferiority of the experimental versus the reference treatment, we have to choose how the retention of effect hypothesis as well as assay sensitivity should be tested. Additionally, the sample size as well as its allocation have to be determined. In Section 2, we motivate and specify the statistical model. We start this section by introducing mixed Poisson distributions as one possibility to model overdispersed count data and establishing the negative binomial distribution as a mixed Poisson distribution. Subsequently, to motivate the statistical model, we discuss examples of endpoints in clinical trials in chronic obstructive pulmonary disease (COPD) and multiple sclerosis (MS) which are commonly modelled as overdispersed count data. Finally, we define the statistical model for this thesis through negative binomially distributed endpoints in Section 2.3. Sections 3 and 4 are about the statistical results and concepts we apply when deducing tests for assay sensitivity and the retention of effect hypothesis. More precisely, in Section 3, we establish the maximum-likelihood estimators for the parameters of the negative binomial model as well as describe the concept of restricted maximum-likelihood estimation. Afterwards, in Section 4, we introduce the types of statistical tests which we will be applied to the hypotheses defined above. Firstly, we describe Wald-type tests as tests whose test statistics are asymptotically standard normally distributed at the boundary of the hypothesis. Secondly, we focus on exact as well as asymptotic permutation tests.

With the knowledge of Sections 3 and 4, we establish different Wald-type tests as well as an exact permutation test for the assay sensitivity and compare the actual level of significance of these tests by Monte-Carlo simulations in Section 5. We will see that the actual level of significance of the Wald-type tests depend among others on the sample size allocation and that not all Wald-type tests are appropriate to test assay sensitivity. Especially, the permutation test outperforms the Wald-type tests concerning being neither liberal nor conservative.

Section 6 deals with testing the retention of effect hypothesis for negative binomially distributed endpoints and planning the sample size for these tests. To test the retention of effect hypothesis, we introduce different Wald-type tests using results from Mielke (2010) in Section 6.1. The Wald-type tests differ in how the variance for the test statistic is estimated. Thereto, we describe an unrestricted and a restricted maximum-likelihood as well as a sample variance estimator. Additionally, in Section 6.2, we establish an asymptotic permutation test by the central limit theorem for conditional permutation distributions from Janssen (1997). After establishing the tests, in Section 6.3 we focus on planning a trial which aims to test the retention of effect hypothesis with a certain power for a fixed alternative. Thereto, we state sample size formulas and as well as the sample size allocation maximizing the power which has been established by Mielke (2010). Additionally, we introduce different restrictions for the sample size and extend the theory of power maximizing allocation with respect to these restrictions. We extend the theory of allocating the sample size by maximizing the power with respect to restrictions about the sample size allocation. Since the properties of the tests and the results for planning the sample size only hold asymptotically, we study the finite sample size properties with Monte-Carlo simulations in Section 6.4. Firstly, we study the actual level of significance and the power of the different tests. These simulations show that the Wald-type test with a restricted maximum-likelihood variance estimator performs best for the considered scenarios. Since the Wald-type tests with a maximum-likelihood variance estimator is based on the assumption of negative binomially distributed observations, we study how robust the test are concerning differently distributed observations. We see that the mentioned Wald-type tests are not robust and become liberal if the distribution changes. Most of the publications about the retention of effect hypothesis claim that the power of the test procedure is approximately the power of the retention of effect hypothesis. We verify the assertion in the case of negative binomially distributed endpoints in the end of Section 6.4. We conclude this thesis with discussing the key results and giving an outlook for further research on this topic in Section 7. In Appendix A we state the technical proofs of this thesis.

2 Three-arm Trials and Overdispersed Count Data

In this section, we introduce mixed Poisson distributions as one possibility to model overdispersed count data. Especially, we define three mixed Poisson distributions: the negative binomial distribution, the Poisson–lognormal distribution, and the Poisson–inverse-Gaussian distribution. Subsequently, we consider clinical trials in multiple sclerosis and chronic obstructive pulmonary disease as examples of trials with endpoints commonly modelled as overdispersed count data. Taking these examples into account, we specify the statistical model as three-armed trials with negative binomially distributed observations.

2.1 Mixed Poisson Distributions

The Poisson distribution is a discrete single–parameter distribution with probability mass function

$$\mathbb{P}_\lambda(X = x) = \frac{\lambda^x \exp(-\lambda)}{x!} \mathbb{1}_{\mathbb{N}_0}(x),$$

with rate $\lambda > 0$ and $\mathbb{1}_{\mathbb{N}_0}(\cdot)$ the indicator function on the natural numbers $\mathbb{N}_0$. The expectation as well as the variance of X are equal to the rate λ. However, if data from an overdispersed distribution, i.e. data from a distribution with variance exceeding expectation, is modelled as Poisson distributed, a too small variance is assumed. As a consequence, if statistical tests for Poisson distributed data are applied to overdispersed count data, the actual level of significance of these tests may be larger than the nominal level. The problem of wrongly modelled count data in clinical trials is addressed in several publications. For instance, Keene et al. (2008b) showed that the analysis of count data in Calverley et al. (2003) with a Poisson regression does not take the overdispersion of the data into account. As a consequence, in this thesis we focus on overdispersed count data. Different approaches of modelling overdispersed count have been suggested by Chapter 2.3, 2.4, and 2.6 of Winkelmann (2003). In this thesis, we consider mixed Poisson distributions as one possibility to model overdispersed count data. Mixed Poisson distributions are Poisson distributions whose rates are assumed to be random, i.e. different observed values of a mixed Poisson distributed random variable are basically observations of Poisson distributed random variables with different rates. Depending on the distribution of the rate, we obtain different mixed Poisson distributions but, as we will see in the next subsection, all mixed Poisson distributions are overdispersed. For further information about mixed Poisson distribution, we refer to Karlis and Xekalaki (2005) who reviewed mixed Poisson distributions, especially, summarized properties and listed publications about different choices for the distribution of the Poisson rate. Here, we regard three different mixed Poisson distributions, namely a Poisson–gamma mixture, which is commonly known as the negative binomial distribution, a Poisson–inverse-Gaussian mixture, and a Poisson–lognormal mixture distribution. The choice of the different mixed Poisson distributions are motivated by clinical trials in multiple sclerosis (MS) and chronic obstructive pulmonary disease (COPD), confer Sections 2.2.1 and 2.2.2.

As mention above, mixed Poisson distributions assume the rate of a Poisson distribution to be random. We start by stating this definition more precisely and proving some properties of mixed Poisson distributions.

Definition 2.1 (Mixed Poisson distribution)**.** A random variable X is distributed according to a mixed Poisson distribution if a random variable $Z > 0$ exists, such that the conditional random variable $X|Z$ is Poisson distributed with rate Z, i.e.

$$X|Z \sim \mathrm{Pois}(Z).$$

We denote the random variable Z as the mixing variable and its distribution as the mixing distribution.

Lemma 2.2. *Let X be a mixed Poisson distribution, and be $f_Z(\cdot)$ the density of the mixing variable $Z > 0$ with respect to a probability measure μ. Then, the probability mass function of X is given by*

$$\mathbb{P}(X = x) = \int_{\mathbb{R}} \mathbb{P}(X = x|Z = z) f_Z(z)\mu(\mathrm{dz}) = \mathbb{1}_{\mathbb{N}_0}(x) \int_{\mathbb{R}} \frac{z^x}{x!} e^{-z} f_Z(z)\mu(\mathrm{dz}).$$

To show that mixed Poisson distributions are always overdispersed, we calculate their expectation and variance by means of the laws of total expectation and total variance.

Theorem 2.3. *Let Z be an arbitrary mixing distribution with existing first and second moment. Furthermore, let X be a mixed Poisson distributed random variable with mixing variable Z. Then, the expectation and variance of X are*

$$\mathbb{E}[X] = \mathbb{E}[Z],$$
$$\mathrm{Var}[X] = \mathbb{E}[Z] + \mathrm{Var}[Z].$$

Proof. The assertions follows immediately from the laws of total expectation and total variance as well as from property that the expectation and variance of a Poisson distributed random variable are equal to the rate. □

Hence, the expectation and the variance of a mixed Poisson distribution are determined by the expectation and the variance of the mixing distribution. Furthermore, a mixed Poisson distribution is always overdispersed.
In the following, we introduce three different mixed Poisson distributions.

Definition 2.4 (Negative binomial distribution)**.** A random variable X is called negative binomially distributed with parameters $\lambda, \phi > 0$ if it has the probability mass function

$$\mathbb{P}(X = x) = \frac{\Gamma(x + {}^{1}\!/\!_{\phi})}{\Gamma({}^{1}\!/\!_{\phi})x!} \left(\frac{1}{1 + \phi\lambda}\right)^{{}^{1}\!/\!_{\phi}} \left(\frac{\lambda\phi}{\lambda\phi + 1}\right)^{x} \mathbb{1}_{\mathbb{N}_0}(x).$$

The next theorem proves that a negative binomial distribution is a Poisson–gamma mixture. Thereto, a random variable Z is said to be gamma distributed with parameters $\alpha, \beta > 0$, if it has the probability density function

$$f_Z(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z} \mathbb{1}_{(0,\infty)}(z)$$

with $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} \mathrm{d}t$ the gamma function.

Theorem 2.5. *Let Z be gamma distributed with parameters $\alpha = 1/\phi$ and $\beta = 1/(\lambda\phi)$. In addition, let the random variable $X|Z$ be Poisson distributed with rate Z. Then, the random variable X is negative binomially distributed with rate λ and shape parameter ϕ.*

Proof. The assertion has been proved with shape parameter $\alpha = 1/\phi$ on pages 35 and 36 in Winkelmann (2003). □

According to 26.1.31 in Abramowitz and Stegun (1970), the expectation and the variance of a gamma distributed random variable Z are α/β and α/β^2, respectively. Hence, from Theorem 2.3 we obtain immediately the expectation and the variance of a negative binomially distributed random variable.

Corollary 2.6. *The expectation and the variance of the negative binomially distributed random variable X with rate λ and shape parameter ϕ are given by $\mathbb{E}[X] = \lambda$ and $\mathrm{Var}[X] = \lambda(1 + \lambda\phi)$, respectively.*

In addition to being a mixed Poisson distribution, the negative binomial distribution can also be motivated by means of a Bernoulli process.

Remark 2.7. Let $Y_1, Y_2, \dots$ be independent and identically Bernoulli distributed random variables with success probability $p \in (0,1)$, i.e. $\mathbb{P}(Y_1 = 1) = p$ and $\mathbb{P}(Y_1 = 0) = 1 - p$. Furthermore, let $k, r \in \mathbb{N}$ be a natural numbers. Then, the random variable X which describes the number of failures k until the r-th success of the Bernoulli process occurred is called negative binomially distributed and its probability mass function is given by

$$\mathbb{P}(X = k) = \binom{k+r-1}{r-1} p^r (1-p)^k.$$

The negative binomial distribution as in Definition 2.4 is obtained by the property $\Gamma(n) = (n-1)!$ as well as the substitutions $r = \phi$ and $p = \lambda/(r+\lambda)$.

After describing the negative binomial distribution, we introduce two additional mixed Poisson distributions. As mentioned above, the choice of these distributions is motivated by the examples in Section 2.2.1 and 2.2.2.

Definition 2.8 (Poisson–inverse-Gaussian distribution)**.** A random variable X is distributed according to a Poisson–inverse-Gaussian distribution with parameters $\lambda, \theta > 0$ if its probability mass function is given by

$$\mathbb{P}(X = x) = \left(\frac{2\theta}{\pi}\right)^{1/2} \frac{1}{x!} e^{\frac{\theta}{\lambda}} \left(\frac{\theta}{2\left(1 + \frac{\theta}{2\lambda^2}\right)}\right)^{\frac{2x-1}{4}} K_{x-1/2}\left(\sqrt{2\theta\left(1 + \frac{\theta}{2\lambda^2}\right)}\right) \mathbb{1}_{\mathbb{N}_0}(x)$$

with $K_\nu(x)$ defined as

$$K_\nu(x) := \frac{\pi}{2} \frac{I_{-\nu}(x) - I_\nu(x)}{\sin(\nu\pi)}$$

and

$$I_\nu(x) := \sum_{m=0}^{\infty} \frac{(x/2)^{\nu+2m}}{m!\Gamma(m+\nu+1)}.$$

the modified Bessel-function of the first kind.

The next theorem states that the Poisson–inverse-Gaussian distribution is a mixed Poisson distribution.

Theorem 2.9. *Let Z be an inverse-Gaussian distributed random variable with parameters $\lambda, \theta > 0$, i.e. Z has the probability density function*

$$f_Z(z) = \left(\frac{\theta}{2\pi z^3}\right)^{1/2} \exp\left(\frac{-\theta(z-\lambda)^2}{2\lambda^2 z}\right) \mathbb{1}_{(0,\infty)}(z).$$

Furthermore, let $X|Z$ be Poisson distributed with rate Z. Then, X is Poisson–inverse-gaussian distributed with parameters $\lambda, \theta > 0$.

Proof. The assertion has been proved by Holla (1967). □

According to equation (7) in Holla (1967), the expectation and the variance of a Poisson–inverse-Gaussian distribution are given by λ and $\lambda(1 + \lambda^2/\theta)$, respectively.
Last but not least, we define the Poisson–lognormal distribution.

Definition 2.10 (Poisson–lognormal distribution)**.** The distribution of a random variable X is called Poisson–lognormal distribution with parameters μ and σ^2 if its probability mass function is given by

$$\mathbb{P}(X = x) = \frac{\mathbb{1}_{\mathbb{N}_0}(x)}{\sqrt{2\pi\sigma^2}x!} \int_0^\infty z^{x-1} \exp\left(-\frac{(\ln(z)-\mu)^2}{2\sigma^2} - z\right) \mathrm{d}z.$$

For the integral no closed-form expression is known.

Bulmer (1974) proves that the Poisson–lognormal distribution is a mixed Poisson distribution with a lognormally distributed mixing distribution.

Theorem 2.11. *Let Z be a lognormally distributed random variable with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, i.e. Z has the density function*

$$f_Z(z) = \frac{1}{\sqrt{2\pi\sigma^2}z} \exp\left(\frac{-(\ln(z)-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{(0,\infty)}(z).$$

If $X|Z$ is Poisson distributed with rate Z, X is Poisson–lognormally distributed with parameters μ and σ^2.

Furthermore, the expectation and variance of a Poisson-lognormally distributed random variable are given by $\exp(\mu + \sigma^2/2)$ and $\exp(\mu + \sigma^2/2) + (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2)$, respectively.
After describing mixed Poisson distributions as one possibility to model overdispersed count data, we motivate the choices of the introduced mixed Poisson distributions in the next subsection.

2.2 Motivational examples

In the following, we cite clinical trials with patients suffering from chronic obstructive pulmonary disease (COPD) as well as multiple sclerosis (MS) as examples for clinical trials with endpoints commonly modeled as overdispersed count data. The subsection is split into two parts, one for each disease, starting with COPD. Both parts start with a description of the symptoms and relevant endpoints in clinical trials. Then, publications modelling the introduced endpoints as mixed Poisson and in particular as negative binomially distributed are cited. By taking several other publications into account, we demonstrate that three-arm trials matter as a design for these diseases. Last but not least, to get an impression about the expectation, the variance as well as the amount of overdispersion commonly observed, we have a closer look at the results of Calverley et al. (2003) and Fox et al. (2012). Within the framework of this thesis, the examples are important, since they motivate the parameters, such as expectation and variance, for which we compare the statistical methods. Hence, the results from Calverley et al. (2003) and Fox et al. (2012) will be analyzed as detailed as possible.

2.2.1 Chronic obstructive pulmonary disease

COPD denotes several lung diseases and frequent symptoms are coughing, sputum production and shortness of breath. Furthermore, common causes of COPD are smoking, air pollution, and occupational exposure. The progress of COPD is characterized by exacerbations, an sudden worsening and lasting of the symptoms (confer Boehringer Ingelheim Pharma GmbH & Co. KG (2013)). Therefore, an important part of COPD therapy is the prevention of exacerbations and in clinical trials exacerbation are an widely used endpoint. The distribution of the number of exacerbations per patient and year has been subject of various publications, confer Suissa (2006), Keene et al. (2007), Keene et al. (2008a), Keene et al. (2008b), and Aaron et al. (2008). Summarized, these publications reveal that the number of exacerbation per patient are overdispersed and in the cases considered the negative binomial distribution was recommended to model the number of exacerbations. In addition, the fact that the design with an experimental treatment as well as an active and a placebo control matters as a design for clinical trials in COPD is affirmed by the number of publications with this particular design, for instance Donohue et al. (2002), Celli et al. (2003), and Brusasco et al. (2006).

As an example for a placebo controlled study in COPD with active control groups, we regard the so-called TRISTAN study published by Calverley et al. (2003). The TRISTAN study is a large clinical study including 1465 patients which compares the effects of salmeterol, fluticasone, a combination of both, and a placebo in treating COPD patients. Even if this trial is not three-armed but four-armed, it is chosen as an example for a trial with endpoints commonly modelled as overdispersed count data, since the distribution of the data was analysed in detail by Keene et al. (2007). They recommended to use a negative binomial distribution to model the exacerbation counts. Assuming negative binomially distributed observations, Table 1 states the exacerbation rates for the different groups of the TRISTAN study. The source for this table is Table II in Keene et al. (2007).

Table 1: Exacerbation rates and number of recruited patients per group of the TRISTAN study.

	Placebo	Salmeterol	Fluticasone	Combination
N	361	371	374	356
Exacerbation rate	1.71	1.28	1.25	1.16

Tables 1 states that between 356 and 374 patients per group had been recruited for the TRISTAN study. In addition, the rates are between 1.16 and 1.71. The placebo group has the largest rate and the group treated with the combination of Solmeterol and Fluticasone has the smallest rate, i.e. the combinational treatment is the most effective one.

Furthermore, when fitting the negative binomial distribution to the data of the TRISTAN study, Keene et al. (2007) assumed that the shape parameter ϕ is equal among the groups. The shape parameter was estimated as $\hat{\phi} = 0.46$ and a 95%-confidence interval is given by $[0.34, 0.60]$. In particular, the confidence interval for the shape parameter states that, if the assumption of negative binomially distributed observations is true, the observations are overdispersed with a confidence of at least 95%. However, it has not been analysed whether the assumption of the shape parameter being equal in all groups is appropriate. We will discuss this assumption and possible model extension in Section 7.

We conclude this example by comparing the negative binomial, the Poisson–inverse-Gaussian and the Poisson–lognormal distribution for expectations and variances motivated by Table 1. To this, we choose the rate of the negative binomial distribution to be $\lambda = 1.3$ and the corresponding shape parameter ϕ to be either 0.3 or 0.7. The parameters for the Poisson–inverse-Gaussian as well as the Poisson–lognormal distribution are defined such that the distributions have the same expectation and variance as the negative binomial distribution.

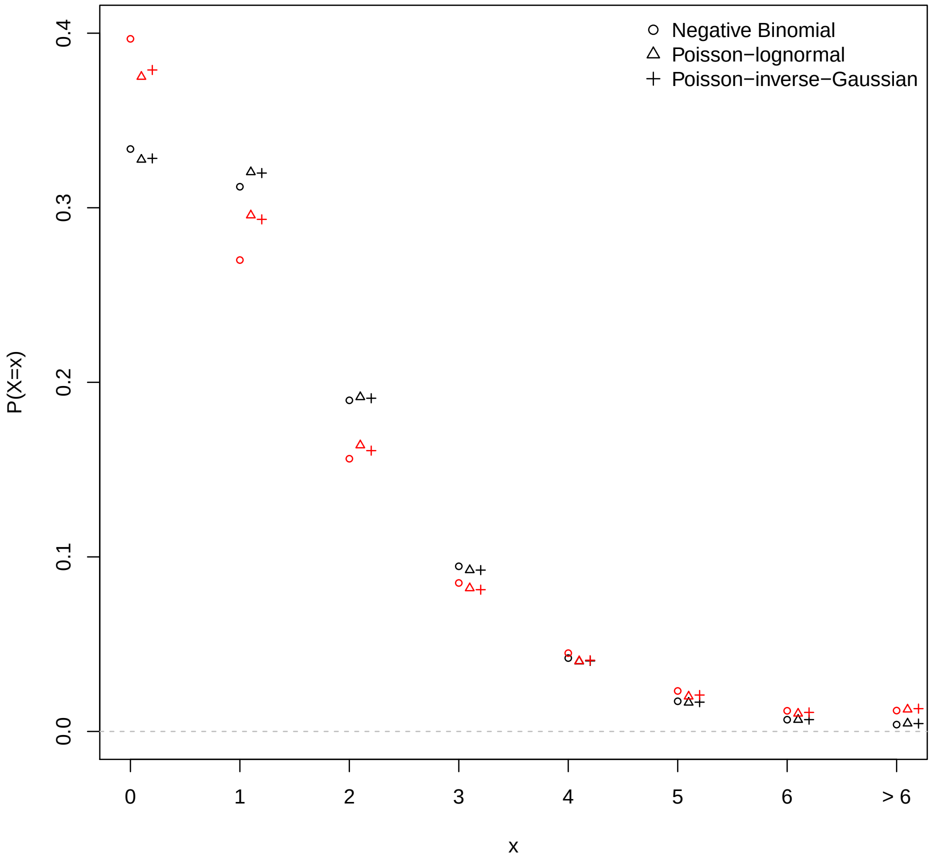


Figure 1: Probability mass functions of the negative binomial, the Poisson–inverse-Gaussian and the Poisson–lognormal distribution with expectation 1.3 and variance 1.807 (black) and 2.483 (red).

Figure 1 shows that for both variances, the probability mass functions decrease in x. First of all, we focus on the distributions with variance 1.807 represented by the black symbols. For a fixed x, the probabilities of the different distributions are nearly the same. The maximal absolute difference between the probabilities for a fixed x is less than 0.08%. However, the negative binomial distribution differs slightly but not more than 0.9% and 0.8% from the Poisson-lognormal and the Poisson-inverse-Gaussian distribution, respectively. Qualitatively, for the distributions with variance 2.483, i.e. the distributions displayed by the red symbols, almost the same holds. However, in this case the Poisson–lognormal and the Poisson–inverse-Gaussian distribution differ more ($< 0.4\%$). In addition, the probability of the negative binomial distribution differs approximately 2.33% and 2.56% from the probabilities of the Poisson–inverse-Gaussian and the Poisson–lognormal distribution, respectively. The largest difference between the probabilities occur at $x = 0$ and $x = 1$ characterized by the negative binomial distribution having more mass at $x = 0$, but lesser mass at $x = 1$. For the other x-values, the negative binomial and the other two mixed Poisson distributions differ by less than 0.8%. Summarizing, for an expectation of 1.3 and variances of 1.807 and 2.483 the probability mass functions have the same decreasing shape. In particular for a fixed variance, the distributions only differ slightly and, therefore, we expect that for these parameters the statistical tests we establish for negative binomially distributed observations are robust concerning a change of the distribution to a Poisson–lognomal or a Poisson–inverse-Gaussian distribution.

2.2.2 Multiple sclerosis

Multiple sclerosis (MS) is an inflammatory disease of the central nervous system with many different symptoms, for instance cognitive impairment, loss of vision, problems with mobility and balance, as well as muscle weakness and stiffness, confer Compston and Coles (2008). For phase II clinical trials in relapse-remitting MS an important endpoint is the number of new or enlarging T_2-weighted hyperintense lesions. Here, T_2-*weighted* denotes the type of magnetic resonance imaging (MRI) used. Hyperintense lesions are the damaged parts of the brain and spinal cord and therefore, small numbers of lesions are desirable. Comparing the mean number of new and enhancing lesions of two groups of patients on different treatments gives information about the efficacy differences of the treatments. Concerning the mentioned and further information about the use of MRI in MS trials, we refer to Chapter 16 in Cohen and Rudick (2003).

Modeling the lesion counts has been part of several publications, confer Sormani et al. (1999), Sormani et al. (2001), Van den Elskamp et al. (2009), Francois et al. (2012). The mentioned publications compare the goodness-of-fit of several distributions for given data sets of lesions counts, especially the fit of the negative binomial distribution is analysed. The data sets differ in number of patients and in particular in the patients disease progression. Summarizing, in most cases the negative binomial distribution has the best goodness-of-fit among the considered distributions and is therefore appropriate to model lesion counts of patients suffering from MS. However, Van den Elskamp et al. (2009) stated that under certain study conditions such as a short follow-up time or activity at baseline, the lesions counts could be modeled using a Poisson–inverse-Gaussian or a Poisson–lognormal distribution. Especially, nearly all of the overdispersed distributions outperform the Poisson distribution. After clarifying that methods for analysing mixed Poisson distributed and in particular negative binomially distributed data are of importance, we expose the need for methods to analyse observations of these distributions within a three-armed trial. Box 1 in Nicholas and Friede (2012) stated that one option for trials in MS are three-arm trials including both an active and a placebo control group. Furthermore, even in future only actively controlled trials to show non-inferiority and superiority are difficult due to the "lack of clear evidence of an effect on a progressive outcome" according to page 1080 in Nicholas and Friede (2012). Therefore, clinical trials in MS are one motivation to study three-arm trials with overdispersed count data.

To receive an impression about the scale of the lesion count means for the different groups within a clinical trial, we consider Fox et al. (2012) as an example. Fox et al. (2012) compare the efficacy and safety of the active agent BG-12 at two different doses with placebo but also includes glatiramer acetate as an active control, i.e. the trial includes three different substances. Even though the trial was not designed to compare the active groups concerning non-inferiority or superiority, it indicates the scale of the lesion counts. Hereinafter, we refer to this study by its name CONFIRM. The following table cites the results for the endpoint *new or enlarging T_2-weighted hyperintense lesions at 2 years* from Table 2 in Fox et al. (2012).

Table 2: Adjusted mean number of new or enlarging T_2-weighted hyperintense lesions at two years, a 95%-confidence interval for the adjusted mean and the corresponding number of patients N of the different treatment groups.

	Placebo	Twice-Daily BG-12	Thrice-Daily BG-12	Glatiramer Acetate
N	139	140	140	153
Adjusted mean no. of lesions	17.4	5.1	4.7	8.0
95% confidence interval	[13.5-22.4]	[3.9-6.6]	[3.6-6.2]	[6.3-10.2]

Compared to the rates in the example motivated by the TRISTAN study stated in Table 1, the adjusted means, which can be interpreted as rates of negative binomial distributions, are much larger, in particular the rate for the placebo group. However, the number of patients in the different groups is less than half of the group sample sizes of the TRISTAN study.

The adjusted means and the confidence intervals have been calculated by a negative binomial regression. Unfortunately, Fox et al. (2012) do not state estimators for the shape parameter neither whether the negative binomial regression assumed the same shape parameter for the different groups. However, to get an impression about the quantity of the shape parameter, we approximate the confidence intervals by

$$[\lambda_{LCL}, \lambda_{UCL}] := \left[\exp\left(\log(\hat{\lambda}) - q_{0.975}\sqrt{\frac{\hat{\phi}+1/\hat{\lambda}}{n}}\right), \exp\left(\log(\hat{\lambda}) + q_{0.975}\sqrt{\frac{\hat{\phi}+1/\hat{\lambda}}{n}}\right)\right]$$

with n the number of observations, $\hat{\lambda}$ the adjusted mean number of lesions, and $q_{0.975}$ the 97.5%-quantile of a standard normal distribution. This method of calculating the confidence interval is a transformed confidence interval for the logarithmized rate $\log(\lambda)$ calculated through a normal approximation. To obtain the magnitude for the shape parameter ϕ, we equate the boundaries of the confidence intervals for the adjusted rate from Table 2 with the formula for λ_{LCL} and λ_{UCL}, respectively. Subsequently, we solve the resulting equation with respect to $\hat{\phi}$. Therefore, we obtain two approximative values for the shape parameter for each treatment.

Table 3: Approximations of the shape parameter ϕ for the different groups of the CONFIRM study.

	Placebo	Twice-Daily BG-12	Thrice-Daily BG-12	Glatiramer Acetate
$\hat{\phi}_1$	2.273	2.427	2.378	2.148
$\hat{\phi}_2$	2.251	2.227	2.583	2.226

The approximations of the negative binomial shape parameter for the different treatments are between 2.148 and 2.583. Of course, there is some degree of uncertainty in the approximations, nevertheless, Table 3 indicates at least the magnitude of the shape parameters. Compared to the exacerbations in COPD, the lesions counts in Phase II trials with patients suffering from MS are much more overdispersed.

Motivated by the means from Table 2, we compare the mixed Poisson distributions introduced in Section 2.1. Thereto, we choose the expectation to be 8. Motivated by Table

3, the shape parameter ϕ is determined as 2.5 which results in a variance of 168. We do not compare different shape parameters because the effect of the shape parameter is qualitatively the same as in Figure 1.

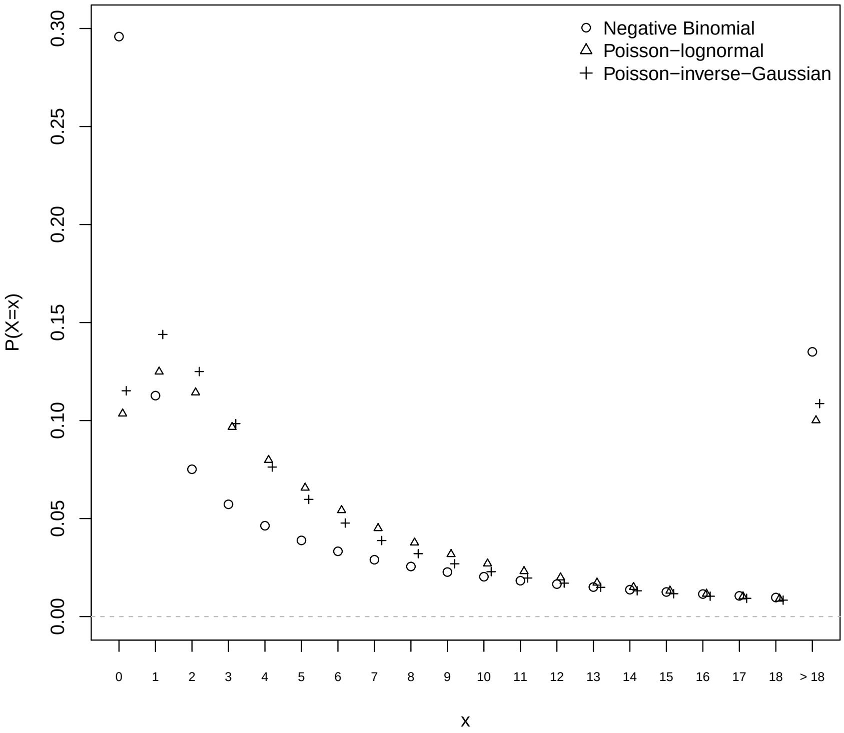


Figure 2: Probability mass functions for the negative binomial, the Poisson–inverse-Gaussian and the Poisson–lognormal distribution with expectation 8 and variance 168.

Figure 2 shows the probability mass functions of the negative binomial, the Poisson-lognormal and the Poisson–inverse-Gaussian distribution with expectation 8 and variance 168 for a range of $x = 1, \dots, 18$. The probability for $x > 18$ are summed up and $P(X = x) < 1\%$ holds for all x larger than 18 and for the different distributions. The probability mass function for the negative binomial distribution is decreasing with a maximum at $x = 0$ and in contrast, the Poisson–lognormal and the Poisson–inverse-Gaussian distribution are unimodal with modus at $x = 1$. In comparison to the other two mixed Poisson distributions, the negative binomial distribution has more mass on $x = 0$. However, the negative binomial distribution has fewer mass on the intervals $\{1, 2, \dots, 12\}$ and $\{1, 2, \dots, 16\}$ compared to the Poisson–inverse-Gaussian and Poisson-lognormal distribution, respectively. The negative binomial distribution has more mass on $x > 18$, but the mass is distributed such that the other two distributions have more mass on the tail. More precisely, the Poisson–inverse-Gaussian and the Poisson–lognormal distribution have more mass on a single x than the negative binomial distribution for $x \geq 76$ and $x \geq 95$, respectively. In particular, the mass on $x \geq 95$ is 0.29% (Poisson–lognormal), 0.30% (Poisson–inverse-Gaussian), and 0.15% (negative binomial). Summarizing, for the current setting, the three mixed Poisson distributions differ much more than for the parameters motivated by the TRISTAN study. Particularly, all three distributions have a very long tails. Due to the difference of the distributions, we do not expect that parametric tests established for the negative binomial distribution are robust concerning the corresponding distribution of the Poisson mixing variable.

Concluding, we considered two examples of clinical trials where endpoints are commonly modelled as overdispersed count data and revealed the need of statistical methods for planning and analyzing three-arm clinical trials with overdispersed count data. Especially, for the lesions counts in clinical trials with MS and the exacerbations in trials with COPD, the negative binomial distribution is recommended to model these outcomes. Furthermore, it is a common assumption in publications about the statistical planning and analysis of trials with these endpoints that the shape parameter is the same for all groups, confer Aban et al. (2009), Friede and Schmidli (2010), and Zhu and Lakkis (2013).

2.3 Statistical Model

Motivated by the examples for clinical trials in COPD and MS, we develop in this thesis the theory of three-arm trials for negative binomially distributed observations and take the other mixed Poisson distributions only into account to study how sensitive the tests based on negative binomially distributed observations are.

For $k = E, R, P$ and $i = 1, \ldots, n_k$, let $X_{k,i}$ be the observations within the experimental treatment (E), reference treatment (R), or placebo (P) group. We assume that the observations are independent and distributed according a negative binomial distribution with different rates but an identical shape parameter, i.e.

$$X_{k,i} \sim \text{NegBin}(\lambda_k, \phi)$$

with $\lambda_k > 0$ and $\phi \geq 0$. Initially, we defined a negative binomial distribution only for a shape parameter ϕ larger than zero but hereafter, we allow the case $\phi = 0$ as an extension of the negative binomial distribution by the Poisson distribution. This extension is well-defined because a negative binomial distribution converges in probability to a Poisson distribution for $\phi \to 0$.

The expectation and the variance of the negative binomially distributed random variable $X_{k,i}$ are given by λ_k and $\lambda_k(1 + \lambda_k \phi) =: \sigma_k^2$, respectively. Furthermore, the parameter space of our statistical model is given by

$$\Theta := \{(\lambda_E, \lambda_R, \lambda_P, \phi) : \lambda_E, \lambda_R, \lambda_P \in \mathbb{R}_+, \phi \in \mathbb{R}_{\geq 0}\} \subset \mathbb{R}^4.$$

With the random variables of the different groups, we define the random vectors

$$\begin{aligned} \mathbf{X_{k,n_k}} &:= (X_{k,1}, \dots, X_{k,n_k}), \qquad k = E, R, P, \\ \mathbf{X_n} &:= (\mathbf{X_{E,n_E}}, \mathbf{X_{R,n_R}}, \mathbf{X_{P,n_P}}) \,. \end{aligned}$$

In this thesis, we consider that the expectations λ_k denote the treatment efficacies and the placebo response as well as that smaller values are desirable. Hence, the hypothesis for assay sensitivity and the retention of effect hypothesis are defined as stated in Section 1.

3 Parameter Estimation

In this subsection, we study different methods of estimating the rates, the shape parameter and the variances σ_k^2, $k = E, R, P$, under the model introduced in Section 2.3. Firstly, we establish the maximum-likelihood estimators for the parameters and show their consistency as well as their asymptotic normality. Secondly, we introduce the idea of restricted maximum-likelihood estimation. For the restricted estimator we only prove the consistency but do not calculate any asymptotic distribution because this estimator is used exclusively to estimate the variance of a test statistic. The estimators and their properties are taken into account when deducing different hypothesis tests in Sections 5 and 6.

