


Semi-parametric analysis of overdispersed count and metric data with varying follow-up times: Asymptotic theory and small sample approximations

Frank Konietschke¹  | Tim Friede²  | Markus Pauly³

¹Department of Mathematical Sciences, University of Texas at Dallas, Dallas, TX, USA

²Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

³Institute of Statistics, Ulm University, Ulm, Germany

Correspondence

Frank Konietschke, Department of Mathematical Sciences, University of Texas at Dallas, Dallas, TX 75080, USA.

Email: fxx141230@utdallas.edu

Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Number: DFG-PA 2409/3-1; FP7 Health, Grant/Award Number: FP HEALTH 2013 - 602144

Abstract

Count data are common endpoints in clinical trials, for example magnetic resonance imaging lesion counts in multiple sclerosis. They often exhibit high levels of overdispersion, that is variances are larger than the means. Inference is regularly based on negative binomial regression along with maximum-likelihood estimators. Although this approach can account for heterogeneity it postulates a common overdispersion parameter across groups. Such parametric assumptions are usually difficult to verify, especially in small trials. Therefore, novel procedures that are based on asymptotic results for newly developed rate and variance estimators are proposed in a general framework. Moreover, in case of small samples the procedures are carried out using permutation techniques. Here, the usual assumption of exchangeability under the null hypothesis is not met due to varying follow-up times and unequal overdispersion parameters. This problem is solved by the use of studentized permutations leading to valid inference methods for situations with (i) varying follow-up times, (ii) different overdispersion parameters, and (iii) small sample sizes.

KEYWORDS

permutation methods, resampling, studentized statistics

1 | INTRODUCTION

Metric data and especially count data are common endpoints in clinical trials. Examples include relapses and magnetic resonance imaging (MRI) lesion counts in relapsing-remitting multiple sclerosis (MS), exacerbations in chronic obstructive pulmonary disease (COPD), and hospitalizations in heart failure. For several of these the negative binomial distribution has been suggested to be an appropriate model accounting for between-patient heterogeneity in event rates manifesting in overdispersion, that is variances exceeding the means. For instance, Wang, Meyerson, Tang, and Qian (2009) suggested the negative binomial model for the analyses of relapses, and Sormani et al. (1999, 2001, 2005) and Van den Elskamp, Knol, Uitdehaag, and Barkhof (2009) for various types of MRI lesion counts in MS. Based on two large-scale COPD trials, Keene, Calverley, Jones, Vestbo, and Anderson (2008) assessed various models and recommended the negative binomial model for application. In the situations described above, commonly analyses methods (e.g. PROC GENMOD in SAS) are applied based on large sample properties of underlying Maximum-Likelihood-Estimates (MLE) and the assumption of a common overdispersion parameter across treatment

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2018 The Authors. *Biometrical Journal* Published by WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

groups. Such distributional assumptions, however, can hardly be verified; especially in case of small to moderate sample sizes (Aban, Cutter, & Mavinga, 2009). Even if the distribution is correctly specified the MLEs of the overdispersion parameters are biased (Link & Sauer, 1997; Lord, 2006; Paul & Islam, 1995; Saha, 2011; Saha & Paul, 2005) that may lead to wrong conclusions. Moreover, it is quite common that varying follow-up times occur, see for example, Chen et al. (2013), McCullagh and Nelder (1989). All of the above-mentioned characteristics may not only be shared by count data, but also by metric data measured on an arbitrary scale. Simultaneously accommodating all of these complications in an accurate statistical inference method in a unified way is a rather challenging task. To the best of our knowledge no suitable methods currently exist that can simultaneously handle heteroscedastic data (counts) with varying follow-up times.

It is the aim of the present paper to develop valid inference procedures for the analysis of such data in general models allowing for possibly time-varying follow-up times and different overdispersion parameters in a nonparametric way. This is accomplished by newly derived unbiased estimators (based on the methods of moments) for the (count) rates and their variances. The rigorous study of their large sample properties then leads to asymptotically correct tests and confidence intervals for treatment effects using critical values from the standard normal distribution.

With small samples the use of normal quantiles for inference can lead to liberal or conservative decisions whereas permutation tests offer an opportunity to derive quantiles from appropriate reference distributions. In particular, the application of studentized permutation procedures is tempting since they have been shown to control the type- I -error rate very accurately in various situations (Chung & Romano, 2013; Chung & Romano, 2016; Janssen, 1997; Konietzschke & Pauly, 2014; Pauly, Brunner, & Konietzschke, 2015). The problem in this particular situation is that with varying follow-up times and unequal overdispersion parameters the usual assumption of independently identically distributed (*iid*) observations in the groups is not met. This issue can be solved by applying more general theorems on permutation statistics by Janssen and Pauls (2003) and Janssen (2005) and Pauly (2011). Even though data may not be exchangeable under the null hypothesis, the derived permutation methods are asymptotically correct in that they control the type I error rate or the coverage probability for hypothesis tests and confidence intervals, respectively.

The paper is organized as follows: The statistical model and point estimates are given in Section 2. Unbiased variance estimators are provided in Section 3. In Section 4, test procedures and confidence intervals are derived. Permutation-based small sample size approximations and simulation results are presented in Section 5. Finally, two illustrative data examples are analyzed in Section 6. The paper closes with a discussion of the proposed methods in Section 7. All proofs are given in the supplement to this paper.

2 | STATISTICAL MODEL, POINT ESTIMATES, AND MULTIVARIATE NORMALITY

We consider a general semi-parametric two-sample layout with independent random variables X_{ik} with

$$E(X_{ik}) = t_{ik} \lambda_i \text{ and } Var(X_{ik}) = \sigma_{ik}^2, \quad i = 1, 2, \quad k = 1, \dots, n_i. \quad (1)$$

Here, the index i represents the treatment groups ($i = 1$ control, and $i = 2$ treatment), and k the subject within treatment group i with individual follow-up time t_{ik} , and $\lambda_i > 0$ the expectation of group i . Note that the variance σ_{ik}^2 may depend on t_{ik} , for example if X_{ik} follows a Negative Binomial distribution (in this special case $\sigma_{ik}^2 = t_{ik} \lambda_i + t_{ik}^2 \lambda_i^2 \phi_i$), a Poisson distribution ($\sigma_{ik}^2 = t_{ik} \lambda_i$), or an Exponential distribution (here $\sigma_{ik}^2 = t_{ik}^2 \lambda_i^2$). We further assume that the fourth moments exist and are bounded, that is $\sup_{k \geq 1} E(X_{ik}^4) \leq C_0 < \infty$ for a constant $C_0 > 0$ and $i = 1, 2$.

