# Testing noninferiority in three-armed clinical trials based on the likelihood ratio statistics

**A. Munk**

*Department of Mathematical Stochastics, University Göttingen*
*e-mail:* munk@math.uni-goettingen.de

**M. Mielke**

*Department of Mathematical Stochastics, University Göttingen*
*e-mail:* mmielke@math.uni-goettingen.de

**G. Skipka**

*Institute for Quality and Efficiency in Health Care*
*e-mail:* guido.skipka@iqwig.de

**G. Freitag**

*Department of Mathematical Stochastics, University Göttingen*
*e-mail:* freitag@math.uni-goettingen.de

**Abstract:** Clinical noninferiority trials with three (or more) groups recently have received much attention, e.g. due to the fact that regulatory agencies often require that a placebo group has to be evaluated in addition to a new experimental drug and an active control. We discuss the likelihood ratio tests for binary endpoints and various noninferiority hypotheses. We find that, depending on the particular hypothesis, either the LR test reduces asymptotically to the intersection union test, or to a test which follows asymptotically a mixture of generalized $\chi^2$-distributions. The performance of this asymptotic is investigated and an exact modification is given. It is shown that this test considerably outperforms multiple testing methods with respect to power, such as the Bonferroni adjustment. The methods are illustrated with a cancer study where antiemetic agents were compared. Finally, we discuss the extension of the results to other settings, such as normal endpoints.

**AMS 2000 subject classifications:** Primary 62F05; secondary 62P10, 62H15, 62F30.
**Keywords and phrases:** Exact methods, Multiple testing, Noninferiority, Order restricted inference, Therapeutic equivalence, Three-arm clinical trials.

## 1. Introduction

Clinical trials to show therapeutic equivalence of two treatments are well established since more than a decade (Hesketh *et al.* 1996; Diehm, Trampisch, Lange & Schmidt 1996; Tebbe *et al.* 1998; Chouela *et al.* 1999; Dammann *et al.* 2000). Here in most cases, a new therapy (e.g. a new treatment or a new dose of a drug) has to be shown to be not relevantly inferior to an active control with respect to a primary clinical endpoint. Closely related to this are superiority trials, where the aim is to show a relevant superiority of the new treatment compared to a standard (Chan 1998; Chuang-Stein 2001; Dunnett & Tamhane 1997; Greco *et al.* 1996; Gustafsson *et al.* 1996; Röhmel & Mansmann 1999). For binary outcomes, and we will deal mainly with this situation in this paper, a considerable amount of statistical methods have been developed since the pioneering work of Dunnett & Gent (1977). This includes asymptotic procedures (Blackwelder 1982; Rodary, Com-Nougue & Tournade 1989; Farrington & Manning 1990) as well as exact methods, which aim for keeping the nominal level exactly for finite sample sizes (see e.g. Chan 1998; Röhmel & Mansmann 1999; Martín Andrés & Herranz Tejedor 2004).

*Preprint submitted to *The Canadian Journal of Statistics (23 March 2006), Accepted 24 April 2007.*

Less work is available on showing noninferiority in three-arm clinical trials, which, however, has become a task of great practical interest, due to the fact that it is often required to include an additional placebo group to guarantee the assay sensitivity, i.e. the ability of a trial to evaluate the efficacy of the new treatment. This is important in cases where one cannot rely on the so-called constancy assumption, which would allow the use of historical data for estimating the effect of the active control treatment. This issue is also highlighted in Rothmann, Li, Chen, Chi, Temple & Tsou (2003) and by recent guidelines (Committee for Proprietary Medicinal Products 1998a,b, 2001, 2002a,b). In general, methods for three or more samples are based on multiple testing procedures, such as the intersection union principle. See e.g. Wiens & Iglewicz (1999) and Tang & Tang (2004) for an account of methods for the assessment of noninferiority with binary data, or Pigeot, Schäfer, Röhmel & Hauschke (2003) for the case of normal responses.

In this paper we will exploit the likelihood ratio (LR) test for the case of multiple samples. We focus primarily on the case of three samples and binary outcomes, however in Section 7 we will briefly discuss extensions to other settings. The paper is organized as follows. The asymptotic theory for general null hypotheses of the type

$$H^U : \vartheta_3 \geq h_1(\vartheta_1) \quad \text{or} \quad \vartheta_3 \geq h_2(\vartheta_2) \tag{1.1}$$

or of the type

$$H^I : \vartheta_3 \geq h_1(\vartheta_1) \quad \text{and} \quad \vartheta_3 \geq h_2(\vartheta_2) \tag{1.2}$$

will be considered in the next section. Here the parameter $\vartheta_i$ represents a failure rate or a success probability under treatment $i$, $i = 1, ..., 3$. Many clinical problems can be expressed by these types of hypotheses. For specific choices of $h_1$ and $h_2$ this includes e.g. hypotheses on the differences, the relative risks or the odds ratios of the parameters.