3.1 Maximum-Likelihood Estimation

We now establish the maximum-likelihood estimators of the rates and the shape parameters as well as their basic properties. One characteristic of the model introduced in Section 2.3 is that the shape parameter ϕ is equal for the different groups and in consequence, the shape parameter is estimated using the observations from all groups.
For the sake of readability, in what follows, we denote $\zeta := (\lambda_E, \lambda_R, \lambda_P, \phi)$. The log-likelihood function $\log l(\zeta|\mathbf{X_n})$ is given by

$$\sum_{k\in\{E,R,P\}} \sum_{i=1}^{n_k} \left[\log\Gamma\left(X_{k,i} + \frac{1}{\phi}\right) - \left(\frac{1}{\phi} + X_{k,i}\right) \log\left(1+\phi\lambda_k\right) + X_{k,i}\log\left(\phi\lambda_k\right) \right.$$
$$\left. - \log(X_{k,i}!) - \log\Gamma\left(\frac{1}{\phi}\right)\right].$$

Noting $\Gamma(z) = (z-1)\Gamma(z-1)$, we obtain the equality

$$\log\left(\Gamma\left(x + \frac{1}{\phi}\right)\right) = \log\Gamma\left(\frac{1}{\phi}\right) + \sum_{i=0}^{x-1} \log\left(i + \frac{1}{\phi}\right), \quad x \geq 0.$$

With $k = E, R, P$, the last equation and the definition $X_{k,\cdot} := \sum_{i=1}^{n_k} X_{k,i}$ yield the following representation of the log-likelihood function $\log l(\zeta|\mathbf{X_n})$

$$\sum_{k\in\{E,R,P\}} X_{k,\cdot}\log(\phi\lambda_k) - \left(\frac{n_k}{\phi} + X_{k,\cdot}\right)\log(1+\phi\lambda_k) + \sum_{i=1}^{n_k}\sum_{j=0}^{X_{k,i}-1} \log\left(j + \frac{1}{\phi}\right) - \log(X_{k,i}!).$$

The maximum-likelihood estimator $\hat{\eta}$ of the parameter η is defined as the maximizer of the log-likelihood function, i.e.

$$\hat{\zeta} := \left(\hat{\lambda}_E, \hat{\lambda}_R, \hat{\lambda}_P, \hat{\phi}\right) := \arg\max_{\zeta \in \Theta} \log l\left(\zeta | \mathbf{X_n}\right).$$

By differentiating the log-likelihood function with respect to the parameter λ_k, equating the resulting derivation to zero, and solving the equation with respect to λ_k, we obtain that the group mean is the unique maximum-likelihood estimator $\hat{\lambda}_k$ for the rate λ_k, i.e.

$$\hat{\lambda}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{k,i}.$$

Due to the independence and identical distribution of the entries of $\mathbf{X_{k,n_k}}$, the maximum-likelihood estimator $\hat{\lambda}_k$ is an unbiased estimator for the rate λ_k. The maximum-likelihood estimator $\hat{\phi}$ is a solution of the equation

$$G(\phi) := \sum_{k \in \{E,R,P\}} n_k \log(\phi \hat{\lambda}_k + 1) - \sum_{i=1}^{n_k} \sum_{j=0}^{X_{k,i}-1} \frac{\phi}{1 + j\phi} \overset{!}{=} 0$$

with respect to ϕ. Since there is no closed form expression known for the solution, the estimator $\hat{\phi}$ has to be calculated iteratively. The Theorem 3.1 makes a point about the existence of the estimator $\hat{\phi}$.

Theorem 3.1. *The maximum-likelihood estimator $\hat{\phi}$ for the shape parameter ϕ exists and is larger than zero if the inequality*

$$\sum_{k=E,R,P} \frac{1}{n_k} \sum_{i=1}^{n_k} (X_{k,i} - \hat{\lambda}_k)^2 - \hat{\lambda}_k > 0$$

holds.

Proof. We prove the sufficient condition for the existence of the maximum-likelihood estimator $\hat{\phi}$, which is the solution of the equation $G(\phi) = 0$, analogously to the proof of the existence of the maximum-likelihood estimator for the shape parameter in the case of independent and identically distributed random variables by Aragón et al. (1992). The idea is to extend the input of $G(\cdot)$ to negative values. In doing so we obtain a function which is continuous in a small neighborhood of zero and for all positive inputs. Then, we show that $\phi = 0$ is a local minimum of $G(\phi)$ with $G(0) = 0$ and that for large ϕ the value of $G(\phi)$ is smaller than zero. Hence, the function $G(\phi)$ has to be zero at least once for $\phi > 0$. The first and second derivation of $G(\cdot)$ are given by

$$G'(\phi) = \sum_{k=E,R,P} \sum_{i=1}^{n_k} \left(- \sum_{j=0}^{x_{k,i}-1} \frac{1}{(j\phi + 1)^2} \right) + \frac{\hat{\lambda}_k}{1 + \phi\hat{\lambda}_k},$$

$$G''(\phi) = \sum_{k=E,R,P} \sum_{i=1}^{n_k} \left(2 \sum_{j=0}^{x_{k,i}-1} \frac{j}{(j\phi + 1)^3} \right) - \frac{\hat{\lambda}_k^2}{(1 + \phi\hat{\lambda}_k)^2}.$$

It holds that $G(0) = G'(0) = 0$ as well as

$$G''(0) = \sum_{k=E,R,P} \frac{1}{n_k} \sum_{i=1}^{n_k} (X_{k,i} - \hat{\lambda}_k)^2 - \hat{\lambda}_k.$$

As assumed, $G''(0)$ is larger than 0 and therefore, $G(\cdot)$ has a local minimum at zero. Last but not least, for large values of ϕ the values of $G(\phi)$ is smaller than zero because

$$\lim_{\phi\to\infty} \frac{G(\phi)}{\phi} = \sum_{k=E,R,P} \frac{\#\Big\{i|X_{k,i} > 0\Big\}}{n_k} < 0$$

holds. □

Theorem 3.1 states that the estimator $\hat{\phi}$ exists if the sum of the sample variances from the different groups is larger than the sum of the group means. Such a statement was expected because the shape parameter determines the amount of overdispersion and overall, the data is overdispersed if the sum of the sample variances if larger than the sum of the means. For the case of independent and identically negative binomially distributed random variables, Aragón et al. (1992) proves that a sample variance larger than the mean is both a sufficient and necessary condition for the existence and the uniqueness of the maximum-likelihood estimator for the shape parameter. The proof explicitly takes advantage of that only one mean exists and therefore, the same approach does not work in our setting. Furthermore, our setting is a special case of the negative binomial regression in Lawless (1987) but even if the maximum-likelihood estimator for the shape parameter is established, to our knowledge, the uniqueness has not been proven, yet. However, for the cases considered, the maximum-likelihood estimator $\hat{\phi}$ has always been unique and the shape of the log-likelihood function $\log l(\eta|\mathbf{X_n})$ in ϕ has the same shape as the log-likelihood function for independent and identically negative binomially distributed random variables in ϕ.

As mentioned above, the maximum-likelihood estimator $\hat{\lambda}_k$ for the rate λ_k is unbiased.

However, we do not expect the maximum-likelihood estimator $\hat{\phi}$ for the shape parameter to be unbiased because the maximum-likelihood estimator for the shape parameter is even in the case of independent and identically distributed random variables not unbiased, confer Saha and Paul (2005) who calculated the bias of the maximum-likelihood estimator for the shape parameter of independent and identically negative binomially distributed random variables.

The next theorem states the consistency and asymptotic normality of the maximum-likelihood estimator $\hat{\zeta}$.

Theorem 3.2. *The maximum-likelihood estimator $\hat{\zeta}$ is a consistent estimator for the parameter vector ζ and it is asymptotically normal distributed in the sense of:*

$$\sqrt{n}\left(\hat{\zeta}-\zeta\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \begin{pmatrix}\frac{\sigma_E^2}{w_E} & 0 & 0 & 0\\ 0 & \frac{\sigma_R^2}{w_R} & 0 & 0\\ 0 & 0 & \frac{\sigma_P^2}{w_P} & 0\\ 0 & 0 & 0 & \Sigma_{4,4}\end{pmatrix}\right)$$

with $\mathcal{D}$ denoting convergence in distribution and

$$\Sigma_{4,4} = \phi^4\left(\sum_{k=E,R,P} w_k\left(\sum_{j=0}^{\infty}(\phi^{-1}+j)^{-2}\mathbb{P}(Y_{k,1}\geq j) - \frac{\phi\lambda_k}{\lambda_k+\phi^{-1}}\right)\right)^{-1}.$$

Proof. Lawless (1987) proved the asymptotic normality of the maximum-likelihood estimators for the negative binomial regression model but used another parametrization for the parameter of interest. However, our model is a special case of the negative binomial regression and the results can be adapted easily by means of the delta method. The consistency in mentioned in Appendix A in Lawless (1987). □

In addition, asymptotically, the maximum-likelihood estimators for a rate and the shape parameter are independent. As we see in the next sections, for several tests we have to estimate the variance σ_k^2, $k = E, R, P$. The next corollary states that the maximum-likelihood estimator for σ_k^2 is obtained by plugging in the corresponding maximum-likelihood estimators for the rate and the shape parameter.

Corollary 3.3. *For $k = E, R, P$, the maximum-likelihood estimator $\hat{\sigma}_k^2$ for the variance*

$\sigma_k^2 = \mathrm{Var}(X_{k,1}) = \lambda_k(1 + \lambda_k \phi)$ *is given by*

$$\hat{\sigma}_k^2 = \hat{\lambda}_k(1 + \hat{\lambda}_k \hat{\phi}).$$

Furthermore, the estimator $\hat{\sigma}_k^2$ is consistent for the variance σ_k^2.

Proof. Due to the functional invariance of maximum-likelihood estimators, the variance estimator $\hat{\sigma}_k^2$ is a plug-in estimator. In addition, the consistence of the estimator follows from the continuous mapping theorem. □

Supported by Monte-Carlo simulations and the fact that the squared estimator $\hat{\lambda}_k^2$ and the estimator for the shape parameter $\hat{\phi}$ are biased and assumed to be biased, respectively, we suppose that the maximum-likelihood estimators for the variances are biased, too.

3.2 Restricted Maximum-Likelihood Estimation

Next, we describe a concept of maximum-likelihood estimation whereby the estimators are restricted to a subspace of the parameter space. Sometimes, for example when estimating the variance for a Wald-type test restricted to the null hypothesis it is required to estimate the rates and the shape parameter with restriction to an inequality $g(\zeta) \geq 0$, with $g : \mathbb{R}^4 \to \mathbb{R}$ a continuous function. Therefore, in the following, we define the restricted parameter space and the maximum-likelihood estimator for the parameters restricted to this space. Then, we study the calculation as well as asymptotic properties of the restricted estimators.
Let Θ_g be the parameter space restricted to $g(\zeta) \geq 0$, i.e.

$$\Theta_g := \left\{ \zeta \in \Theta : g(\zeta) \geq 0 \right\}.$$

Then, the restricted maximum-likelihood estimators are defined by

$$\hat{\zeta}_{RML} := (\hat{\lambda}_{E,RML}, \hat{\lambda}_{R,RML}, \hat{\lambda}_{P,RML}, \hat{\phi}_{RML}) := \underset{\zeta \in \Theta_g}{\arg\max} \log l(\zeta | \mathbf{X_n}).$$

Whether the restricted maximum-likelihood estimator exists and is unique depends on the restricted parameter space. Hereinafter, we assume that the estimator exists and is unique. If the unrestricted maximum-likelihood estimators fulfil the condition $g(\hat{\zeta}) \geq 0$, the restricted maximum-likelihood estimators coincide with the unrestricted ones. Otherwise, the restricted estimators are located at the boundary of the restricted parameter space, i.e.

they meet the condition $g(\hat{\zeta}_{RML}) = 0$. Then, the restricted estimators can be calculated by

$$\hat{\zeta}_{RML} = \underset{g(\zeta)=0}{\arg\max} \log l(\zeta|\mathbf{X_n}).$$

The uniqueness as well as whether a closed form expression for $\hat{\zeta}_{RML}$ exists depends on the function $g(\cdot)$. The next theorem states the restricted maximum-likelihood estimator for the variance σ_k^2.

Theorem 3.4. *For $k = E, R, P$, the restricted maximum-likelihood estimator $\hat{\sigma}^2_{k,RML}$ is given by*

$$\hat{\sigma}^2_{k,RML} = \hat{\lambda}_{k,RML}(1 + \hat{\lambda}_{k,RML}\, \hat{\phi}_{RML}).$$

If the true parameter vector is located in the restricted parameter space , i.e. $\zeta \in \Theta_g$, the restricted maximum-likelihood estimator $\hat{\sigma}^2_{k,RML}$ is a consistent estimator for σ_k^2.

Proof. As for the unrestricted maximum-likelihood estimator for the variance, due to the functional invariance of maximum-likelihood estimators, $\hat{\sigma}^2_{k,RML}$ is obtained by plugging in the corresponding estimators for the rate and the shape parameter. The consistency of the restricted maximum-likelihood variance estimator follows if the restricted maximum-likelihood estimators for the rates and the shape parameter are consistent. The consistency of the restricted estimator $\hat{\zeta}_{RML}$ follows from the consistency of the unrestricted estimator $\hat{\zeta}$ because the function g, which defines the restricted parameter space, is continuous. □

However, the consistency of the restricted maximum-likelihood estimators does not hold if the true parameter vector is not located in the restricted parameter space, i.e. $\zeta \in \Theta \backslash \Theta_g$. Nevertheless, under certain to be specified conditions, the restricted estimators converge almost surely to a parameter vector located in Θ_g. Sufficient conditions for the almost surely convergence of the restricted estimator $\hat{\zeta}_{RML}$ has been proved on page 20f. in Mielke (2010). In the following we recapitulate these results. Thereto, let $\zeta \in \Theta$ be an arbitrary parameter vector and let $\zeta_0 := (\lambda_{E,0}, \lambda_{R,0}, \lambda_{P,0}, \phi) \in \Theta$ be the true parameter vector. Furthermore, let $c = (c_E, c_R, c_P)$ be a vector of weights. Then, we define the weighted Kullback–Leibler divergence between ζ and ζ_0 by

$$K(\zeta_0, \zeta, c) := \sum_{k=E,R,P} c_k K((\lambda_{k,0}, \phi_0), (\lambda_k, \phi)) \tag{3.1}$$

with

$$K((\lambda_{k,0},\phi_0),(\lambda_k,\phi)) := \mathbb{E}_{(\lambda_{k,0},\phi_0)}\left[\log\left(\mathbb{P}_{(\lambda_{k,0},\phi_0)}(X=\cdot)\right) - \log\left(\mathbb{P}_{(\lambda_k,\phi)}(X=\cdot)\right)\right]$$

the usual Kullback–Leibler divergence measuring the difference between two probability distributions. In addition, we define ζ_{Θ_g} as the minimizer of the KL-divergence $K(\zeta_0,\zeta,(w_E,w_R,w_P))$ with respect to $\zeta \in \Theta_g$, i.e.

$$\zeta_{\Theta_g} := \arg\min_{\zeta\in\Theta_g} K(\zeta_0,\zeta,(w_E,w_R,w_P)).$$

Theorem 3.5 states sufficient conditions for the almost surely convergence of the restricted estimator $\hat{\zeta}_{RML}$ to ζ_{Θ_g}.

Theorem 3.5. ***Condition 1:*** *If the true parameter is located in the restricted parameter space, $\zeta_0 \in \Theta_g$, and none of the three groups vanishes asymptotically, i.e. $\lim_{n\to\infty} n_k/n = w_k \in (0,1)$, the argument ζ_{Θ_g} which minimized the Kullback-Leibler divergence is well-defined.*
Condition 2: *Any sequence of parameter vectors in the restricted space $\zeta^{(n)} \in \Theta_g$ which limit is located in the closure of the parameter space but not in the parameter space itself, i.e. $\lim_{n\to\infty}\zeta^{(n)} \in \overline{\Theta}\backslash\Theta$ or which length converges to infinity, $\lim_{n\to\infty}\|\zeta^{(n)}\| = \infty$, has a mass of zero:*

$$\lim_{n\to\infty}\prod_{k=E,R,P}\mathbb{P}_{(\lambda_k^{(n)},\phi^{(n)})}(X_{k,1}=\cdot) = 0 \qquad \mathbb{P}_{(\lambda_{k,0},\phi_0)} - a.s.$$

If Condition 1 and 2 hold, the restricted maximum-likelihood estimator $\hat{\zeta}_{RML}$ converges almost surely to the minimizer ζ_{Θ_g} of the Kullback–Leibler divergence.

We refer to Theorem 3.5 in later sections when we establish Wald-type tests with restricted maximum-likelihood variance estimators. Summarizing, we calculated the maximum-likelihood estimator $\hat{\zeta}$ for the parameter vector ζ and, additionally, we proved its consistency and asymptotic normality in Theorem 3.2. With the consistency of $\hat{\zeta}$, we concluded that the maximum-likelihood variance estimator $\hat{\sigma}_k^2$ with $k = E, R, P$ is consistent, too. Moreover, we described the idea of restricted maximum-likelihood estimators and showed that the restricted maximum-likelihood estimators for the parameter vector ζ and the variance of a negative binomial distribution are consistent if the true parameter is located in the restricted parameter space. In addition, we stated that under certain conditions the restricted estimator $\hat{\zeta}_{RML}$ converges almost surely to the minimizer of a Kullback–Leibler divergence. The calculation of the restricted estimators depends on the function g and will therefore be discussed later on.

4 Hypothesis Testing

In the following, we introduce the notation of hypothesis testing for a one-sided hypothesis H_0 as well as the idea of Wald-type and permutation tests. Thereto, in this section, let $\mathbf{Y_n} := (Y_{n,1}, \dots, Y_{n,n}) \sim F_{n,\eta}$ be a vector of n random variables with an unknown parameter $\eta \in \mathbb{R}^d$ and probability measure $\mathbb{P} \equiv \mathbb{P}_\eta$. With $g : \mathbb{R}^d \to \mathbb{R}$ a continuous function, we define the one-sided hypothesis

$$H_0 : g(\eta) \leq 0 \qquad \text{versus} \qquad H_1 : g(\eta) > 0.$$

Furthermore, let $T_n : \mathbb{R}^n \to \mathbb{R}$ be a test statistic mapping the random vector $\mathbf{Y_n}$ to a real number and let the output $T_n(\mathbf{Y_n})$ be distributed according to G_η. In addition, to define a hypothesis test for H_0, we assume that the cumulative distribution function of G_η is increasing in $g(\eta)$. We define a level α test Ψ_n as a function mapping the random vector $\mathbf{Y_n}$ into the set $[0, 1]$, i.e.

$$\begin{aligned} \Psi_n : \mathbb{R}^n &\to [0,1] \\ \mathbf{Y_n} &\mapsto \begin{cases} 1 & T_n(\mathbf{Y_n}) > c_n \\ \gamma_n & T_n(\mathbf{Y_n}) = c_n \\ 0 & T_n(\mathbf{Y_n}) < c_n \end{cases} \end{aligned}$$

with $c_n \in \mathbb{R}$ and $\gamma_n \in [0, 1]$ being defined such that for all η with $g(\eta) = 0$ the equation

$$\mathbb{E}_{\mathbb{P}} \left[\Psi_n(\mathbf{Y_n})\right] = \mathbb{P}(T_n(\mathbf{Y_n}) > c_n) + \gamma_n \cdot \mathbb{P}(T_n(\mathbf{Y_n}) = c_n) \overset{!}{=} \alpha \tag{4.1}$$

holds, with $\alpha \in (0, 1)$ the so-called level of significance. In practice, the hypothesis test Ψ_n is usually defined with γ_n to be zero and the outcome $\Psi_n = 1$ is interpreted as a rejection of the hypothesis H_0 with level of significance α. In other words, if the hypothesis H_0 is true, the probability of a false rejection, which is referred to as the type I error, is at most α. Here, the assumption that the distribution function of G_η is increasing in $g(\eta)$ assures that the type I error rate is smaller than α if $g(\eta) < 0$ holds. Besides the type I error rate, the type II error rate is defined as the probability of not rejecting the hypothesis H_0 if the alternative H_1 is true. The probability of the complementary event, i.e. the probability of rejecting the hypothesis if the alternative is true, is called the power of a test.

For a given distribution $F_{n,\eta}$, it is not always possible to construct an appropriate test statistic T_n for the hypothesis H_0 such that calculating the parameters c_n and γ_n is feasible. In particular, the construction is not possible for the hypotheses we study in this thesis. As a consequence, we introduce two ways of constructing asymptotic tests. As the name implies, the idea of asymptotic tests is to determine the parameters c_n and γ_n with respect to the asymptotic distribution of the test statistic T_n, if calculating them for the actual distribution is not feasible. Hence, an asymptotic level α test Ψ_n is defined such that the limit

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}}\left[\Psi_n(\mathbf{Y_n})\right] \leq \alpha$$

holds for all η with $g(\eta) \leq 0$ and, in particular, equality holds for all η with $g(\eta) = 0$.
In this thesis, we introduce Wald-type and permutation tests as approaches to construct hypothesis tests. Wald-type tests are asymptotic tests assuming an asymptotically standard normally distributed test statistic. In contrast, we introduce permutation tests as exact tests assuming exchangeable random variables and, if possible, extend them to asymptotic tests in case the assumption of exchangeable random variables is not given. Random variables are called exchangeable if the joint distribution of the random variables is invariant to permutations.

4.1 Wald-type Tests

The idea of Wald-type tests is to define a test statistic T_n for H_0 through the maximum-likelihood estimator for $g(\eta)$ which is asymptotically standard normally distributed if $g(\eta) = 0$ holds. Then, the Wald-type test Ψ_n^{Wald} is defined with the parameter c_n as the $(1-\alpha)$-quantile $q_{1-\alpha}$ of a standard normal distribution and the parameter γ_n as zero. Pioneering work on Wald-type tests has been done by Wald (1943). However, in the following, we outline the definition of Wald-type tests as introduced by Engle (1984).
Let the random vector $\mathbf{Y_n}$ and the hypothesis H_0 be defined as above and in addition let $\hat{g}(\mathbf{Y_n}|\eta)$ be a consistent maximum-likelihood estimator for $g(\eta)$ which is asymptotically standard normally distributed in the sense of

$$\sqrt{n}\Big(\hat{g}(\mathbf{Y_n}|\eta) - g(\eta)\Big) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2).$$

To obtain a test statistic T_n^{Wald} which is asymptotically standard normally distributed for $g(\eta) = 0$, we divide the term $\sqrt{n}\hat{g}(\mathbf{Y_n}|\eta)$ by an, at least under H_0, consistent estimator

$\hat{\sigma}^2(\mathbf{Y_n})$ for the variance σ^2, i.e.

$$T_n^{Wald}(\mathbf{Y_n}) := \sqrt{n}\frac{\hat{g}(\mathbf{Y_n}|\eta)}{\sqrt{\hat{\sigma}^2(\mathbf{Y_n})}} \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1) \text{ for } g(\eta) = 0.$$

With this, we define the Wald-type test for the hypothesis H_0 by

$$\Psi_n^{Wald}(\mathbf{Y_n}) := \begin{cases} 1 & T_n^{Wald}(\mathbf{Y_n}) > q_{1-\alpha} \\ 0 & T_n^{Wald}(\mathbf{Y_n}) \leq q_{1-\alpha} \end{cases}.$$

Of course, for $g(\eta) = 0$ holds that the asymptotic level of significance is equal to α, thus

$$\lim_{n\to\infty} \mathbb{E}_{\mathbb{P}}\left[\Psi_n^{Wald}(\mathbf{Y}_n)\right] = \lim_{n\to\infty} \mathbb{P}\left(T_n^{Wald}(\mathbf{Y_n}) > q_{1-\alpha}\right) = \alpha.$$

Since the distribution function of the test statistic T_n^{Wald} is increasing in $g(\eta)$, the Wald-type test Ψ_n^{Wald} is an asymptotic level α test.

4.2 Permutation Tests

In this subsection, we introduce two types of permutation tests, an exact test as well as an asymptotic test. The exact permutation test assumes exchangeable random variables, i.e. the joint distribution of the random variables is invariant to permutations. In contrast, the asymptotic permutation test does not assume exchangeable random variables.

Let the random vector $\mathbf{Y_n}$ and the hypothesis H_0 be defined as before. Generalized, exact permutation tests for the hypothesis H_0 base on the assumption that the entries of the random vector $\mathbf{Y}_n$ are exchangeable for $g(\eta) = 0$. Hence, at the boundary of the hypothesis H_0, the distribution of an arbitrary, for H_0 appropriate test statistic $T_n^{Perm}(\mathbf{Y_n})$ does not change if the random vector $\mathbf{Y_n}$ is permuted. Meaning, for $g(\eta) = 0$ the equality in distribution

$$T_n^{Perm}(\mathbf{Y_n}) \overset{\mathcal{D}}{=} T_n^{Perm}(\tau_n(\mathbf{Y_n}))$$

holds with $\tau_n(\mathbf{Y_n})$ an uniformly distributed random variable on the space of permutations of $\mathbf{Y_n}$ and, thus, each permutation of $\mathbf{Y_n}$ has the probability $1/n!$. We denote the random variable $\tau_n(\mathbf{Y_n})$ as the uniformly distributed permutation and, hereinafter, let $\tilde{\mathbb{P}}$ be the probability measure of $\tau_n(\mathbf{Y_n})$. Additionally, the probability measure $\tilde{\mathbb{P}}$ is assumed to be independent of $\mathbb{P}$. Having this in mind, we define a (one-sided) permutation test.

Definition 4.1 (One-sided exact permutation test)**.** Let $T_n^{Perm}(\mathbf{Y_n})$ be a test statistic which is appropriate to test the hypothesis H_0 and let $\tau_n(\mathbf{Y_n})$ be an uniformly distributed permutation. With the previously introduced notation, we define a permutation test for the hypothesis H_0 as a function mapping the vector $\mathbf{Y_n}$ into the set $[0, 1]$, i.e.

$$\Psi_n^{Perm} : \mathbb{R}^n \to [0,1]$$
$$\mathbf{Y_n} \mapsto \begin{cases} 1 & T_n^{Perm}(\mathbf{Y_n}) > c_n^{Perm} \\ \gamma_n^{Perm} & T_n^{Perm}(\mathbf{Y_n}) = c_n^{Perm} \\ 0 & T_n^{Perm}(\mathbf{Y_n}) < c_n^{Perm} \end{cases}$$

with $c_n^{Perm} \equiv c_n^{Perm}(\mathbf{Y_n})$ and $\gamma_n^{Perm} \equiv \gamma_n^{Perm}(\mathbf{Y_n})$ such that the equation

$$\begin{aligned}&\mathbb{E}_{\tilde{\mathbb{P}}}\left[\Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))|\mathbf{Y_n}\right] \\ =&\tilde{\mathbb{P}}\left(T_n^{Perm}(\tau_n(\mathbf{Y_n})) > c_n^{Perm}\middle|\mathbf{Y_n}\right) + \gamma_n^{Perm} \cdot \tilde{\mathbb{P}}\left(T_n^{Perm}(\tau_n(\mathbf{Y_n})) = c_n^{Perm}\middle|\mathbf{Y_n}\right) \overset{!}{=} \alpha\end{aligned}$$

holds.

Due to the independence of the probability measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$, conditioning on $\mathbf{Y_n}$ is not mandatory. However, it clarifies that the permutation test Ψ_n^{Perm} or to be precise its constants c_n^{Perm} and γ_n depend on the realizations of $\mathbf{Y_n}$ as well as on the distribution of $T_n^{Perm}(\tau_n(\mathbf{Y_n}))\Big|\mathbf{Y_n}$ but not on the distribution of $\mathbf{Y_n}$. The next theorem proves that the permutation test as defined above is an exact test at the boundary of the hypothesis.

Theorem 4.2. *Let the random vector $\mathbf{Y_n}$, the hypothesis H_0, and the uniformly distributed permutation $\tau_n(\mathbf{Y_n}))$ be defined as above and let the probability measures $\mathbb{P}$ and $\tilde{\mathbb{P}}$ be independent. Furthermore, let the entries of the random vector $\mathbf{Y_n}$ be exchangeable if $g(\eta) = 0$ holds. Then, the permutation test Ψ_n^{Perm} as defined in Definition 4.1 is an exact level α test at the boundary ∂H_0 of the hypothesis, i.e. for $g(\zeta) = 0$.*

Proof. We prove that $\mathbb{E}_{\mathbb{P}}\left[\Psi_n^{Perm}(\mathbf{Y_n})\right] = \alpha$ holds if $g(\eta)$ is equal to zero. Hence, in the following, we assume that η fulfills the equation $g(\eta) = 0$. Due to the exchangeablity of the random variables, the equality in distribution

$$\Psi_n^{Perm}(\mathbf{Y_n}) \overset{\mathcal{D}}{=} \Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))$$

holds. Hence, the law of total expectation yields the equation

$$\mathbb{E}_{\mathbb{P}}\left[\Psi_n^{Perm}(\mathbf{Y_n})\right] = \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}|\tilde{\mathbb{P}}}\left[\Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))\middle|\mathbf{Y_n}\right]\right].$$

Since the probability measures are independent, the measure $\mathbb{P}|\tilde{\mathbb{P}}$ is equal to $\mathbb{P}$ and the expectations can be switched. Taking the property $\mathbb{E}_{\tilde{\mathbb{P}}}\left[\Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))\right] = \alpha$ into account yield the assertion

$$\mathbb{E}_{\mathbb{P}}\left[\Psi_n^{Perm}(\mathbf{Y_n})\right] = \mathbb{E}_{\tilde{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}}\left[\Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))\right]\right] = \mathbb{E}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}\left[\Psi_n^{Perm}(\tau_n(\mathbf{Y_n}))\right]\right] = \mathbb{E}_{\mathbb{P}}\left[\alpha\right] = \alpha.$$

□

Whether the actual level of significance of the permutation test Ψ_n^{Perm} is less than α for $g(\eta) < 0$ depends on the test statistic. However, hereinafter, we denote the permutation test as an exact test but we keep in mind that, primarily, the test is exact at the boundary of the hypothesis.

So far, we have defined an exact permutation test with parameters c_n^{Perm} and γ_n determined by the conditional permutation distribution of the test statistic given the realizations of $\mathbf{Y_n}$. However, the conditional distribution does mostly not correspond to any known distribution but can be approximated by Monte-Carlo simulations. Thus, in practice, the parameters c_n^{Perm} and γ_n are determined through simulations. For the exact permutation test from Definition 4.1, we assumed exchangeable random variables at the boundary of H_0 and, of course, this assumption does not always hold. In particular, for the retention of effect hypothesis the corresponding random variables are not exchangeable. Therefore, a permutation test as introduced above is in general not a level α test. However, Janssen (1997) established an asymptotic permutation test for non-i.i.d. random variables which does not assume exchangeable random variables under H_0. Next, we introduce the idea and summarize the main results of Janssen (1997).

Basically, Janssen's asymptotic permutation test considers an appropriate, and at the boundary of H_0 asymptotically standard normally distributed test statistic T_n^{Perm}. Intuitively and as for the Wald-type test, a one-sided asymptotic level α test for H_0 is obtained by rejecting the hypothesis if and only if the test statistic is larger than the $(1-\alpha)$-quantile $q_{1-\alpha}$ of a standard normal distribution. However, if the distribution of the test statistic T_n^{Perm} converges slowly to a standard normal distribution, the actual level of the test can differ clearly from α, especially, if the sample size is small. Therefore, instead of using a quantile of a standard normal distribution, the asymptotic permutation test rejects H_0

if and only if the test statistic T_n^{Perm} is larger than the quantile $c_{1-\alpha}^{Perm}$ of the conditional permutation distribution of $T_n^{Perm}(\tau_n(\mathbf{Y_n}))|\mathbf{Y_n}$. In other words, the idea is to approximate the distribution of the test statistic not through a normal distribution but with means of a permutation distribution. The crucial point why this definition yields an asymptotic level α test is that, under certain to be specified conditions, the permutation quantile $c_{1-\alpha}^{Perm}$ converges to the quantile $q_{1-\alpha}$.

More precisely, let $\mathbf{Y_n}$ and H_0 be defined as above and for each $n \in \mathbb{N}$, let $(c_{n,i})_{i\leq n}$ be a sequence of real numbers, which is taken into account to define the test statistic. As mentioned previously, the test statistic for the permutation test is defined such that it is asymptotically standard normally distributed. Janssen (1997) defined the test statistic through the linear statistic $\sum_{i=1}^n c_{n,i}Y_{n,i}$. The definition of the test statistic by means of a linear statistic allows to apply certain central limit theorems to show the asymptotic normality of the test statistic. We obtain the test statistic for the asymptotic permutation test by studentizing the linear statistic, i.e. divide it by an estimator $\hat{\sigma}_{Perm}(\mathbf{Y_n})$ of its standard deviation which needs to be specified. Thus, the test statistic is given by

$$T_n^{Perm}(\mathbf{Y_n}) := \frac{\sum_{i=1}^n c_{n,i}Y_{n,i}}{\hat{\sigma}_{Perm}(\mathbf{Y_n})}. \tag{4.2}$$

In particular, the coefficients $(c_{n,i})_{i\leq n}$ are chosen such that the resulting test statistic fits the hypothesis and that the asymptotic normality of the test statistic holds.

With the definition of the test statistic in (4.2), we next define the asymptotic permutation test.

Definition 4.3. Let the random vector $\mathbf{Y_n}$ and the hypothesis H_0 be defined as before. Furthermore, let the test statistic $T_n^{Perm}(\mathbf{Y_n})$ be defined as in (4.2). Then, we define the asymptotic permutation test Ψ_n^{Perm} for the hypothesis H_0 by

$$\begin{aligned} \Psi_n^{Perm} &: \mathbb{R}^n \to [0,1] \\ \mathbf{Y_n} &\mapsto \begin{cases} 1 & T_n^{Perm}(\mathbf{Y_n}) > c_{1-\alpha}^{Perm} \\ 0 & T_n^{Perm}(\mathbf{Y_n}) \leq c_{1-\alpha}^{Perm} \end{cases}, \end{aligned}$$

with $c_{1-\alpha}^{Perm} \equiv c_{1-\alpha}^{Perm}(\mathbf{Y_n})$ the $(1-\alpha)$-quantile of the conditional permutation distribution of the test statistic given observations of $\mathbf{Y_n}$, i.e.

$$c_{1-\alpha}^{Perm} := \min\left\{c \in \mathbb{R} : \tilde{\mathbb{P}}\left(T_n^{Perm}\left(\tau_n(\mathbf{Y_n}) > c\right)\Big|\mathbf{Y_n}\right) \leq \alpha\right\}.$$

To assure that the permutation test defined in Definition 4.3 is an asymptotic level α test, it suffices to show that $T_n^{Perm}(\mathbf{Y_n})$ is asymptotically standard normally distributed and that the conditional permutation distribution $T_n^{Perm}(\tau_n(\mathbf{Y_n}))|\mathbf{Y_n}$ converges to a standard normal distribution. The central limit theorem for conditional permutation distributions, confer Theorem 3.3 in Janssen (1997), proves that the conditional permutation distribution converges to a standard normal distribution and states sufficient conditions for this convergence.