The design is allowed to be completely heteroscedastic, that is every observation might have a different expectation and variance. All statistical procedures for the analysis of *iid* observations are inappropriate for statistical inference in model (1). Let $N = \sum_{i=1}^2 n_i$ denote the total sample size, $T_i = \sum_{k=1}^{n_i} t_{ik}$ the total follow-up times in group i , $i = 1, 2$, and let $T = \sum_{i=1}^2 T_i$ denote the total follow-up times across both treatment groups. The unknown rate parameters λ_i can be estimated without bias by

$$\hat{\lambda}_i = \frac{1}{T_i} \sum_{k=1}^{n_i} X_{ik} \quad (2)$$

and can be interpreted as a weighted mean of the data. The variance of $\hat{\lambda}_i$ is given by

$$\sigma_i^2 = Var(\hat{\lambda}_i) = \frac{1}{T_i^2} \sum_{k=1}^{n_i} \sigma_{ik}^2. \quad (3)$$

For the derivation of asymptotic results for the rate estimates (2), the following mild regularity conditions on sample sizes and follow-up times are required:

$$t_{ik} \in [L, U] \text{ where } 0 < L < U < \infty, \quad (4)$$

$$N \rightarrow \infty \text{ such that } \frac{n_i}{N} \rightarrow \kappa_i \in (0, 1), \quad (5)$$

$$T \rightarrow \infty \text{ such that } \frac{T_i}{T} \rightarrow \tilde{\kappa}_i \in (0, 1), \quad (6)$$

$$\frac{1}{T_i} \sum_{k=1}^{n_i} \sigma_{ik}^2 \rightarrow \tilde{\tau}_i^2 \in (0, \infty), \text{ as } T_i \rightarrow \infty. \quad (7)$$

Assumption (4) ensures that the follow-up times appear on a fixed time interval of interest, while Assumptions (5)–(7) guarantee the existence of limiting variances of the point estimates, see Theorem 2.1 below. In particular, it follows immediately, that the estimator $\hat{\lambda}_i$ is consistent as $n_i \rightarrow \infty$ and $T_i \rightarrow \infty$, respectively. However, the variance σ_i^2 defined in (3) represents an unknown weighted sequence of the quantities σ_{ik}^2 , which depends on both the follow-up times and sample sizes. Thus, it cannot be represented by model constants. In order to derive inference methods for the general hypothesis $H_0 : h(\lambda_1, \lambda_2) = \theta_0$, however, the estimator needs to be multiplied by adequate known coefficients, such that σ_i^2 converges to a specific variance constant, which is, asymptotically, unaffected by the follow-up times and sample sizes. The result along with the multivariate normality of the estimator $\hat{\lambda} = (\hat{\lambda}_1, \hat{\lambda}_2)'$ of $\lambda = (\lambda_1, \lambda_2)'$ are given in the next theorem.

Theorem 2.1.

(1) Under Assumptions (4), (6), and (7),

$$\sqrt{\frac{T_1 T_2}{T}} (\hat{\lambda} - \lambda) \xrightarrow{D} N(\mathbf{0}, \Sigma), \text{ where } \Sigma = \text{diag} \{ \tilde{\kappa}_2 \tilde{\tau}_1^2, \tilde{\kappa}_1 \tilde{\tau}_2^2 \} \quad (8)$$

is a diagonal limiting covariance matrix.

Note that the diagonal covariance matrix Σ neither depends on the sample sizes n_i , nor on the time-varying coefficients t_{ik} . The matrix, is, however, unknown in practical applications, and needs to be estimated. An unbiased and \mathcal{L}_2 -consistent estimator is derived in the next section.

3 | ESTIMATION OF THE VARIANCE

Moment-based estimators for variances denote, roughly speaking, the squared deviation from the mean. In model (1), however, no uniquely defined mean exists. In particular, the variance σ_i^2 is a sum of variances, and is not defined as a fixed variance constant. Therefore, the usual sample variance moment-based estimator is biased, a rather inappropriate characteristic of a variance estimator. Below, we derive an unbiased and consistent moment-based estimator of σ_i^2 .

Define the random variables $\tilde{Z}_{ik} = X_{ik} - t_{ik} \hat{\lambda}_i$, and note that $E(\tilde{Z}_{ik}) = 0$ for all $i = 1, 2$, and $k = 1, \dots, n_i$. The variables \tilde{Z}_{ik} describe the deviation of X_{ik} to its estimated expectation. An unbiased moment-based estimator can now be derived by considering the squared deviation from \tilde{Z}_{ik} along with a bias correction. Define

$$K_i = \sum_{k=1}^{n_i} \frac{t_{ik}^2}{(T_i - 2t_{ik})T_i} \quad (9)$$

and consider

$$\tilde{\sigma}_i^2 = \frac{1}{(1 + K_i)T_i^2} \sum_{k=1}^{n_i} \frac{T_i}{(T_i - 2t_{ik})} \tilde{Z}_{ik}^2. \quad (10)$$

The estimator $\tilde{\sigma}_i^2$ is not a usual sample variance estimator, since it only involves sums of the follow-up times t_{ik} as weighting factors. However, it describes the mean squared deviation from the observations X_{ik} to their estimated mean $t_{ik}\hat{\lambda}_i$. Further let

$$\hat{\Sigma} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2) = \frac{T_1 T_2}{T} \text{diag}(\tilde{\sigma}_1^2, \tilde{\sigma}_2^2) \tag{11}$$

denote the diagonal matrix with diagonal elements $\frac{T_1 T_2}{T} \tilde{\sigma}_1^2$ and $\frac{T_1 T_2}{T} \tilde{\sigma}_2^2$, respectively. It is shown in the next theorem, that $\tilde{\sigma}_i^2$ is an unbiased estimator of σ_i^2 and that $\hat{\Sigma}$ is \mathcal{L}_2 -consistent.

Theorem 3.1. *For each $i = 1, 2$ the estimator $\tilde{\sigma}_i^2$ is an unbiased estimator of σ_i^2 . Moreover, the estimator $\hat{\Sigma}$ is \mathcal{L}_2 -consistent, that is*

$$\|\hat{\Sigma}\Sigma^{-1} - I_2\|_2^2 \rightarrow 0, T \rightarrow \infty.$$

A detailed proof is given in the supplementary material.

Remark. We note that the variance estimator $\tilde{\sigma}_i^2$ may become negative in “severe” situations, that is if any t_{ik} is way larger than the others. In this case we suggest to use the asymptotically unbiased version

$$\tilde{\sigma}_i^{2*} = \frac{1}{n_i(n_i - 1)} \sum_{k=1}^{n_i} (X_{ik} - t_{ik}\hat{\lambda}_i)^2$$

of $\tilde{\sigma}_i^2$ instead.

The asymptotic normality of the point estimates and the consistent variance estimates can now be used for the derivation of test procedures and confidence intervals.

4 | TEST PROCEDURES AND CONFIDENCE INTERVALS

In this section, different procedures for testing the null hypothesis $H_0 : h(\lambda_1, \lambda_2) = \theta_0$ as well as confidence intervals for the treatment effect $h(\lambda_1, \lambda_2)$ will be discussed, where $h : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ is continuously differentiable in (λ_1, λ_2) . Let $\mathbf{g}(h) = \mathbf{g}(h, \lambda_1, \lambda_2) = (\frac{\partial h}{\partial \lambda_1}, \frac{\partial h}{\partial \lambda_2})'$ denote the gradient of h with estimator $\hat{\mathbf{g}}(h) = \hat{\mathbf{g}}(h, \hat{\lambda}_1, \hat{\lambda}_2) = (\frac{\partial h}{\partial \hat{\lambda}_1}, \frac{\partial h}{\partial \hat{\lambda}_2})'$. It follows from the multivariate delta-method that

$$f \left(h(\hat{\lambda}_1, \hat{\lambda}_2)' - h(\lambda_1, \lambda_2) \right) \xrightarrow{D} N(0, \sigma_h^2), \tag{12}$$

where

$$f = \sqrt{\frac{T_1 T_2}{T}} \text{ and } \sigma_h^2 = (\mathbf{g}(h))' \Sigma \mathbf{g}(h). \tag{13}$$