The type of hypothesis $H^U$ can be used to show that a new treatment is "as effective as" (in the sense of being not relevantly inferior to) both of two standards, or that two new treatments are as effective as a standard. Moreover, the problem of showing both the superiority of a standard treatment as compared to placebo and the noninferiority of a test treatment as compared to the standard can be also formulated by using a null hypothesis of "union type". More precisely, in the latter case we obtain a testing problem

$$\tilde{H}^U : \vartheta_3 \geq \vartheta_1 \quad \text{or} \quad \vartheta_3 \leq h_2(\vartheta_2). \tag{1.3}$$

Hence, the theoretical point of view, problems (1.1) and (1.3) can be treated in essentially the same way. For the sake of brevity we will only describe the results for (1.1) in the following.

In contrast, the null hypothesis $H^I$ in (1.2) is suitable for showing that a new treatment is as effective as one of two standards, or that one of two new treatments is as effective as a standard. For example, Hesketh *et al.* (1996) aim at showing that a new treatment at one of two different doses is as effective as a standard one.

In Section 2 we will show that the LR test of $H^U$ is asymptotically the same as performing independently two LR tests for the single null hypotheses $H^1$ and $H^2$, respectively, where

$$H^1 : \vartheta_3 \geq h_1(\vartheta_1), \qquad H^2 : \vartheta_3 \geq h_2(\vartheta_2), \tag{1.4}$$

which yields the intersection union test (IUT; cf. Berger 1997). Thus, we can immediately use the two-sample results from Munk, Skipka & Stratmann (2005) for this case. For $H^I$ the situation is quite different and a rather complicated asymptotic law results which, among others, depends on the specific functions $h_1$, $h_2$.

For practical purposes, in Section 4 we suggest for large sample sizes an asymptotic modification and for small sample sizes an exact modification of the asymptotic test of $H^I$, which is based on the cumulative likelihood ratios.

In an extensive numerical study in Section 5, it is shown that competitors, such as Bonferroni adjusted tests for $H^I$, are outperformed in terms of power for most parameter settings

in the alternative. This may lead to a considerable reduction of sample size when planning the study. Numerically we found that a reduction of required sample sizes of up to 20% can be achieved in certain settings. The performance of the proposed tests for $H^U$ follows immediately from the results on the two-sample case as presented in Skipka, Munk & Freitag (2004) or Munk, Skipka & Stratmann (2005).

At a first glance, it might be considered as a general drawback of any global test for (1.2), that it does not provide us with the information as to which of the sub-hypotheses in (1.4) can be rejected. However, an application of the closed testing principle will show that a posteriori the pairwise comparisons can be performed in addition to this test, if $H^I$ can be rejected, while still the nominal level $\alpha$ is maintained. This is discussed in Section 3.

Finally, in Section 6 the LR approach is illustrated by means of the data from a cancer trial, where three antiemetic treatments were compared.

The computation of the exact tests is numerically rather involved and SAS code can be obtained from the authors on request. In order to keep the paper readable, all proofs are deferred to the Appendix.

## 2. The likelihood ratio statistics – asymptotic results

Let us consider throughout the following three independent Bernoulli samples $X_{ij} \sim B(1, \vartheta_i)$ with failure rates $\vartheta_i$ and sample sizes $n_i$ $(j = 1, \ldots, n_i,\ i = 1, 2, 3)$. For the functions $h_i : [0, 1] \to [0, 1]$, $i = 1, 2$ in (1.1) and (1.2), we assume that they are strictly isotonic and twice differentiable. This includes hypotheses on the difference $(h^{DI}(\vartheta) = \vartheta + \theta_0)$, for the relative risk $(h^{RR}(\vartheta) = \vartheta\theta_0)$, or for the odds ratio $(h^{OR}(\vartheta) = \frac{\theta_0}{\theta_0+\vartheta^{-1}-1})$.

More general, the $h_i$ might take into account also combinations of different measures of discrepancies as well as different values of $\theta_0$, depending on the underlying response rate (see Röhmel & Mansmann 1999, for a careful discussion of this issue). The threshold parameter $\theta_0$ will subsume the maximum clinically relevant amount, which has to be fixed in advance. The choice of the $h_i$ and of $\theta_0$ is a difficult and important task and will depend on the particular medical application. We will not pursue this issue here, for a careful discussion we refer to Lange (2003) and Lange & Freitag (2005), and the references given there.