Theorem 4.4 (Central limit theorem for conditional permutation distributions)**.** *As before, let $\mathbf{Y_n}$ be a random vector with length n and let $(c_{n,i})_{i\leq n}$ be a sequence of real numbers. Furthermore, suppose that the following conditions hold:*

1. *For each $n \in \mathbb{N}$ the sum of the squared regression coefficients and the regression coefficients is equal to one and zero, respectively:*

$$\sum_{i=1}^{n} c_{n,i}^2 = 1 \quad \forall\, n \in \mathbb{N},$$
$$\sum_{i=1}^{n} c_{n,i} = 0 \quad \forall\, n \in \mathbb{N}.$$

2. *With $\overline{Y}_{n,\cdot} := \sum_{i=1}^n Y_{n,i}/n$ the average, it holds that*

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(Y_{n,i} - \overline{Y}_{n,\cdot})^2 > 0 \qquad \mathbb{P} - a.s.$$

3. *There exists $\tilde{\sigma} > 0$ such that*

$$\frac{1}{\hat{\sigma}^2_{Perm}(\tau_n(\mathbf{Y}_n))}\frac{1}{n}\sum_{i=1}^{n}(Y_{n,i} - \overline{Y}_{n,\cdot})^2 \xrightarrow[n\to\infty]{\mathbb{P}\times\tilde{\mathbb{P}}} \tilde{\sigma}^2.$$

4. *The maximum of the sequence $(c_{n,i})_{i\leq n}$ of real numbers converges to zero:*

$$\max_{1\leq i\leq n} |c_{n,i}| \xrightarrow{n\to\infty} 0.$$

5. *For $d \to \infty$ it holds:*

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(Y_{n,i} - \overline{Y}_{n,\cdot})^2 \mathbb{1}_{[d,\infty)}\left(|Y_{n,i} - \overline{Y}_{n,\cdot}|\right) \to 0 \qquad \mathbb{P} - a.s.$$

Under the assumptions 1.-5., the permutation statistic in (4.2) is asymptotically normal distributed, i.e.:

$$\sup_{t\in\mathbb{R}} \left(\left| \tilde{\mathbb{P}} \left(T_n^{Perm}(\tau_n(\mathbf{Y_n})) \leq t | \mathbf{Y_n} \right) - \Phi\left(\frac{t}{\tilde{\sigma}}\right) \right| \right) \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Proof. Confer proof of Theorem 3.3 in Janssen (1997). □

Therefore, to construct an asymptotic permutation test, we have to determine the sequence $(c_{i,n})_{i\leq n}$ an the estimator $\hat{\sigma}_{Perm}$, which fulfill the corresponding conditions. In conclusion, in this section we stated some basic notations of hypothesis testing and, afterwards, introduced the Wald-type test as an asymptotic test. Wald-type tests are defined through an asymptotically normal distributed test statistic. As an alternative to Wald-type tests, we established an exact and an asymptotic permutation test. The exact permutation test bears on exchangeable random variables and its rejection are is defined by the permutation distribution of the test statistic given the corresponding observations. For the asymptotic permutation test, we did not assume exchangeable random variables, but, in contrast to the asymptotic permutation test, we assumed the test statistic to be asymptotically normal distributed. Analogously to the exact permutation test, we defined the rejection area of the asymptotic permutation test by a quantile of the test statistics permutation distribution. Last but not least, we stated the central limit theorem for conditional permutation distributions which ensures that the defined asymptotic permutation test is an asymptotic level α test.

5 Test for Assay Sensitivity

In the following, we study statistical tests for the assay sensitivity of a three-arm clinical trial. In Section 1, we already mentioned that we consider assay sensitivity as the superiority of the experimental or the reference treatment over placebo which resulted in assay sensitivity being defined as one of the following statistical testing problems

1. $H_0^{EP} : \lambda_E \geq \lambda_P \quad \text{versus} \quad H_1^{EP} : \lambda_E < \lambda_P,$
2. $H_0^{RP} : \lambda_R \geq \lambda_P \quad \text{versus} \quad H_1^{RP} : \lambda_R < \lambda_P,$
3. $H_0^{EP \cup RP} : H_0^{EP} \cup H_0^{RP} \quad \text{versus} \quad H_1^{EP \cap RP} : H_1^{EP} \cap H_1^{RP}.$

The first two hypotheses imply the same statistical problem. Analogously to the test procedure testing both, assay sensitivity and non-inferiority/superiority of the experimental versus the reference treatment, the hypothesis $H_0^{EP \cup RP}$ can be tested by testing the hypothesis H_0^{EP} and H_0^{RP} separately. Hence, without loss of generality, we only study the statistical hypothesis

$$H_0^{EP} : \lambda_E \geq \lambda_P \qquad \text{versus} \qquad H_1^{EP} : \lambda_E < \lambda_P.$$

Thereto, we introduce different Wald-type tests and a permutation test. Wald-type tests are commonly taken into account when comparing two rates of negative binomial distributions, confer Aban et al. (2009), Friede and Schmidli (2010), and Zhu and Lakkis (2013). However, to our knowledge there are no publications applying the permutation test to count data. We end this section by comparing the actual level of the established hypothesis tests with a simulation study for parameter settings motivated by the examples from Section 2.

In this section, we consider neither power, nor sample size planning, nor optimal sample size allocations for the corresponding tests, since the sample size is in general determined for the test procedure.

5.1 Wald-type Tests

In what follows, we construct different Wald-type tests for the hypothesis H_0^{EP}. Basically, the test statistic of the first test is obtained directly by the maximum-likelihood estimators for the rates of a negative binomial distribution. Furthermore, the second test statistic is defined by means of the logarithmized rate estimators to take account of the estimator's skewness. For both types of test statistics, we introduce different consistent variance estimators. At first, we note that the hypothesis H_0^{EP} can be written as $H_0^{EP} : \lambda_P - \lambda_E \leq 0$. As motivated in Section 4, we construct a Wald-type test by means of a consistent, asymptotically normal distributed maximum-likelihood estimator for the parameter of interest $\lambda_P - \lambda_E$. With the consistency and asymptotic normality of the maximum-likelihood estimators $\hat{\lambda}_E$ and $\hat{\lambda}_P$, confer Theorem 3.2, it follows that at the boundary of the hypothesis H_0^{EP}, i.e. for $\lambda_P - \lambda_E = 0$, the asymptotic normality

$$\sqrt{n}(\hat{\lambda}_P - \hat{\lambda}_E) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma^2_{EP})$$

holds with the variance σ^2_{EP} given by

$$\frac{\sigma^2_E}{w_E} + \frac{\sigma^2_P}{w_P} = \frac{\lambda_E(1+\phi\lambda_E)}{w_E} + \frac{\lambda_P(1+\phi\lambda_P)}{w_P}.$$

Here, $\lim_{n\to\infty} n_k/n = w_k \in (0,1)$ holds, meaning none of the groups vanished asymptotically. Hence, with $\hat{\sigma}^2_{EP}$ an under H_0^{EP} consistent estimator for the variance, we define the first Wald-type test statistic for H_0^{EP} by

$$T^{EP}_{n,Wald}(\mathbf{X_n}) := \sqrt{n}\frac{\hat{\lambda}_P - \hat{\lambda}_E}{\sqrt{\hat{\sigma}^2_{EP}}}.$$

The definition of the test statistic $T^{EP}_{n,Wald}$ results in the Wald-type test

$$\Psi^{EP}_{n,Wald}(\mathbf{X_n}) := \begin{cases} 1 & T^{EP}_{n,Wald}(\mathbf{X_n}) \geq q_{1-\alpha} \\ 0 & T^{ER}_{n,Wald}(\mathbf{X_n}) < q_{1-\alpha} \end{cases}.$$

However, the question remains how to estimate the variance σ^2_{EP} consistently. By taking the results from Section 3 into account, we establish three different appropriate estimators for the variance σ^2_{EP}. First of all, we deduce the unrestricted maximum-likelihood estimator for σ^2_{EP} from the maximum-likelihood estimators for σ^2_k, $k = E, P$.

Theorem 5.1 (Unrestricted maximum-likelihood estimator for the variance σ^2_{EP})**.** *For $k = E, P$, let $\hat{\sigma}^2_k$ be the maximum-likelihood estimator for the variance σ^2_k as in Corollary 3.3. Then, the maximum-likelihood estimator $\sigma^2_{EP,ML}$ for the variance σ^2_{EP} is given by*

$$\hat{\sigma}^2_{EP,ML} = \frac{\hat{\sigma}^2_E}{w_E} + \frac{\hat{\sigma}^2_P}{w_P}.$$

The estimator $\hat{\sigma}^2_{EP,ML}$ is consistent for the variance σ^2_{EP}.

Proof. Due to the functional invariance of maximum-likelihood estimators, $\hat{\sigma}^2_{EP,ML}$ is obtained by plugging in the corresponding maximum-likelihood estimators $\hat{\sigma}^2_k$. The consistency follows, since the sum of two consistent estimators is consistent for the limits of the two estimators. □

Analogously to the unrestricted maximum-likelihood estimator for the variance σ^2_{EP}, we define the restricted maximum-likelihood estimator for σ^2_{EP} by restricting the parameter estimators to the hypothesis H_0^{EP}. Estimating the variance restricted to the hypothesis can be advantageous in the sense of that the test statistic converges faster to its asymptotic distribution. As mentioned before, if the unrestricted maximum-likelihood estimators are located in the hypothesis, i.e. if $\hat{\lambda}_P - \hat{\lambda}_E \leq 0$ holds, the restricted maximum-likelihood estimators coincide with the unrestricted ones. On the other hand, we have to calculate the restricted ones by maximizing the likelihood function with respect to the boundary $\lambda_E = \lambda_P$ of the hypothesis. The next theorem states conditions for the solution of this problem.

Theorem 5.2. *If the maximum-likelihood estimators are not located in the hypothesis, the restricted maximum-likelihood estimators for the rates are given by*

$$\begin{aligned}\hat{\lambda}_{E,RML} &= \hat{\lambda}_{P,RML} = \frac{1}{n_E + n_P} \sum_{k=E,P} \sum_{i=1}^{n_k} X_{k,i},\\ \hat{\lambda}_{R,RML} &= \hat{\lambda}_R.\end{aligned}$$

Furthermore, the restricted maximum-likelihood estimator $\hat{\phi}_{RML}$ of the shape parameter is given as the maximizer of the log-likelihood function $\log l(\hat{\lambda}_{E,RML}, \hat{\lambda}_{R,RML}, \hat{\lambda}_{P,RML}, \phi|\mathbf{X_n})$ *with respect to ϕ and it is a solution of the equation*

$$G(\phi) := \sum_{k=E,R,P} \left(\frac{n_k}{\phi^2} \log\left(1 + \phi\hat{\lambda}_{k,RML}\right) - \sum_{i=1}^{n_k} \sum_{j=0}^{X_{k,i}-1} \frac{1}{j\phi^2 + \phi} \right) \overset{!}{=} 0 \tag{5.1}$$

with respect to ϕ.

The results from Theorem 5.2 follow immediately from the derivation of the log-likelihood function restricted to $\lambda_P = \lambda_E$. In Theorem 5.2 we proved that the restricted maximum-likelihood estimators for the rates λ_E and λ_P are equal to the mean of the observations from both groups. In addition, the estimator for the rate λ_R, which is not part of the hypothesis

H_0^{EP}, is equal to the unrestricted maximum-likelihood estimator $\hat{\lambda}_R$. Analogously to the unrestricted maximum-likelihood estimator for the shape parameter in Section 3, there is no closed form expression known for the restricted estimator $\hat{\phi}_{RML}$ and we cannot prove that Equation (5.1) has a unique solution. However, for the cases considered, the solution has been unique. Moreover, in the cases considered, it can be shown graphically that the function $G(\cdot)$ has the same shape as the corresponding function for independent and identically distributed random variables which solution is unique. With the same arguments as for the unrestricted maximum-likelihood variance estimator, the restricted one is a plug-in estimator. As mentioned in Section 3, the restricted maximum-likelihood estimators are consistent under H_0^{EP}. Thus, the restricted maximum-likelihood variance estimators are consistent, which is also stated in the next theorem.

Theorem 5.3 (Restricted maximum-likelihood estimator for σ^2_{EP})**.** *The maximum-likelihood estimator for the variance σ^2_{EP} with restriction to the hypothesis H_0^{EP} is given by*

$$\hat{\sigma}^2_{EP,RML} = \frac{\hat{\lambda}_{E,RML}(1+\hat{\phi}_{RML}\hat{\lambda}_{E,RML})}{w_E} + \frac{\hat{\lambda}_{P,RML}(1+\hat{\phi}_{RML}\hat{\lambda}_{P,RML})}{w_P}$$

with $\hat{\lambda}_{E,RML}, \hat{\lambda}_{P,RML}$, and $\hat{\phi}_{RML}$ the restricted maximum-likelihood estimators for the rates and the shape parameter, respectively. Under the hypothesis H_0^{EP}, the restricted maximum-likelihood estimator $\hat{\sigma}^2_{EP,RML}$ is consistent.

In case of independent and identically negative binomially distributed random variables, the maximum-likelihood estimator for the shape parameter is biased. Therefore, we expect that the unrestricted and the restricted maximum-likelihood estimator for the shape parameter as well as for the variance σ^2_{EP} are also biased. Additionally, this assertion is supported by Monte-Carlo simulations. As as consequence, we next estimate the variance σ^2_{EP} unbiased by means of the sample variance of the active treatment group E and the placebo group P.

Definition 5.4. Let $\hat{\sigma}^2_k$ be the sample variance of the observations from group $k = E, P$, i.e.

$$\hat{\sigma}^2_{k,SV} := \frac{1}{n_k - 1}\sum_{i=1}^{n_k}\left(X_{k,i} - \overline{X}_{k,\cdot}\right)^2.$$

Then, the sample variance estimator $\hat{\sigma}^2_{EP,SV}$ for the variance σ^2_{EP} is given by

$$\hat{\sigma}^2_{EP,SV} := \frac{\hat{\sigma}^2_{E,SV}}{w_E} + \frac{\hat{\sigma}^2_{P,SV}}{w_P}.$$

Theorem 5.5. *The sample variance estimator $\hat{\sigma}^2_{EP,SV}$ is an unbiased and consistent estimator for σ^2_{EP}.*

Proof. To prove the unbiasedness it suffices to show that the sample variance $\hat{\sigma}^2_{k,SV}$ is unbiased which is of course given because the random variables $X_{k,i}$ are independent and identically distributed for $i = 1, \dots, n_k$.
The consistency follows if the variance estimators $\hat{\sigma}^2_{k,SV}$ are consistent. By means of the algebraic formula for the variance, the estimator $\hat{\sigma}^2_{k,SV}$ can be rearranged to a function of the first and second sample moment. Since the sample moments are consistent estimators for the corresponding moments, the consistency of the sample variance $\hat{\sigma}^2_{k,SV}$ follows immediately. □

The distribution of the maximum-likelihood estimator $\hat{\lambda}_k$ is positively skewed and depending on the skewness, the test statistic $T^{EP}_{n,Wald}$ converges slowly against a standard normal distribution. To act contrary to the positive skew, we consider the logarithmized maximum-likelihood estimators $\log(\hat{\lambda}_k)$, $k = E, P$. Theorem 5.6 proves that the logarithmized maximum-likelihood estimator is consistent and asymptotically normal distributed.

Theorem 5.6. *The logarithmized maximum-likelihood estimator $\log(\hat{\lambda}_k)$, $k = E, P$ is a consistent estimator for $\log(\lambda_k)$ and, additionally, it is asymptotically normal distributed in the sense of*

$$\sqrt{n}\left(\log(\hat{\lambda}_k) - \log(\lambda_k)\right) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}\left(0, \frac{\sigma^2_{k,\log}}{w_k}\right)$$

with

$$\sigma^2_{k,\log} = \frac{\sigma^2_k}{\lambda^2_k} = \phi + \frac{1}{\lambda_k}.$$

Proof. Since the logarithm $\log(\cdot)$ is a continuous function, the consistency of the logarthmized rate estimator follows from the consistency of $\hat{\lambda}_k$ for λ_k and the continuous mapping theorem. The asymptotic normality follows by means of the delta method, which is studied detailed in Chapter 5.5.4 in Casella and Berger (2002). □

Therefore, we define a Wald-type test statistic by

$$T^{EP,\log}_{n,Wald}(\mathbf{X_n}) := \sqrt{n}\frac{\log(\hat{\lambda}_P) - \log(\hat{\lambda}_E)}{\sqrt{\hat{\sigma}^2_{EP,\log}}}$$

where $\hat{\sigma}^2_{EP,\log}$ denotes an under H_0^{EP} consistent estimator for the variance

$$\sigma^2_{EP,\log} := \frac{\sigma^2_{E,\log}}{w_E} + \frac{\sigma^2_{P,\log}}{w_P}.$$

The asymptotic normality of the test statistic $T^{EP,\log}_{n,Wald}$ at the boundary of the hypothesis H_0^{EP} follows from Theorem 5.6. Thus, we define a Wald-type test for the hypothesis H_0^{EP} by

$$\Psi^{EP,log}_{n,Wald}(\mathbf{X_n}) := \begin{cases} 1 & T^{EP,\log}_{n,Wald}(\mathbf{X_n}) > q_{1-\alpha} \\ 0 & T^{EP,\log}_{n,Wald}(\mathbf{X_n}) \leq q_{1-\alpha} \end{cases}.$$

In Section 4.1, we stated that Wald-type tests are defined through maximum-likelihood estimators for the parameters of the hypothesis. It should be mentioned that the definition of the Wald-type test $\Psi^{EP,\log}_{n,Wald}$ does not contradict this definition of Wald-type tests because $\log(\hat{\lambda}_k)$ is the maximum-likelihood estimator of $\log(\lambda_k)$ and the hypothesis H_0^{EP} is equivalent to $\log(\lambda_P) - \log(\lambda_E)$. Analogously to the estimators for the variance σ^2_{EP}, the variance $\sigma^2_{EP,\log}$ can be estimated by an unrestricted or a restricted maximum-likelihood as well as by a sample variance based estimator for the variance $\sigma^2_{EP,\log}$. Since the different variance estimators are obtained by plugging in the corresponding estimators for the rates, the shape parameter or the sample variance, respectively, we omit stating them. However, it should be mentioned that in contrast to the sample variance estimator $\hat{\sigma}^2_{EP,SV}$, the sample variance estimator for the variance $\sigma^2_{EP,\log}$ is biased since the reciprocal of the squared maximum-likelihood estimator of a rate is not an unbiased estimator for the reciprocal of the squared rate.

To conclude, we established two approaches for Wald-type tests for the hypothesis H_0^{EP} where one takes the difference of the rates and the other one the difference of the logarithmized rates into account. In addition, we introduced different ways to estimate the variance for the corresponding test statistics. At least asymptotically, both tests result in the same decision.

5.2 Permutation test

The Wald-type tests, which has been established in the last subsection, are asymptotic tests and, as a consequence, their actual level of significance is not guaranteed to be α. However, at the boundary of the hypothesis H_0^{EP}, i.e. for $\lambda_E = \lambda_P$, the random variables

from the experimental treatment group E and the placebo group P are exchangeable, since they are independent and identically distributed. Hence, in this subsection, we construct a permutation test for H_0^{EP} which is exact at the boundary of the hypothesis.
A permutation test statistic for H_0^{EP} can obviously defined through the difference of the estimated rates of the different groups, i.e.

$$\tilde{T}_{n,Perm}^{EP}\left(\mathbf{X}_{E,n_E}, \mathbf{X}_{P,n_P}\right) := \hat{\lambda}_P - \hat{\lambda}_E.$$

Even if this test statistic is appropriate to construct an exact permutation test, we define the test statistic for the permutation test as

$$T_{n,Perm}^{EP}\left(\mathbf{X}_{E,n_E}, \mathbf{X}_{P,n_P}\right) := \sqrt{n}\frac{\hat{\lambda}_P - \hat{\lambda}_E}{\sqrt{\frac{\hat{\sigma}^2_{P,SV}}{w_P} + \frac{\hat{\sigma}^2_{E,SV}}{w_E}}}.$$

This test statistic corresponds to the Wald-type test statistic $T_{n,Wald}^{EP}$ with the variance estimated by the sample variances. Lemma 4.1 in Janssen (1997) proves that the permutation distribution of $T_{n,Perm}^{EP}\left(\tau_{n_E+n_P}\left(\mathbf{X}_{E,n_E}, \mathbf{X}_{P,n_P}\right)\right)$ conditioned on $\left(\mathbf{X}_{E,n_E}, \mathbf{X}_{P,n_P}\right)$ converges asymptotically to a standard normal distribution regardless of whether the random variables of the different groups are exchangeable at the boundary of the hypothesis. Hence, even if the assumption of exchangeability is not fulfilled, the permutation test for H_0^{EP} based on $T_{n,Perm}^{EP}$ is at least asymptotically exact. The random variables are not exchangeable, if, for instance, the shape parameter is not equal among the two groups. By means of the statistic $T_{n,Perm}^{EP}$, the one-sided permutation test $\Psi_{n,Perm}^{EP}$ is defined analogously to Definition 4.1. Since the assumptions of Theorem 4.2 hold, it follows that the permutation test $\Psi_{n,Perm}^{EP}$ is an exact test at the boundary of the hypothesis H_0^{EP}. Moreover, Monte-Carlo simulations showed that the level of significance is less than α in the interior of the null hypothesis, i.e. for $\lambda_E > \lambda_P$. As mentioned before, in practice and especially for the simulation study in the next subsection, we define the parameter $\gamma_{n,Perm}$ to be zero and, hence, the quantile $c_{n,Perm}$ is given by

$$c_{n,Perm} = \operatorname{argmin}\left\{c \in \mathbb{R} : \tilde{\mathbb{P}}\left(T_{n,Perm}^{EP}(\tau_{n_E+n_P}(\mathbf{X_{E,n_E}}, \mathbf{X_{P,n_P}})) > c \middle| \mathbf{X_{E,n_E}}, \mathbf{X_{P,n_P}}\right) \leq \alpha\right\}.$$

In consequence, if $\lambda_E = \lambda_P$ holds, the actual level of the resulting permutation test is possibly not equal to α but slightly smaller. However, this definition avoids test outcomes which cannot interpreted uniquely as a rejection or non-rejection of the hypothesis. An exact calculation of the quantile $c_{n,Perm}$ for a given vector of observations $(\mathbf{X_{E,n_E}}, \mathbf{X_{P,n_P}})$ is in general not feasible due to a large number of possible permutations of the observations. More precisely, the number of possible allocations of n observations into two groups with n_1 and n_2 observations, respectively, is given by $\binom{n}{n_1}$. Therefore, we approximate the quantile $c_{n,Perm}$ by Monte-Carlo methods.

5.3 Simulation study

So far, we established a wide range of tests for the hypothesis H_0^{EP} but especially for the Wald-type tests, we do not know the actual level. Therefore, in the following we compare the different tests for the hypothesis H_0^{EP} by Monte-Carlo simulations. For this purpose, we have to determine the rates λ_E, λ_R, λ_P and the shape parameter ϕ for the corresponding negative binomial distributions. Although we only compare the effect of one active treatment with the placebo response, we have to specify the parameter for the second active treatment because the shape parameter is estimated taking the observations of all groups into account. In addition, the sample sizes n_E, n_R, and n_P have2 to be fixed. We motivate the choices of the different parameters by the examples for clinical trials discussed in Sections 2.2.1 and 2.2.2. Additionally, we construct the different tests with a level of significance $\alpha = 0.05$ and run M Monte-Carlo simulations. In the following, let $\hat{\alpha}_{act} = \hat{\alpha}_{act}(M)$ be the approximation of the actual level α_{act} of a test. We choose the number of Monte-Carlo simulations to be $M = 20{,}000$. The number M is motivated by a statistical test assessing whether the simulated actual level is equal to α. More precisely, since a statistical test Ψ is Bernoulli distributed with success probability equal to the actual level of significance α_{act}, the number of rejected hypothesis $M\hat{\alpha}_{act}(M)$ is binomial distributed with number of trials and success probability equal to the number of simulations and the actual level, respectively. Hence, the rejection area of an asymptotic two-sided test for the hypothesis $H_0 : \alpha_{act} = \alpha$ with level of significance 0.05 is given by

$$\begin{aligned} &\left[0, \alpha - q_{0.975}\sqrt{\frac{\alpha(1-\alpha)}{M}}\right] \cup \left[\alpha + q_{0.975}\sqrt{\frac{\alpha(1-\alpha)}{M}}, 1\right] \\ =&[0, 0.04698] \cup [0.05302, 1]. \end{aligned}$$

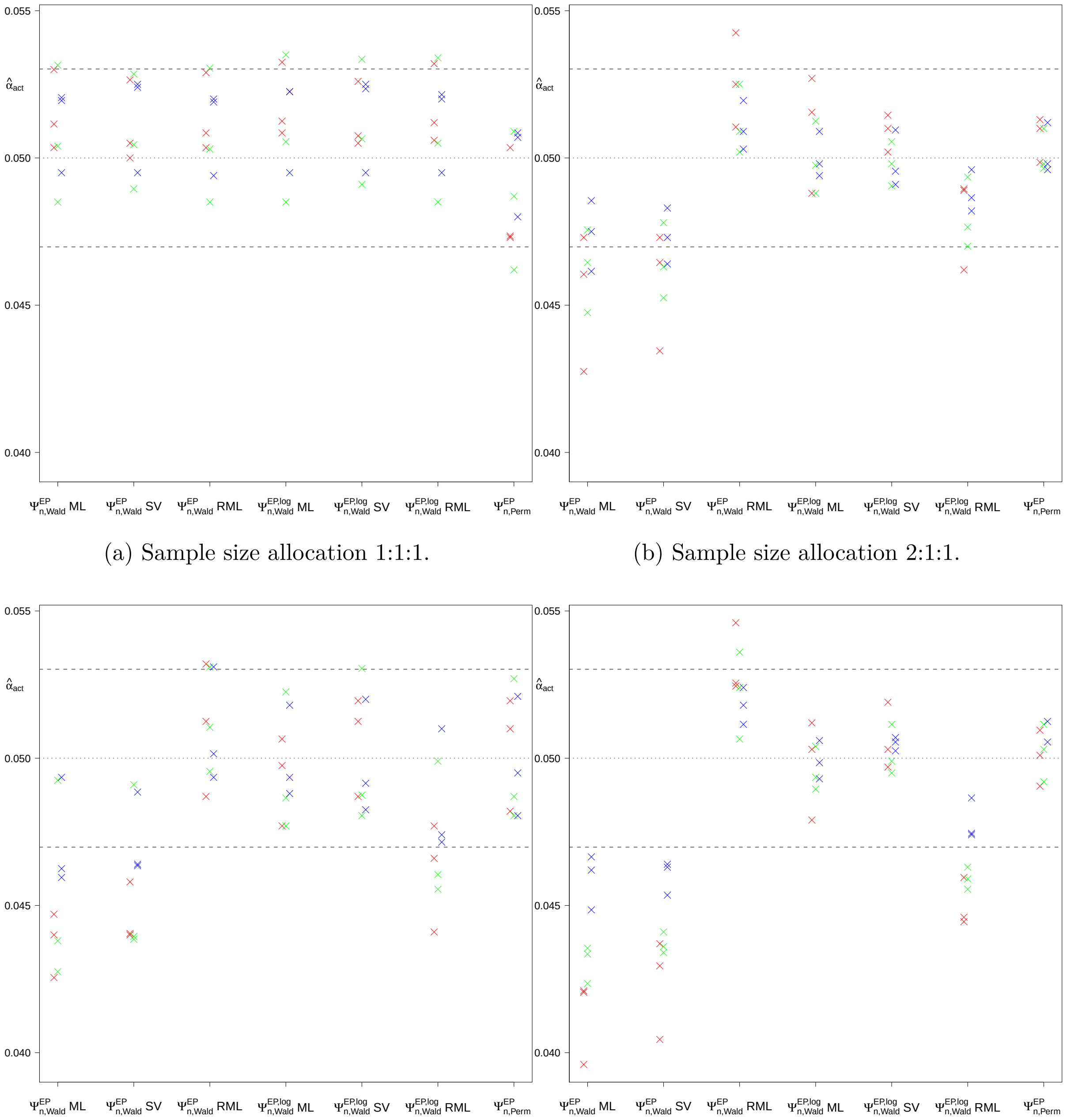
Therefore, if the approximation $\hat{\alpha}_{act}$ of the actual level α_{act} is contained in the set above, the actual level of the corresponding test is significantly different from $\alpha = 0.05$. In this section's graphics, the boundaries of the rejection area are plotted as dashed grey lines.

From a practical point of view, if the actual level of a test deviates from α, we regard an upward deviation as more worse then a downward deviation. This is due because a liberal tests means that we falsely assume the experimental treatment to be more effective than the placebo more times than planned. On the other hand, if a test is conservative, the error probability of a falsely rejected hypothesis is less then considered but at least a treatment is not incorrectly considered as effective more often than intended. It is worth mentioning that in some cases a small inflation of the actual level is tolerated, confer Section 4.4 in Friede et al. (2007), who tolerated a deviation of ±10%. Next, we define the parameter setting for the first Monte-Carlo study by means of the example we discussed in Section 2.2.1.

Definition 5.7 (Parameter setting motivated by the TRISTAN study in COPD)**.**

$$\begin{aligned} \lambda_E = \lambda_R = \lambda_P = 1.71 \\ \phi \in \{0.3, 0.5, 0.7\}, \\ n \in \{550, 1100, 2200\}, \\ n_E : n_R : n_P \in \{1{:}1{:}1,\ 2{:}1{:}1,\ 2{:}2{:}1,\ 3{:}2{:}1\}. \end{aligned}$$

The choice of the rates is based on the exacerbation rates from Table 1. Since we simulate the actual level at the boundary of the hypothesis, we have to choose at least the rate λ_E to be equal to λ_P. The rate λ_R which potentially affects the actual level due to an influence on the variance estimation is also defined equally to the placebo rate λ_P for the sake of simplicity. The different shape parameters are motivated by the confidence interval $[0.34, 0.6]$ for the shape parameter ϕ of the TRISTAN study. With the choice of the shape parameter as above we cover the range of the confidence interval. Last but not least, the exacerbations rates in Table 1 were calculated with results from about 360 patients per group which corresponds to a three-arm study with approximately 1080 patients. To cover a wide range, we define the total number of observations in Definition 5.7 as 550, 1100 and 2200. The sample size allocations are common allocations in three-arm clinical trials, confer Pigeot et al. (2003), who take the allocations 1:1:1, 2:2:1, and 3:2:1 for simulations into account. Additionally, we consider the allocation 2:1:1, which is one further example for the idea of allocating more patients to the experimental treatment group than to the other groups to obtain much information about the experimental treatment. The next figure shows the results of the Monte-Carlo study.



(a) Sample size allocation 1:1:1.

(b) Sample size allocation 2:1:1.

(c) Sample size allocation 2:2:1.

(d) Sample size allocation 3:2:1.

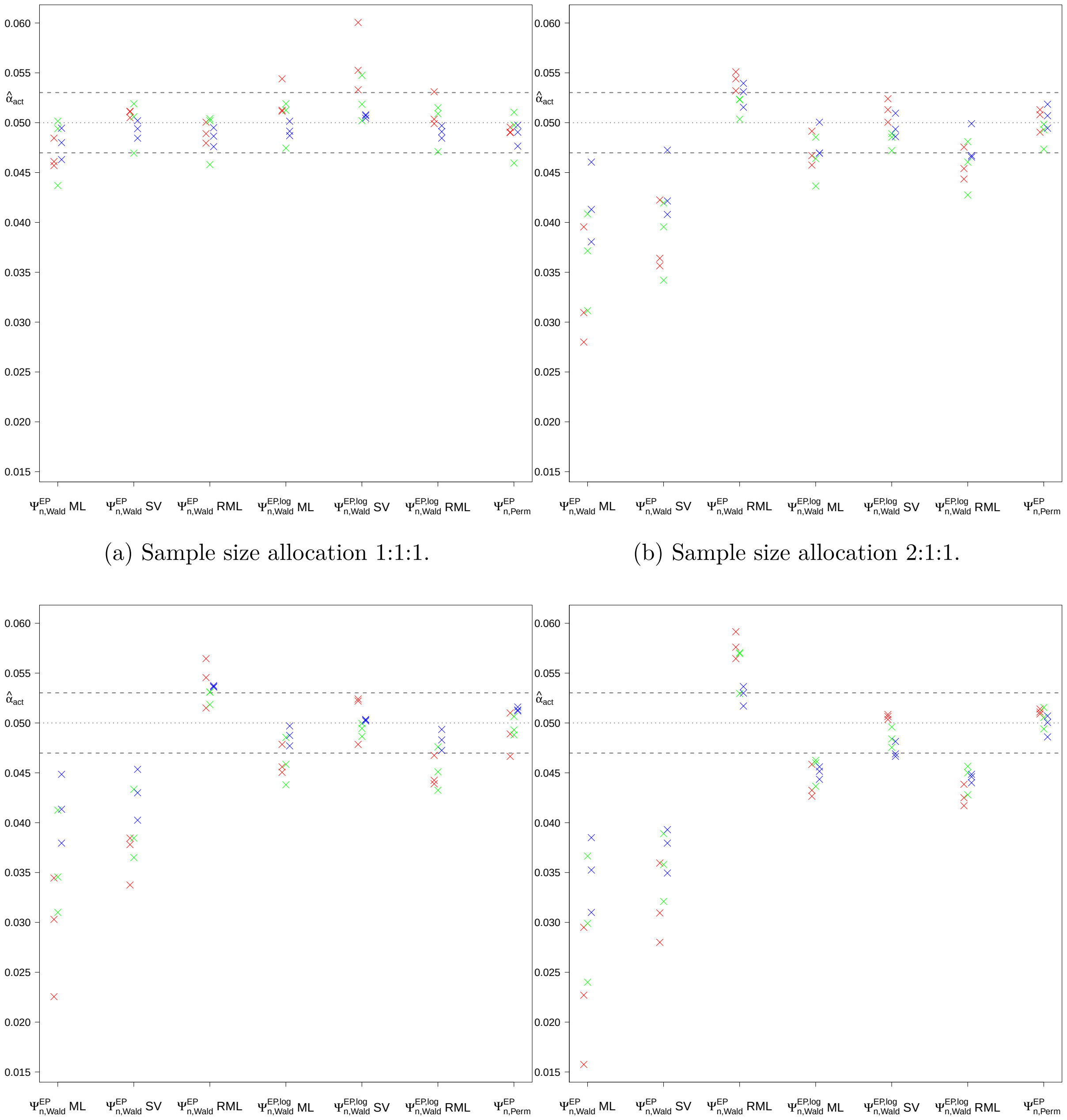
Figure 3: Actual level of different tests for H_0^{EP} by sample size allocation. The points for a sample size of 550 are red, for one of 1100 are green, and for one of 2200 are blue. Actual levels between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations *ML*, *SV*, *RML* denote whether the variance of the Wald-type test has been estimated by the unrestricted maximum-likelihood estimator, the sample variance estimator or the restricted maximum-likelihood estimator, respectively. The values for the shape parameter are not marked differently.

Altogether, the permutation test $\Psi_{n,Perm}^{EP}$ performs best and is recommended for usage, i.e. there is no trend for being liberal or conservative detectable. In the following, we regard the different sample size allocations and state which test are appropriate for usage besides the permutation test. For the sample size allocation 1:1:1, the actual levels of the Wald-type tests tend to be slightly liberal, since some points are above or at least near the upper grey line. However, this seems to be a trend and none of the Wald-type tests can be regarded as inappropriate for the current parameter combinations. Among each other, the results for the allocation 2:1:1, 2:2:1, and 3:2:1 are qualitatively the same. The Wald-type test $\Psi_{n,Wald}^{EP}$ with a restricted maximum-likelihood variance estimator is not appropriate to test the hypothesis H_0^{EP} because it tends to be liberal. The other tests can be applied but the Wald-type test $\Psi_{n,Wald}^{EP}$ with an unrestricted maximum-likelihood variance estimator or a sample variance estimator as all as the Wald-type test $\Psi_{n,Wald}^{EP,\log}$ with a restricted maximum-likelihood variance estimator are conservative or, at least, tend to be conservative.
Next, we define the parameter setting motivated by the example for a clinical trial in MS from Section 2.2.2.

Definition 5.8 (Parameter setting motivated by the CONFIRM study in MS)**.**

$$\begin{aligned}\lambda_E = \lambda_R = \lambda_P = 17.4,\\ \phi \in \{1,2,3\},\\ n \in \{215,430,860\},\\ n_E : n_R : n_P \in \{1{:}1{:}1,\ 2{:}1{:}1,\ 2{:}2{:}1,\ 3{:}2{:}1\}.\end{aligned}$$

We calculated the sample sizes n in Definition 5.8 the same way as the sample size in Definition 5.7. In contrast, the shape parameter is not motivated by a confidence interval but by the approximations for the shape parameter from Table 3. Figure 4 shows the actual levels of the different tests for the parameter setting stated in Definition 5.8.



(a) Sample size allocation 1:1:1.

(b) Sample size allocation 2:1:1.

(c) Sample size allocation 2:2:1.

(d) Sample size allocation 3:2:1.

Figure 4: Actual level of different tests for H_0^{EP} by sample size allocation. The points for a sample size of 215 are red, for one of 430 are green, and for one of 860 are blue. Actual level between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations *ML*, *SV*, *RML* denote whether the variance of the Wald-type test has been estimated by the unrestricted maximum-likelihood estimator, the sample variance estimator or the restricted maximum-likelihood estimator, respectively. The values for the shape parameter are not marked differently.