The variance σ_h^2 is unknown, and must be estimated in practical applications. However, σ_h^2 is a linear combination of the individual variances σ_i^2 , respectively. It follows immediately, that a consistent estimator is given by

$$\hat{\sigma}_h^2 = (\hat{\mathbf{g}}(h))' \hat{\Sigma} \hat{\mathbf{g}}(h). \tag{14}$$

Based on the asymptotic normality of $f(h(\hat{\lambda}_1, \hat{\lambda}_2)' - h(\lambda_1, \lambda_2))$ and Slutsky's Theorem, it thus follows that

$$T_{(h)}(\theta) = f \frac{\left(h(\hat{\lambda}_1, \hat{\lambda}_2) - \theta \right)}{\hat{\sigma}_h} \xrightarrow{D} N(0, 1) \tag{15}$$

where $\theta = h(\lambda_1, \lambda_2)$. For large sample sizes, the null hypothesis $H_0 : h(\lambda_1, \lambda_2) = \theta_0$ will be rejected at a two-sided significance level α , if $|T_{(h)}(\theta_0)| \geq z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ -quantile of the standard normal distribution. Asymptotic

$(1 - \alpha)$ -confidence intervals for θ are obtained from

$$P \left(\theta \in \left[h \left(\hat{\lambda}_1, \hat{\lambda}_2 \right) \pm \frac{z_{1-\alpha/2} \hat{\sigma}_h}{f} \right] \right) \rightarrow 1 - \alpha. \quad (16)$$

5 | SMALL SAMPLE APPROXIMATIONS AND SIMULATION RESULTS

Extensive simulations were conducted to investigate the accuracies of the test procedures derived in Section 4 for small sample sizes with regard to (i) controlling the type-1 error rate at the nominal significance level ($\alpha = 5\%$), (ii) their powers to detect certain alternatives $H_1 : h(\lambda_1, \lambda_2) \neq \theta_0$, and (iii) the coverage probabilities of the corresponding confidence intervals in (16). All simulations were conducted with R environment, version 2.15.2. (R Development Core Team, 2010), each with $n_{sim} = 10,000$ simulation runs.

In all simulations, we focus on testing the hypothesis

$$H_0^{(L)} : h_L(\lambda_1, \lambda_2) = \log(\lambda_1/\lambda_2) = 0 \text{ vs. } H_1^{(L)} : \log(\lambda_1/\lambda_2) \neq 0, \quad (17)$$

corresponding to the function $h(\lambda_1, \lambda_2) = L(\lambda_1, \lambda_2) = \log(\lambda_1/\lambda_2)$. The test statistic is given by

$$T_{(L)} = f \frac{\log(\hat{\lambda}_1/\hat{\lambda}_2)}{\sqrt{\hat{\sigma}_1^2/\hat{\lambda}_1^2 + \hat{\sigma}_2^2/\hat{\lambda}_2^2}}, \quad (18)$$

which yield to asymptotically valid tests $\psi_f = \mathbf{1}\{|T_{(L)}| \leq z_{1-\alpha/2}\}$ for $H_0^{(L)}$. Moreover, confidence intervals can be derived from (16), respectively. Simulation studies indicate, however, that the statistic $T_{(L)}$ in (18) tends to result in rather liberal conclusions for small sample sizes ($n_i \leq 20$). Therefore, we propose a studentized permutation approach to approximate its sampling distribution for small sample sizes. This will be explained in the next section.

5.1 | A studentized permutation approach

Permutation tests are widely known to be robust and exact level α tests when the data are exchangeable. Exchangeability implies, however, that variances across the groups are identical. As mentioned above, the data are allowed to be completely heteroscedastic in model (1). Roughly speaking, a usual permutation test would fail to test the null hypotheses formulated above. However, asymptotic permutation tests can be obtained, if appropriate *studentized statistics* are permuted, which will now be briefly explained: It turns out that the test statistic $T_{(L)}$ follows, asymptotically, a standard normal distribution under the null hypothesis. A permutation or resampling test would now lead to accurate results (at least asymptotically), if the conditional permutation distribution of the test statistic $T_{(L)}$, say F^* , would generally mimic the null distribution of the test statistic. That is, both distributions should at least coincide asymptotically. If that is the case, critical values (or P -values) could be computed from the permutation distribution instead of the standard normal distribution for making inferences. Therefore, the goal of the following investigations is to show that the permutation distribution of $T_{(L)}$, F^* , is indeed the standard normal distribution. In order to do so, some notations and ideas about the permutation schemes are necessary:

Let $\mathbf{X} = (X_{11}, \dots, X_{1n_1}, X_{21}, \dots, X_{2n_2})'$ denote the pooled sample, and let $\mathbf{t} = (t_{11}, \dots, t_{1n_1}, t_{21}, \dots, t_{2n_2})'$ denote the corresponding vector of the pooled follow-up times t_{ik} . For a fixed, but random permutation π of $(1, \dots, N)$, let $\mathbf{X}^\pi = (X_{11}^\pi, \dots, X_{1n_1}^\pi, X_{21}^\pi, \dots, X_{2n_2}^\pi)'$ and $\mathbf{t}^\pi = (t_{11}^\pi, \dots, t_{1n_1}^\pi, t_{21}^\pi, \dots, t_{2n_2}^\pi)'$ denote the permuted data and corresponding follow-up times, respectively.

Permuting \mathbf{X} and \mathbf{t} using the same random permutation π , the permuted values X_{ik}^π and t_{ik}^π are not necessarily independent, which is a rather (at least technically) undesirable property in this context. We therefore propose to permute \mathbf{X} and \mathbf{t} independently. This is similar to two sample problems with right-censored survival data, where it is also recommended that the permuted failure times do not occur in general with their corresponding censoring indicators, see Janssen and Mayer (2003) as well as Brendel, Janssen, Meyer, and Pauly (2014). To this end, we consider another random permutation π' of $(1, \dots, N)$ that is independent of π and calculate the permuted estimators $\hat{\lambda}_i^{(\pi, \pi')} = \hat{\lambda}_i(\mathbf{X}^\pi, \mathbf{t}^{\pi'})$ and $\hat{\sigma}_h^{2(\pi, \pi')} = \hat{\sigma}_h^2(\mathbf{X}^\pi, \mathbf{t}^{\pi'})$. Note that the possible number of random permutation is considerably increased when permuting both \mathbf{X} and \mathbf{t} independently.

It turns out that the distribution of the test statistic $f h(\hat{\lambda}_1, \hat{\lambda}_2)$ differs in the general model (1) from its permutation distribution, and a valid level α test can not be achieved in this setup. Therefore, we consider the distribution of the test statistic $T_{(h)}$ defined in (15) and of the studentized quantity

$$T_{(h)}^{(\pi, \pi')} = f \frac{h(\hat{\lambda}_1^{(\pi, \pi')}, \hat{\lambda}_2^{(\pi, \pi')})}{\hat{\sigma}_h^{(\pi, \pi')}}. \tag{19}$$

The conditional limiting distribution of $T_{(h)}^{(\pi, \pi')}$ given the data \mathbf{X} will be derived in the next theorem.