Analogously to the scenarios motivated by the TRISTAN study, the permutation test $\Psi_{n,Perm}^{EP}$ can be recommended for application for the sample size allocations considered. Besides the permutation test, for the allocation 1:1:1, all Wald-type tests except the test $\Psi_{n,Wald}^{EP,\log}$ with a restricted maximum-likelihood variance estimator are appropriate for usage. The test last mentioned tends to be liberal for the smaller sample sizes. For the other sample size allocations, the results are qualitatively the same among each other. More precisely, the Wald-type test $\Psi_{n,Wald}^{EP}$ with the restricted maximum-likelihood variance estimator is liberal and, thus, not appropriate for usage. The other tests can be applied but only the permutation test $\Psi_{n,Perm}^{EP}$ and the Wald-type test $\Psi_{n,Wald}^{EP,\log}$ with the sample variance estimator are not conservative.

To conclude, for the considered parameter combinations motivated by the examples in Section 2 for trials in COPD and MS, the permutation test $\Psi_{n,Perm}^{EP}$ performs all in all the best concerning a small deviation of the actual level from $\alpha = 0.05$. For both parameter settings holds that especially the sample size allocation affects the actual level of the Wald-type tests. However, the magnitude of the influence depends on the test as well as the parameter setting. In particular, the Wald-type tests $\Psi_{n,Wald}^{EP,\log}$ seem to be more robust concerning an unbalanced allocation than the Wald-type tests $\Psi_{n,Wald}^{EP}$. Additionally, for some of the sample size allocations, the actual level of the Wald-type tests from Figure 3 and Figure 4 differ. Hence, depending on the sample size allocation, the parameter setting has an influence on the actual level of Wald-type tests.

Lastly, it should be emphasized that in practice, the sample size allocation as well as the sample size itself are known and therefore, a statistical test do not have to perform well for all allocations and sample sizes.

6 Retention of Effect Hypothesis

In this section, we describe several Wald-type tests and an asymptotic permutation test for the retention of effect hypothesis

$$H_0^{RET}: \lambda_P - \lambda_E \leq \Delta(\lambda_P - \lambda_R) \qquad \text{versus} \qquad H_1^{RET}: \lambda_P - \lambda_E > \Delta(\lambda_P - \lambda_R),$$

with $\Delta \in (0,1)$ the prespecified non-inferiority margin. Furthermore, we calculate functions approximating the power of the Wald-type tests and the permutation test. Subsequently, from these power approximating functions we obtain sample size formulas, i.e. formulas approximating the sample size necessary to test the retention of effect hypothesis with power $1-\beta$ for a given parameter vector $\zeta_{H_1^{RET}}$ located in the alternative and a given sample size allocation (w_E, w_R, w_P). Since the sample size allocation influences the power, we suggest different asymptotically optimal sample size allocations, too. Thereby, optimality is regarded in terms of a larger power. We conclude this section by studying the finite sample size properties of the hypothesis tests for H_0^{RET}. Thereto, we simulate the actual levels and the power of these the tests. Additionally, since some of the Wald-type tests take explicitly into account that the observations are negative binomially distributed, we analyze how sensitive the tests are concerning deviations from the assumed distribution. Finally, we show by means of Monte-Carlo simulations that the sample size for the test procedure can be planned through the tests for the retention of effect hypothesis.

It should be mentioned that the theory which will be established in this section also holds for $\Delta \in (1, \infty)$, i.e. we also could test superiority instead of non-inferiority. The margin $\Delta = 1$ corresponds to testing superiority of the experimental over the reference treatment without taking the placebo into account. We exclude this case in the following because the theory of the different tests certainly holds but not the theory for the optimal sample size allocations.

6.1 Wald-type Tests

In the following, we introduce Wald-type tests for the retention of effect hypothesis H_0^{RET} and analogously to the Wald-type tests for the assay sensitivity, we establish an unrestricted and a restricted maximum-likelihood as well as a sample variance based estimator for the variance within the test statistic.

First of all, with the parameter $\eta := (1-\Delta)\lambda_P + \Delta\lambda_R - \lambda_E$, we rearrange the retention of effect hypothesis to

$$H_0^{RET} : \eta \leq 0 \qquad \text{versus} \qquad H_1^{RET} : \eta > 0.$$

According to Section 4.1, we have to estimate the parameter η with a consistent, at the boundary of the hypothesis asymptotically standard normally distributed maximum-likelihood estimator to establish the Wald-type tests for the retention of effect hypothesis. With the functional invariance of maximum-likelihood estimators we obtain the maximum-likelihood estimator $\hat{\eta} := (1-\Delta)\hat{\lambda}_P + \Delta\hat{\lambda}_R - \hat{\lambda}_E$ for the parameter η. From the properties of the maximum-likelihood estimators for the rates from Theorem 3.2, it follows that the estimator $\hat{\eta}$ meets the requirements.

Theorem 6.1. *Let none of the groups $k = E, R, P$ vanish asymptotically, i.e. the convergence $\lim_{n\to\infty} n_k/n = w_k \in (0,1)$ holds. Then, the maximum-likelihood estimator $\hat{\eta}$ is a consistent estimator for the parameter of interest η and the asymptotic normality*

$$\sqrt{n}\,(\hat{\eta} - \eta) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0, \sigma_{RET}^2)$$

holds. The variance σ_{RET}^2 is given by

$$\begin{aligned} \sigma_{RET}^2 &:= \frac{\sigma_E^2}{w_E} + \Delta^2\frac{\sigma_R^2}{w_R} + (1-\Delta)^2\frac{\sigma_P^2}{w_P} \\ &= \frac{\lambda_E(1+\lambda_E\phi)}{w_E} + \Delta^2\frac{\lambda_R(1+\lambda_R\phi)}{w_R} + (1-\Delta)^2\frac{\lambda_P(1+\lambda_P\phi)}{w_P}. \end{aligned} \tag{6.1}$$

Let $\hat{\sigma}_{RET}^2$ be an estimator for the variance σ_{RET}^2 which is consistent under the hypothesis H_0^{RET}. Then, we define a Wald-type test statistic for the retention of effect hypothesis by

$$T_{n,Wald}^{RET}(\mathbf{X_n}) := \sqrt{n}\frac{\hat{\eta}}{\sqrt{\hat{\sigma}_{RET}^2}}. \tag{6.2}$$

The asymptotic normality of $T_{n,Wald}^{RET}$ at the boundary of H_0^{RET} follows immediately from Theorem 6.1. Therewith, we define the Wald-type test for the retention of effect hypothesis by

$$\Psi_{n,Wald}^{RET}(\mathbf{X_n}) := \begin{cases} 1 & T_{n,Wald}^{RET}(\mathbf{X_n}) \geq q_{1-\alpha} \\ 0 & T_{n,Wald}^{RET}(\mathbf{X_n}) < q_{1-\alpha}. \end{cases}$$

After establishing the Wald-type test $\Psi_{n,Wald}^{RET}$ for the retention of effect hypothesis, next, we study consistent estimators $\hat{\sigma}_{RET}^2$ for the variance σ_{RET}^2.

As before, the maximum-likelihood variance estimator $\hat{\sigma}_{RET,ML}^2$ is obtained by replacing the variances σ_k^2, $k = E, R, P$, in Formula (6.1) by the corresponding maximum-likelihood estimators from Corollary 3.3. Due to the continuous mapping theorem, the consistency of the estimator $\hat{\sigma}_{RET,ML}^2$ follows from the consistency of the maximum-likelihood estimators for the variances σ_k^2 shown in Corollary 3.3. We assume that the estimator $\hat{\sigma}_{RET,ML}^2$ is biased because the maximum-likelihood estimators for the group variances are expected to be biased. Additionally, Monte-Carlo simulations support the assertion of a biased estimator $\hat{\sigma}_{RET,ML}^2$.

In what follows, we establish the maximum-likelihood variance estimator $\hat{\sigma}_{RET,RML}^2$ restricted to the hypothesis H_0^{RET}. Thereto, we calculate the corresponding restricted estimators for the rates and the shape parameter. As mentioned before, the restricted and the unrestricted maximum-likelihood estimators coincide if the unrestricted ones are located in the hypothesis, i.e. if $\hat{\eta} \leq 0$ holds. In contrast, the restricted maximum-likelihood estimators are calculated by maximizing the log-likelihood function restricted to

$$\eta = 0 \qquad \Longleftrightarrow \qquad \lambda_P = \frac{\lambda_E - \Delta\lambda_R}{1 - \Delta}$$

if the unrestricted estimators are not located in the hypothesis. Necessary conditions for the maximizer of the restricted log-likelihood function

$$\log l_{H_0^{RET}}(\lambda_E, \lambda_R, \phi | \mathbf{X_n}) := \frac{n_E}{\phi} \log\left(\frac{1}{1 + \lambda_E\phi}\right) + X_{E,\cdot} \log\left(\frac{\lambda_E\phi}{1 + \lambda_E\phi}\right) + \frac{n_R}{\phi} \log\left(\frac{1}{1 + \lambda_R\phi}\right)$$
$$+ X_{R,\cdot} \log\left(\frac{\lambda_R\phi}{1 + \lambda_R\phi}\right) + \frac{n_P}{\phi} \log\left(\frac{1}{1 + \frac{\lambda_E - \Delta\lambda_R}{1-\Delta}\phi}\right) + X_{P,\cdot} \log\left(\frac{\frac{\lambda_E - \Delta\lambda_R}{1-\Delta}\phi}{1 + \frac{\lambda_E - \Delta\lambda_R}{1-\Delta}\phi}\right)$$
$$+ \sum_{k \in \{E,R,P\}} \sum_{i=1}^{n_k} \sum_{j=0}^{X_{k,i}-1} \log\left(j + \frac{1}{\phi}\right)$$

are given by equations resulting from equating the partial derivatives of the restricted log-likelihood function with zero:

$$\begin{aligned}\frac{\partial \log l_{H_0^{RET}}(\lambda_E,\lambda_R,\phi|\mathbf{X_n})}{\partial \lambda_E} &= -\frac{n_P+\phi x_{P,\cdot}}{1-\Delta+(\lambda_E-\Delta\lambda_R)\phi}+\frac{x_{P,\cdot}}{\lambda_E-\Delta\lambda_R}+\frac{x_{E,\cdot}-\lambda_E n_E}{\lambda_E(\lambda_E\phi+1)}\overset{!}{=}0,\\ \frac{\partial \log l_{H_0^{RET}}(\lambda_E,\lambda_R,\phi|\mathbf{X_n})}{\partial \lambda_R} &= \frac{\Delta(n_P+\phi x_{P,\cdot})}{1-\Delta+(\lambda_E-\Delta\lambda_R)\phi}-\frac{\Delta x_{P,\cdot}}{\lambda_E-\Delta\lambda_R}+\frac{x_{R,\cdot}-\lambda_R n_R}{\lambda_R(\lambda_R\phi+1)}\overset{!}{=}0,\\ \frac{\partial \log l_{H_0^{RET}}(\lambda_E,\lambda_R,\phi|\mathbf{X_n})}{\partial \phi} &= \frac{X_{E,\cdot}-n_E\lambda_E}{\phi(\phi\lambda_E+1)}+\frac{X_{R,\cdot}-n_R\lambda_R}{\phi(\phi\lambda_R+1)}+\frac{X_{P,\cdot}-n_P\lambda_P}{\phi(\phi\lambda_P+1)}\\ &\quad+\frac{n_R\log(\phi\lambda_R+1)+n_E\log(\phi\lambda_E+1)+n_P\log(\phi\lambda_P+1)}{\phi^2}\\ &\quad-\sum_{k\in\{E,R,P\}}\sum_{i=1}^{n_k}\sum_{j=0}^{X_{k,i}-1}\frac{1}{j\phi^2+\phi}\overset{!}{=}0.\end{aligned}$$

For the restricted rate estimators $\hat{\lambda}_{k,RML}$, $k=E,R$, as well as for the restricted shape estimator $\hat{\phi}_{RML}$ no closed form expression is known. The restricted maximum-likelihood estimator $\hat{\lambda}_{P,RML}$ for λ_P is given by

$$\hat{\lambda}_{P,RML}=\frac{\hat{\lambda}_{E,RML}-\Delta\hat{\lambda}_{R,RML}}{1-\Delta}.$$

Additionally, it is not known whether the restricted estimators exist nor if they are unique when they exist. However, in all cases considered the restricted log-likelihood function has a unique maximum for the parameter spaces considered. As mentioned in Section 3, the restricted maximum-likelihood estimators are consistent if the true parameter is located in the hypothesis H_0^{RET}.

Finally, we introduce the sample variance based estimator for the variance σ^2_{RET} which is unbiased. With the sample variance $\hat{\sigma}^2_{k,SV}$ of the observations from group $k=E,R,P$ introduced in Definition 5.4, we define the sample variance based estimator

$$\hat{\sigma}^2_{RET,SV}:=\frac{\hat{\sigma}^2_{E,SV}}{w_E}+\Delta^2\frac{\hat{\sigma}^2_{R,SV}}{w_R}+(1-\Delta)^2\frac{\hat{\sigma}^2_{P,SV}}{w_P}$$

for the variance σ^2_{RET}. Both, the unbiasedness and the consistency of the estimator $\hat{\sigma}^2_{RET,SV}$ follow from the properties of the sample variances.

Remark 6.2. The Wald-type test $\Psi^{RET}_{n,Wald}$ with the variance estimated by $\hat{\sigma}^2_{RET,SV}$ does not assume the parametric model introduced in Section 2.3. A nonparametric model such that the first, second, and fourth moments of the random variables $X_{k,i}$ exist as well as that the random variables are independent suffices for the Wald-type test $\Psi_{n,Wald}^{RET}$ with the sample variance estimator to be an asymptotic level α test for the retention of effect hypothesis H_0^{RET}. In this case, the rates for the hypothesis H_0^{RET} are the expectation of the random variables, i.e. $\lambda_k = \mathbb{E}[X_{k,i}]$ with $i = 1, \ldots, n_k$ and $k = E, R, P$.

6.2 Permutation test

In this section, we introduce an asymptotic permutation test for the retention of effect hypothesis H_0^{RET}. In Section 4.2, we proved that we can construct an exact permutation test for a certain hypothesis if the corresponding random variables are exchangeable at the boundary of this hypothesis. However, at the boundary of the retention of effect hypothesis, i.e. for $\eta = 0$, the entries of the random vector $\mathbf{X_n}$ are not exchangeable and hence we construct an asymptotic permutation test. In Equation (4.2), we defined a test statistic to construct an asymptotic permutation test as

$$T_{n,Perm}^{RET}(\mathbf{X_n}) := \frac{\sum_{i=1}^{n} c_{n,i} X_{n,i}}{\hat{\sigma}_{Perm}(\mathbf{X_n})}$$

with $\hat{\sigma}_{Perm}(\mathbf{X_n})$ an estimator for the standard deviation of $\sum_{i=1}^{n} c_{n,i} X_{n,i}$. In addition, we cited in Theorem 4.4 the central limit theorem for conditional permutation distribution which guarantees that under certain conditions a permutation test defined through T_n^{Perm} is an asymptotic test. Thus, in the following, we define the coefficients $(c_{i,n})_{i \leq n}$ and the variance estimator $\hat{\sigma}_{Perm}(\mathbf{X_n})$ such that the test statistic $T_{n,Perm}^{RET}$ fulfills the assumptions of Theorem 4.4 and is therefore appropriate for defining an asymptotically exact permutation test $\Psi_{n,Perm}^{RET}$ as in Definition 4.3. Thereto, we first of all define for each $n \in \mathbb{N}$ the scheme of regression coefficients $(c_{n,i})_{i \leq n}$ by

$$c_{n,i} := \sqrt{\frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta - 1)^2 n_E n_R}} \times \begin{cases} -\frac{1}{n_E} & i = 1, \ldots, n_E \\ \frac{\Delta}{n_R} & i = n_E + 1, \ldots, n_E + n_R \\ \frac{1-\Delta}{n_P} & i = n_E + n_R + 1, \ldots, n \end{cases}. \tag{6.3}$$

The variance σ_{Perm}^2 of the weighted sum $\sum_{i=1}^{n} c_{n,i} X_{n,i}$ is given by

$$\operatorname{Var}\left[\sum_{i=1}^{n} c_{n,i} X_{n,i}\right] = \frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta - 1)^2 n_E n_R} \left(\frac{\sigma_E^2}{n_E} + \Delta^2 \frac{\sigma_R^2}{n_R} + (1 - \Delta)^2 \frac{\sigma_P^2}{n_P}\right).$$

Therefore, we define an estimator $\hat{\sigma}^2_{Perm}(\mathbf{X_n})$ for the variance σ^2_{Perm} by

$$\hat{\sigma}^2_{Perm}(\mathbf{X_n}) := \frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (1-\Delta)^2 n_E n_R} \left(\frac{\hat{\sigma}^2_{E,SV}}{n_E} + \Delta^2 \frac{\hat{\sigma}^2_{R,SV}}{n_R} + (1-\Delta)^2 \frac{\hat{\sigma}^2_{P,SV}}{n_P} \right)$$

with $\hat{\sigma}^2_{k,SV}$ the sample variance estimator for σ^2_k as in Definition 5.4. We defined the test statistic for the permutation test $\Psi^{RET}_{n,Perm}$ as above because it corresponds to the way Janssen (1997) defined the test statistic for an asymptotic permutation test and it simplifies the proof of the asymptotic normality of the conditional permutation distribution of the test statistic. However, simple rearrangements show that the test statistic $T^{RET}_{n,Perm}$ is equal to the test statistic of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the sample variance estimator. Theorem 6.3 proves that the test statistic $T^{RET}_{n,Perm}$ fulfills the assumptions of the central limit theorem for conditional permutation distributions.

Theorem 6.3. *Let $\mathbb{P}$ denote the probability measure of $\mathbf{X_n}$ and let $\tau_n(\mathbf{X_n})$ be a random variable whose realizations are the permutations of $\mathbf{X_n}$. The random variable $\tau_n(\mathbf{X_n})$ is assumed to be uniformly distributed on the space of permutations of the vectors with length n and its probability measure is denoted as $\tilde{\mathbb{P}}$. Both probability measures are assumed to be independent. Moreover, none of the groups vanishes asymptotically, i.e. $\lim_{n\to\infty} n_k/n = w_k \in (0,1)$. Then, with the definition of $c_{n,i}$ and $\hat{\sigma}^2_{Perm}$ as before the asymptotic normality of $T^{RET}_{n,Perm}$ in the sense of*

$$\sup_{t\in\mathbb{R}} \left(\left| \tilde{\mathbb{P}} \left(T^{RET}_{n,Perm}(\tau_n(\mathbf{X_n})) \leq t | \mathbf{X_n} \right) - \Phi(t) \right| \right) \xrightarrow[n\to\infty]{\mathbb{P}} 0$$

holds.

The proof is stated in Appendix A. From the asymptotic normality proved in Theorem 6.3, it follows that the asymptotic permutation test $\Psi^{RET}_{n,Perm}$ defined through the test statistic $T^{RET}_{n,Perm}$ with Definition 4.3 is an asymptotic level α test. Remark 6.4 discusses the assumptions for the permutation test $\Psi^{RET}_{n,Perm}$.

Remark 6.4. In the proof of Theorem 6.3, concerning the distribution of the random variables $X_{k,i}$ with $i = 1, \ldots, n_k$ and $k = E, R, P$, we only took into account that the fourth moments $\mathbb{E}[X^4_{k,i}]$ are bounded and that the random variables are uncorrelated, confer the inequality in (A.4). Hence, the permutation test $\Psi^{RET}_{n,Perm}$ is a non-parametric asymptotic test if the random variables $X_{k,i}$ are independent and the fourth moment exists.

We end this subsection with a corollary proving that the power of the permutation test approaches one if n becomes large.

Corollary 6.5. *The asymptotic power of the permutation test $\Psi^{RET}_{n,Perm}$ is one.*

Proof. As mentioned before, the test statistic of the permutation test corresponds to the test statistic of the Wald-type test with the sample variance estimator

$$T^{RET}_{n,Perm}(\mathbf{X_n}) = \sqrt{n}\frac{\hat{\eta}}{\sqrt{\hat{\sigma}^2_{RET,SV}}}.$$

The maximum-likelihood estimator $\hat{\eta}$ converges in probability to the true parameter η. In case that the true parameter is part of the alternative, $\eta > 0$ holds. Since the variance estimator $\hat{\sigma}^2_{RET,SV}$ is a consistent estimator for σ^2_{RET}, the test statistic $T^{RET}_{n,Perm}$ converges to infinity if the true parameter is located in the alternative. Since additionally the quantile for the permutation test converges to the quantile of a standard normal distribution, the permutation test has asymptotic power one. □

6.3 Sample Size Formula and Optimal Sample Size Allocation

When planning the sample size of a trial, one often fixes a parameter $\zeta_{H_1^{RET}}$ in the alternative H_1^{RET} and determines the sample size such that the rate of not detecting the corresponding effect $\eta_{H_1^{RET}}$ is less than a prespecified parameter $\beta \in (0,1)$. In other words, the sample size is calculated such that the trial has at least a power of $1-\beta$ for the parameter $\zeta_{H_1^{RET}}$. The choice of the parameter $\zeta_{H_1^{RET}}$ depends for instance on the assumed effect $\eta_{H_1^{RET}}$ as well as on the assumed variances in the different groups. The assumed sizes of the variances might be informed by estimates from other studies or historical values. However, determining the parameter $\zeta_{H_1^{RET}}$ is mostly difficult especially if no information about the variance or the shape parameter are available. We discuss the difficulties of sample size planning and other possible solutions such as adaptive designs in Section 7.

In this subsection, we study the concept of determining the sample size for a given parameter vector in the alternative and how to allocate this sample size between the different groups.

6.3.1 Sample Size Formula

In the following, we calculate formulas to approximate the sample size $n_{1-\beta}$ for which the different Wald-type tests and the permutation test have a power of $1-\beta$ for a given parameter vector $\zeta_{H_1^{RET}}$ located in the alternative of the retention of effect hypothesis. When establishing the different sample size formulas, we distinguish the Wald-type test with the restricted maximum-likelihood variance estimator from the other Wald-type tests as well as the permutation test, since the restricted maximum-likelihood variance estimator is not consistent under the alternative.

First of all, we establish an approximative sample size formula for the Wald-type test $\Psi_{n,Wald}^{RET}$ with the variance estimated by the sample variance or the unrestricted maximum-likelihood estimator and for the permutation test $\Psi_{n,Perm}^{RET}$. Let Ψ_n^{RET} be one of the last mentioned tests and $\zeta_{H_1^{RET}}$ the true parameter. By means of the asymptotic normality of the maximum-likelihood estimator $\hat{\eta}$, it follows that the asymptotic normality

$$\sqrt{n}\frac{\hat{\eta} - \eta_{H_1^{RET}}}{\sqrt{\hat{\sigma}^2_{RET}}} \xrightarrow{n \to \infty} \mathcal{N}(0,1)$$

holds with $\hat{\sigma}^2_{RET}$ the corresponding consistent variance estimator. Hence, we approximate the power of the test Ψ_n^{RET} by

$$\begin{aligned} &\mathbb{E}_{\zeta_{H_1^{RET}}}\left[\Psi_n^{RET}(\mathbf{X_n})\right] \approx \mathbb{P}_{\zeta_{H_1^{RET}}}\left(T_n^{RET}(\mathbf{X_n}) \geq q_{1-\alpha}\right) \\ =&\mathbb{P}_{\zeta_{H_1^{RET}}}\left(\sqrt{n}\frac{\hat{\eta} - \eta_{H_1^{RET}}}{\sqrt{\hat{\sigma}^2_{RET}}} \geq q_{1-\alpha} - \sqrt{n}\frac{\eta_{H_1^{RET}}}{\sqrt{\hat{\sigma}^2_{RET}}}\right) \approx \Phi\left(\sqrt{n}\frac{\eta_{H_1^{RET}}}{\sqrt{\sigma^2_{RET}}} - q_{1-\alpha}\right) \end{aligned} \tag{6.4}$$

with $\Phi(\cdot)$ the cumulative distribution function of the standard normal distribution. For the Wald-type tests, equality holds for the first approximation. However, in case of the permutation test, we approximated the quantile for the rejection area by the quantile of a standard normal distribution, since the quantile of the conditional permutation distribution converges to the quantile of a standard normal distribution, confer Theorem 6.3.

For the sample size approximation $n_{1-\beta}$, we obtain the formula

$$n_{1-\beta}(\zeta_{H_1^{RET}}) = (q_{1-\alpha} + q_{1-\beta})^2 \frac{\sigma^2_{RET}}{\eta^2_{H_1^{RET}}}. \tag{6.5}$$

As mentioned above, for the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator $\hat{\sigma}^2_{RET,RML}$ the sample size formula (6.5) is not appropriate because the restricted variance estimator is not consistent for the variance σ^2_{RET} if the true parameter is located in the alternative H_1^{RET}. However, with $\sigma^2_{RET,RML}$ denoting the limit of the restricted maximum-likelihood variance estimator whose calculation will be discussed later on, the asymptotic normality

$$\sqrt{n}\frac{\hat{\eta}-\eta_{H_1^{RET}}}{\sqrt{\hat{\sigma}^2_{RET,RML}}}\xrightarrow[n\to\infty]{\mathcal{D}}\mathcal{N}\left(0,\frac{\sigma^2_{RET}}{\sigma^2_{RET,RML}}\right)$$

holds. Analogously to the derivation of the approximation of the power function for the other hypothesis tests, we approximate the power of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator by

$$\mathbb{E}_{\zeta_{H_1^{RET}}}\left[\Psi^{RET}_{n,Wald}(\mathbf{X_n})\right]\approx\Phi\left(\sqrt{n}\frac{\eta_{H_1^{RET}}}{\sqrt{\sigma^2_{RET}}}-q_{1-\alpha}\frac{\sqrt{\sigma^2_{RET,RML}}}{\sqrt{\sigma^2_{RET}}}\right).\tag{6.6}$$

Hence, in case of a restricted variance estimator, we approximate the sample size $n_{1-\beta}$ for the Wald-type test $\Psi^{RET}_{n,Wald}$ by

$$n_{1-\beta}(\zeta_{H_1^{RET}})=\left(q_{1-\alpha}\frac{\sqrt{\sigma^2_{RET,RML}}}{\sqrt{\sigma^2_{RET}}}+q_{1-\beta}\right)^2\frac{\sigma^2_{RET}}{\eta^2_{H_1^{RET}}}.\tag{6.7}$$

It remains to calculate the limit $\sigma^2_{RET,RML}$ of the restricted maximum-likelihood variance estimator $\hat{\sigma}^2_{RET,RML}$ if the true parameter is located in the alternative. Thereto, we calculate the limit of the restricted maximum-likelihood estimators for the rates and the shape parameter. In Theorem 3.5, we stated conditions such that these estimators converge almost surely against the parameter which minimizes the Kullback-Leiber divergence defined in (3.1). In the following, we discuss these conditions under our model but primarily, we introduce some notations. Let $\Theta_{\partial H_0}$ be the parameter space Θ restricted to the boundary ∂H_0^{RET} of the retention of effect hypothesis, i.e.

$$\Theta_{\partial H_0}:=\Big\{(\lambda_E,\lambda_R,\lambda_P,\phi)\in\Theta|\lambda_E-\Delta\lambda_R+(\Delta-1)\lambda_P=0\Big\}.$$

Moreover, Θ_{H_1} denotes the parameter space of the alternative H_1^{RET} and we use the notations $\zeta\in\Theta_{\partial H_0}$ as well as $\zeta_{H_1^{RET}}=(\lambda_{E,1},\lambda_{R,1},\lambda_{P,1},\phi_1)\in\Theta_{H_1}$. The first condition from Theorem 3.5 claims that the parameter ζ_{RML} minimizing the Kullback-Leibler divergence $K(\zeta_{H_1^{RET}},\zeta,w)$ with respect to ζ is well defined. Under the model from Section 2.3, the

Kullback-Leibler divergence $K(\zeta_{H_1^{RET}}, \zeta, w)$, confer (3.1), is given by

$$
\begin{aligned}
&\sum_{k=E,R,P} w_k \mathbb{E}_{(\lambda_{k,1},\phi_1)} \left[\log\left(\mathbb{P}_{(\lambda_{k,1},\phi_1)}(X=\cdot)\right) - \log\left(\mathbb{P}_{(\lambda_k,\phi)}(X=\cdot)\right)\right] \\
&= \sum_{k=E,R,P} w_k \mathbb{E}_{(\lambda_{k,1},\phi_1)} \left[\log\left(\frac{\Gamma\left(X+\frac{1}{\phi_1}\right)}{\Gamma\left(X+\frac{1}{\phi}\right)}\right) - \log\left(\frac{\Gamma\left(\frac{1}{\phi_1}\right)}{\Gamma\left(\frac{1}{\phi}\right)}\right) + X\log\left(\frac{\lambda_{k,1}\phi_1}{\lambda\phi}\right)\right. \\
&\qquad\qquad \left. + \left(X+\frac{1}{\phi}\right)\log\left(1+\lambda\phi\right) - \left(X+\frac{1}{\phi_1}\right)\log\left(1+\lambda_{k,1}\phi_1\right)\right] \\
&= \sum_{k=E,R,P} w_k \left(\mathbb{E}_{(\lambda_{k,1},\phi_1)}\left[\log\left(\frac{\Gamma\left(X+\frac{1}{\phi_1}\right)}{\Gamma\left(X+\frac{1}{\phi}\right)}\right)\right] - \log\left(\frac{\Gamma\left(\frac{1}{\phi_1}\right)}{\Gamma\left(\frac{1}{\phi}\right)}\right) + \lambda_{k,1}\log\left(\frac{\lambda_{k,1}\phi_1}{\lambda\phi}\right)\right. \\
&\qquad\qquad \left. + \left(\lambda_{k,1}+\frac{1}{\phi}\right)\log\left(1+\lambda\phi\right) - \left(\lambda_{k,1}+\frac{1}{\phi_1}\right)\log\left(1+\lambda_{k,1}\phi_1\right)\right).
\end{aligned}
$$

However, no closed form expression exists for the remaining expectation and thus it has to be approximated numerically. Moreover, we were not able to prove that the Kullback-Leibler divergence $K(\zeta_{H_1^{RET}}, \zeta, w)$ has a (unique) global minimum with respect to ζ. It should be noted that the divergence is neither convex nor quasiconvex which follows from choosing the sample size allocation $w = (1/3, 1/3, 1/3)$, the parameter in the alternative $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.71, 2)$, and the margin $\Delta = 43/55$ to determine the hypothesis. Both inequalities do not hold for $\zeta_1 = (4.4, 1, 16.58333, 8.2)$, $\zeta_2 = (6.6, 7.9, 1.941667, 8.1)$ and $t = 0.51$. When calculating the Kullback-Leibler divergence, we applied the approximation

$$
\mathbb{E}_{(\lambda_{k,1},\phi_1)}\left[\log\left(\frac{\Gamma\left(X+\frac{1}{\phi_1}\right)}{\Gamma\left(X+\frac{1}{\phi}\right)}\right)\right] \approx \sum_{x=0}^{10{,}000} \log\left(\frac{\Gamma\left(X+\frac{1}{\phi_1}\right)}{\Gamma\left(X+\frac{1}{\phi}\right)}\right)\mathbb{P}_{(\lambda_{k,1},\phi_1)}(X=x).
$$

To approximate the expectation, we only take the values up to $x = 10{,}000$ into account because for the cases considered with reasonable $\lambda_{k,1}$, i.e. $\lambda_{k,1} \leq 50$, considering more terms of the sum does not change in the sense of computational accuracy. Concerning the existence and the uniqueness of the minimizer ζ_{RML}, for the cases considered, it existed and has been unique on a reasonable choice of the parameter space.

Additionally, we have to show that Condition 2 of the theorem is fulfilled. Thereto, any sequence of parameter vectors in the restricted space $\zeta^{(n)} \in \Theta_{\partial H_0}$ which limit is located in the closure of the parameter space but not in the parameter space itself, i.e. $\lim_{n\to\infty} \zeta^{(n)} \in \overline{\Theta}\backslash\Theta$

has a mass of zero, i.e.

$$\lim_{n\to\infty} \prod_{k=E,R,P} \mathbb{P}_{(\lambda_k^{(n)},\phi^{(n)})}(X_{k,1} = \cdot) = 0 \qquad \mathbb{P}_{\zeta_{H_1^{RET}}} - a.s.$$

However, if we regard a sequence $\zeta^{(n)}$ with limit $(\lambda_E, \lambda_R, \lambda_P, \phi) = (0, 1, 1, 1) \in \overline{\Theta}\backslash\Theta$, the probability function $\mathbb{P}_{(\lambda_E^{(n)},\phi^{(n)})}(X_{E,1} = \cdot)$ converges to one at $X_{E,1} = 0$ and the limit is not $\mathbb{P}_{\zeta_{H_1^{RET}}}$ almost surely zero. Nevertheless, the limit of the restricted maximum-likelihood estimators can be calculated by minimizing the Kullback-Leibler divergence, since in the proof of the convergence of the restricted maximum-likelihood estimators to the minimizer of the Kullback-Leibler divergence, the second condition only ensures that the minimum is located in a compact set. By restricting the parameter space to an arbitrary large but compact set this condition is still fulfilled and such a restriction has no effect in practice.

6.3.2 Optimal Sample Size Allocations

In the following, we establish different optimal sample size allocations for the statistical tests of the retention of effect hypothesis. As the optimality criteria we consider a maximal power for a fixed sample size n. Furthermore, when maximizing the power with respect to the sample size allocation, we introduce several restrictions for the allocation. These restrictions are be motivated by ethical as well as practical reasons.

Since we do not know the exact power function of the Wald-type tests and the permutation test, we consider the approximative power functions (6.4) and (6.6) to determine the effect of the sample size allocation on the power. These power functions become more accurate if the sample size n increases. Besides the sample size n, the approximative sample size formulas depend on the effect $\eta_{H_1^{RET}}$, the quantile $q_{1-\alpha}$, and especially the variance σ^2_{RET}. If the variance is estimated with the restricted maximum-likelihood variance estimator, the power approximation additionally depends on $\sigma^2_{RET,RML}$.

Maximizing the approximative power function (6.4) with respect to the sample size allocation $w = (w_E, w_R, w_P)$ corresponds to minimizing the variance $\sigma^2_{RET}(w_E, w_R, w_P)$ which is given by

$$\sigma^2_{RET}(w_E, w_R, w_P) = \frac{\sigma_E^2}{w_E} + \Delta^2\frac{\sigma_R^2}{w_R} + (1-\Delta)^2\frac{\sigma_P^2}{w_P} \tag{6.8}$$

with respect to (w_E, w_R, w_P). At least for large sample sizes, the allocation minimizing (6.8) also minimizes the approximation (6.6) because the term

$$q_{1-\alpha}\frac{\sqrt{\sigma^2_{RET,RML}(w_E, w_R, w_P)}}{\sqrt{\sigma^2_{RET}(w_E, w_R, w_P)}}$$

is negligible compared to $\sqrt{n}\eta_{H_1^{RET}}/\sqrt{\sigma^2_{RET}}$. With w_{opt} denoting the optimal sample size allocation, we obtain the minimization problem

$$\begin{aligned} w_{opt} := \arg\min \quad & \sigma^2_{RET}(w_E, w_R, w_P) \\ \text{s.t.} \quad & w_E + w_R + w_P = 1 \\ & (w_E, w_R, w_P) \in (0,1)^3. \end{aligned}$$

According to Section 4.1.1 in Mielke (2010), the optimal allocation w_{opt} is given by

$$w_{opt} = \left(\frac{\sigma_E}{\sigma_E + \Delta\sigma_R + |1-\Delta|\sigma_P}, \frac{\Delta\sigma_R}{\sigma_E + \Delta\sigma_R + |1-\Delta|\sigma_P}, \frac{|1-\Delta|\sigma_P}{\sigma_E + \Delta\sigma_R + |1-\Delta|\sigma_P}\right).$$

The optimal sample size for the group $k = E, R, P$ depends on the (assumed) standard deviation σ_k as well as on the margin Δ and the standard deviations for the other groups. More precisely, if the standard deviation of one group increases, the optimal sample size for this group increases as well. Depending on the variances and the margin Δ, the optimal sample size allocation w_{opt} can yield a rather small sample size for one group resulting in a lack of information about this group. In particular, if the variance in the placebo group is much larger than the variances in the active treatment group and the margin Δ is not near to one, the sample size in the placebo group becomes large. Especially, the sample size in the placebo group becomes larger than the sample sizes in the active treatment groups. However, due to ethical reasons it is sometimes not feasible that the sample size in the placebo group is larger than the sample size in the treatment groups. Therefore, it is reasonable to demand that the fractions w_E and w_R are at least as large as w_P. Additionally, to avoid that too few patients are allocated to a group, we introduce a lower bound $m \in (0, 1/3]$ for the fraction w_P. Too few patients in one group are sometimes not desirable because they result in a lack of information, which may be required for other aims of the study. Considering these additional restraints, the optimal allocation $w_{opt,m}$ is the solution of the minimization problem

$$\begin{aligned} w_{opt,m} := \arg\min \quad & \sigma^2_{RET}(w_E, w_R, w_P) \\ \text{s.t.} \quad & f_1(w_E, w_R, w_P) := w_P - w_E \le 0 \\ & f_2(w_E, w_R, w_P) := w_P - w_R \le 0 \\ & f_3(w_E, w_R, w_P) := m - w_P \le 0 \\ & h(w_E, w_R, w_P) := w_E + w_R + w_P - 1 = 0 \\ & (w_E, w_R, w_P) \in (0,1)^3. \end{aligned} \tag{6.9}$$

Theorem 6.6, whose proof is stated in Appendix A, reveals that $w_{opt,m}$ exists and is unique.