Theorem 5.1. *Let $T_{(h)}^{(\pi, \pi')}$ as given in (19) and denote by $\Phi(x)$ the standard normal distribution function. If $\sigma_L^2 > 0$, then we have convergence under the null as well as under the alternative with*

$$\sup_{x \in \mathbb{R}} \left| P\left(T_{(L)}^{(\pi, \pi')} \leq x\right) - \Phi(x) \right| \xrightarrow{P} 0.$$

Theorem 5.1 states that the limiting standard normal distribution of $T_{(L)}^{(\pi, \pi')}$ does not depend on the distribution of the data, particularly, it is achieved for arbitrary $h(\lambda_1, \lambda_2) = \theta_0$, that is it even holds under the alternative.

Let $\psi_f^{(\pi, \pi')} = 1\{T_{(h)} \leq z_{\alpha/2}^{(\pi, \pi')}\} + 1\{T_{(h)} \geq z_{1-\alpha/2}^{(\pi, \pi')}\}$, where $z_{\alpha/2}^{(\pi, \pi')}$ denotes the $\alpha/2$ -quantile from the studentized permutation distribution of $T_{(L)}$. In the next theorem, we will show that both the conditional and unconditional tests are asymptotically equivalent, which means, that both tests have, asymptotically, the same power to detect certain alternatives.

Theorem 5.2. *Suppose that the assumptions of Theorem 5.1 are fulfilled.*

1. *Under the null hypothesis $H_0 : h(\lambda_1, \lambda_2) = 0$, the studentized permutation test $\psi_f^{(\pi, \pi')}$ is asymptotically exact at α level of significance, that is $E(\psi_f^{(\pi, \pi')}) \rightarrow \alpha$, and asymptotically equivalent to ψ_f , that is*

$$E\left(\left|\psi_f^{(\pi, \pi')} - \psi_f\right|\right) \rightarrow 0, f \rightarrow \infty.$$

2. *The permutation test $\psi_f^{(\pi, \pi')}$ is consistent, that is we have convergence*

$$E(\psi_f^{(\pi, \pi')}) \rightarrow \alpha 1\{h(\lambda_1, \lambda_2) = 0\} + 1\{h(\lambda_1, \lambda_2) \neq 0\}, f \rightarrow \infty.$$

In particular, Theorem 5.1 states that the distributions of the pivotal quantity $T_{(h)}$ and of the studentized permutation statistic $T_{(h)}^{(\pi, \pi')}$ asymptotically coincide. Under the assumptions of Theorem 5.1, approximate $(1 - \alpha)$ -confidence intervals for θ can be obtained from

$$P\left(\theta \in \left[h(\hat{\lambda}_1, \hat{\lambda}_2) - \frac{z_{1-\alpha/2}^{(\pi, \pi')}}{f} \hat{\sigma}_h, h(\hat{\lambda}_1, \hat{\lambda}_2) - \frac{z_{\alpha/2}^{(\pi, \pi')}}{f} \hat{\sigma}_h \right] \right) \rightarrow 1 - \alpha. \tag{20}$$

5.2 | Simulation results

In a negative binomial- $NB(t_{ik} \lambda_i, \phi_i)$ -model we investigate the empirical control of the preassigned type-1 error rate at the usual two-sided significance level $\alpha = 5\%$ of the statistic $T_{(L)}$ in (18) using the standard normal approximation as given in (15), and the permutation test using the quantiles of the conditional distribution of $T_{(h)}^{(\pi, \pi')}$ in (19) as critical values. As a further competing procedure, we estimate the variances σ_i^2 using maximum likelihood methods. In this $NB(t_{ik} \lambda_i, \phi_i)$ -model the variance σ_i^2 is given by the weighted sequence of the quantities $t_{ik} \lambda_i + t_{ik}^2 \lambda_i^2 \phi_i$, respectively. An intuitive plug-in estimation approach is achieved by replacing the unknown parameter λ_i by $\hat{\lambda}_i$ from above and ϕ_i by a consistent maximum-likelihood estimator (ML) $\hat{\phi}_i$, for example by using

$$\hat{\sigma}_i^2 = \frac{1}{T_i^2} \sum_{k=1}^{n_i} \left\{ t_{ik} \hat{\lambda}_i + t_{ik}^2 \hat{\lambda}_i^2 \hat{\phi}_i \right\}, \tag{21}$$

TABLE 1 Simulated designs, where $m \in \{0, 5, 10, 20, 25\}$ and $\mathbf{n}_1 = (7, 7)'$, $\mathbf{n}_2 = (7, 15)'$

Setting	$\lambda_1 = \lambda_2$	Sizes	Overdisp.	Interpretation
1	1.5	$\mathbf{n} = \mathbf{n}_1 + m$	$\phi = \phi_1$	Balanced/equal overdispersion
2	1.5	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_1$	Unbalanced/equal overdispersion
3	1.5	$\mathbf{n} = \mathbf{n}_1 + m$	$\phi = \phi_2$	Balanced/unequal overdispersion
4	1.5	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_2$	Unbalanced/unequal overdispersion (positive pairing)
5	1.5	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_3$	Unbalanced/unequal overdispersion (negative pairing)
6	10	$\mathbf{n} = \mathbf{n}_1 + m$	$\phi = \phi_4$	Balanced/equal overdispersion
7	10	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_4$	Unbalanced/equal overdispersion
8	10	$\mathbf{n} = \mathbf{n}_1 + m$	$\phi = \phi_5$	Balanced/unequal overdispersion
9	10	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_5$	Unbalanced/unequal overdispersion (positive pairing)
10	10	$\mathbf{n} = \mathbf{n}_2 + m$	$\phi = \phi_6$	Unbalanced/unequal overdispersion (negative pairing)

Here $\phi_1 = (\phi_1, \phi_2)' = (0.3, 0.3)'$, $\phi_2 = (\phi_1, \phi_2) = (0.3, 0.5)'$, $\phi_3 = (\phi_1, \phi_2) = (0.5, 0.3)'$, $\phi_4 = (\phi_1, \phi_2)' = (3, 3)'$, $\phi_5 = (\phi_1, \phi_2) = (3, 5)'$, and $\phi_6 = (\phi_1, \phi_2) = (5, 3)'$ denote vectors of overdispersion parameters and $\mathbf{n}_i + m$ means that every component of \mathbf{n}_i , that is each group size, is increased by m .

see, for example Schneider, Schmidli, and Friede (2013). This estimation approach, however, has the disadvantage that neither $\hat{\lambda}_i^2$ nor $\hat{\phi}_i$ are unbiased estimators of λ_i^2 or ϕ_i , respectively, resulting in biased variance estimators. The variance estimators $\hat{\sigma}_i^2$ used in $T_{(L)}$ are finally replaced by $\hat{\sigma}_i^2$, and the corresponding Wald-statistic, which is asymptotically equivalent to the Likelihood-ratio test, denoted by LRT.

5.2.1 | Type-1 error rate simulations

We explore the behavior of the test statistics for smaller and larger effect rates λ_1 and $\lambda_2 \in \{1.5, 10\}$ as well as smaller and larger overdispersion parameters ϕ_1 and $\phi_2 \in \{0.3, 0.5, 3, 5\}$.

All simulation designs are motivated by the examples presented in Section 6. A major assessment criterion for the accuracy of the procedures is their behavior when increasing sample sizes are combined with increasing variance parameter constellations (positive pairing) or with decreasing variances (negative pairing). We investigate balanced situations with sample size vector $\mathbf{n}_1 = (n_1, n_2)' = (7, 7)$ and unbalanced situations with sample size vector $\mathbf{n}_2 = (n_1, n_2) = (7, 15)'$. The sample sizes are increased by adding a constant m to the components of the vectors \mathbf{n}_1 or \mathbf{n}_2 , respectively. The different simulation settings are displayed in Table 1. Each simulation setting $\mathbf{n} = \mathbf{n}_s(m) = (n_1 + m, n_2 + m)'$ represents a different design with an increasing sample size m , where $s = 1, 2$, see Table 1.

Data were generated from $X_{ik} \sim NB(t_{ik} \lambda_i, \phi_i)$, where t_{ik} denotes the realization from a uniformly distributed random variable $T_{ik} \sim U(1, 2)$, respectively. For each simulation setting, the same generated follow-up times t_{ik} were used for the $n_{sim} = 10,000$ simulation runs, but they were newly generated for each design. The number of random permutations was set to $n_{perm} = 10,000$. The simulated type-1 error rates for a significance level $\alpha = 5\%$ assuming uniformly distributed follow-up times are displayed in Figure 1.

It turns out that in case of small effect rates ($\lambda_1 = \lambda_2 = 1.5$) and small overdispersion parameters the statistics $T_{(L)}$ based on the normal approximation as well as the LRT statistics based on ML tend to be slightly liberal. It can be readily seen from Figure 1 that the permutation tests control the type-1 error rate best, even for extremely small sample sizes. In case of larger effect rates and overdispersion parameters the distribution of the data is much more skewed. In these situations the procedures $T_{(L)}$ based on the normal approximation and ML tend to considerably overreject the null hypothesis $H_0^{(L)}$. Remarkably, the estimated type-1 error rates are even larger than 20% and 10%, respectively in Settings 6–10 (see Figure 1). In comparison, the permutation technique greatly improves the finite sample performance of all asymptotic procedures, and is therefore recommended in practical applications.

In order to investigate the impact of the underlying distributions of the follow-up times, we resimulate the same designs with exponentially distributed follow-up times $T_{ik} \sim Exp(2) + 1$. The results are displayed in the supplementary material. It can be seen that the shape of the underlying follow-up times distributions slightly affect the behavior of the statistics in all scenarios. This is intuitively clear, since the different follow-up times particularly influence the variance of the effect estimators, and increase the variance with wider ranging follow-up times or certain amount of skewness. Therefore, all procedures tend to be slightly more liberal when wide ranging follow-up times and small sample sizes are apparent. This can be particularly seen by the permutation test. The liberality, disappears with increasing sample sizes.

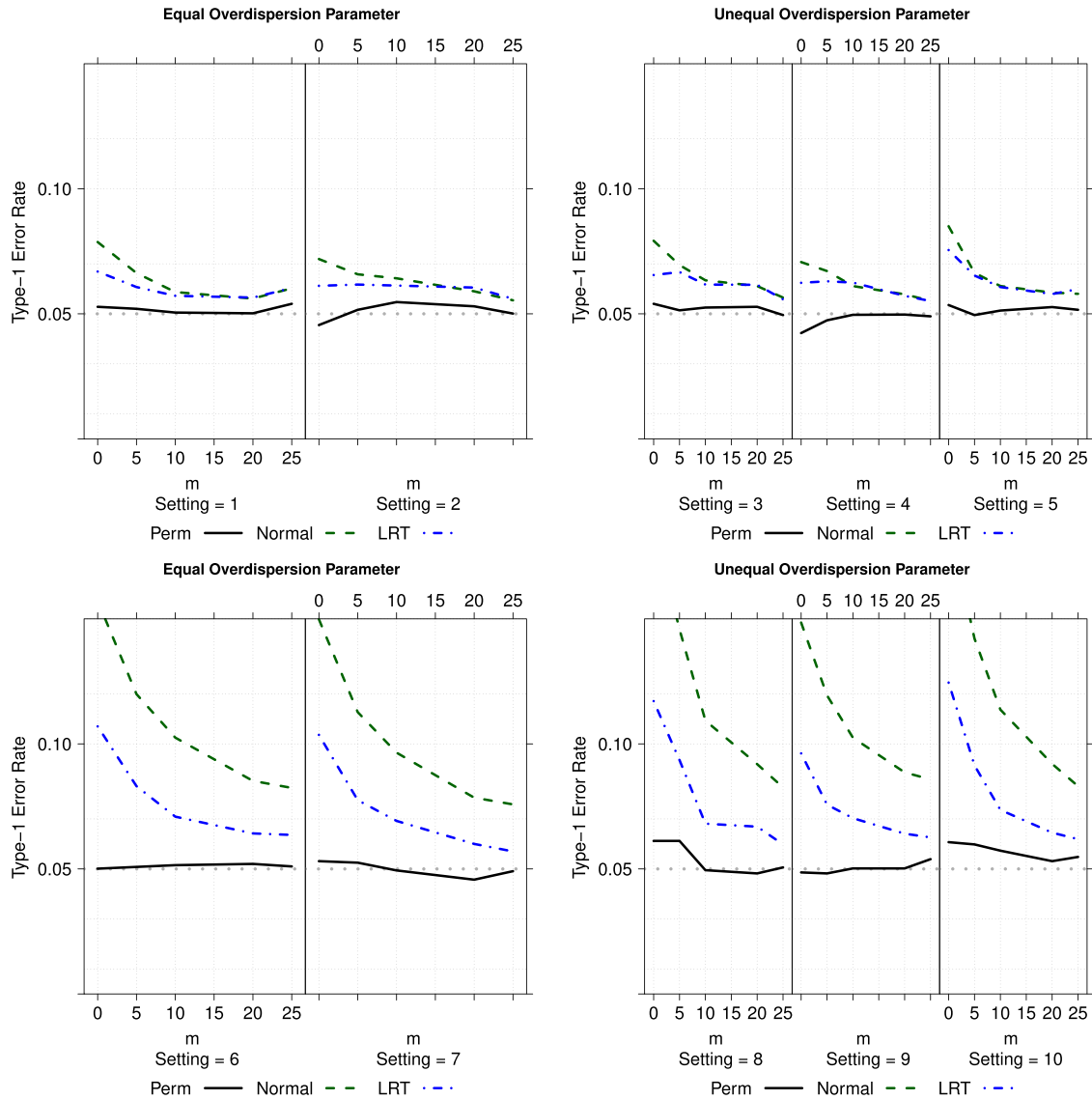


FIGURE 1 Type-I error level ($\alpha = 5\%$) simulation results (y-axis) of the statistics $T_{(L)}$ in (18), permutation test $T_{(h)}^{(\pi, \pi')}$ in (19) and ML-based statistics for different distributions, sample size increments $m \in \{0, 5, 10, 15, 20, 25\}$ (x-axis), where t_{ik} denote the realizations from $T_{ik} \sim U(1, 2)$. The simulation settings are described in Table 1

5.2.2 | Power comparisons

The type-1 error rate simulation results presented in Section 5.2.1 indicate a quite liberal behavior of the methods $T_{(L)}$ and ML-based statistics under certain parameter constellations and small sample sizes. All methods tend to accurate conclusions with large sample sizes. The liberality of these methods increases the “power” of the methods to detect alternatives in small sample size settings. In an additional simulation study, not presented here, it turned out, that with large sample sizes, that is when all competing methods are accurate, their powers are all very similar.