Theorem 6.6. *The minimization problem* (6.9) *has a unique solution which can be calculated by the Karush-Kuhn-Tucker (KKT) conditions and will be stated in the proof.*

In addition to the previously mentioned restraints for the sample size allocation, often the sample sizes in the active treatment groups shall be equal, i.e. $w_E = w_R$. Confer Section 3.3 in Pigeot et al. (2003) who calculated the optimal sample size in a three-arm trials with normally distributed endpoints and equal variances for the different groups. In this case, the optimal sample size allocation is the solution of the minimisation problem (6.10).

$$\begin{aligned} w_{opt,E=R} := \arg\min \quad & \sigma^2_{RET}(w_E, w_E, w_P) \\ \text{s.t.} \quad & f_1(w_E, w_P) := w_P - w_E \le 0 \\ & f_2(w_E, w_P) := m - w_P \le 0 \\ & h(w_E, w_P) := 2w_E + w_P - 1 = 0 \\ & w_E, w_P \in (0,1) \end{aligned} \tag{6.10}$$

Theorem 6.7. *The optimization problem* (6.10) *has a unique solution which will be calculated in the proof.*

Analogously to the proof of Theorem 6.6, the proof of Theorem 6.7 is stated Appendix A. So far, we stated three different optimal sample size allocations which all depend on a prespecified parameter vector $\zeta_{H_1^{RET}}$ located in the alternative. However, if the sample size is planned for more than one alternative or no certain alternative is specified, (Mielke, 2010, Section 4.1.3) recommends the use of the rule of thumb $n_E : n_R : n_P = 1 : \Delta : (1 - \Delta)$ which corresponds to $w_{rot} = (1/2, \Delta/2, (1 - \Delta)/2)$. The rule of thumb is motivated by resulting in a smaller variance σ^2_{RET} than the allocations 1:1:1 and 2:2:1 for certain ratios σ^2_P/σ^2_T, see Theorems 3 and 4 in Mielke (2010). Theoretical comparison of the introduced optimal sample size allocations is difficult because there is not one explicit expression for the solution $w_{opt,m}$. However, when defining the parameter setting for the simulation of the power in the next subsection, we compare the different optimal sample size allocations for several examples.

6.4 Simulation Study

In the forgoing subsections, we described the asymptotic theory of the different Wald-type tests $\Psi^{RET}_{n,Wald}$ and the permutation test $\Psi^{RET}_{n,Perm}$ for the retention of effect hypothesis. In this subsection, we study the finite sample size properties of the tests. Therefore, we simulate the actual level of significance as well as the power of the different statistical tests for the retention of effect hypothesis. Afterwards, for scenarios where multiple tests can be applied, we additionally compare their power. Analogously to the simulations in Section 5, we define the parameter vector $(\lambda_E, \lambda_R, \lambda_P, \phi, n_E, n_R, n_P)$ by taking the examples from Section 2 into account. Moreover, we discuss the optimal sample size allocation for these parameter settings.

As mentioned before, we motivate the choice of the rates, the shape parameter and the sample sizes by the examples from Sections 2.2.1 and 2.2.2. The non-inferiority margin is defined such that the parameters for the simulation of the actual level are located at the boundary of the hypothesis. As well, we apply the same sample size allocations as for the simulations in Section 5.3 and the rule of thumb $1 : \Delta : (1 - \Delta)$. Additionally, we allocate the sample size by the optimal allocations calculated in the last subsection. To this, we specify the variances σ^2_E, σ^2_R, and σ^2_P by determining the rates and the shape parameter. Thereto, choosing a parameter vector $\zeta_{H_0^{RET}} = (\lambda_E, \lambda_R, \lambda_P, \phi)$ which is located in the hypothesis H_0^{RET} is not reasonable because the optimal allocations aim to maximize the power. Thus, for a given parameter vector $\zeta_{H_0^{RET}}$ used to simulate the actual level, we specify a parameter vector $\zeta_{H_1^{RET}} = (\lambda_{E,1} = \lambda_R, \lambda_R, \lambda_P, \phi)$ located in the alternative. More precisely, the vector $\zeta_{H_1^{RET}}$ is defined with the parameters λ_R, λ_P, and ϕ as in the vector $\zeta_{H_0^{RET}}$ and the rate of the experimental treatment group $\lambda_{E,1}$ is equal to the rate λ_R. In other words, for parameters located in the alternative we determine that the experimental and the reference treatment are equally effective. When simulating the power for the parameter vector $\zeta_{H_1^{RET}}$, we also use this vector to calculate the optimal sample size allocation. A disadvantage of this approach is that eventually we simulate the power for the optimal sample size allocations which has been calculated with the true parameter, i.e. when comparing the power of the tests for different sample size allocations some might take advantage of information which are not accessible in practice. As mentioned before, we discuss the problem of unknown nuisance parameters in planning the sample size allocation of a trial in Section 7.

In the following, we determine the parameter vectors $\zeta_{H_0^{RET}}$ and $\zeta_{H_1^{RET}}$, the non-inferiority margin Δ and the sample sizes by means of the examples from Sections 2.2.1 and 2.2.2 and calculate the optimal sample size allocations. Motivated by the results of Calverley et al. (2003) stated in Table 1, we choose the rates located in the hypothesis H_0^{RET} to be $\lambda_E = 1.28$, $\lambda_R = 1.16$, and $\lambda_P = 1.71$. Hence, we obtain the vectors $\zeta_{H_0^{RET}} = (1.28, 1.16, 1.17, \phi)$ and $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.17, \phi)$. The margin Δ is given by $\Delta = (\lambda_P - \lambda_E)/(\lambda_P - \lambda_R) = 43/55$. As in Definition 5.7, the shape parameters and the sample sizes are determined as $\phi =$ 0.3, 0.5, 0.7 and $n =$ 550, 1100, 2200, respectively. Table 4 lists the optimal sample size allocations w_{opt}, $w_{opt,m}$, and $w_{opt,E=R}$. For the calculation of the optimal allocations $w_{opt,m}$ and $w_{opt,E=R}$, we choose the lower boundary m of the sample size for the placebo group to be $m = 10\%$. The choice of m is arbitrary.

Table 4: Different optimal sample size allocations for the tests of the retention of effect hypothesis calculated with the parameter $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.17, \phi)$ and the non-inferiority margin $\Delta = 43/55$.

Shape parameter ϕ	$w_{opt} \equiv w_{opt,m}$	$w_{opt,E=R}$
$\phi = 0.3$	$(0.4849, 0.3791, 0.1360)$	$(0.4324, 0.4324, 0.1352)$
$\phi = 0.5$	$(0.4834, 0.3779, 0.1387)$	$(0.4311, 0.4311, 0.1378)$
$\phi = 0.7$	$(0.4823, 0.3771, 0.1407)$	$(0.4301, 0.4301, 0.1398)$

Table 4 shows that for the current setting the optimal sample size allocation w_{opt} and $w_{opt,m}$ are equal. Furthermore, the effect of the shape parameter on the sample size allocation is negligible. The sample size proportion of the placebo group is nearly the same for all allocations. However, the sample size proportions for the active groups differ between the optimal allocations in about 5%. The rule of thumb $1 : \Delta : (1-\Delta)$ results in the sample size allocation $(0.5, 0.3909, 0.1091)$ which differs only slightly from the sample size allocations w_{opt}. Hence, the optimal allocations as well as the rule of thumb allocated the sample size unbalanced and such that the experimental treatment group has the largest sample size and the placebo group the smallest. The next definition summarizes the choices of the parameters.

Definition 6.8 (Parameter setting motivated by TRISTAN study in COPD)**.**

$$\begin{aligned}
\zeta_{H_0^{RET}} &= (1.28, 1.16, 1.17, \phi),\\
\zeta_{H_1^{RET}} &= (1.16, 1.16, 1.17, \phi),\\
\Delta &= 43/55,\\
\phi &\in \{0.3, 0.5, 0.7\},\\
n &\in \{550, 1100, 2200\},\\
n_E : n_R : n_P &\in \{1{:}1{:}1,\ 2{:}1{:}1,\ 2{:}2{:}1,\ 3{:}2{:}1, 1 : \Delta : (1-\Delta), w_{opt}, w_{opt,E=R}\}.
\end{aligned}$$

Below, we define the parameter setting motivated by the CONFIRM study Fox et al. (2012) in MS discussed in Section 2.2.2. With respect to the mean number of lesions in Table 2, we define the rates $\lambda_E = 8$, $\lambda_R = 5.1$, and $\lambda_P = 17.4$. As before, the non-inferiority margin Δ is defined such that the parameters are located at the boundary of the hypothesis, i.e. $\Delta = 94/123$. The shape parameters and the sample sizes are defined as in Definition 5.8. Table 5 states the different optimal sample size allocations.

Table 5: Different optimal sample size allocations for the tests of the retention of effect hypothesis calculated with the parameter $\zeta_{H_1^{RET}} = (5.1, 5.1, 17.4, \phi)$ and the non-inferiority margin $\Delta = 94/123$.

ϕ	w_{opt}	$w_{opt,m}$	$w_{opt,E=R}$
$\phi = 1$	$(0.3967, 0.3032, 0.3001)$	$(0.3967, 0.3032, 0.3001)$	$(0.3509, 0.3509, 0.2982)$
$\phi = 2$	$(0.3933, 0.3005, 0.3062)$	$(0.3933, 0.3034, 0.3034)$	$(0.3478, 0.3478, 0.3043)$
$\phi = 3$	$(0.3920, 0.2996, 0.3084)$	$(0.3920, 0.3040, 0.3040)$	$(0.3467, 0.3467, 0.3065)$

For the sample size allocations stated in Table 5, the shape parameter effects the allocation only slightly. Moreover, the differences between the sample size allocations w_{opt} and $w_{opt,m}$ are negligible. The sample size proportion of the placebo group are almost equal for the different allocations. The proportions for the active treatment groups differ in about 5% between the allocations w_{opt} and $w_{opt,m}$ on the one side and $w_{opt,E=R}$ on the other side. The rule of thumb sample size allocation is given by $w_{rot} = (0.5, 0.3821, 0.1179)$, i.e. the allocation w_{rot} is clearly different from the optimal allocations and more unbalanced. Compared to the optimal allocations stated in Table 4, the optimal allocation from Table 5 are much more balanced. In particular, the allocation $w_{opt,E=R}$ only differs slightly from the allocation 1:1:1.

The next definition summarizes the parameter setting.

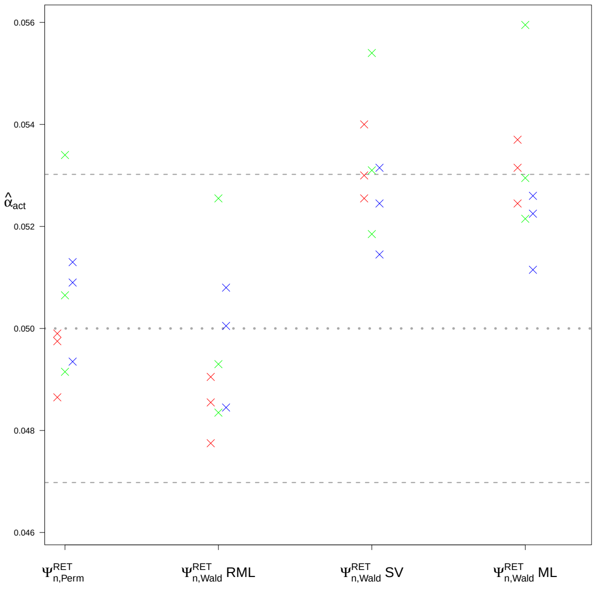
Definition 6.9 (Parameter setting motivated by CONFIRM study in MS)**.**

$$\begin{aligned}
\zeta_{H_0^{RET}} &= (8, 5.1, 17.4, \phi),\\
\zeta_{H_1^{RET}} &= (5.1, 5.1, 17.4, \phi),\\
\Delta &= 94/123,\\
\phi &\in \{1, 2, 3\},\\
n &\in \{215, 430, 860\},\\
n_E : n_R : n_P &\in \{1{:}1{:}1,\ 2{:}1{:}1,\ 2{:}2{:}1,\ 3{:}2{:}1, 1 : \Delta : (1-\Delta), w_{opt}, w_{opt,E=R}\}.
\end{aligned}$$

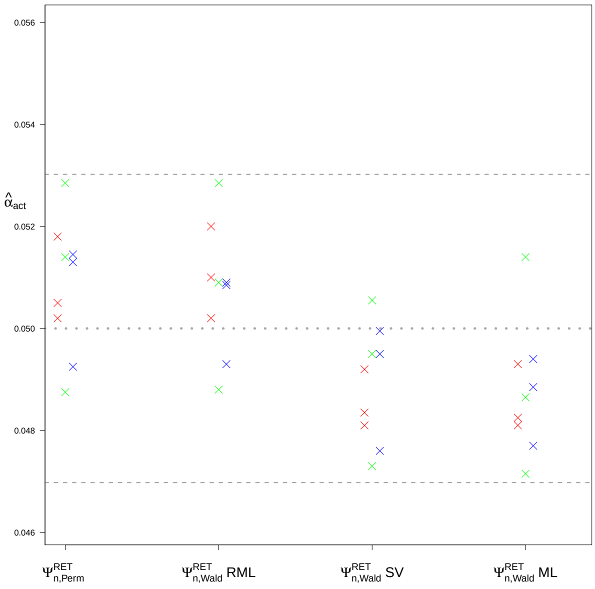
After defining two parameter settings for the Monte-Carlo simulations, we study the actual level and the power of the different tests.

6.4.1 Actual Level

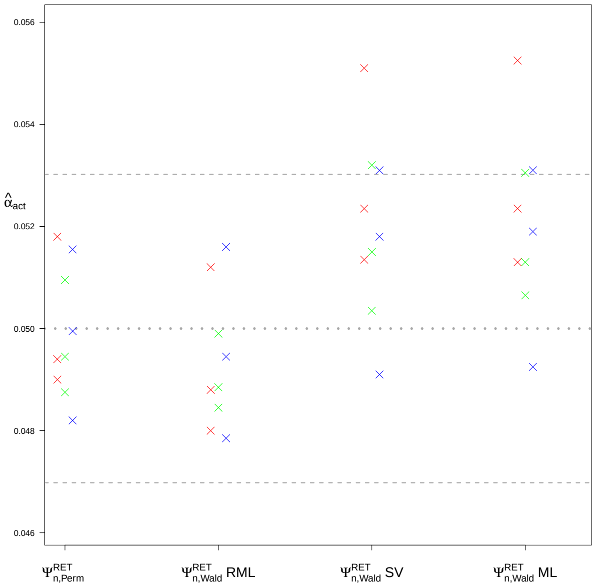
Our aim below is to study the actual level of significance α_{act} of the Wald-type tests $\Psi_{n,Wald}^{RET}$ with the variance estimated by the sample variance or a maximum-likelihood estimator as well as the actual level of the permutation test $\Psi_{n,Perm}^{RET}$. The first two figures show the actual levels for the parameters stated in Definition 6.8. However, the results for the rule of thumb allocation w_{rot} are not shown, since they are qualitatively similar to the results for the sample size allocation 3:2:1. As before, the results base on M = 20,000 Monte-Carlo simulations and the quantile, which determines the rejection area of the permutation test, relies on 20,000 random permutation. The Wald-type tests and the permutation test are constructed with a level of significance $\alpha = 0.05$. Hence, with M = 20,000 it follows that a two-sided level 0.05 test rejects the hypothesis that the actual level α_{act} is equal to $\alpha = 0.05$ if the simulated actual level $\hat{\alpha}_{act}$ is not contained in [0.04698, 0.05302]. In the corresponding figures, the boundaries of the interval are shown as dashed grey lines. As argued in Section 5.3, a Wald-type or permutation test for the retention of effect hypothesis is only appropriate for usage if it is not liberal. Last but not least, we only study the actual levels of the tests constructed with $\alpha = 0.05$ but possible effects of α could be part of further research.



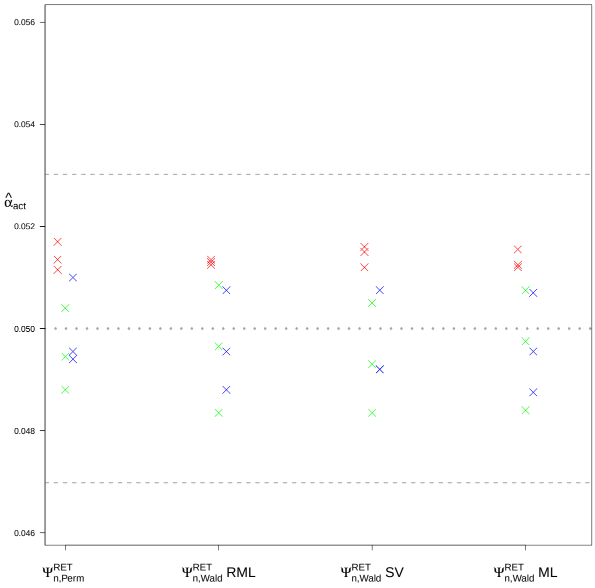
(a) Sample size allocation 1:1:1.



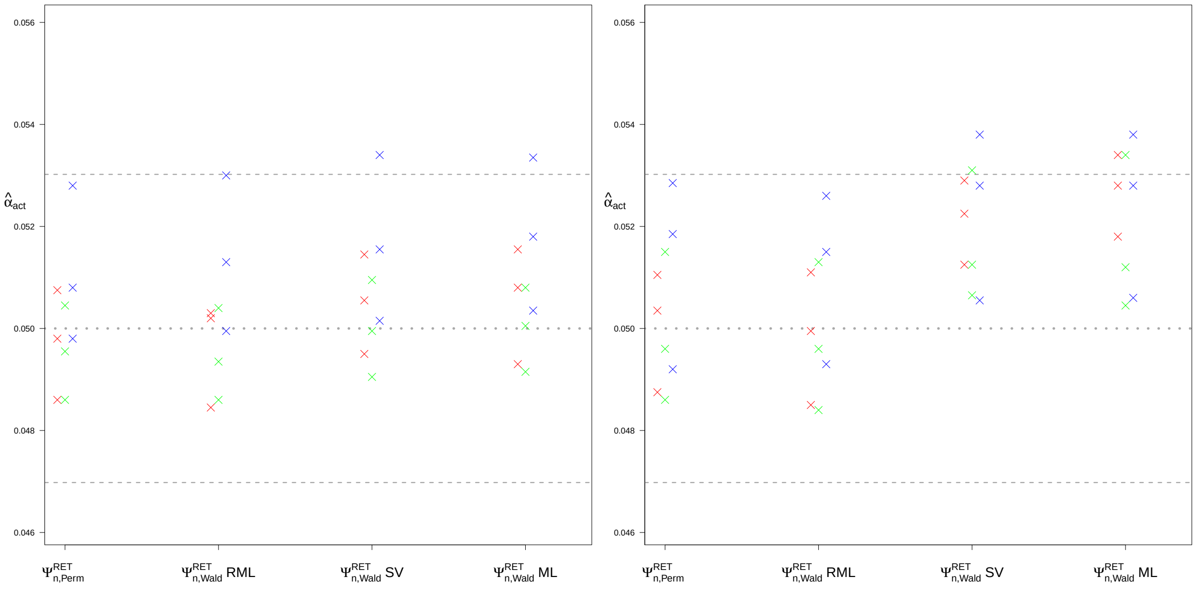
(b) Sample size allocation 2:1:1.



(c) Sample size allocation 2:2:1.



(d) Sample size allocation 3:2:1.



(e) Sample size allocation $w_{opt,m}$.

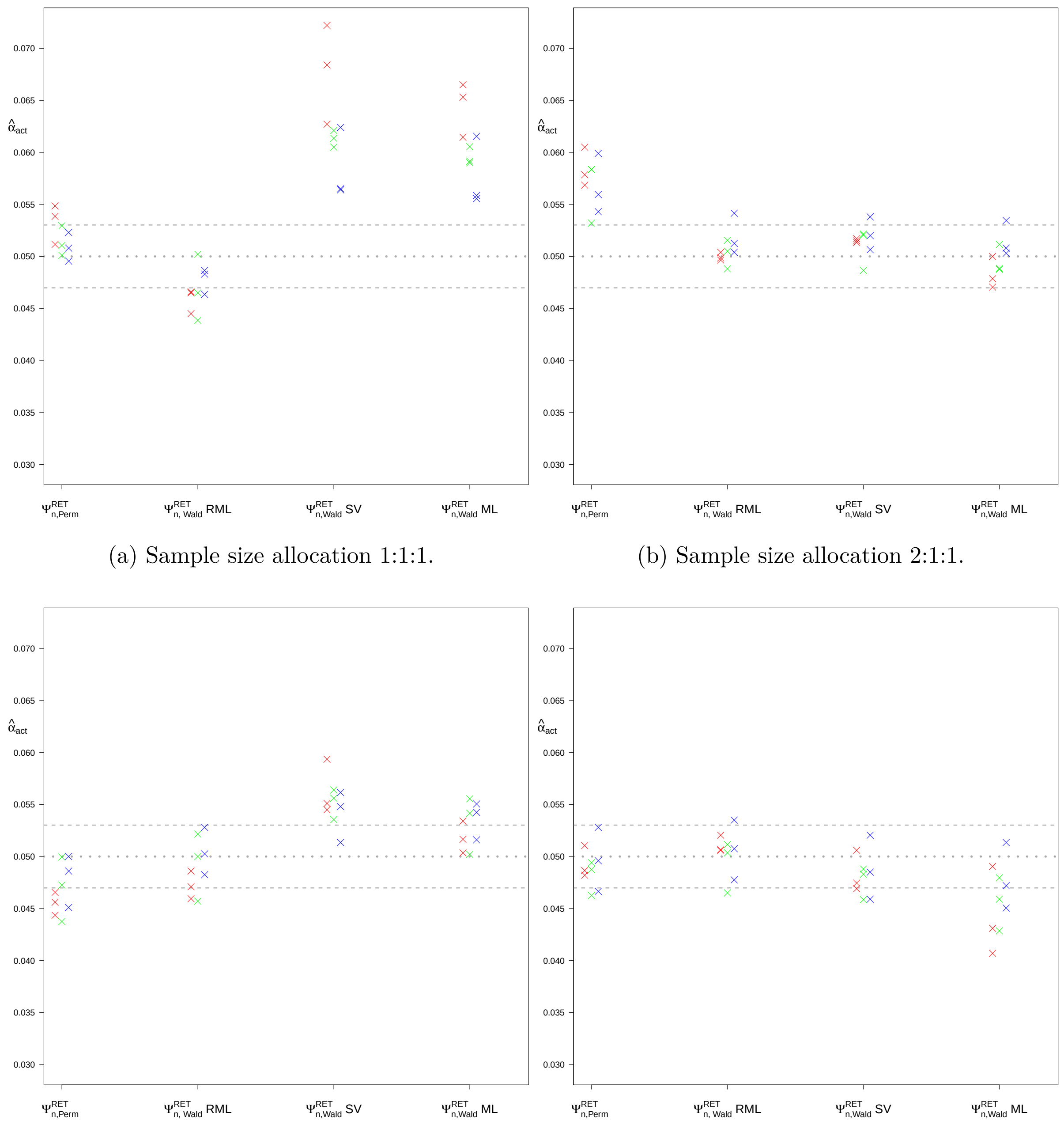
(f) Sample size allocation $w_{opt,E=R}$.

Figure 5: Actual level of different tests for H_0^{RET} by sample size allocation for the parameters from Definition 6.8. The points for a sample size of 550 are red, for one of 1100 are green, and for one of 2200 are blue. Actual levels between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations *ML*, *SV*, *RML* indicate the variance estimator of the Wald-type tests. The shape parameters are not distinguished.

Figure 5 shows that the permutation test $\Psi^{RET}_{n,Perm}$ and the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator perform well, i.e. the actual levels of these tests are only in a few cases not between the dashed grey lines, and are therefore recommended for application. Moreover, the sample size and its allocation does not affect the actual level significantly. In contrast, the performance of the Wald-type tests $\Psi^{RET}_{n,Wald}$ with the sample variance estimator or the unrestricted maximum-likelihood variance estimator depends on the sample size allocation. For the allocations 1:1:1, 2:2:1, and $w_{opt,E=R}$ these tests tend to be liberal, i.e. they are not appropriate to test the retention of effect hypothesis H_0^{RET}. However, for the allocation 2:1:1, 3:2:1, and $w_{opt,m}$ the Wald-type tests $\Psi^{RET}_{n,Wald}$ with the sample variance and the unrestricted maximum-likelihood estimator to not tend to be liberal and conservative and are suitable tests for H_0^{RET} under the given setting. In case that multiple tests are appropriate for a given sample size allocation, we additionally compare their power later on in this section.

Next, we study the actual level of the different tests for the parameters from Definition 6.9.

The results of the optimal allocation $w_{opt,m}$ are not shown because they are qualitatively the same as for the allocation 1:1:1 which is what we expected, since the sample size allocations do not differ much.

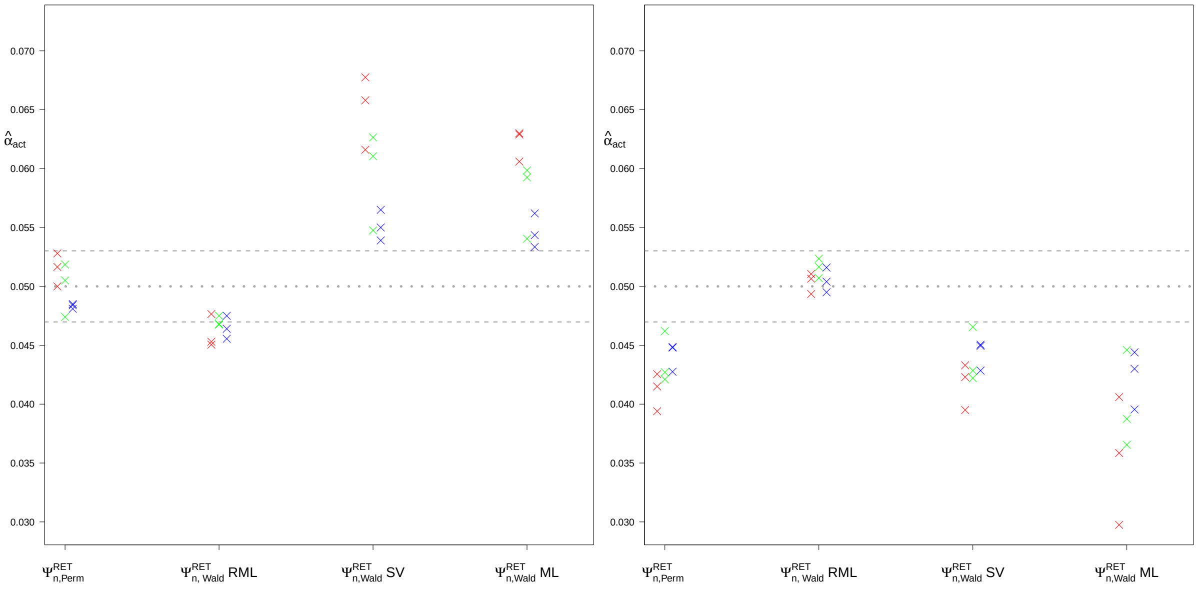


(a) Sample size allocation 1:1:1.

(b) Sample size allocation 2:1:1.

(c) Sample size allocation 2:2:1.

(d) Sample size allocation 3:2:1.



(e) Sample size allocation $w_{opt,E=R}$.

(f) Sample size allocation $1 : \Delta : (1 - \Delta)$.

Figure 6: Actual level of different tests for H_0^{RET} by sample size allocation for the parameters from Definition 6.9. The points for a sample size of 215 are red, for one of 430 are green, and for one of 860 are blue. Actual levels between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations ML, SV, RML indicate the variance estimator of the Wald-type tests. The shape parameters are not distinguished.

Altogether, the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator performs best and can be recommended for application. The influence of both sample size allocation and sample size on the actual level is rather small. For the sample size allocation 1:1:1 the tends to be slightly conservative and for the other sample size allocations, there is no trend detectable. In the following, we analyse for each sample size allocation which tests are applicable besides the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator.

For the sample size allocation 1:1:1 applying the permutation test is only appropriate if the sample size is large enough. The remaining Wald-type tests $\Psi^{RET}_{n,Wald}$ are liberal and thus not appropriate to test the retention of effect hypothesis in this setting. The permutation test $\Psi^{RET}_{n,Perm}$ is the only test which cannot recommended for application for the sample size allocation 2:1:1. If the sample size is allocated according to 2:2:1, the permutation test $\Psi^{RET}_{n,Perm}$ can be applied but is slightly conservative in contrast to the Wald-type test $\Psi^{RET}_{n,Wald}$ with a unrestricted maximum-likelihood or a sample variance estimator which are liberal and should therefore not be applied. For the allocation 3:2:1, all tests are appropriate for usage but the Wald-type test $\Psi_{n,Wald}^{RET}$ with the unrestricted maximum-likelihood estimator tends to be conservative for smaller sample sizes. For the sample size allocation $w_{opt,E=R}$, the permutation tests can be applied and it is in contrast to the Wald-type test $\Psi_{n,Perm}^{RET}$ with the restricted maximum-likelihood variance estimator not conservative. For the sample sizes allocated to the rule of thumb, the remaining tests can be applied but they are conservative. To conclude, besides the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator, the actual levels of the different Wald-type tests and the permutation test are influence by the sample size allocation. Thereto, the influence on the actual levels is much stronger for the parameters from Definition 6.9 than for the parameters from Definition 6.8. Whether this is due to the sample sizes, rates or shape parameters cannot be ascertained with the present Monte-Carlo simulations. Moreover, the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator performed well for all considered sample size allocations if the parameters were defined as in Definition 6.8. However, this does not hold for the other parameter setting. In total, the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator is stable concerning the sample size allocation. Besides, the other tests, i.e. the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type tests $\Psi_{n,Wald}^{RET}$ with the sample variance or the unrestricted maximum-likelihood variance estimator, are clearly liberal in same cases. Finally, it should be mentioned that the number of simulations of the discussed Monte-Carlo results is not sufficiently large enough to detect the influence of the sample size and especially the shape parameter for a fixed rate. An increase of the number of simulations is not feasible due to the computing time of the permutation test. However, at least for the Wald-type tests Monte-Carlo simulations with 100,000 replicates are possible. The results show that the actual levels approach $\alpha = 0.05$ if the shape parameter ϕ decreases or the sample size n increases what was to be expected.

6.4.2 Power

After studying the actual levels of the Wald-type tests $\Psi_{n,Wald}^{RET}$ and the permutation test $\Psi_{n,Perm}^{RET}$, we compare the power of these tests and study how well the power approximations (6.4) and (6.6) fit the actual power. In the following, we only consider the tests which are not liberal for the corresponding sample size allocations because the other tests are not appropriate.

If the approximative power functions from Equations (6.4) and (6.6) are larger than the power itself, the corresponding sample size formulas from Equations (6.5) and (6.7) result in a too small sample size and are therefore not recommended for application. However, if the approximations are smaller than the power, the sample size formulas yield sample sizes resulting in a larger power which is also not desirable but at least the power is not smaller than planned. Firstly, we study the power of the different tests for a parameter setting motivated by the TRISTAN study. As stated in Definition 6.8, we choose the parameter vector $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.71, \phi)$ as well as the non-inferiority margin $\Delta = 43/55$. In contrast to the Monte-Carlo simulations of the actual levels from Section 6.4.1, we only regard the results for the shape parameter $\phi = 0.5$, since they are qualitatively the same for the shape parameters $\phi = 0.3$ and $\phi = 0.7$. Regarding this, it should be mentioned that if the shape parameter ϕ increases for a fixed sample size the power decreases. For a fixed sample size allocation, the tests which are not liberal have nearly the same power with a difference less than the Monte-Carlo error. Since the results for the different sample size allocations do not differ qualitatively, we only show the results for the sample size allocation $w_{opt} \equiv w_{opt,m}$.

The approximative power function for the Wald-type test with the restricted maximum-likelihood variance estimator from (6.6) is a function in the limit $\sigma^2_{RET,RML}$ of the restricted maximum-likelihood variance estimator $\hat{\sigma}^2_{RET,RML}$. The limit $\sigma^2_{RET,RML}$ is given by

$$\frac{\lambda_{E,RML}(1+\lambda_{E,RML}\phi_{RML})}{w_E} + \frac{\lambda_{R,RML}(1+\lambda_{R,RML}\phi_{RML})}{w_R} + \frac{\lambda_{P,RML}(1+\lambda_{P,RML}\phi_{RML})}{w_P}$$

with $\zeta = (\lambda_{E,RML}, \lambda_{R,RML}, \lambda_{P,RML}, \phi_{RML})$ the minimizer of the Kullback-Leibler divergence $K(\zeta, \zeta_{H_1^{RET}}, w)$ from (3.1). The Kullback-Leibler divergence for negative binomially distributed endpoints is stated in Section 6.3.1 and can only be minimized iteratively. By doing so, we obtain $\zeta = (1.222, 1.059, 1.639, 0.503)$ and $\sigma^2_{RET,RML} = 7.893$ for the allocation $w_{opt} = (0.4834, 0.3779, 0.1387)$. In comparison, the limit of the unrestricted variance estimators is equal to $\sigma^2_{RET} = 7.845$, i.e. the variances are similar and we do not expect big differences between the approximative power functions.

As above, the results are based on $M = 20{,}000$ simulations. We simulate the quantile of the conditional permutation distribution with 20,000 random permutations and the tests are defined with the level of significance $\alpha = 0.05$.