5.2.3 | Simulated coverage rates of the confidence intervals

Next we investigate the empirical coverage probabilities of the corresponding confidence intervals. Data were generated by $X_{1k} \sim NB(\lambda_1 t_{1k}, \phi_1), k = 1, \dots, n_1$ and $X_{2k} \sim NB(\lambda_1(1 + \delta)t_{2k}, \phi_2)$ for varying $\delta \in \{0, 0.1, 0.2, \dots, 2\}$, $n_1, n_2 \in \{10, 20\}$, and different overdispersion parameters. For illustration purposes, we only display the results using uniformly distributed follow-up times, different overdispersion parameters $\phi_1 = 3$ and $\phi_2 = 5$ and rate $\lambda_1 = 10$. The results are displayed in Figure 2. It is readily seen that the competing procedures tend to be rather liberal, while the empirical coverage probabilities of the permutation-based confidence intervals are closer to the nominal level of 95%. The quality of the approximation depends on sample sizes and

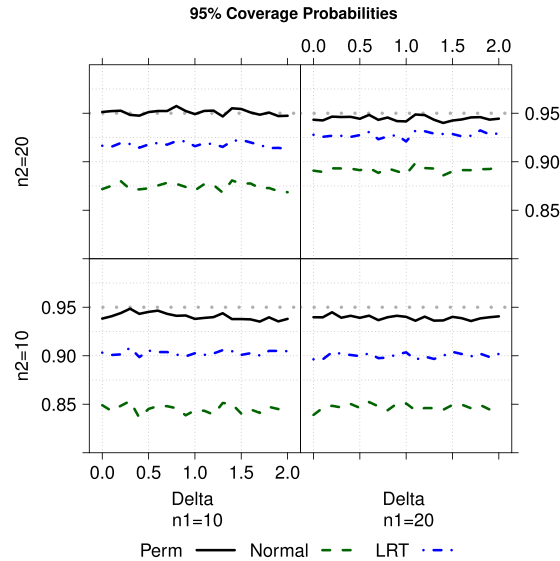


FIGURE 2 Empirical coverage probabilities of nominal 95% confidence intervals of the corresponding confidence intervals given in (16), permutation- based confidence intervals given in (20) and ML-based LRT statistics for different distributions and rate increments $\delta \in \{0, 0.1, \dots, 2\}$ (x-axis) and unequal overdispersion parameters ($\phi_1 = 3, \phi_2 = 5$), where t_{ik} denote the realizations from $T_{ik} \sim U(1, 2)$

TABLE 2 Type-I error level ($\alpha = 5\%$) simulation results of the statistics $T_{(L)}$ in (18) and the permutation test $T_{(L)}^{(\pi, \pi')}$ in (19) using χ^2 -square and exponentially distributed data in different designs, where t_{ik} denote the realizations from $T_{ik} \sim U(1, 2)$

n_1	n_2	$X_{ik} \sim \chi^2_{t_{ik}}$		$X_{ik} \sim \text{Exp}(t_{ik} \cdot 1/2)$	
		$T_{(L)}^{(\pi, \pi')}$	$T_{(L)}$	$T_{(L)}^{(\pi, \pi')}$	$T_{(L)}$
7	7	0.0567	0.1168	0.0440	0.0937
7	15	0.0479	0.1063	0.0373	0.0884
12	12	0.0521	0.0862	0.0500	0.0807
12	20	0.0473	0.0825	0.0364	0.0644
17	17	0.0498	0.0757	0.0355	0.0565
17	25	0.0521	0.0769	0.0540	0.0822
27	27	0.0535	0.0694	0.0854	0.1058
27	35	0.0522	0.0684	0.0454	0.0618
32	32	0.0544	0.0698	0.0469	0.0575
32	40	0.0494	0.0634	0.0526	0.0623

the actual levels of heteroscedasticity across the groups and their allocations. If the larger sample has a smaller variance than the smaller sample ($n_1 = 20, n_2 = 10$), the confidence intervals tend to be slightly liberal for small samples. However, this issue vanishes with increasing sample sizes.

5.2.4 | Simulation results for general metric data

As mentioned in the Introduction and in the description of model (1), data is not required to be count data and thus, numerical investigations of the behavior of the studentized permutation test are intriguing. We therefore investigate the empirical control of the type-1 error rate of the studentized permutation test $T_{(L)}^{(\pi, \pi')}$ in completely heteroscedastic designs with metric data following exponential or χ^2 -distributions. The method will be compared with $T_{(L)}$ using the standard normal approximation. Exponentially distributed variables were generated by $X_{ik} \sim \text{Exp}(t_{ik} \cdot 1/2)$, $i = 1, 2, k = 1, \dots, n_i$, and χ^2 -variables were generated by $X_{ik} \sim \chi^2_{t_{ik}}$, respectively.

The results are displayed in Table 2 and show that the studentized permutation approach controls the nominal type-1 error rate very well and greatly improves the standard normal approximation.

6 | TWO ILLUSTRATIVE EXAMPLES

Pediatric MS with disease onset under the age of 16 is uncommon and qualifies as a rare disease. Differences in clinical presentation before and after puberty have been reported (Huppke et al., 2014). Randomized controlled trials in pediatric MS have been very rare (Unkel et al., 2016), but are becoming more common now (Rose & Müller, 2016). We consider a randomized controlled trial assessing efficacy and safety of interferon beta-1a compared to no treatment in pediatric MS reported by Pakdaman, Fallah, Sahraian, Pakdaman, and Meysamie (2006). In this trial, 16 patients were randomized to verum or control. Relapse rates and new T2 lesions were both considered as endpoints. The estimated rates and overdispersion parameters are given in Table 3. As a second example, we consider the Acyclovir trial reported by Lycke et al. (1996). In this experiment, Acyclovir treatment was used in a randomized, double-blind, placebo-controlled clinical trial with parallel groups to test the hypothesis that herpes virus infections are involved in the pathogenesis of MS. In total, $N = 60$ adult patients were recruited, whereas $n_1 = n_2 = 30$ were randomized to placebo or active treatment, respectively. The data (relapse counts) can be found in Figure 1 in the original publication (Lycke et al., 1996). As a secondary analysis of this trial, the relapse counts from patients that showed a progressive course during the trial were excluded from the statistical analysis. In this situation, patients have different follow-up times and estimators must be weighted accordingly.

The estimated rates and overdispersions being defined as variance-to-mean ratios are given in Table 3. It can be readily seen from Table 3 that the overdispersion parameters seem to differ between the treatment groups, and even underdispersed counts are apparent. The effect of the different overdispersion parameters on the behavior of the statistical methods has been analyzed in detail in extensive simulation studies in Section 5.2.

Both motivating examples discussed above used over- and underdispersed counts as outcomes. Here, we present the results based on standard methods including normal approximation and maximum-likelihood as well as the new developed methods. The test statistic being used is given by

$$T_{NB} = \frac{\log(\hat{\lambda}_1 / \hat{\lambda}_2)}{\sqrt{\frac{\hat{\sigma}_1^{2(c,P)}}{\hat{\lambda}_1^{2(c)}} + \frac{\hat{\sigma}_2^{2(c,P)}}{\hat{\lambda}_2^{2(c)}}}}, \tag{22}$$

where

$$\hat{\sigma}_i^2 = \frac{1}{T_i^2} \sum_{k=1}^{n_i} \left\{ t_{ik} \hat{\lambda}_i + t_{ik}^2 \hat{\lambda}_i^2 \hat{\phi} \right\},$$

denotes the estimated variance of the effect estimator using a MLE estimator of the overdispersion parameter ϕ , which is assumed to be identical across both treatment groups.