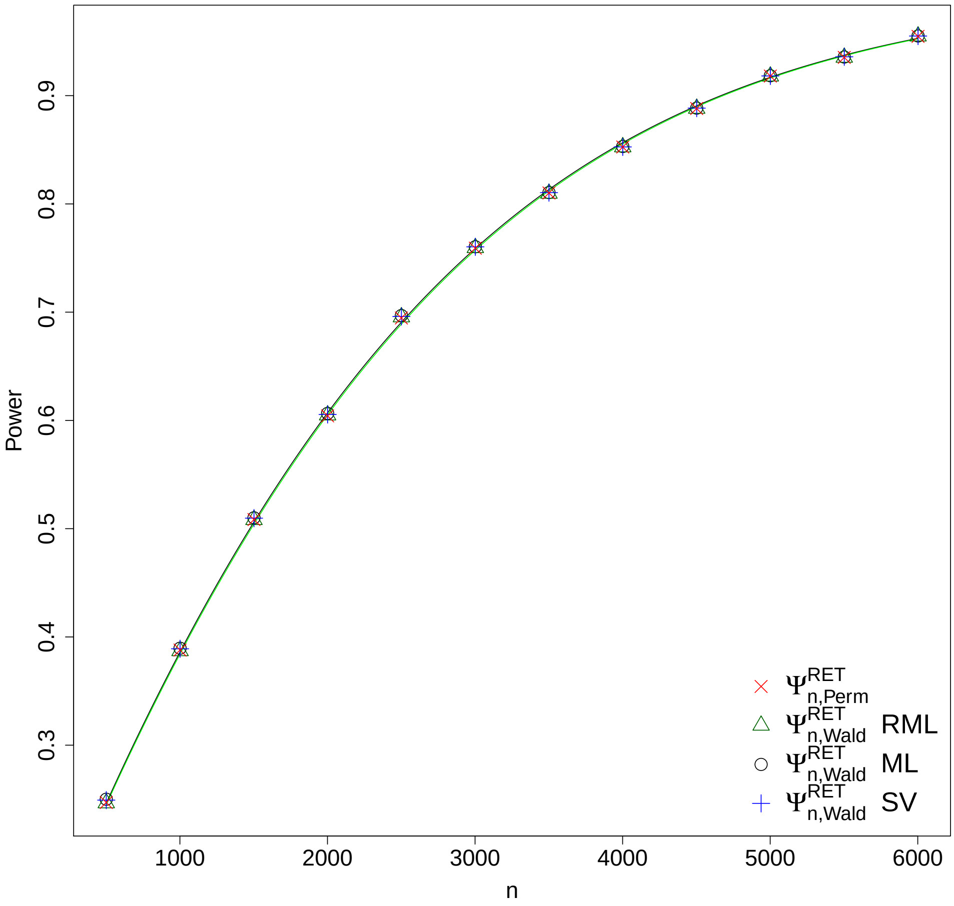


Figure 7: Power of different tests for H_0^{RET} with rates $\lambda_{E,1} = \lambda_R = 1.16$ as well as $\lambda_P = 1.71$, non-inferiority margin $\Delta = 94/123$, shape parameter $\phi = 0.5$, and sample size allocation w_{opt}. The black and the green line are the power approximations from Equations (6.4) and (6.6), respectively.

Figure 7 displays that the power of the different Wald-type tests $\Psi_{n,Wald}^{RET}$ and the permutation test $\Psi_{n,Perm}^{RET}$ is almost the same. This also holds for the power approximations (6.4) and (6.6) which differ at most by 0.007. Moreover, both formulas approximate the power well. Thus, the corresponding sample size formulas can be used for sample size planning. In particular, since the different Wald-type tests and the permutation test are neither conservative nor liberal and have the same power for the sample size allocation $w_{opt,m}$, none of them is superior for the considered setting. Since the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator is recommended for all sample size allocations, we compare the power of the mentioned test for the different allocations. Instead of the actual power we plot the power approximation, since the difference between them is rather small and in particular not significant.

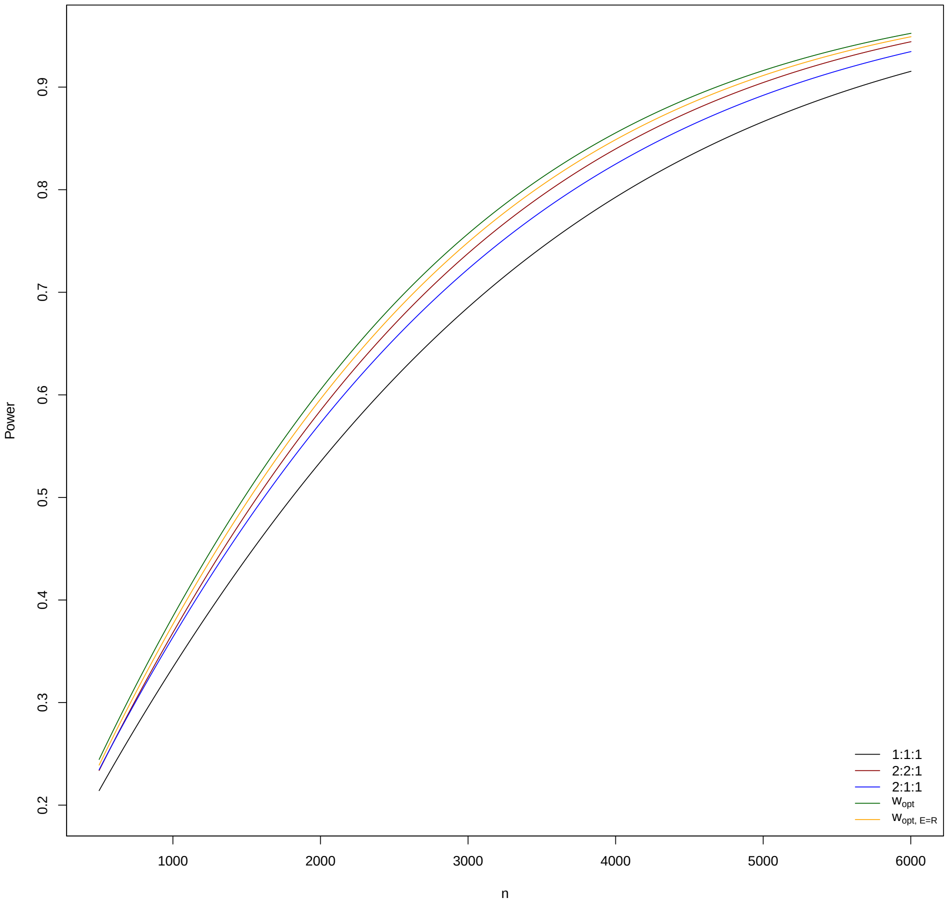
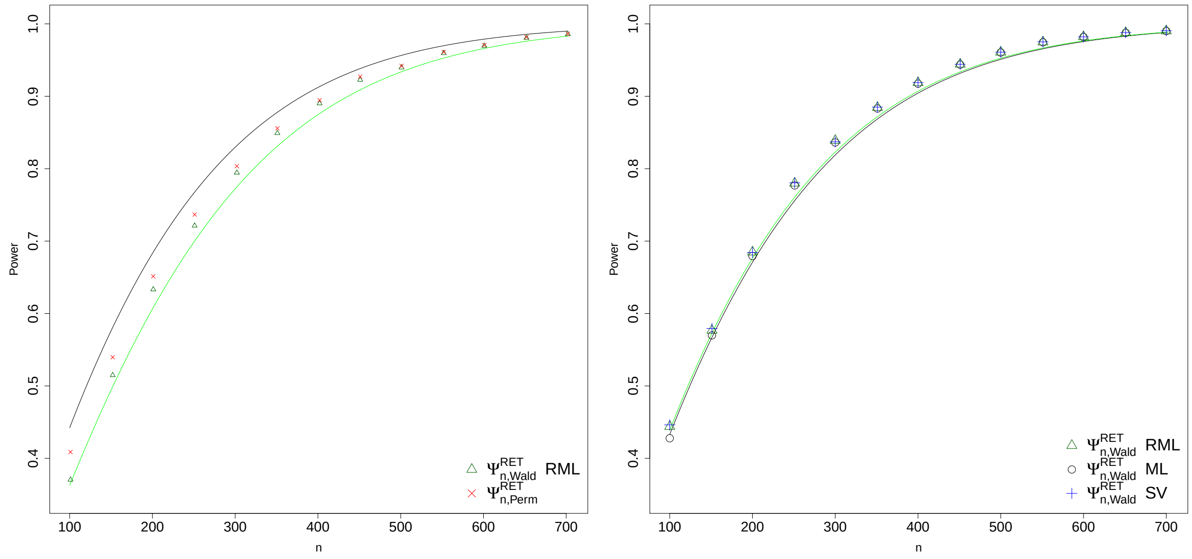


Figure 8: Power of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator for rates $\lambda_{E,1} = \lambda_R = 1.16$ as well as $\lambda_P = 1.71$, non-inferiority margin $\Delta = 94/123$, and shape parameter $\phi = 0.5$ by sample size allocation.

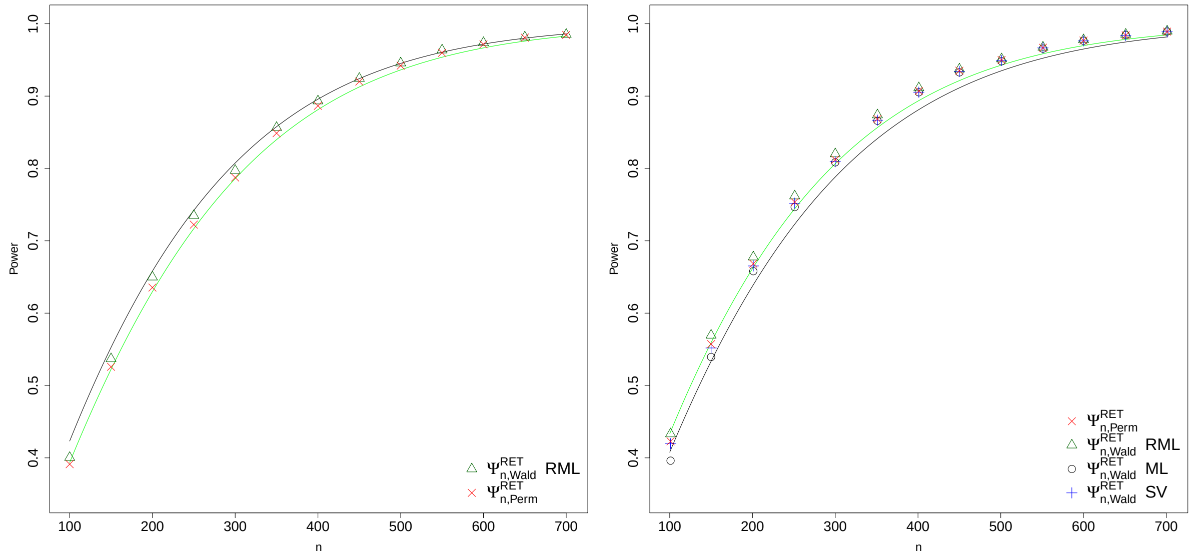
The power approximations for the sample size allocation 3:2:1 and $1 : \Delta : (1 - \Delta)$ are not shown in Figure 8, since they are almost identically to the power approximation for the optimal sample size allocation w_{opt}. The differences are at most 0.0024 which is negligible, since this is smaller than the deviations of the actual power from the approximations. In Figure 8 we see that the power of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator is maximized by the sample size allocation w_{opt} among the allocations considered, what was to be expected. Additionally, the rule of thumb results in a very good approximation of the allocation w_{opt}. In particular, around half of the patients should be allocated to the experimental treatment group. Additionally, the power decreases as the sample size becomes balanced. For the, from a practical point of view, most important range for the power of 0.7-0.95, the power for the sample size allocations 1:1:1 and w_{opt} differ at most by 0.071.

In what follows, we study the power of the Wald-type tests $\Psi^{RET}_{n,Wald}$ and the permutation test $\Psi^{RET}_{n,Perm}$ for the parameter setting motivated by the example for a clinical trial in MS discussed in Section 2.2.2. As mentioned at the beginning of this subsection, we choose the rates $\lambda_{E,1} = \lambda_R = 5.1$ as well as $\lambda_P = 17.4$ and the non-inferiority margin $\Delta = 94/123$. We omit the graphics of the results for the shape parameters $\phi = 1$ and $\phi = 3$ and only have a closer look at the results for the shape parameter $\phi = 2$ because the outcomes are qualitatively the same. Moreover, we leave out the graph for the rule of thumb sample size allocation because the results are qualitatively equal to the results for the sample size allocation 3:2:1. As seen in Figure 6, for the sample size allocations 1:1:1 and $w_{opt,m}$ the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator is the only non-liberal test and therefore we omit the corresponding graphs, too. For these allocations, the power of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted maximum-likelihood variance estimator is slightly larger than the approximation, i.e. the sample size formula (6.7) is not recommended for sample size planning.



(a) Sample size allocation $w_{opt,E=R}$.

(b) Sample size allocation 2:1:1.



(c) Sample size allocation 2:2:1.

(d) Sample size allocation 3:2:1.

Figure 9: Power of different tests for H_0^{RET} by sample size allocation. The rates are $\lambda_{E,1} = \lambda_R = 5.1$ as well as $\lambda_P = 17.4$ and the non-inferiority margin is given by $\Delta = 94/123$. The black and the green line are the power approximations from Equations (6.4) and (6.6), respectively.

Figure 9 shows that for the sample size allocation $w_{opt,E=R}$, the permutation test $\Psi_{n,Perm}^{RET}$ is more powerful than the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator. That was to be expected, since the Wald-type test is slightly conservative. Moreover, the power of the Wald-type test is larger than the power approximation and, therefore, the corresponding sample size formula can applied for sample size planning. On the contrary, the power of the permutation test is smaller than the approximation and, hence, the corresponding sample size formula is not appropriate to plan the sample size. For the sample size allocation 2:1:1, the three Wald-type tests $\Psi_{n,Wald}^{RET}$ have almost the sample same power which is a bit larger than the approximations whose difference is negligible, i.e. the sample size formulas can be applied. Regarding the sample size allocation 2:2:1, the power of the permutation test $\Psi_{n,Perm}^{RET}$ is smaller than the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator. Analogously to the sample size allocation $w_{opt,E=R}$, the power approximation from (6.4) has larger values than the power of the permutation test and, in consequence, the corresponding sample size formulas are not appropriate for usage. In contrast, the power of the Wald-type test is larger than its approximation and as a result, a trial with a sample size from Formula (6.6) will be slightly overpowered. Finally, we study the power of the tests for the sample size allocation 3:2:1. At least for the sample sizes where the different points can be distinguished graphically, the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator has the largest power. The approximative power functions are smaller than the respective actual power values and hence the sample size formulas can be used.

Next, we compare the power of the tests for different sample size allocations. Thereto, for each sample size allocation, we take the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator into account because it is recommended for use for all sample size allocations. Additionally, besides for the sample size allocation $w_{opt,E=R}$, the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator has the largest power. Concerning the test and the sample size allocation with the largest power, the power of the permutation test $\Psi_{n,Perm}^{RET}$ for the sample size allocation $w_{opt,E=R}$ is neither the test with the largest nor the smallest power and is therefore omitted in the next graph. Since the power is mostly distinct from its approximation, we compare the simulated power itself. However, to simplify the comparison, we interpolate the respective points linearly.

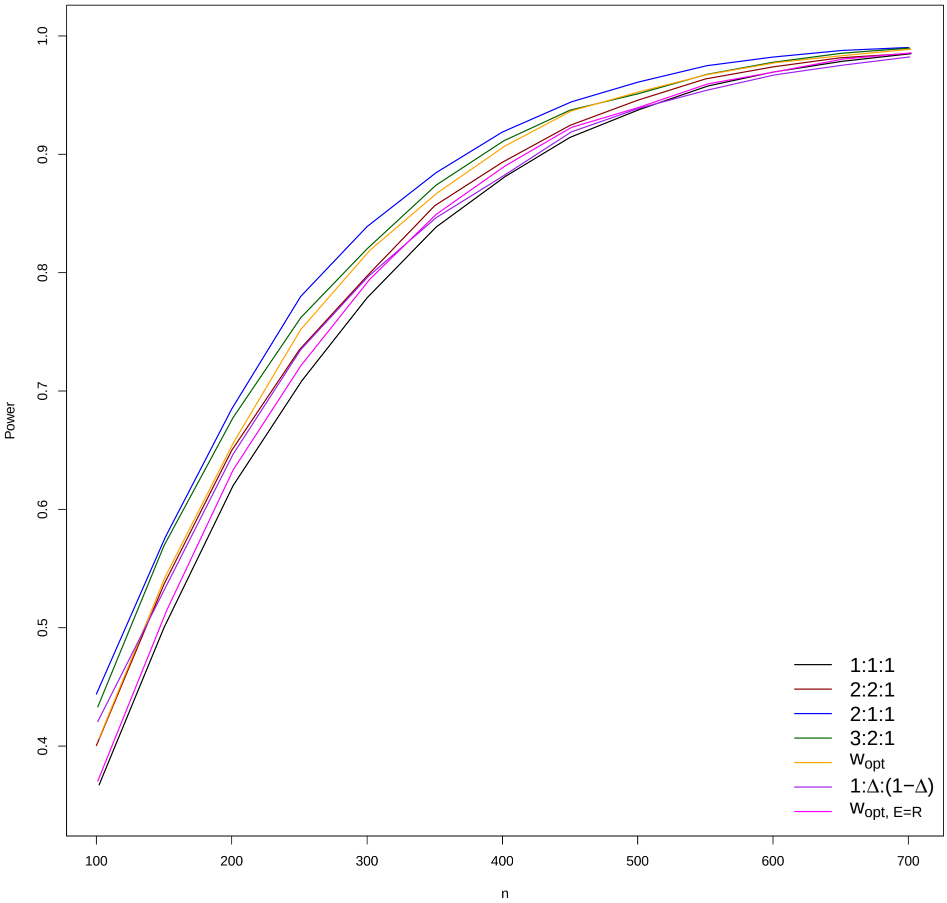


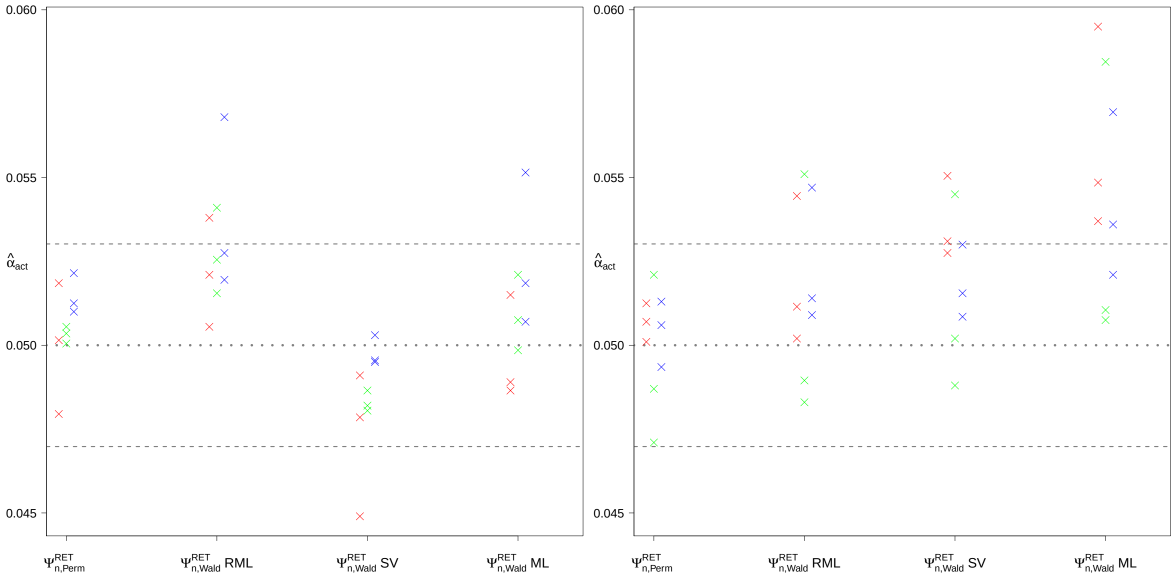
Figure 10: Power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator for rates $\lambda_{E,1} = \lambda_R = 1.16$ as well as $\lambda_P = 1.71$, non-inferiority margin $\Delta = 94/123$, and shape parameter $\phi = 0.5$ by sample size allocation.

The sample size allocation $w_{opt,m}$ is omitted in Figure 10, since the power is nearly identically to the power for the allocation 2:2:1. Figure 10 shows that the sample size allocation w_{opt} does not result in the largest power but the sample size allocation 2:1:1. For most sample sizes, the allocation 1:1:1 yield the smallest power and the rule of thumb is not a good approximation of the allocation with the largest power. For the important range 0.7-0.95 of the power, the difference between the largest and the smallest power for a fixed sample size is 0.0768. In essence, the allocation w_{opt} does not maximize the power and is therefore not optimal. This may be due to the allocation w_{opt} is based on the approximative power function for the Wald-type test $\Psi_{n,Wald}^{RET}$ with an unrestricted variance estimator. Determining an optimal allocation with respect to the approximative power function for

the Wald-type test $\Psi_{n,Wald}^{RET}$ the restricted maximum-likelihood variance estimator could be a part of further research on this topic. This results in an optimization problem which can only be solved numerically, since the power approximation depends on the limit of the restricted maximum-likelihood variance estimator $\sigma_{RET,RML}^2$ and thus on the minimizer of the Kullback-Leibler divergence.
Summarizing, for the first parameter setting the power of the tests which are not liberal is nearly the same and the approximative power functions describe the power well. Among the considered sample size allocation, the allocation w_{opt} maximizes the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator. In contrast, for the second parameter setting, the power of the tests differ and none of the considered tests has the largest power for all considered sample size allocations. However, the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood estimator has only for the allocation $w_{opt,E=R}$ not the largest power, since here the permutation test $\Psi_{n,Perm}^{RET}$ has the largest one. The power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator is for the considered cases not larger than its approximation, i.e. the sample size can be planned with the corresponding sample size formula. However, the power is not maximized for the allocation w_{opt} but for the allocation 2:1:1, i.e. formulas for the optimal sample size allocations from Section 6.3.2 do not hold in this setting.

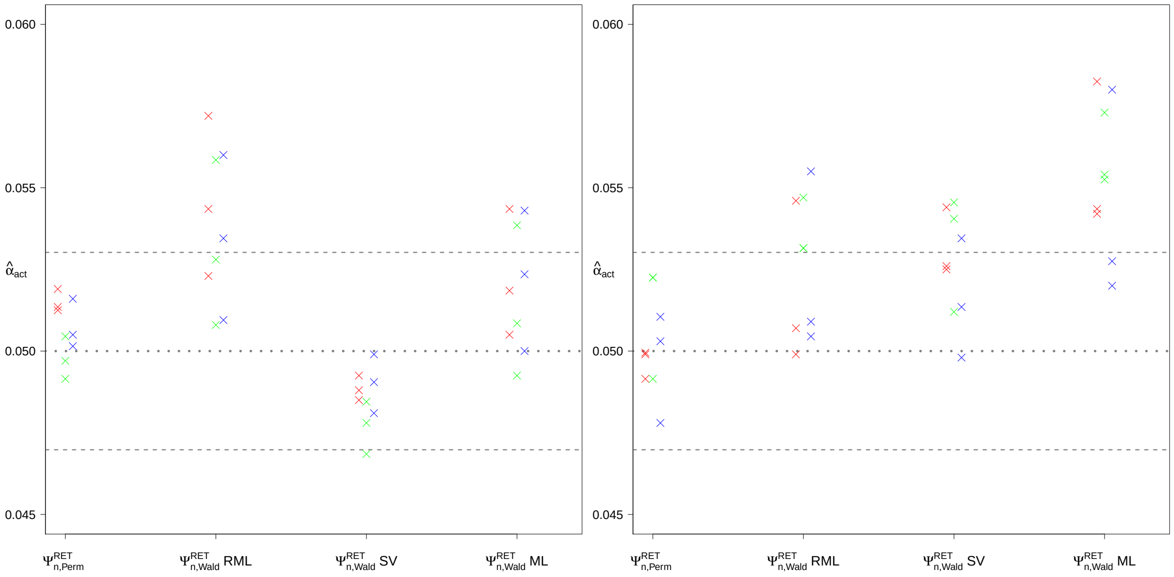
6.4.3 Robustness Concerning Deviations from the Assumed Distribution

The Wald-type tests $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood variance estimator have been established under the assumption of negative binomially distributed random variables. The assumptions of a specific distribution is mostly uncertain and, therefore, we study how sensitive the different Wald-type tests and the permutation test are concerning deviations from the assumed negative binomial distribution stated in Section 2.3. As to that, we assume that the random variables $X_{k,i}$ with $i = 1, \dots, n_k$ and $k = E, R, P$ are Poisson–inverse-Gaussian and Poisson–lognormally distributed as destribed and motivated in Section 2. The expectation and the variance are assumed to be the same as for the Monte-Carlo simulations in Section 6.4.1. Firstly, we simulate the actual levels of the Wald-type tests $\Psi_{n,Wald}^{RET}$ and the permutation test $\Psi_{n,Perm}^{RET}$ for Poisson–inverse-Gaussian and Poisson–lognormally distributed random variables which parameters are chosen analogously to Definition 6.8. As before, the results base on $M = 20{,}000$ Monte-Carlo simulations, the quantile of the conditional permutation distribution on 20,000 random permutations and the test are constructed with the level of significance $\alpha = 0.05$. We expect that the actual levels of the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the sample variance estimator are not significantly different from the actual levels of these tests for negative binomially distributed observations because these tests are not constructed for a specific distribution. Since Figure 1 shows that the differences between the probability function of the negative binomial, the Poisson–inverse-Gaussian and the Poisson–lognormal distribution are rather small, we assume that the actual levels of the Wald-type tests $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood variance estimator are not affected significantly if the observations are not negative binomially distributed. The results for the sample size allocations 2:1:1 and 2:2:1 are exemplary for the effects caused by a different distribution and, thus, we omit the graphs for the other allocations.



(a) Poisson–inverse-Gaussian distribution. Sample size allocation 2:1:1.

(b) Poisson–inverse-Gaussian distribution. Sample size allocation 2:2:1.



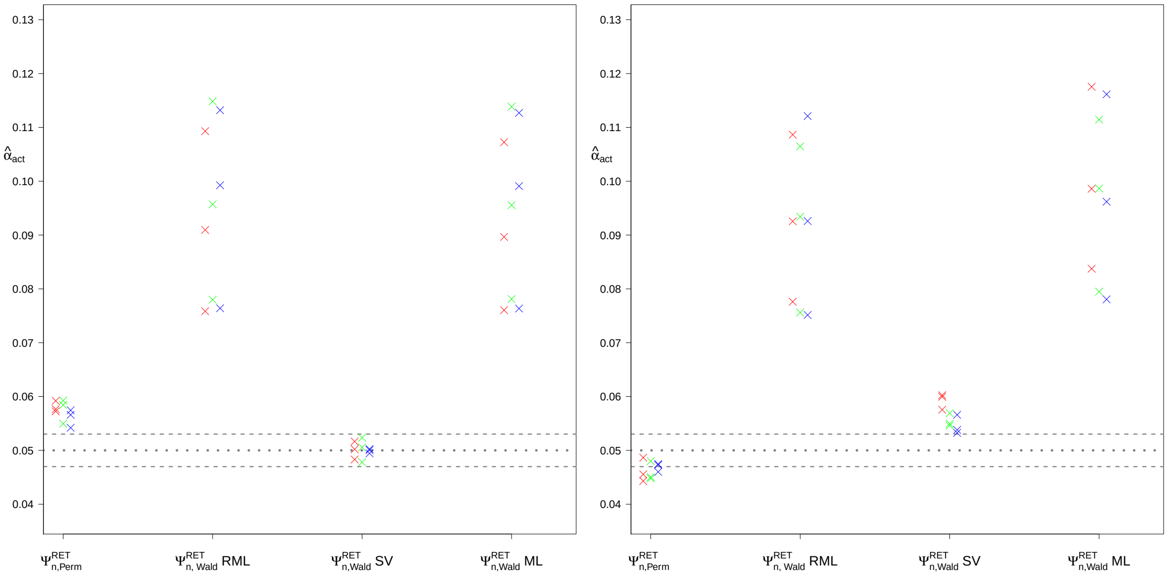
(c) Poisson–lognormal distribution. Sample size allocation 2:1:1.

(d) Poisson–lognormal distribution. Sample size allocation 2:2:1.

Figure 11: Actual level of different tests for H_0^{RET} by sample size allocation for Poisson–inverse-Gaussian and Poisson–lognormally distributed random variables. The points for a sample size of 550 are red, for one of 1100 are green, and for one of 2200 are blue. Actual levels between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations ML, SV, RML indicate the variance estimator of the Wald-type tests. The shape parameters are not marked differently.

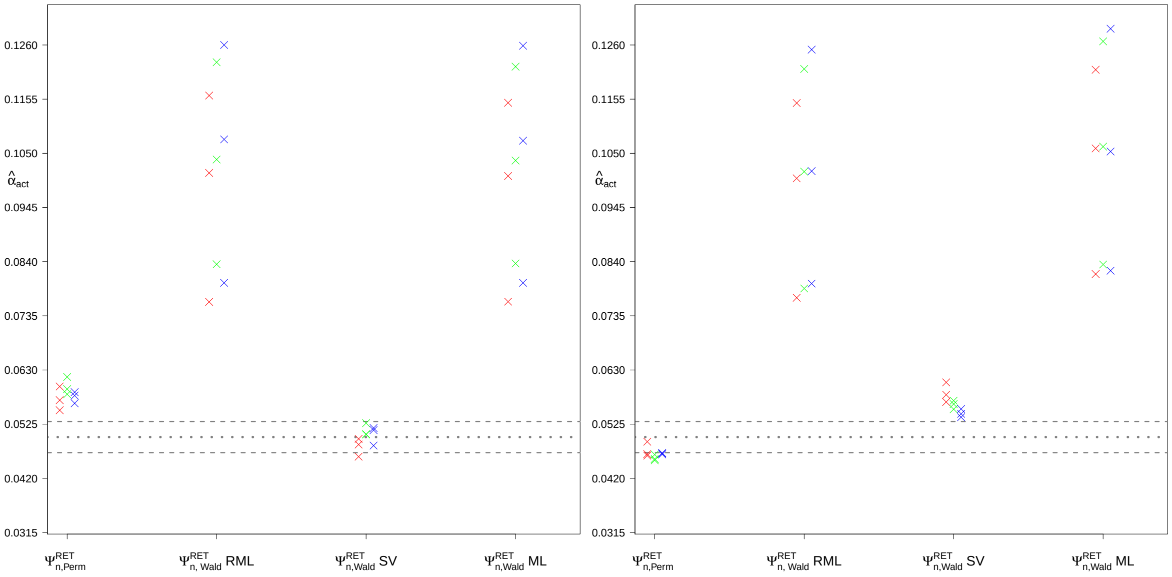
We analyze the results of Figure 11 by comparing them to the corresponding results for negative binomially distributed random variables from Figure 5. As expected, the actual levels of the permutation test $\Psi^{RET}_{n,Perm}$ and the Wald-type test $\Psi^{RET}_{n,Wald}$ with the sample variance estimator only differ within the limits of the Monte-Carlo error from the actual levels for negative binomially distributed observations. In contrast and in particular contrary to our expectations, the actual levels of the Wald-type test $\Psi^{RET}_{n,Wald}$ with the unrestricted or the restricted maximum-likelihood variance estimator increase if the distribution of the random variables changes. Additionally, since the actual levels of the Wald-type test $\Psi^{RET}_{n,Wald}$ with a maximum-likelihood variance estimator differ much for a fixed sample size, the shape parameter seems to affect the inflation. We conclude that the Wald-type test $\Psi^{RET}_{n,Wald}$ with a maximum-likelihood variance estimator is affected by a different distribution but the inflation depends on both, the allocation and the distribution. Since the inflation is small, the Wald-type test $\Psi^{RET}_{n,Wald}$ with a maximum-likelihood variance estimator can in several cases still be recommended for use.

In the following, we study the actual levels of the Wald-type tests $\Psi_{n,Wald}^{RET}$ and the permutation test $\Psi_{n,Perm}^{RET}$ for Poisson–inverse-Gaussian and Poisson–lognormally distributed observations with expectation and variance equal to the setting for the negative binomial distribution from Definition 6.9. As before, for the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the sample variance estimator we do not expect any significant changes of the actual level. However, for the Wald-type test $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood variance estimator an affect on the actual levels is expected, since Figure 2 shows that the probability mass function of the mixed Poisson distributions differ clearly.



(a) Poisson–inverse-Gaussian distribution. Sample size allocation 2:1:1.

(b) Poisson–inverse-Gaussian distribution. Sample size allocation 2:2:1.



(c) Poisson–lognormal distribution. Sample size allocation 2:1:1.

(d) Poisson–lognormal distribution. Sample size allocation 2:2:1.

Figure 12: Actual level of different tests for H_0^{RET} by sample size allocation for Poisson–inverse-Gaussian and Poisson–lognormally distributed random variables. The points for a sample size of 215 are red, for one of 430 are green, and for one of 860 are blue. Actual levels between the lower and upper dashed grey lines do not differ significantly from $\alpha = 0.05$. The abbreviations *ML*, *SV*, *RML* indicate the variance estimator of the Wald-type tests. The shape parameters are not marked differently.

Comparing Figures 6 and 12, the actual levels of the permutation test $\Psi^{RET}_{n,Perm}$ and the Wald-type test $\Psi^{RET}_{n,Wald}$ with the sample variance estimator are not affected by a different mixed Poisson distribution. However, the Wald-type tests $\Psi^{RET}_{n,Wald}$ with the restricted or unrestricted maximum-likelihood variance estimator have actual levels of at least 7.5%, i.e. they become very liberal if the random variables are not negative binomially distributed. To conclude, the permutation test $\Psi^{RET}_{n,Perm}$ and the Wald-type test $\Psi^{RET}_{n,Wald}$ with the sample variance estimator are robust concerning a different mixed Poisson distribution. That was to be expected, since both tests do not bear on a specific distribution, confer Remarks 6.2 and 6.4. However, the Wald-type test $\Psi^{RET}_{n,Wald}$ with the restricted or unrestricted maximum-likelihood variance estimator is affected by a different distribution even if the distribution is nearly the same as the negative binomial distribution and are therefore not appropriate to test the retention of effect hypothesis H_0^{RET} if the random variables are not exactly negative binomially distributed. The more the distribution differs from a negative binomial distribution, the more liberal the Wald-type test $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood variance estimator. Analogously to the actual level, the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the sample variance estimator and the permutation test $\Psi_{n,Perm}^{RET}$ is not affected significantly if the distribution of the endpoint changes.

6.4.4 Power of the Test Procedure

The power of a three-arm non-inferiority trial is in general reported with respect to both assay sensitivity and non-inferiority. Thus, from a theoretical point of view, we have to determine the sample size through the power of the test procedure and not through the power for the retention of effect hypothesis. However, for reasonable alternatives Section 5 in Kieser and Friede (2007) and Section 4.2 in Mielke et al. (2008) show that for binary and censored exponentially distributed endpoints, respectively, the power of the test for the retention of effect hypothesis is approximately the power of the test procedure. Hence, in this section, we study the power of the test procedure and compare it with the power of the test for the retention of effect hypothesis for negative binomially distributed endpoints. Thereto, we only take the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator as a test for the hypothesis H_0^{RET} into account because it has the best overall performance in case of negative binomially distributed observations. Additionally, we define assay sensitivity as the superiority of the experimental as well as the reference treatment over placebo. If the assertion that the power of the test procedure and the test for the retention of effect hypothesis are similar holds for this definition of assay sensitivity, it also holds if assay sensitivity is defined by just one superiority. Since we showed that the permutation $\Psi_{n,Perm}^{EP}$ performed best when testing the superiority of the experimental treatment versus placebo, see Section 5.3, we test superiority of an active treatment over placebo with the permutation test $\Psi_{n,Perm}^{EP}$.
Firstly, we compare the power of the test procedure and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood estimator for the scenario motivated by the TRISTAN study, i.e. for the parameter vector $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.71, 0.5)$ and the non-inferiority margin $\Delta = 43/55$. As for the power simulations for the retention of effect hypothesis, the results for the shape parameters $\phi = 0.3$ and $\phi = 0.7$ are qualitatively the same and therefore omitted. As in Section 6.4.2, we simulated the power for the sample sizes $n \in \{500, 1000, \ldots, 4000\}$. We do not show any graphs because the power curves are the same as in Figure 7. However, Table 6 states the difference between the power of the test procedure and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator for all sample sizes

considered (entire range) as well as for the sample sizes where the power in the practically relevant range $70\% - 95\%$ (relevant range). A difference larger than zero implies that the test for the retention of effect hypothesis is more powerful.