As competing methods, we also analyze the data using both a Negative Binomial Regression- and Poisson Regression using *SAS PROC GENMOD*.

TABLE 3 Estimated rates and overdispersion parameters (Variance / Mean Ratio) for the two example studies

Endpoint	Group	Estimated rate $\hat{\lambda}_i$	Sample variance	Estimated overdispersion
Pediatric MS trial ($N=16$)				
T2 lesions	Control	11.875	13.268	1.117
	Active	10.625	16.839	1.585
Relapses	Control	4.5	6.571	1.460
	Active	2.375	0.268	0.113
Acyclovir trial ($N=60$)				
Relapses	Control	3.133	6.602	2.107
	ACYC	2.067	3.030	1.466
Acyclovir trial ($N=60$; Secondary analysis)				
Relapses	Control	3.205	6.602	2.060
	ACYC	2.118	3.172	1.498

TABLE 4 Statistical analysis of the examples using $h(\lambda_1, \lambda_2) = \log(\lambda_1/\lambda_2)$: Approximate method, Effect ($\log(\hat{\lambda}_1/\hat{\lambda}_2)$), Standard Error (SE), Test Statistic (= Effect / SE), and 95% confidence intervals

Method	Effect	SE	Statistic	P-value	95% CI
T2 lesions					
Normal (15)	0.111	0.174	0.638	0.524	(-0.231; 0.453)
LRT (21)	0.111	0.162	0.686	0.493	(-0.207; 0.429)
LRT.Pool (22)	0.111	0.161	0.691	0.489	(-0.204; 0.427)
Perm (19)	0.111	0.174	0.638	0.545	(-0.269; 0.510)
NB-Reg	0.111	0.161	0.691	0.489	(-0.204; 0.428)
Pois-Reg	0.111	0.149	0.745	0.456	(-0.181; 0.405)
Relapses					
Normal (15)	0.639	0.216	2.964	0.003	(0.216; 1.062)
LRT (21)	0.639	0.302	2.116	0.034	(0.047; 1.231)
LRT.Pool (22)	0.639	0.284	2.254	0.024	(0.083; 1.195)
Perm (19)	0.639	0.216	2.964	0.026	(0.116; 1.162)
NB-Reg	0.639	0.284	2.254	0.024	(0.096; 1.215)
Pois-Reg	0.639	0.284	2.254	0.024	(0.096; 1.215)
Acyclovir relapses					
Normal (15)	0.416	0.215	1.939	0.052	(-0.004; 0.837)
LRT (21)	0.416	0.228	1.824	0.068	(-0.031; 0.863)
LRT.Pool (22)	0.416	0.231	1.805	0.071	(-0.036; 0.868)
Perm (19)	0.416	0.215	1.939	0.054	(-0.007; 0.842)
NB-Reg	0.416	0.231	1.805	0.071	(-0.035; 0.870)
Pois-Reg	0.416	0.164	2.544	0.011	(0.098; 0.741)
Acyclovir relapses (Secondary analysis)					
Normal (15)	0.414	0.218	1.904	0.057	(-0.012; 0.841)
LRT (21)	0.414	0.230	1.798	0.072	(-0.037; 0.866)
LRT.Pool (22)	0.414	0.233	1.781	0.075	(-0.076; 0.845)
Perm (19)	0.414	0.218	1.904	0.062	(-0.022; 0.845)
NB-Reg	0.415	0.233	1.780	0.075	(-0.040; 0.874)
Pois-Reg	0.422	0.165	2.553	0.011	(0.101; 0.750)

Thus, the illustrative examples include constant as well as varying follow-up times, and even the analyses with constant follow-up times still presents a challenge since the sample sizes are with 16 and 60 very and moderately small, and the overdispersion is fairly pronounced, in particular for the MRI lesion counts and relapses. The effect estimates, standard errors, test statistics, P -values as well as 95%-confidence intervals are displayed in Table 4.

It can be readily seen from Table 4, that the estimated standard errors of the effect estimates for the T2 lesions are likely, and therefore all methods results in the same conclusion. Only the estimated standard error being computed via a Poisson-Regression tends to be smaller. This occurs because the Poisson-Regression sets the overdispersion to be zero, by default. A significant effect at 5% level can not be detected with any method ($P > 0.05$). The relapse rates are significantly different at 5%-level of significance. It can be seen, however, that the estimates of the standard errors significantly differ from the moment-based unbiased variance estimators (SE = 0.216 vs. SE = 0.302 using ML). Therefore, the P -values based on ML estimates are larger than using the moments-based estimator and standard normal distribution ($P = 0.003$ vs. $P = 0.034$). However, since sample size is rather small, the permutation approach is the most robust method in this setup, and results in a P -value of $P = 0.026$. Since both over- and underdispersed counts were observed, the ML.Pool, the negative binomial, and poisson regression are tend to provide identical results.

The results obtained for the Acyclovir trial, however, differ significantly. First, both treatment groups show a different overdispersion. Therefore, the SE obtained by a Poisson-Regression is way smaller than with all other methods, and thus results in a significant treatment effect at 5% level of significance. Comparing the other estimation approaches it can be seen that the ML-based estimation approaches (assuming negative binomial distribution) of the SE tend to be larger than the unbiased

methods-of-moments based methods. The largest SE is estimated via ML.Pool (which is identical to a NB-Regression). The estimated standard error based on the unbiased variance estimate is given by $SE = 0.215$. Therefore, the P -values range from 0.052 through 0.071. Due to the moderate sample size of $N = 60$, both the normal and permutation approximation tend to provide similar P -values with $P = 0.052$ and $P = 0.054$, respectively. The secondary analysis of the the Acyclovir trial shows similar results to the above. This occurs because only the relapse counts from four of the 60 patients were excluded from the analysis. However, slightly different effect estimates coming from the Negative Binomial and Poisson Regression can be seen. This occurs, because in case of unequal follow-up times the rates are estimated using maximum likelihood estimation methods, which are not identical to moment (mean-based) methods.

7 | DISCUSSION

In this paper, inference methods for testing hypotheses formulated in terms of the effect rates of overdispersed counts were developed without assuming a specific data distribution and/or different overdispersion parameters. They are based on the asymptotic properties of novel unbiased estimators of the count rates and their variances. In order to provide valid methods for small sample sizes, resampling methods have been derived. Although data is in general not exchangeable, following the ideas of Neuhaus (1993), Janssen (1997, 2005), and Chung and Romano (2013), studentized permutation techniques could be applied. Simulation studies indicate, however, that the procedures control the nominal level reasonably well even with $n_i \approx 5$.

Furthermore, in clinical trials, the computation of confidence intervals for the treatment effects is important, following the ICH E9 guideline for randomized clinical trials: “*Estimates of treatments shall be accompanied by confidence intervals, whenever possible*” (ICH E9 Guideline 1998, chap. 5.5, p. 25). For instance, Saha (2013) investigates different methods for the computation of confidence intervals for the mean difference in the analysis of overdispersed count data (assuming constant follow-up times t_{ik}). In this paper, these procedures were generalized for possibly time-varying and overdispersed count data and equipped with the studentized permutation approach. Extensive simulation studies show that the new methods improve the existing methods in terms of coverage probability and type- I -error rate control. Furthermore, we only considered one possible unbiased estimator of the rates λ_i by $\hat{\lambda}_i = \frac{1}{T_i} \sum_{k=1}^{n_i} X_{ik}$, which is known as a weighted mean estimator. Another unbiased estimator is given by the unweighted mean $\hat{\lambda}_i^{(u)} = \frac{1}{n_i} \sum_{k=1}^{n_i} \frac{X_{ik}}{t_{ik}}$, or least-square based estimators $\hat{\lambda}_i = (\mathbf{t}_i' \mathbf{t}_i)^{-1} \mathbf{t}_i' \mathbf{X}_i$, where $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})'$ and $\mathbf{X}_i = (X_{i1}, \dots, X_{in_i})'$ denote the vectors of follow-up times and response per group i , respectively. Investigating and comparing those estimators and generalizations thereof is tempting and will be subject to future research.

In future investigations, the results shall be extended to more general models allowing for covariates (e.g. for baseline adjustment) and several samples. Furthermore, investigating the overlap of range-preserving confidence intervals for the effects is an interesting attempt for making inferences (Noguchi & Marmolejo-Ramos, 2016).

ACKNOWLEDGMENTS

T.F. is grateful for funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under grant agreement number FP HEALTH 2013 - 602144. Moreover, M.P. was supported by the German Research Foundation project DFG-PA 2409/3-1.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Frank Konietzschke  <https://orcid.org/0000-0002-5674-2076>

Tim Friede  <https://orcid.org/0000-0001-5347-7441>

REFERENCES

- Aban, I. B., Cutter, G. R., & 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, 820–833.
- Brendel, M., Janssen, A., Meyer, C.-D., & Pauly, M. (2014). Weighted logrank permutation tests for randomly right censored life science data. *Scandinavian Journal of Statistics*, 41, 742–761.

- Chung, E. Y., & Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, 41, 484–507.
- Huppke, B., Ellenberger, D., Rosewich, H., Friede, T., Gärtner, J., & Huppke, P. (2014). Clinical presentation of pediatric multiple sclerosis before puberty. *European Journal of Neurology*, 21, 441–446.
- ICH (1998). Statistical Principles for Clinical Trials. Guideline.
- Janssen, A. (1997). Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem. *Statistics in Probability Letters*, 36, 9–21.
- Janssen, A. (2005). Resampling student's t-type statistics. *Annals of the Institute of Statistical Mathematics*, 57, 507–529.
- Janssen, A., & Mayer, C.-D. (2001). Conditional studentized survival tests for randomly censored models. *Scandinavian Journal of Statistics*, 28, 283–293.
- Janssen, A., & Pauls, T. (2003). How do Bootstrap and permutation tests work? *Annals of Statistics*, 31, 768–806.
- Keene, O. N., Calverley, P. M. A., Jones, P. W., Vestbo, J., & Anderson, J. A. (2008). Statistical analysis of exacerbation rates in COPD: TRISTAN and ISOLDE revisited. *European Respiratory Journal*, 32, 17–24.
- Link, W. A., & Sauer, J. R. (1997). Estimation of population trajectories from count data. *Biometrics*, 51, 488–497.
- Lord, D. (2006). Modeling motor vehicle crashes using Poisson-gamma models: examining the effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter. *Accident Analysis & Prevention*, 38, 751–766.
- Lycke, J., Svennerholm, B., Hjelmquist, E., Frisin, L., Badr, G., Andersson, M., & Andersen, O. (1996). Acyclovir treatment of relapsing-remitting multiple sclerosis. *Journal of Neurology*, 243, 214–224.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. London: Chapman and Hall/CRC.
- Neuhaus, G. (1993). Conditional rank tests for the two-sample problem under random censorship. *Annals of Statistics*, 21, 1760–1779.
- Noguchi, K., & Marmolejo-Ramos, F. (2016). Assessing equality of means using the overlap of range-preserving confidence intervals. *The American Statistician*, 70(4), 325–334.
- Pakdaman, H., Fallah, A., Sahraian, M. A., Pakdaman, R., & Meysamie, A. (2006). Treatment of early onset multiple sclerosis with suboptimal dose of interferon beta-1a. *Neuropediatrics*, 37(4), 257–260.
- Paul, S. R., & Islam, A. S. (1995). Analysis of proportions in the presence of over-/under-dispersion. *Biometrics*, 51, 1400–1410.
- Pauly, M. (2011). Weighted resampling of martingale difference arrays with applications. *Electronic Journal of Statistics*, 5, 41–52.
- Pauly, M., Brunner, E., & Konietzschke, F. (2015). Asymptotic permutation tests in general factorial designs. *Journal of the Royal Statistical Society: Series B*, 77, 461–473.
- Rose, K., & Müller, T. (2016). Children with multiple sclerosis should not become therapeutic hostages. *Therapeutic Advances in Neurological Disorders*, 9(5), 389–395.
- Saha, K. K. (2011). Interval estimation of the overdispersion parameter in the analysis of one way layout of count data. *Statistics in Medicine*, 30, 39–51.
- Saha, K. K. (2013). Interval estimation of the mean difference in the analysis of over-dispersed count data. *Biometrical Journal*, 55(1), 114–133.
- Saha, K., & Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics*, 61, 179–185.
- Schneider, S., Schmidli, H., & Friede, T. (2013). Robustness of methods for blinded sample size re-estimation with overdispersed count data. *Statistics in Medicine*, 32, 3623–3635.
- Sormani, M. P., Bruzzi, P., Beckmann, K., Kappos, L., Miller, D. H., Polman, C., & Filippi, M. (2005). The distribution of magnetic resonance imaging response to interferon 1b in multiple sclerosis. *Journal of Neurology*, 252, 1455–1458.
- Sormani, M. P., Bruzzi, P., Miller, D. H., Gasperini, C., Barkhof, F., & 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, 74–80.
- Sormani, M. P., Bruzzi, P., Rovaris, M., Barkhof, F., Comi, G., Miller, D. H., & Filippi, M. (2001). Modelling new enhancing MRI lesion counts in multiple sclerosis. *Multiple Sclerosis*, 7, 298–304.
- Unkel, S., Röver, C., Stallard, N., Benda, N., Posch, M., Zohar, S., & Friede, T. (2016). Systematic reviews in paediatric multiple sclerosis and Creutzfeldt-Jakob disease exemplify shortcomings in methods used to evaluate therapies in rare conditions. *Orphanet Journal of Rare Diseases*, 11, 16.
- Van den Elskamp, I. J., Knol, D. L., Uitdehaag, B. M. J., & Barkhof, F. (2009). The distribution of new enhancing lesion counts in multiple sclerosis: Further explorations. *Multiple Sclerosis*, 15, 42–49.
- Wang, Y. C., Meyerson, L., Tang, Y. Q., & Qian, N. (2009). Statistical methods for the analysis of relapse data in MS clinical trials. *Journal of the Neurological Sciences*, 285, 206–211.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Konietschke F, Friede T, Pauly M. Semi-parametric analysis of overdispersed count and metric data with varying follow-up times: Asymptotic theory and small sample approximations. *Biometrical Journal*. 2019;61:616–629. <https://doi.org/10.1002/bimj.201800027>