Table 6: Difference between the power of the test procedure and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator for the parameter $\zeta_{H_1^{RET}} = (1.16, 1.16, 1.71, 0.5)$ and the non-inferiority margin $\Delta = 43/55$ by sample size allocation.

Allocation	Entire range	Relevant range
1:1:1	2.47%	0%
2:1:1	5.55%	0%
2:2:1	3.0%	0%
3:2:1	4.555%	0%
w_{opt}	4.795%	0%
$w_{opt,E=R}$	3.97%	0%
w_{rot}	5.53%	0%

Table 6 shows that the power of the test procedure and the power of the Wald-type test differ at most by 5.55% but within the relevant range for the power, the difference of the power functions is zero. Thus, the sample size for a three-arm non-inferiority trial can be planned through the test for the retention of effect hypothesis.

Analogously to the parameter motivated by the TRISTAN study, we compare the power for the parameter $\zeta_{H_1^{RET}} = (5.1, 5.1, 17.4, 2)$ and the non-inferiority margin $\Delta = 94/123$ motivated by the CONFIRM study. Exemplary, we analyse the results for $\phi = 2$ but it should be mentioned that the results for the shape parameters $\phi = 1$ and $\phi = 3$ are qualitatively the same. As for the power simulation in Section 6.4.2, the sample size is chosen as $n \in \{50, 100, \dots, 700\}$. For the shape of the power curve see Figure 9.

Table 7: Difference between the power of the test procedure and the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator for the parameter $\zeta_{H_1^{RET}} = (5.1, 5.1, 17.4, 2)$ and the non-inferiority margin $\Delta = 94/123$ by sample size allocation.

Allocation	Entire range	Relevant range
1:1:1	1.675%	0%
2:1:1	5.28%	0.105%
2:2:1	2.43%	0.08%
3:2:1	4.63%	0.0245%
w_{opt}	2.505%	0.01%
$w_{opt,E=R}$	1.75%	0%
w_{rot}	5.055%	0.64%

For all sample sizes considered, the power functions differ at most by 5.28%. In contrast to the results from Table 6, the differences between the power functions are not zero for the relevant range. However, they are less than 0.7% and, as before, the sample size formula for the Wald-type test can be applied to plan the sample size of a three-arm non-inferiority trial.

7 Conclusion and Discussion

The aim of this thesis was to develop tests for the retention of effect hypothesis and assay sensitivity as well as to derive formulas for the sample size and its allocation. The major result of this thesis is that the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator is appropriate to test the retention of effect hypothesis H_0^{RET} if the observations are negative binomially distributed and the shape parameter is equal among the groups. However, this Wald-type test is sensitive to deviations from the assumed distribution. Moreover, we showed that in this setting assay sensitivity should be tested with a permutation test. In the following, we summarize and discuss the major results of this thesis in more detail and subsequently give an outlook regarding further research on the retention of effect hypothesis.

In Section 2.3, we defined that the random variables $X_{k,i}$ with $i = 1, \ldots, n_k$ and $k = E, R, P$ are negative binomially distributed with rates λ_k and shape parameter ϕ. To test assay sensitivity we introduced different Wald-type tests as well as an exact permutation test in Section 5. The Monte-Carlo simulations of the tests' actual levels of significance in Section 5.3 show that the performance of most of the Wald-type tests depend on the sample size allocation as well as on the size of the rates and the shape parameter. However, the actual level of the permutation test is not influenced significantly by these factors and thus recommended to test assay sensitivity.

In Section 6, we dealt with tests for the retention of effect hypothesis H_0^{RET}. We established the different Wald-type tests in Section 6.1 and an asymptotic permutation test in Section 6.2. In Section 6.4.1, we compared the actual levels of significance of the different tests in Monte-Carlo simulations. It became apparent that overall the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator performs best, i.e. the test is not liberal and at most slightly conservative. Moreover, this test is robust over various scenarios of rates, shape parameters, sample sizes and sample size allocations. Depending on these factors, the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type test $\Psi_{n,Wald}^{RET}$ with an unrestricted variance estimator are also appropriate for application. Since the Wald-type test $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood variance estimator has been constructed with the assumption of negative binomially distributed random variables, we studied in Section 6.4.3 how the actual levels change if the observations are Poisson–lognormally or Poissson–inverse-Gaussian distributed. The actual level of the Wald-type test $\Psi_{n,Wald}^{RET}$ with a maximum-likelihood estimator gets affected by deviations from the assumed distribution, i.e. the test becomes liberal. The magnitude of the inflation depends on the distribution as well as on the expectation, variance, sample size, and sample size allocation. However, the actual levels of the permutation test $\Psi_{n,Perm}^{RET}$ and the Wald-type test $\Psi_{n,Wald}^{RET}$ with a sample variance estimator are not affected significantly.

To plan the different tests for the retention of effect hypothesis, we introduced power approximations for the different tests and established the sample size formulas in Section 6.3.1. Both the power approximations and the sample size formulas are motivated by the asymptotic normality of the tests. In Section 6.3.2, we calculated sample size allocations which maximize the power for a fixed sample size. Since we do not know the actual power of the tests, we calculated the optimal allocations such that they maximize the power approximations, i.e. if the power approximations are inaccurate, the calculated allocations do not necessarily maximize the power. The comparison of the power for the non-liberal tests in Section 6.4.2 showed that for the scenarios motivated by the TRISTAN study the corresponding tests have almost the same power for a fixed sample size allocation. This could be due to the large sample size, i.e. to obtain a power of 70% a sample size of about 3,000 is required. Moreover, the formulas (6.4) and (6.6) approximate the power well and thus the sample size formulas 6.5 and 6.7 are appropriate to determine the sample size for the corresponding tests of the retention of effect hypothesis. In Figure 8 we compared the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator for different sample size allocations. Among the considered allocations, the power is maximized for the allocation w_{opt}, which has been defined in (6.9), and the power decreases as the allocation becomes balanced. In other words, the sample size which maximizes the power approximation for the tests with an unrestricted variance estimator also maximizes the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with a restricted maximum-likelihood variance estimator. For the scenarios motivated by the CONFIRM study a sample size n of about 300 is required for a power of 70%, i.e. the sample size is much smaller than in the first example. For some allocations the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with an unrestricted variance estimator and the permutation test $\Psi_{n,Perm}^{RET}$ is smaller than its approximation (6.4). In these cases the sample size formula (6.5) is not recommended for usage. However, the approximation (6.6) of the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator fits well and hence the corresponding sample size formula (6.7) can be applied. In Figure 10 we saw that for this setting the optimal sample size allocations from Section 6.3.2 do not maximize the power of the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator. Summarizing, among the considered tests and in case of negative binomially distributed observations the Wald-type test $\Psi_{n,Wald}^{RET}$ with the restricted maximum-likelihood variance estimator performs best and thus is recommended for use.

In the following, we discuss some open questions of this thesis and give an outlook about future research on the planning and assessing of three-arm non-inferiority trials. In particular, we focus on how the model can be extended for instance to include covariates. Additionally, we give ideas about dealing with that in the sample size calculation as well in the optimal allocations knowledge about the unknown variances σ_E^2, σ_R^2, and σ_P^2 is required.
In this thesis, we defined non-inferiority through the retention of effect hypothesis

$$H_0^{RET} : (\lambda_P - \lambda_E) \leq \Delta(\lambda_P - \lambda_R) \qquad \text{versus} \qquad H_1^{RET} : (\lambda_P - \lambda_E) > \Delta(\lambda_P - \lambda_R)$$

with $\Delta \in (0, 1)$ the prespecified clinical relevance. In Section 1, we motivated the retention of effect hypothesis by the non-inferiority hypothesis $H_0 : \lambda_E \geq \lambda_R + \delta$ with the prespecified clinical relevance $\delta := f(\lambda_P - \lambda_R)$. However, Hida and Tango (2011) proposed not to define non-inferiority through the retention of effect hypothesis because the definition of δ would contradict to δ being a prespecified margin since it is defined by means of rates. Instead of testing superiority of the reference treatment over placebo and non-inferiority defined through the retention of the effect hypothesis, Hida and Tango (2011) suggested to test the inequality

$$\lambda_E < \lambda_R + \Delta < \lambda_P$$

with two separate hypothesis tests. More precisely, it has been suggested that relevant superiority of the reference treatment over the placebo and non-inferiority of the experimental and the reference treatment should be tested. For further discussions of this proposal we refer to Röhmel and Pigeot (2011). However, establishing tests and deriving formulas for the sample size and its allocation for the hypotheses proposed could be part of further research.
From a theoretical point of view, it remains to be proved that the maximum-likelihood estimator $\hat{\phi}$ is unique. Moreover, the existence and uniqueness of the maximum-likelihood estimators restricted to the retention of effect hypothesis H_0^{RET}, which have been introduced in Section 6.1, has not been proved yet. Last but not least it has not been shown yet that the minimizer of the Kullback-Leibler divergence is unique for the model stated in Section 2.3.

A crucial assumption for the model in Section 2.3 was that the shape parameter ϕ is equal among the groups. Thus, a logical extension of the model is to allow unequal shape parameters. To extend the model in this way, we just have to replace the estimator $\hat{\phi}$ by the corresponding maximum-likelihood estimator $\hat{\phi}_k$ with $k = E, R, P$. The asymptotic theory of the different Wald-type tests for assay sensitivity and the retention of effect hypothesis as well as the asymptotic theory of the permutation test for the retention of effect hypothesis still holds. However, the permutation test for assay sensitivity is not exact any more, since the random variables are not exchangeable at the boundary of the hypothesis.

In a clinical trial, for each patient a number of baseline characteristics are gathered, for instance gender, age, weight, or the measurement of the endpoint at baseline. When included in the final analysis, the are referred to as covariates. To allow the inclusion of covariates, Lawless (1987) described the negative binomial regression. Basically, the negative binomial regression assumes that the rate of a negative binomial distribution is a function in the covariates. In the following we briefly introduce negative binomial regression and state how it can be applied to test the retention of effect hypothesis. Using the notation of Lawless (1987), let Y be an observation and $\mathbf{x} \in \mathbb{R}^m$ a vector of covariates then the negative binomial regression model is given by

$$\mathbb{P}(Y = y|\mathbf{x}) = \frac{\Gamma\left(y + \frac{1}{\phi}\right)}{y!\Gamma\left(\frac{1}{\phi}\right)} \left(\frac{\phi\mu(\mathbf{x})}{1 + \phi\mu(\mathbf{x})}\right)^y \left(\frac{1}{1 + \phi\mu(\mathbf{x})}\right)^{\frac{1}{\phi}}$$

for $y \in \mathbb{N}_0$. As before, ϕ is the shape parameter. With $T > 0$ and $\beta \in \mathbb{R}^m$, the function $\mu(\mathbf{x})$ is defined by

$$\mu(\mathbf{x}) := T \exp\left(\mathbf{x}'\boldsymbol{\beta}\right).$$

The expectation and the variance of Y conditioned on $\mathbf{x}$ is given by $\mathbb{E}[Y|\mathbf{x}] = \mu(\mathbf{x})$ and $\mathrm{Var}[Y|\mathbf{x}] = \mu(\mathbf{x})(1+\phi\mu(\mathbf{x}))$. Next, we apply the negative binomial regression to the retention of effect hypothesis. Thereto, let Y_i and $\mathbf{x_i}$ with $i = 1, \ldots, n$ be the observation and the vector of covariables for the i-th patient, respectively. Without restricting the generality of the results, we set $T = 1$. The treatment group of the i-th patient is determined by the first three entries of $\mathbf{x}$, i.e. $(x_{i,1}, x_{i,2}, x_{i,3}, \ldots) = (1, 0, 0, \ldots)$ corresponds to the experimental treatment group, $(x_{i,1}, x_{i,2}, x_{i,3}, \ldots) = (0, 1, 0, \ldots)$ to the reference treatment group, and $(x_{i,1}, x_{i,2}, x_{i,3}, \ldots) = (0, 0, 1, \ldots)$ to the placebo group. It should be mentioned that in general two indicators are sufficient to model three groups if β_1 is defined as one. However, we use three indicators since it results in the rates, which describe the treatment efficacies as well as the placebo response, given by $\log(\beta_1)$, $\log(\beta_2)$, and $\log(\beta_3)$, respectively. Thereby, the retention of effect hypothesis is given by

$$\begin{aligned} & H_0^{RET}: \log(\beta_3) - \log(\beta_1) \leq \Delta(\log(\beta_3) - \log(\beta_2)) \\ \text{versus} \quad & H_1^{RET}: \log(\beta_3) - \log(\beta_1) > \Delta(\log(\beta_3) - \log(\beta_2)). \end{aligned}$$

Lawless (1987) proved that the maximum-likelihood estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are asymptotically normal distributed. Hence, the theory of the Wald-type test with a maximum-likelihood variance estimator for the retention of effect hypothesis from Mielke (2010) also holds for this setting. However, the theory of the Wald-type test with a sample variance estimator and the theory of the asymptotic permutation test cannot be applied easily. This is mainly because the sample variance is not a consistent estimator for the maximum-likelihood estimator of $\log(\beta_j)$ with $j = 1, 2, 3$ and the permutation test as introduced by Janssen (1997) is only applicable if the effect can be estimated the corresponding sample mean. Summarizing, if we assume that the link function $\mu(\mathbf{x}) = T \exp(\mathbf{x}'\boldsymbol{\beta})$ holds, the statistical model and the Wald-type test with a maximum-likelihood variance estimator can easily be extended to covariates.

The sample size formulas and optimal sample size allocations introduced in Section 6.3 require knowledge about the variances σ_E^2, σ_R^2, and σ_P^2. Even if the sample size is planned based on variances from similar studies, it might be too small which results in an underpowered or if it is too large in an overpowered trial. However, adaptive designs provide a solution to this problems. In Section 2 in Gallo et al. (2006), an adaptive design has been defined as

> "... a clinical study design that uses accumulating data to decide how to modify aspects of the study as it continues, without undermining the validity and integrity of the trial. The goal of adaptive designs is to learn from the accumulating data and to apply what is learned as quickly as possible."

Additionally, Gallo et al. (2006) discussed issues and opportunities of adaptive designs. Similar definitions of an adaptive design are given on page 10 in CHMP (2007) as well as in Section III.A in FDA (2010). Moreover, CHMP (2007) and FDA (2010) defined regulatory requirements for adaptive designs. Concerning the sample size planning, we focus on designs for clinical trials with sample size review as one possibility of adaptive designs. Thereto, we describe two different approaches. Both approaches include a small study which is part of the actual trial but also affects the design of the trial. Firstly, we describe a design with a sample size review based on the estimation of nuisance parameters. Here, the small study on which the sample size review is based is called internal pilot study. Secondly, we extend this design to two stage designs where the sample size is reviewed as well as an interim analysis is performed. In this case, the small study is denoted as stage I. Hereinafter, we denote the sample size of the internal pilot study and stage I with n_0.

Regarding the design where only the sample size is reviewed, the unknown variances are estimated with the results from the internal pilot study and the estimates in turn are used to reestimate the sample size $\hat{n}_{1-\beta}$ for the whole trial. To complete the trial $\max(0, \hat{n}_{1-\beta} - n_0)$ patients have to be recruited. For general discussions about the purpose of internal pilot studies confer Wittes and Brittain (1990). For a review of sample size reestimation procedures confer Friede and Kieser (2006) and in particular for sample size reestimation in a two-arm non-inferiority trial with negative binomially distributed endpoints see Friede and Schmidli (2010). To our knowledge, there are no publications studying such a design for three-arm trials. Therefore, in the following, we discuss ideas how the sample size can be reestimated in three-arm non-inferiority trials with negative binomially distributed endpoints and non-inferiority defined by the retention of effect hypothesis H_0^{RET}. Thereby, we restrict ourselves to the sample size formula (6.5), since the approach follows analogously for the formula (6.7). Additionally, we assume that we are in a scenario where the sample size formula works. Then, the sample size of trial is calculated by

$$n_{1-\beta} = (q_{1-\alpha} + q_{1-\beta})^2 \frac{\sigma_{RET}^2}{\eta_{H_1^{RET}}^2}$$

with $\eta_{H_1^{RET}} = (1-\Delta)\lambda_{P,H_1^{RET}} + \Delta\lambda_{R,H_1^{RET}} - \lambda_{E,H_1^{RET}}$ being the assumed effect and σ_{RET}^2 the assumed variance given by

$$\sigma_{RET}^2 = \frac{\sigma_E^2}{w_E} + \frac{\sigma_R^2}{w_R} + \frac{\sigma_P^2}{w_P}.$$

After recruiting n_0 patients, the variance σ_{RET}^2 in the sample size formula is replaced by its estimate $\hat{\sigma}_{RET}^2$. Thus, the reestimated sample size is given by

$$\hat{n}_{1-\beta} = (q_{1-\alpha} + q_{1-\beta})^2 \frac{\hat{\sigma}_{RET}^2}{\eta_{H_1^{RET}}^2}.$$

As mentioned above, $\max(0, \hat{n}_{1-\beta} - n_0)$ patients need to be recruited and, afterwards, the final analysis is done with the results from both the internal pilot study and the second stage.

After focusing on designs where only the sample size is reviewed, we discuss approaches for designs with an effect based sample size reestimation and an interim analysis potentially allowing early rejection of the null hypothesis. Thereto, an approach for sample size and effect review in three-arm non-inferiority trials with normally distributed endpoints and homogeneous variances has been proposed by Schwartz and Denne (2006). However, to our knowledge, there are no publications addressing this approach in three-arm non-inferiority trials with negative binomially distributed endpoints. In Schwartz and Denne (2006), the idea is to replace the variance σ^2_{RET} in the sample size formula as above but, additionally, substitute the effect $\eta_{H_1^{RET}}$ by an unbiased estimator. More precisely, if we assume that the rates in the experimental and the reference treatment group are equal under the fixed alternative, we obtain the effect

$$\eta_{H_1^{RET}} = (1-\Delta)(\lambda_{P,H_1^{RET}} - \lambda_{R,H_1^{RET}}).$$

Now, an unbiased estimator $\hat{\eta}_{H_1^{RET}}$ for the effect $\eta_{H_1^{RET}}$ is obtained by estimating the rates unbiased with the corresponding sample mean. Replacing the effect and the variance in the sample size formula by its estimates yield the reestimated sample size

$$\hat{n}_{1-\beta} = (q_{1-\alpha} + q_{1-\beta})^2 \frac{\hat{\sigma}^2_{RET}}{\hat{\eta}^2_{H_1^{RET}}}.$$

If the estimated effect $\hat{\eta}_{H_1^{RET}}$ is small, the sample size $\hat{n}_{1-\beta}$ becomes large. Additionally, a small estimated effect also indicates that the trial may lack assay sensitivity. Thereto, as an interim analysis, it is feasible to test assay sensitivity and even non-inferiority as well as extend the trial to a superiority trial. Such a sequential design has been introduced for three-arm non-inferiority trials with binary distributed observation by Li and Gao (2010). In a generalized setting such adaptive designs and their regulatory as well as statistical issues have been studied by Koyama et al. (2005).

For further research on this topic one has to study whether a sample size reestimation and an interim analysis affects the actual level of significance and the power of the study. Last but not least, to recalculate the optimal sample size allocation, one replaces the assumed variances σ^2_E, σ^2_R, and σ^2_P by its estimates. Of course, the formula for optimal sample size allocations has to be appropriate if resulting allocation should yield valid results.

A Appendix

Proof of Theorem 6.3. To prove the assertion, we show that the conditions 1.-5. of Theorem 4.4 are fulfilled and thereto we proceed analogously to the proof of Lemma 4.1 in Janssen (1997). For the sake of a convenient notation, we denote the i-th entry of the vector of random variables $\mathbf{X_n} = (\mathbf{X_{E,n_E}}, \mathbf{X_{R,n_R}}, \mathbf{X_{P,n_P}})$ as $X_{n,i}$. Without loss of generality, we assume that the expectation of $\frac{1}{n}\sum_{i=1}^{n} X_{n,i}$ is equal to zero, since the statistic $T_{n,Perm}^{RET}(\mathbf{X_n})$ is invariant under the same shift for each $X_{n,i}$. Of course, the shift does not effect the variance of the random variables.

1. Basic calculations show that the sum of the $c_{n,i}$ and the sum of the $c_{n,i}^2$ is zero and one, respectively, i.e.

$$\begin{aligned}\sum_{i=1}^{n} c_{n,i} =& \sqrt{\frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R}} \times \\ & \left(-\sum_{i=1}^{n_E}\frac{1}{n_E} + \Delta\sum_{i=n_E+1}^{n_E+n_R}\frac{1}{n_R} + (1-\Delta)\sum_{i=n_E+n_R+1}^{n_E+n_R+n_P}\frac{1}{n_P}\right) = 0,\\ \sum_{i=1}^{n} c_{n,i}^2 =& \frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R} \times \\ & \left(\sum_{i=1}^{n_E}\frac{1}{n_E^2} + \Delta^2\sum_{i=n_E+1}^{n_E+n_R}\frac{1}{n_R^2} + (\Delta-1)^2\sum_{i=n_E+n_R+1}^{n_E+n_R+n_P}\frac{1}{n_P^2}\right)\\ =& \frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R}\left(\frac{1}{n_E} + \Delta^2\frac{1}{n_R} + (\Delta-1)^2\frac{1}{n_P}\right) = 1.\end{aligned}$$

2. In the following, we prove that the limes inferior of the sample variance of $\mathbf{X_n}$ is $\mathbb{P}$-almost surely positive, i.e.

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(X_{n,i} - \overline{X}_{n,\cdot})^2 > 0 \qquad \mathbb{P} - a.s.$$

With the assumption that the expectation of the average $\overline{X}_{n,\cdot}$ is zero and the strong law of large numbers, the average $\overline{X}_{n,\cdot}$ converges almost surely to zero. By means of the continuous mapping theorem, the squared average $\overline{X}_{n,\cdot}^2$ converges almost surely to zero. With the property that the sum of three sequences of random variables converges almost surely if each of the sequences converges almost surely as well as with the strong law of large numbers, the average of the squared random variables converges almost surely:

$$\frac{1}{n}\sum_{i=1}^{n} X_{n,i}^2 \xrightarrow{n\to\infty} w_E(\sigma_E^2+\lambda_E^2)+w_R(\sigma_R^2+\lambda_R^2)+w_P(\sigma_P^2+\lambda_P^2) \qquad \mathbb{P}\text{-a.s.}$$

Hence, with the algebraic formula of the sample variance

$$\frac{1}{n}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_{n,\cdot})^2 \xrightarrow{n\to\infty} w_E(\sigma_E^2+\lambda_E^2)+w_R(\sigma_R^2+\lambda_R^2)+w_P(\sigma_P^2+\lambda_P^2)>0 \qquad \mathbb{P}\text{-a.s}$$

3. To prove the convergence

$$\frac{1}{\hat{\sigma}_{Perm}^2(\tau_n(\mathbf{X_n}))}\frac{1}{n}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_{n,\cdot})^2 \xrightarrow[\mathbb{P}\times\tilde{\mathbb{P}}]{n\to\infty} 1, \tag{A.1}$$

we show that the variance estimators $\hat{\sigma}_{Perm}^2(\tau_n(\mathbf{X_n}))$ and $\frac{1}{n}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_{n,\cdot})^2$ converge in $\mathbb{P}\times\tilde{\mathbb{P}}$-probability to the same limit, i.e.

$$\frac{1}{n}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_{n,\cdot})^2 \xrightarrow[\mathbb{P}\times\tilde{\mathbb{P}}]{n\to\infty} w_E(\sigma_E^2+\lambda_E^2)+w_R(\sigma_R^2+\lambda_R^2)+w_P(\sigma_P^2+\lambda_P^2), \tag{A.2}$$

$$\hat{\sigma}_{Perm}^2(\tau_n(\mathbf{X_n})) \xrightarrow[\mathbb{P}\times\tilde{\mathbb{P}}]{n\to\infty} w_E(\sigma_E^2+\lambda_E^2)+w_R(\sigma_R^2+\lambda_R^2)+w_P(\sigma_P^2+\lambda_P^2). \tag{A.3}$$

The assertion A.2 follows immediately from 2., since we showed $\mathbb{P}$-a.s. convergence which implies $\mathbb{P}\times\tilde{\mathbb{P}}$-a.s. convergence which in turn yield convergence in $\mathbb{P}\times\tilde{\mathbb{P}}$-probability.

To prove the convergence of the variance estimator $\hat{\sigma}_{Perm}^2(\tau_n(\mathbf{X_n}))$, we decompose it by means of the algebraic formula for the sample variance, i.e.

$$\hat{\sigma}_{Perm}^2(\tau_n(\mathbf{X_n})) = W_{n,1}-W_{n,2}^2-W_{n,3}^2-W_{n,4}^2$$

with

$$\begin{aligned}
W_{n,1} :=& \frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R} \left(\frac{1}{n_E(n_E-1)} \sum_{i=1}^{n_E} X_{n,\tau(i)}^2 \right. \\
&\left. + \frac{\Delta^2}{n_R(n_R-1)} \sum_{i=n_E+1}^{n_E+n_R} X_{n,\tau(i)}^2 + \frac{(\Delta-1)^2}{n_P(n_P-1)} \sum_{i=n_E+n_R+1}^{n_E+n_R+n_P} X_{n,\tau(i)}^2 \right) \\
W_{n,2} :=& \sqrt{\frac{n_R n_P}{(n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R) n_E(n_E-1)}} \sum_{i=1}^{n_E} X_{n,\tau(i)} \\
W_{n,3} :=& \sqrt{\frac{\Delta^2 n_E n_P}{(n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R) n_R(n_R-1)}} \sum_{i=n_E+1}^{n_E+n_R} X_{n,\tau(i)} \\
W_{n,4} :=& \sqrt{\frac{(\Delta-1)^2 n_E n_R}{(n_R n_P + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R) n_P(n_P-1)}} \sum_{i=n_E+n_R+1}^{n_E+n_R+n_P} X_{n,\tau(i)}.
\end{aligned}$$

Thereby, $X_{n,\tau(i)}$ denotes the i-th entry of the vector $\tau_n(\mathbf{X_n})$. We prove the convergence of the variance estimator $\tilde{\sigma}^2_{Perm}\left(\tau_n\left(\mathbf{X_n}\right)\right)$ by showing that $W_{n,2}, W_{n,3}$, and $W_{n,4}$ converge in $\mathbb{P} \times \tilde{\mathbb{P}}$-probability to zero as well as that $W_{n,1}$ converges to the limit stated in (A.3). The proof for the convergence of $W_{n,2}, W_{n,3}$, and $W_{n,4}$ are similar and therefore, we only regard $W_{n,2}$. Let $\varepsilon > 0$ be an arbitrary real number, Markov's inequality yield

$$\begin{aligned}
(\mathbb{P} \times \tilde{\mathbb{P}}) \left(|W_{n,2}| \geq \varepsilon\right) &= (\mathbb{P} \times \tilde{\mathbb{P}}) \Big(\Big| W_{n,2} - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] \Big| \geq \varepsilon \Big) \\
\leq& (\mathbb{P} \times \tilde{\mathbb{P}}) \Big(\Big| W_{n,2} - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] + \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] \Big| \geq \varepsilon \Big) \\
\leq& (\mathbb{P} \times \tilde{\mathbb{P}}) \Big(\Big| W_{n,2} - \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] \Big| + \Big| \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] \Big| \geq \varepsilon \Big) \\
\leq& \frac{1}{\Big(\varepsilon - \Big|\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}]\Big|\Big)^2} \operatorname{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[W_{n,2}\right].
\end{aligned}$$

Later on, we show that $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] = 0$ holds. Due to the independence of $\mathbb{P}$ and $\tilde{\mathbb{P}}$ and the law of total variance, we obtain

$$\begin{aligned}
\operatorname{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[W_{n,2}\right] =& \mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\operatorname{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[W_{n,2}\middle|\mathbf{X_n}\right]\right] + \operatorname{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[W_{n,2}\middle|\mathbf{X_n}\right]\right] \\
=& \mathbb{E}_{\mathbb{P}}\left[\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\middle|\mathbf{X_n}\right]\right] + \operatorname{Var}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\middle|\mathbf{X_n}\right]\right] \\
=& \mathbb{E}_{\mathbb{P}}\left[\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right] + \operatorname{Var}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right].
\end{aligned}$$

Hence, $W_{n,2}$ converges in probability to zero if the expectation $E_{\mathbb{P}}\left[\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right]$ and the variance $\operatorname{Var}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right]$ converge to zero as n tends to infinity. For the sake of readability, we define

$$\kappa := \frac{n_R n_P}{(n_R n_P + \Delta^2 n_E n_P + (\Delta - 1)^2 n_E n_R) n_E (n_E - 1)}.$$

Taking into account that $X_{n,\tau(i)}$ and $X_{n,\tau(1)}$ have the same distribution as well as that with probability ${}^1\!/_n$ the random variable $X_{n,\tau(1)}$ is equal to the random variable $X_{n,i}$, $i = 1, \dots, n$, for the expectation $\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]$ holds

$$\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right] = \sqrt{\kappa} \sum_{i=1}^{n_E} \mathbb{E}_{\tilde{\mathbb{P}}}\left[X_{n,\tau(i)}\right] = \sqrt{\kappa}\, n_E \mathbb{E}_{\tilde{\mathbb{P}}}\left[X_{n,\tau(1)}\ \right] = \sqrt{\kappa}\, n_E \frac{1}{n} \sum_{i=1}^{n} X_{n,i}.$$

Thus, $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}] = 0$ follows immediately from the independence of $\mathbb{P}$ and $\tilde{\mathbb{P}}$ as well as the assumption that the average $\overline{X}_{n,\cdot}$ has expectation zero. With $\kappa = \kappa(n) \in \mathcal{O}({}^1\!/_{n^2})$, we obtain

$$\lim_{n\to\infty} \operatorname{Var}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right] \leq \lim_{n\to\infty} \kappa \frac{n_E}{n} \max_{1\leq i\leq n} \operatorname{Var}_{\mathbb{P}}[X_{n,i}] = 0. \tag{A.4}$$

Moreover, to prove that the expectation $\mathbb{E}_{\mathbb{P}}\left[\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]\right]$ converges to zero, we rearrange the variance

$$\begin{aligned}
\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right] &= \kappa \operatorname{Var}_{\tilde{\mathbb{P}}}\left[\sum_{i=1}^{n_E} X_{n,\tau(i)}\right] = \kappa\, \mathbb{E}_{\tilde{\mathbb{P}}}\left[\left(\sum_{i=1}^{n_E} X_{n,\tau(i)} - n_E \overline{X}_{n,\cdot}\right)^2\right] \\
=&\kappa\, \mathbb{E}_{\tilde{\mathbb{P}}}\left[\sum_{i,j=1}^{n_E} X_{n,\tau(i)} X_{n,\tau(j)} - 2 n_E \overline{X}_{n,\cdot} \sum_{i=1}^{n_E} X_{n,\tau(i)} + n_E^2 \overline{X}_{n,\cdot}^2\right] \\
=&\kappa\, \mathbb{E}_{\tilde{\mathbb{P}}}\left[\sum_{\substack{i,j=1,\\ i\neq j}}^{n_E} X_{n,\tau(i)} X_{n,\tau(j)} + \sum_{i=1}^{n_E} X_{n,\tau(i)}^2\right] - \kappa\, n_E^2 \overline{X}_{n,\cdot}^2.
\end{aligned}$$

For $i \neq j$ and $i' \neq j'$, the random variables $X_{n,\tau(i)} X_{n,\tau(j)}$ and $X_{n,\tau(i')} X_{n,\tau(j')}$ are identically distributed with respect to $\tilde{\mathbb{P}}$ and with probability $1/(n(n-1))$ the random variable $X_{n,\tau(1)}X_{n,\tau(2)}$ is equal to $X_{n,i}X_{n,j}$. Thus, we obtain

$$\begin{aligned}
\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right] =&\kappa\,\frac{n_E(n_E-1)}{n(n-1)}\sum_{\substack{i,j=1,\\ i\neq j}}^{n} X_{n,i}X_{n,j} + \kappa\frac{n_E}{n}\sum_{i=1}^{n} X_{n,i}^2 - \kappa\, n_E^2\overline{X}_{n,\cdot}^2 \\
=&\kappa\,\frac{n_E(n_E-1)}{n(n-1)}\left(n^2\overline{X}_{n,\cdot}^2 - \sum_{i=1}^{n} X_{n,i}^2\right) + \kappa\frac{n_E}{n}\sum_{i=1}^{n} X_{n,i}^2 - \kappa\, n_E^2\overline{X}_{n,\cdot}^2 \\
=&\kappa\,\left(\frac{n_E}{n} - \frac{n_E(n_E-1)}{n(n-1)}\right)\sum_{i=1}^{n} X_{n,i}^2 + \kappa\left(\frac{n_E(n_E-1)n}{(n-1)} - n_E^2\right)\overline{X}_{n,\cdot}^2 \\
=&\kappa\,\frac{n_E(n-n_E)}{n(n-1)}\sum_{i=1}^{n} X_{n,i}^2 - \kappa\frac{n_E(n-n_E)}{n-1}\overline{X}_{n,\cdot}^2 \\
=&\kappa\frac{n_E(n-n_E)}{n(n-1)}\sum_{i=1}^{n}(X_{n,i}-\overline{X}_{n,\cdot})^2.
\end{aligned}$$

We already proved that the sample variance converges $\mathbb{P}$ almost surely and with $k = k(n) \in \mathcal{O}(1/n^2)$, it follows that the expectation $\mathbb{E}_{\mathbb{P}}[\operatorname{Var}_{\tilde{\mathbb{P}}}\left[W_{n,2}\right]]$ converges to zero as n approaches infinity. Thus, $W_{n,2}$ and analogously $W_{n,3}$ as well as $W_{n,4}$ converge in $\mathbb{P} \times \tilde{\mathbb{P}}$-probability to zero.
To prove the convergence of the variance estimator $\hat{\sigma}^2_{Perm}\left(\tau_n(\mathbf{X_n})\right)$, it remains to show that $W_{n,1}$ converges in probability to $W_1 := w_E(\sigma_E^2+\lambda_E^2)+w_R(\sigma_R^2+\lambda_R^2)+w_P(\sigma_P^2+\lambda_P^2)$. Analogously to $W_{n,2}$, let $\varepsilon > 0$ be an arbitrary real number and n sufficiently large such that $|\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,1}] - W_1| < \varepsilon$, Markov's inequality yield

$$(\mathbb{P}\times\tilde{\mathbb{P}})\left(|W_{n,1}| \geq \varepsilon\right) \leq \frac{1}{\left(\varepsilon - \left|\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,1}] - W_1\right|\right)^2}\operatorname{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}\left[W_{n,1}\right].$$

Before showing that the right side converges to zero as n approaches infinity, we simplify the notation of $W_{n,1}$. Thereto, we define the sequence $(d_{n,i})_{i\leq n}$ as

$$d_{n,i} := \frac{n_E n_R n_P}{n_P n_R + \Delta^2 n_P n_R + (\Delta-1)^2 n_E n_R} \times \begin{cases} \frac{1}{n_E(n_E-1)} & i = 1,\ldots,n_E \\ \frac{\Delta^2}{n_R(n_R-1)} & i = n_E+1,\ldots,n_E+n_R \\ \frac{(\Delta-1)^2}{n_P(n_P-1)} & i = n_E+n_R+1,\ldots,n \end{cases}$$

and with that, $W_{n,1}$ is equal to $\sum_{i=1}^{n} d_{n,i}X_{n,\tau(i)}^2$. For the sums $d_{n,\cdot} := \sum_{i=1}^{n} d_{n,i}$ and

$\sum_{i=1}^{n} d_{n,i}^2$, we have the asymptotic properties $\lim_{n\to\infty} d_{n,\cdot} = 1$ which follows from

$$\begin{aligned}\lim_{n\to\infty} d_{n,\cdot} &= \lim_{n\to\infty} \frac{n_E n_R n_P}{n_P n_R + \Delta^2 n_E n_P + (\Delta-1)^2 n_E n_R} \left(\frac{1}{n_E - 1} + \frac{\Delta^2}{n_R - 1} + \frac{(\Delta-1)^2}{n_P - 1} \right) \\ &= \lim_{n\to\infty} \frac{n_E n_R n_P}{(n_E-1)(n_R-1)(n_P-1)} \\ &\quad \times \frac{(n_R-1)(n_P-1) + \Delta^2 (n_E-1)(n_P-1) + (\Delta-1)^2 (n_R-1)(n_P-1)}{n_P n_R + \Delta^2 n_P n_E + (\Delta-1)^2 n_E n_R},\end{aligned}$$

as well as $\lim_{n\to\infty} \sum_{i=1}^{n} d_{n,i}^2 = 0$ which follows from

$$\begin{aligned}\lim_{n\to\infty} \sum_{i=1}^{n} d_{n,i}^2 &= \lim_{n\to\infty} \left(\frac{n_E n_R n_P}{n_P n_R + \Delta^2 n_P n_E + (\Delta-1)^2 n_E n_R} \right)^2 \\ &\quad \times \left(\frac{1}{n_E^2 (n_E-1)^2} + \frac{\Delta^4}{n_R^2 (n_R-1)^2} + \frac{(\Delta-1)^4}{n_P^2 (n_P-1)^2} \right) \\ &= \lim_{n\to\infty} \left(1 + \Delta^2 \frac{n_E}{n_R} + (\Delta-1)^2 \frac{n_E}{n_P} \right)^{-2} \\ &\quad \times \left(\frac{n_E^2}{n_E (n_E-1)^2} + \frac{\Delta^4 n_E^2}{n_R (n_R-1)^2} + \frac{(\Delta-1)^4 n_E^2}{n_P (n_P-1)^2} \right).\end{aligned}$$

For both limits, we took into account that none of the three groups vanish asymptotically, i.e. $\lim_{n\to\infty} n_k/n = w_k \in (0,1)$. Due to the independence of $\mathbb{P}$ and $\tilde{\mathbb{P}}$, the expectation of $W_{n,1}$ with respect to $\mathbb{P} \times \tilde{\mathbb{P}}$ is given by

$$\begin{aligned}\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,1}] &= \sum_{i=1}^{n} d_{n,i} \mathbb{E}_{\mathbb{P}} \left[\mathbb{E}_{\tilde{\mathbb{P}}} \left[X_{n,\tau(i)}^2 \right] \right] = d_{n,\cdot} \mathbb{E}_{\mathbb{P}} \left[\frac{1}{n} \sum_{i=1}^{n} X_{n,i}^2 \right] \\ =& d_{n,\cdot} \frac{1}{n} \left(n_E (\sigma_E^2 + \lambda_E^2) + n_R (\sigma_R^2 + \lambda_R^2) + n_P (\sigma_P^2 + \lambda_P^2) \right).\end{aligned}$$

It follows that the expectation $\mathbb{E}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,1}]$ converges to W_1. As for $\mathrm{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,2}]$, the variance of $W_{n,1}$ is equal to

$$\mathrm{Var}_{\mathbb{P}\times\tilde{\mathbb{P}}}[W_{n,1}] = \mathbb{E}_{\mathbb{P}} \left[\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}] \right] + \mathrm{Var}_{\mathbb{P}} \left[\mathbb{E}_{\tilde{\mathbb{P}}}[W_{n,1}] \right].$$

Since the forth moment of $X_{n,i}$ with $i = 1, \ldots, n$ is bounded, for the second term follows

$$\mathrm{Var}_{\mathbb{P}} \left[\mathbb{E}_{\tilde{\mathbb{P}}}[W_{n,1}] \right] = d_{n,\cdot}^2 \, \mathrm{Var}_{\mathbb{P}} \left[\frac{1}{n} \sum_{i=1}^{n} X_{n,i}^2 \right] \leq d_{n,\cdot}^2 \frac{1}{n} \max_{1\leq i\leq n} \mathbb{E}_{\mathbb{P}} \left[X_{n,i}^4 \right].$$

Hence, the variance $\mathrm{Var}_{\mathbb{P}}\left[\mathbb{E}_{\tilde{\mathbb{P}}}[W_{n,1}]\right]$ converges to zero as n approaches infinity. It remains to prove that $E_{\mathbb{P}}\left[\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}]\right]$ converges to zero. Thereto, we calculate the variance $\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}]$ and for the sake of readability we omit the limits of the sums.

$$\begin{aligned}
\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}] &= \mathbb{E}_{\tilde{\mathbb{P}}}\left[\left(\sum_i d_{n,i}X^2_{n,\tau(i)} - \mathbb{E}_{\tilde{\mathbb{P}}}\left[\sum_i d_{n,i}X^2_{n,\tau(i)}\right]\right)^2\right] \\
=&\mathbb{E}_{\tilde{\mathbb{P}}}\left[\left(\sum_i d_{n,i}X^2_{n,\tau(i)} - d_{n,\cdot}\frac{1}{n}\sum_i X^2_{n,i}\right)^2\right] \\
=&\mathbb{E}_{\tilde{\mathbb{P}}}\left[\sum_{i,j} d_{n,i}d_{n,j}X^2_{n,\tau(i)}X^2_{n,\tau(j)} - 2d_{n,\cdot}\left(\sum_i d_{n,i}X^2_{n,\tau(i)}\right)\frac{1}{n}\sum_i X^2_{n,i} + \left(d_{n,\cdot}\frac{1}{n}\sum_i X^2_{n,i}\right)^2\right] \\
=&\mathbb{E}_{\tilde{\mathbb{P}}}\left[\sum_{i\neq j} d_{n,i}d_{n,j}X^2_{n,\tau(i)}X^2_{n,\tau(j)} + \sum_i d^2_{n,i}X^4_{n,\tau(i)}\right] - \left(d_{n,\cdot}\frac{1}{n}\sum_i X^2_{n,i}\right)^2.
\end{aligned}$$

With the same arguments as before, we calculate the expectation and obtain the variance

$$\begin{aligned}
&\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}] \\
=&\left(\sum_{i\neq j} d_{n,i}d_{n,j}\right)\frac{1}{n(n-1)}\sum_{i\neq j}X^2_{n,i}X^2_{n,j} + \left(\sum_i d^2_{n,i}\right)\frac{1}{n}\sum_i X^4_{n,\tau(i)} - \left(d_{n,\cdot}\frac{1}{n}\sum_i X^2_{n,i}\right)^2.
\end{aligned}$$

The first term of the variance $\mathrm{Var}_{\tilde{\mathbb{P}}}[W_{n,1}]$ is equal to

$$\begin{aligned}
&\frac{d^2_{n,\cdot} - \sum_i d^2_{n,i}}{n(n-1)}\left(\left(\sum_i X^2_{n,i}\right)^2 - \sum_i X^4_{n,i}\right) \\
=&\frac{1}{n(n-1)}\left(d^2_{n,\cdot}\left(\sum_i X^2_{n,i}\right)^2 - \left(\sum_i d^2_{n,\cdot}\right)\sum_i X^4_{n,i} - \left(\sum_i d^2_{n,i}\right)\left(\sum_i X^2_{n,i}\right)^2\right) \\
&+\frac{1}{n(n-1)}\left(\sum_i d^2_{n,i}\right)\sum_i X^4_{n,i}.
\end{aligned}$$

Hence, with $\overline{d}_{n,\cdot} := d_{n,\cdot}/n$ the variance $\operatorname{Var}_{\tilde{\mathbb{P}}}[W_{n,1}]$ can be rearranged to

$$\begin{aligned}
&\operatorname{Var}_{\tilde{\mathbb{P}}}[W_{n,1}] \\
=&d_{n,\cdot}^2 \left(\frac{\left(\sum_i X_{n,i}^2\right)^2}{n(n-1)} - \frac{\sum_i X_{n,i}^4}{n(n-1)} - \frac{\left(\sum_i X_{n,i}^2\right)^2}{n^2} \right) + \left(\sum_i d_{n,i}^2\right) \left(-\frac{\left(\sum_i X_{n,i}^2\right)^2}{n(n-1)} + \frac{\sum_i X_{n,i}^4}{n-1} \right) \\
=&\sum_i (d_{n,i} - \overline{d}_{n,\cdot})^2 \frac{1}{n-1} \sum_j \left(X_{n,j}^2 - \frac{1}{n} \sum_i X_{n,i}^2 \right)^2 .
\end{aligned}$$

Since the forth moment of $X_{n,i}$ with $i = 1, \dots, n$ is bounded, the term

$$\frac{1}{n-1} \sum_j \left(X_{n,j}^2 - \frac{1}{n} \sum_i X_{n,i}^2 \right)^2$$

converges in $\mathbb{P}$-probability to a finite limit. Since $\sum_i (d_{n,i} - \overline{d}_{n,\cdot})^2$ converges to zero, it follows that the expectation $\mathbb{E}_{\mathbb{P}}[\operatorname{Var}_{\tilde{\mathbb{P}}}[W_{n,1}]$ also converges to zero. Therefore, $W_{n,1}$ converges in $\mathbb{P} \times \tilde{\mathbb{P}}$-probability to W_1.

4. In the following, we prove that the maximum of the absolute values of the coefficients $(c_{n,i})_{i \leq n}$ approaches zero if n tends to infinity. As before, we take into account that none of the three groups vanishes asymptotically:

$$\begin{aligned}
&\max_{1 \leq i \leq n} |c_{n,i}| \\
=&\sqrt{\frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta - 1)^2 n_E n_R}} \max \left\{ \frac{1}{n_E}, \frac{\Delta}{n_R}, \frac{|\Delta - 1|}{n_P} \right\} \\
\leq&\sqrt{\frac{n_E n_R n_P}{n_R n_P + \Delta^2 n_E n_P + (\Delta - 1)^2 n_E n_R}} \left(\frac{1}{n_E} + \frac{\Delta}{n_R} + \frac{|\Delta - 1|}{n_P} \right) \\
=&\sqrt{\frac{1}{1 + \Delta^2 \frac{n_E}{n_R} + (\Delta - 1)^2 \frac{n_E}{n_P}}} \left(\frac{1}{\sqrt{n_E}} + \frac{\Delta \sqrt{n_E}}{n_R} + \frac{|\Delta - 1| \sqrt{n_E}}{n_P} \right) \xrightarrow{n \to \infty} 0.
\end{aligned}$$

5. Next, we prove the convergence

$$\lim_{d \to \infty} \limsup_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \left(X_{n,i} - \overline{X}_{n,\cdot} \right)^2 \mathbb{1}_{[d,\infty)}(|X_{n,i} - \overline{X}_{n,\cdot}|) = 0 \qquad \mathbb{P}\text{-a.s.}$$

With $\limsup_{n \to \infty} (a_n + b_n) \leq \limsup_{n \to \infty} a_n + \limsup_{n \to \infty} b_n$ and the algebraic formula for the variance, we obtain $\mathbb{P}$ almost surly the inequality

$$\lim_{d\to\infty} \limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \left(X_{n,i} - \overline{X}_{n,\cdot}\right)^2 \mathbb{1}_{[d,\infty)}(|X_{n,i} - \overline{X}_{n,\cdot}|)$$
$$\leq \lim_{d\to\infty} \left(\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_{n,i}^2 \mathbb{1}_{[d,\infty)}(|X_{n,i} - \overline{X}_{n,\cdot}|) + \limsup_{n\to\infty} \left(-\overline{X}_{n,\cdot}^2 \mathbb{1}_{[d,\infty)}(|X_{n,i} - \overline{X}_{n,\cdot}|)\right)\right)$$

Since $\overline{X}_{n,\cdot}$ converges $\mathbb{P}$ almost surly to zero, the second limes superior is zero for each d. Due to the strong law of large number which holds because the $\mathbb{E}[X_{n,i}^2]$ are bounded, for each d, the first limes superior is equal to

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_{n,i}^2 \mathbb{1}_{[d,\infty)}(|X_{n,i} - \overline{X}_{n,\cdot}|) = \sum_{k=E,R,P} w_k \mathbb{E}\left[X_{k,1}^2 \mathbb{1}_{[d,\infty)}(|X_{k,1})\right].$$

Since the expectation $\mathbb{E}[X_{k,1}^2]$ exists for $k = E, R, P$, the expectation $E\left[X_{k,1}^2 \mathbb{1}_{[d,\infty)}(|X_{k,1})\right]$ converges to zero as d approaches infinity.

Since the test statistic $T_{n,Perm}^{RET}$ fulfills the points 1.-5. from Theorem 4.4, the assertion holds. □

Proof of Theorem 6.6. The function $x \mapsto {}^1/_x$, $x > 0$, is strictly convex and since $\sigma_E^2, \Delta^2\sigma_R^2, (1-\Delta)^2\sigma_P^2 > 0$ holds, the function $\sigma_{RET}^2(w_E, w_R, w_P)$ is a strictly convex function and a local minimum is also a unique global minimum. Since the function $f_i, i = 1, 2, 3$, are continuously differentiable convex functions and $h(\cdot)$ is affine, the KKT conditions stated below are sufficient conditions for a local minimum. Hence, we can solve the minimization problem (6.9) by finding vectors $w^* \in \mathbb{R}^3$ which fulfill the KKT conditions

$$\begin{aligned} f_i(w^*), \leq 0 \qquad & i = 1, 2, 3, \\ h(w^*) = 0, & \\ \lambda_i \geq 0, \qquad & i = 1, 2, 3, \\ \lambda_i f_i(w^*) = 0, \qquad & i = 1, 2, 3, \\ \triangledown \sigma_{ER}^2(w^*) + \lambda_1 \triangledown f_1(w^*) + \lambda_2 \triangledown f_2(w^*) + \lambda_3 \triangledown f_3(w^*) + \mu \triangledown h(w^*) = 0 & \end{aligned}$$

with $\lambda \in \mathbb{R}^3$ and $\mu \in \mathbb{R}$. Hence, the fifth KKT condition is equal to the following system of linear equations

$$
\begin{aligned}
&-\frac{\sigma_E^2}{w_E^2} && -\lambda_1 && &&+\mu = 0\\
&-\frac{\Delta^2\sigma_R^2}{w_R^2} && && -\lambda_2 &&+\mu = 0\\
&-\frac{(1-\Delta)^2\sigma_P^2}{w_P^2} && +\lambda_1 && +\lambda_2 -\lambda_3 &&+\mu = 0.
\end{aligned}
$$

We calculate the optimal sample size allocation $w_{opt,m}$ by distinguishing the eight different cases such that the formula

$$\Big(\lambda_1 = 0 \vee f_1(w) = 0\Big) \wedge \Big(\lambda_2 = 0 \vee f_2(w) = 0\Big) \wedge \Big(\lambda_3 = 0 \vee f_3(w) = 0\Big)$$

is true, i.e. the forth KKT condition is fulfilled. For each case, we obtain requirements such that the KKT conditions hold. Hence, if an allocation w^* fulfils the resulting requirement, it also fulfils the KKT conditions and is therefore the unique solution $w_{opt,m}$.

1. $f_1(w) = f_2(w) = f_3(w) = 0$.
   Thus, $w_E = w_R = w_P = m$ holds. Due to $h(w) = 0$, it follows that m has to be 1/3. Therefore, if $m = 1/3$, the optimal sample size allocation is given by $w_E = w_R = w_P = 1/3$.

2. $f_1(w) = f_2(w) = \lambda_3 = 0$.
   With $h(w) = 0$, it follows that the equality $w_E = w_R = w_P = 1/3$ holds. Since we assumed $m \leq 1/3$, the condition $f_3(w) \leq 0$ is fulfilled. The inequalities $\lambda_1, \lambda_2 \geq 0$ as well as the fifth KKT condition holds if the inequalities

   $$
   \begin{aligned}
   \Delta^2\sigma_R^2 + (1-\Delta)^2\sigma_P^2 &\geq 3\sigma_E^2,\\
   \sigma_E^2 + (1-\Delta)^2\sigma_P^2 &\geq 3\sigma_R^2
   \end{aligned}
   $$

   are true. In addition, these are sufficient conditions for $w_E = w_R = w_P = 1/3$ being the solution $w_{opt,m}$.

3. $f_1(w) = \lambda_2 = \lambda_3 = 0$.
   The condition $f_1(w) = 0$ yield $w_E = w_P$ and with $h(w) = 0$, we obtain $w_R = 1 - 2w_E$. Additionally, $f_2(w) \leq 0$ and $f_3(w) \leq 0$ result in $w_P \in [m, 1/3]$. The fifth KKT

condition has the solutions

$$w_{P,\pm} = \begin{cases} \frac{\pm\sqrt{2\Delta^2\sigma_R^2(\sigma_E^2+(1-\Delta)^2\sigma_P^2)}+2(\sigma_E^2+(1-\Delta)^2\sigma_P^2)}{2(2\sigma_E^2-\Delta^2\sigma_R^2+2(1-\Delta)^2\sigma_P^2)} & \Delta^2\sigma_R^2 \neq 2(\sigma_E^2+(1-\Delta)^2\sigma_P^2) \\ \frac{1}{4} & \Delta^2\sigma_R^2 = 2(\sigma_E^2+(1-\Delta)^2\sigma_P^2) \end{cases}$$

and $\lambda_1 \geq 0$ results in

$$\frac{\Delta^2\sigma_R^2}{(1-2w_P)^2} \geq \frac{\sigma_E^2}{w_P^2}$$

Hence, if one of the $w_{P,\pm}$ is contained in the interval $[m, 1/3]$ and fulfills the inequality stated last, it determines the optimal allocation $w_{opt,m}$.

4. $\lambda_1 = \lambda_2 = \lambda_3 = 0$.
   In this case the optimal solution $w_{opt,m}$ is equal to the unrestricted optimal sample size allocation w_{opt} if for the optimal allocation w_{opt} the restriction $w_E, w_R \geq w_P \geq m$ holds.

5. $\lambda_1 = \lambda_2 = f_3(w) = 0$.
   The equalities $f_3(w) = 0$ and $h(w) = 0$ result in $w_P = m$ and $w_E = 1 - w_R - m$, respectively. Further, the inequalities $f_1(w), f_2(w) \leq 0$ yield $w_E, w_R \in [m, 1-2m]$. Additionally, due to the fifth KKT condition, for w_R holds

   $$w_{R,\pm} = \begin{cases} \frac{\pm\sqrt{\sigma_E^2\Delta^2\sigma_R^2(m-1)^2}+\Delta^2\sigma_R^2(m-1)}{\sigma_E^2-\Delta^2\sigma_R^2} & \sigma_E^2 \neq \Delta^2\sigma_R^2 \\ \frac{1-m}{2} & \sigma_E^2 = \Delta^2\sigma_R^2 \end{cases}$$

   and it follows that $\lambda_3 \geq 0$ is equal to

   $$\frac{\Delta^2\sigma_R^2}{w_{R,\pm}^2} \geq \frac{(1-\Delta)^2\sigma_P^2}{m^2}.$$

   Therefore, if one of the $w_{R,\pm}$ fulfils the corresponding conditions, it determines the optimal solution $w_{opt,m}$.

6. $f_1(w) = \lambda_2 = f_3(w) = 0$.
   The equation $f_1(w) = f_3(w) = 0$ yield $w_E = w_P = m$. With $h(w) = 0$, we obtain

$w_R = 1 - 2m$. These allocations are the optimal solution $w_{opt,m}$ if the inequalities

$$\begin{aligned} \frac{\Delta^2 \sigma_R^2}{(1-m)^2} &\geq \frac{\sigma_E^2}{m^2}, \\ 2\frac{\Delta^2 \sigma_R^2}{(1-m)^2} &\geq \frac{\sigma_E^2 + (1-\Delta)^2 \sigma_P^2}{m^2} \end{aligned}$$

hold, since this assures the the fifth KKT condition as well as $\lambda_1, \lambda_3 \geq 0$ are fulfilled.

7. $\lambda_1 = f_2(w) = f_3(w) = 0$.
   From $f_2(w) = f_3(w) = 0$ and $h(w) = 0$, we obtain $w_R = w_P = m$ as well as $w_E = 1 - 2m$. Moreover, the fifth KKT condition and $\lambda_2, \lambda_3 \geq 0$ results in the inequalities

$$\begin{aligned} \frac{\sigma_E^2}{(1-m)^2} &\geq \frac{\Delta^2 \sigma_R^2}{m^2}, \\ 2\frac{\sigma_E^2}{(1-2m)^2} &\geq \frac{\Delta^2 \sigma_R^2 + (1-\Delta)^2 \sigma_P^2}{m^2}. \end{aligned}$$

8. $\lambda_1 = f_2(w) = \lambda_3 = 0$.
   The equality $f_2(w) = 0$ yield $w_R = w_P$ and from $h(w) = 0$ follows that $w_E = 1 - 2w_R$ holds. Moreover, $f_1(w) \leq 0$ and $f_3(w) \leq 0$ result in $w_R \in [m, 1/3]$ and from the fifth KKT condition we obtain

$$w_{R,\pm} = \begin{cases} \frac{\pm\sqrt{2\sigma_E^2(\Delta^2\sigma_R^2+(1-\Delta)^2\sigma_P^2)}-2(\Delta^2\sigma_R^2+(1-\Delta)^2\sigma_P^2)}{2(\sigma_E^2-2(\Delta^2\sigma_R^2+(1-\Delta)^2\sigma_P^2))} & \sigma_E^2 \neq 2((1-\Delta)^2\sigma_P^2 + \Delta^2\sigma_R^2) \\ \frac{1}{4} & \sigma_E^2 = 2((1-\Delta)^2\sigma_P^2 + \Delta^2\sigma_R^2) \end{cases}.$$

   Last but not least, $\lambda_2 \geq 0$ yield

$$\frac{\sigma_E^2}{(1-2w_{R,\pm})^2} \geq \frac{\Delta^2 \sigma_R^2}{w_{R,\pm}^2}.$$

Summing up, we obtain the solution $w_{opt,m}$ by finding a vector $w = (w_E, w_R, w_P)$ which fulfills one of the conditions stated in the items 1.-8. □

Proof of Theorem 6.7. Firstly, we rearrange the minimization problem (6.10) to obtain a one-dimensional optimization problem

$$\begin{aligned} w_{opt,E=R} := \arg\min \quad & \sigma^2_{RET}(w_E, w_E, 1-2w_E) := f(w_E) \\ \text{s.t.} \quad & w_E \in \left[m, \frac{1-m}{2}\right]. \end{aligned}$$

The function f is strictly convex on the interval $(0, 1/2)$ and hence, it has a unique global minimum on the interval $[m, (1-m)/2]$ which will hereafter denoted as the domain of $f(\cdot)$. To determine this minimum, we calculate the root of the derivative of $f(\cdot)$. If the root w^*_E is contained in the domain of $f(\cdot)$, it is the global minimum. However, if it is not in the domain, the function f has a minimum at the boundary of $[m, (1-m)/2]$. More precisely, if $w^*_E < m$ holds, the minimum of $f(w_E)$ is at $w^*_E = m$, and if $w^*_E > (1-m)/2$ holds, $(1-m)/2$ is the minimum of the function f on $[m, (1-m)/2]$.
The function $f(\cdot)$ and its first derivative $f'(\cdot)$ are given by

$$\begin{aligned} f(w_E) &= \frac{\sigma^2_E + \Delta^2\sigma^2_R}{w_E} + \frac{(1-\Delta)^2\sigma^2_P}{1-2w_E}, \\ f'(w_E) &= -\frac{\sigma^2_E + \Delta^2\sigma^2_R}{w^2_E} + 2\frac{(1-\Delta)^2\sigma^2_P}{(1-2w_E)^2}. \end{aligned}$$

Equating the first derivative with zero and rearranging the resulting equation yield

$$\left(2(1-\Delta)^2\sigma^2_P - 4(\sigma^2_E + \Delta^2\sigma^2_R)\right) w^2_E + 4(\sigma^2_E + \Delta^2\sigma^2_R)w_E - (\sigma^2_E + \Delta^2\sigma^2_R) = 0. \tag{A.5}$$

To calculate the w_E solving the equation, we differentiate the cases $(1-\Delta)^2\sigma^2_P = 2(\sigma^2_E + \Delta^2\sigma^2_R)$ and $(1-\Delta)^2\sigma^2_P \neq 2(\sigma^2_E + \Delta^2\sigma^2_R)$. For the first case, $w^*_E = 1/4$ solves Equation (A.5). For the second case, the solution of Equation (A.5) with restriction to $w_E \in (0, 1/2)$ is given by

$$w^*_E = \frac{-2(\sigma^2_E + \Delta^2\sigma^2_R) + \sqrt{2(1-\Delta)^2\sigma^2_P(\sigma^2_E + \Delta^2\sigma^2_R)}}{2(1-\Delta)^2\sigma^2_P - 4(\sigma^2_E + \Delta^2\sigma^2_R)}.$$

As mentioned before, $(w^*_E, w^*_E, 1-2w^*_E)$ is the solution of minimization problem (6.10) if $w^*_E \in [m, (1-m)/2]$ holds. If w^*_E is smaller than m, the solution of (6.10) is given by $w^*_E = (m, m, 1-2m)$ and if w^*_E is larger than $(1-m)/2$, we obtain the solution $w^*_E = ((1-m)/2, (1-m)/2, m)$. □

References

Aaron, S. D., Fergusson, D., Marks, G. B., Suissa, S., Vandemheen, K. L., Doucette, S., Maltais, F., Bourbeau, J. F., Goldstein, R. S., Balter, M., et al. (2008). Counting, analysing and reporting exacerbations of copd in randomised controlled trials. *Thorax*, 63(2):122–128.

Aban, I. B., Cutter, G. R., and Mavinga, N. (2009). Inferences and power analysis concerning two negative binomial distributions with an application to mri lesion counts data. *Computational statistics & data analysis*, 53(3):820–833.

Abramowitz, M. and Stegun, I. (1970). Handbook of mathematical functions.

Aragón, J., Eberly, D., and Eberly, S. (1992). Existence and uniqueness of the maximum likelihood estimator for the two-parameter negative binomial distribution. *Statistics & probability letters*, 15(5):375–379.

Boehringer Ingelheim Pharma GmbH & Co. KG (2013). Was ist copd? http://www.copd-aktuell.de/copd-erkrankung/was-ist-copd.htm. Accessed: 2013-07-26.

Brusasco, V., Hodder, R., Miravitlles, M., Korducki, L., Towse, L., and Kesten, S. (2006). Health outcomes following treatment for 6 months with once daily tiotropium compared with twice daily salmeterol in patients with copd. *Thorax*, 58(5):399–404.

Bulmer, M. (1974). On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, pages 101–110.

Calverley, P., Pauwels, R., Vestbo, J., Jones, P., Pride, N., Gulsvik, A., Anderson, J., and Maden, C. (2003). Combined salmeterol and fluticasone in the treatment of chronic obstructive pulmonary disease: a randomised controlled trial. *The Lancet*, 361(9356):449–456.

Casella, G. and Berger, R. (2002). Statistical inference. duxbury. *Pacific Grove, California, USA*.

Celli, B., Halpin, D., Hepburn, R., Byrne, N., Keating, E., and Goldman, M. (2003). Symptoms are an important outcome in chronic obstructive pulmonary disease clinical trials: results of a 3-month comparative study using the breathlessness, cough and sputum scale (bcss). *Respiratory medicine*, 97:S35–S43.

CHMP (2007). Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design. http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline /2009/09/WC500003616.pdf. Accessed: 2013-10-17.

Cohen, J. and Rudick, R. (2003). *Multiple Sclerosis Therapeutics*. Martin Dunitz.

Compston, A. and Coles, A. (2008). Multiple sclerosis. *The Lancet*, 372(9648):1502–1517.

D'Agostino, R. B., Massaro, J. M., and 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.

Donohue, J. F., van Noord, J. A., Bateman, E. D., Langley, S. J., Lee, A., Witek, T. J., Kesten, S., and Towse, L. (2002). A 6-month, placebo-controlled study comparing lung function and health status changes in copd patients treated with tiotropium or salmeterol. *CHEST Journal*, 122(1):47–55.

Engle, R. F. (1984). Wald, likelihood ratio, and lagrange multiplier tests in econometrics. *Handbook of econometrics*, 2:775–826.

FDA (2010). Adaptive design clinical trials for drugs and biologics (draft). http://www.fda.gov/downloads/Drugs/.../Guidances/ucm201790.pdf. Accessed: 2013-10-17.

Fleming, T. R. (2008). Current issues in non-inferiority trials. *Statistics in medicine*, 27(3):317–332.

Fox, R. J., Miller, D. H., Phillips, J. T., Hutchinson, M., Havrdova, E., Kita, M., Yang, M., Raghupathi, K., Novas, M., Sweetser, M. T., et al. (2012). Placebo-controlled phase 3 study of oral bg-12 or glatiramer in multiple sclerosis. *New England Journal of Medicine*, 367(12):1087–1097.

Francois, M., Peter, C., and Gordon, F. (2012). Dealing with excess of zeros in the statistical analysis of magnetic resonance imaging lesion count in multiple sclerosis. *Pharmaceutical Statistics*, 11(5):417–424.

Friede, T. and Kieser, M. (2006). Sample size recalculation in internal pilot study designs: a review. *Biometrical Journal*, 48(4):537–555.

Friede, T., Mitchell, C., and Müller-Velten, G. (2007). Blinded sample size reestimation in non-inferiority trials with binary endpoints. *Biometrical Journal*, 49(6):903–916.

Friede, T. and Schmidli, H. (2010). Blinded sample size reestimation with negative binomial counts in superiority and non-inferiority trials. *Methods of Information in Medicine*, 49(6):618.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug development—an executive summary of the phrma working group. *Journal of biopharmaceutical statistics*, 16(3):275–283.

Hasler, M., Vonk, R., and 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.

Hida, E. and Tango, T. (2011). On the three-arm non-inferiority trial including a placebo with a prespecified margin. *Statistics in medicine*, 30(3):224–231.

Hill, A. B. (1994). The continuing unethical use of placebo controls. *N Engl J Med*, 331:394–398.

Hinde, J. and Demétrio, C. G. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, 27(2):151–170.

Holla, M. (1967). On a poisson-inverse gaussian distribution. *Metrika*, 11(1):115–121.

ICH (2000). Choice of control group and related issues in clinical trials (e10). http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products /Guidelines/Efficacy/E10/Step4/E10_Guideline.pdf. Accessed: 2013-08-27.

ICH (2010). Guidance for industry - non-inferiority clinical trials (draft). http://www.fda.gov/downloads/Drugs/.../Guidances/UCM202140.pdf. Accessed: 2013-10-16.

Janssen, A. (1997). Studentized permutation tests for non-iid hypotheses and the generalized behrens-fisher problem. *Statistics & probability letters*, 36(1):9–21.

Karlis, D. and Xekalaki, E. (2005). Mixed poisson distributions. *International Statistical Review*, 73(1):35–58.

Keene, O., Calverley, P., Jones, P., Vestbo, J., and Anderson, J. (2008a). Statistical analysis of copd exacerbations. *European Respiratory Journal*, 32(5):1421–1422.

Keene, O., Calverley, P., Jones, P., Vestbo, J., and Anderson, J. (2008b). Statistical analysis of exacerbation rates in copd: Tristan and isolde revisited. *European Respiratory Journal*, 32(1):17–24.

Keene, O. N., Jones, M. R., Lane, P. W., and Anderson, J. (2007). Analysis of exacerbation rates in asthma and chronic obstructive pulmonary disease: example from the tristan study. *Pharmaceutical Statistics*, 6(2):89–97.

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

Koch, A. and Röhmel, J. (2004). Hypothesis testing in the "gold standard" design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics*, 14(2):315–325.

Kombrink, K., Munk, A., and 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*.

Koyama, T., Sampson, A. R., and Gleser, L. J. (2005). A framework for two-stage adaptive procedures to simultaneously test non-inferiority and superiority. *Statistics in medicine*, 24(16):2439–2456.

Lawless, J. F. (1987). Negative binomial and mixed poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lewis, J. A., Jonsson, B., Kreutz, G., Sampaio, C., and van Zwieten-Boot, B. (2002). Placebo-controlled trials and the declaration of helsinki. *The Lancet*, 359(9314):1337–1340.

Li, G. and Gao, S. (2010). A group sequential type design for three-arm non-inferiority trials with binary endpoints. *Biometrical Journal*, 52(4):504–518.

Mielke, M. (2010). *Maximum Likelihood Theory for Retention of Effect Non-inferiority Trials*. PhD thesis, Georg-August-Universität Göttingen.

Mielke, M., Munk, A., and 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.

Munzel, U. (2009). Nonparametric non-inferiority analyses in the three-arm design with active control and placebo. *Statistics in medicine*, 28(29):3643–3656.

Nicholas, R. and Friede, T. (2012). Considerations in the design of clinical trials for relapsing multiple sclerosis. *Clinical Investigation*, 2(11):1073–1083.

Pigeot, I., Schäfer, J., Röhmel, J., and 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.

Röhmel, J. and Pigeot, I. (2011). Statistical strategies for the analysis of clinical trials with an experimental treatment, an active control and placebo, and a prespecified fixed non-inferiority margin for the difference in means. *Statistics in Medicine*, 30(26):3162–3164.

Rothmann, M., Li, N., Chen, G., Chi, G. Y., Temple, R., and Tsou, H.-H. (2003). Design and analysis of non-inferiority mortality trials in oncology. *Statistics in medicine*, 22(2):239–264.

Saha, K. and Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61(1):179–185.

Schwartz, T. A. and Denne, J. S. (2006). A two-stage sample size recalculation procedure for placebo-and active-controlled non-inferiority trials. *Statistics in medicine*, 25(19):3396–3406.

Snapinn, S. M. et al. (2000). Noninferiority trials. *Curr Control Trials Cardiovasc Med*, 1(1):19–21.

Sormani, M., Bruzzi, P., Miller, D., Gasperini, C., Barkhof, F., and Filippi, M. (1999). Modelling mri enhancing lesion counts in multiple sclerosis using a negative binomial model: implications for clinical trials. *Journal of the Neurological Sciences*, 163(1):74–80.

Sormani, M., Bruzzi, P., Rovaris, M., Barkhof, F., Comi, G., Miller, D., Cutter, G., and Filippi, M. (2001). Modelling new enhancing MRI lesion counts in Multiple Sclerosis. *Multiple Sclerosis*, 7(5):298–304.

Stucke, K. and Kieser, M. (2012). A general approach for sample size calculation for the three-arm 'gold standard'non-inferiority design. *Statistics in Medicine*, 31(28):3579–3596.

Suissa, S. (2006). Statistical treatment of exacerbations in therapeutic trials of chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 173(8):842–846.

Temple, R. and Ellenberg, S. S. (2000). Placebo-controlled trials and active-control trials in the evaluation of new treatments. part 1: ethical and scientific issues. *Annals of Internal Medicine*, 133(6):455–463.

Van den Elskamp, I., Knol, D., Uitdehaag, B., and Barkhof, F. (2009). The distribution of new enhancing lesion counts in Multiple Sclerosis: further explorations. *Multiple Sclerosis*, 15(1):42–49.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3):426–482.

Winkelmann, R. (2003). *Econometric analysis of count data*. Springer.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72.

WMA (2008). Declaration of helsinki - ethical principles for medical research involving human subjects. http://www.wma.net/en/30publications/10policies/b3/. Accessed: 2013-10-16.

Zhu, H. and Lakkis, H. (2013). Sample size calculation for comparing two negative binomial rates. *Statistics in Medicine*.

Acknowledgements

I would like to express my sincere gratitude to Prof. Dr. Tim Friede and Prof. Dr. Axel Munk for making this master thesis possible. In addition, I would like to thank Prof. Dr. Tim Friede for the useful comments, remarks and guidance during the time of research and writing.

Furthermore, I would like to thank Novartis Vaccines and Diagnostics GmbH for the financial support during my master thesis and, especially, the *Statistics and Mathematical Modelling* group for the great working atmosphere during the last two years. In particular, I am very grateful to Dr. Walter Hoyer and Alexander Pysik for their constant encouragement.

Finally, I would like to thank my colleagues at the *Department of Medical Statistics* for their helpfulness and the great working atmosphere.

Declaration of Authorship

I hereby declare that this master thesis is the product of my own independent work. All content and ideas drawn directly or indirectly from external sources are indicated as such. The thesis has not been submitted to any other examining body and has not been published.

Göttingen, den 22.10.2013