Let $\vartheta = (\vartheta_1, \vartheta_2, \vartheta_3)^\top$ and $x = (x_1, x_2, x_3)^\top$, $x_i = \sum_{j=1}^{n_i} x_{ij}$; $i = 1, 2, 3$, then the likelihood is given as

$$L_x(\vartheta) = \prod_{i=1}^{3} \binom{n_i}{x_i} \vartheta_i^{x_i} \left(1 - \vartheta_i\right)^{n_i - x_i} \quad . \tag{2.1}$$

Further, let $\hat{\vartheta} = (x_i/n_i)_{i=1,2,3}$ the vector of the unconstrained maximum likelihood estimators (MLE), whereas the restricted MLE to a hypothesis $H$ (RMLE) will be denoted as

$$\hat{\vartheta}^* \in \{\arg\max_{\vartheta \in H} L_x(\vartheta)\} \ .$$

In the following we investigate the likelihood ratio statistic

$$T = T(x) := 2[\log L_x(\hat{\vartheta}) - \log L_x(\hat{\vartheta}^*)] \ . \tag{2.2}$$

for the hypotheses in (1.1) and (1.2).

### *2.1. The hypothesis $H^U$*

We will show in this section that for $H^U$ in (1.1) the LR statistic is the maximum of the two-sample LR statistics for $H^1$ and $H^2$, respectively (cf. (1.4)). Further, we will see that asymptotically, the three sample case for $H^U$ will be reduced to the two-sample case, which has been treated in Munk, Skipka & Stratmann (2005). From there we require the following

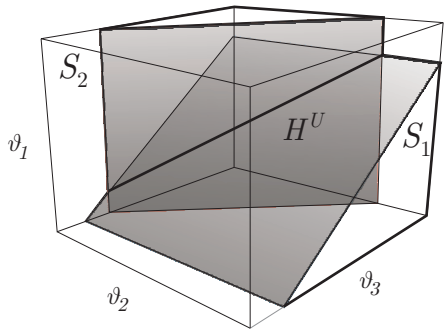


FIG 1. *Boundaries of the null spaces for $H^U$ for the difference.*

result.

THEOREM 1. *Let $\Theta_0 = \{\vartheta \in [0,1]^2 : \vartheta_1 \geq h(\vartheta_2)\}$ and $\Theta_0^h = \{\vartheta \in [0,1]^2 : \vartheta_1 = h(\vartheta_2)\}$. Assume that $X_1 \sim B(n_1, \vartheta_1)$ and $X_2 \sim B(n_2, \vartheta_2)$ are independent, where $n_1, n_2 \geq 1$. Let $h : [0,1] \to [0,1]$ be continuous and increasing, and not identically 1. Denote the two-sample likelihood as*

$$L_{x,2}(\vartheta) = \binom{n_1}{x_1} \vartheta_1^{x_1} (1-\vartheta_1)^{n_1-x_1} \binom{n_2}{x_2} \vartheta_2^{x_2} (1-\vartheta_2)^{n_2-x_2} .$$

*Then*
*a) the MLE restricted to $\Theta_0$ exists and is given as $\hat{\vartheta}^* = \hat{\vartheta}$ (the unrestricted MLE) if $\hat{\vartheta} \in \Theta_0$ and if $\hat{\vartheta} \notin \Theta_0$ as*

$$\hat{\vartheta}^* = \left\{ \arg \max_{\{\vartheta : \vartheta_1 = h(\vartheta_2)\}} L_{x,2}(\vartheta) \right\} \subseteq \Theta_0^h, \tag{2.3}$$

*i.e. the RMLE is attained on the boundary curve $\Theta_0^h$ of $\Theta_0$.*
*b) If further $h \in C^{(1)}[0,1]$, we have for $\vartheta_1 = h(\vartheta_2)$ and for any solution $\hat{\vartheta}^*$*

$$-2 \ln \lambda \xrightarrow{\mathcal{D}} U \sim \tfrac{1}{2} + \tfrac{1}{2} F_{\chi_1^2},$$

*as $\min\{n_1, n_2\} \to \infty$, s.t. $\frac{n_1}{n_2} \to c \in (0, \infty)$, where $\lambda = L_{x,2}(\hat{\vartheta}) / L_{x,2}(\hat{\vartheta}^*)$ denotes the likelihood ratio and $F_{\chi_1^2}$ the c.d.f. of the square of a standard normal random variable.*

We mention that for many common measures of discrepancy $h$ the two-sample RMLE can be computed explicitly (Miettinen & Nurminen 1985; Skipka, Munk & Freitag 2004), otherwise numerical methods can be used. Theorem 1 allows one to simplify the computational effort significantly, because the maximum of the likelihood $L_{x,2}$ over the two-dimensional set $\Theta_0$ reduces to maximization over a curve, where $\vartheta_1 = h(\vartheta_2)$. Furthermore, sufficient conditions on $h$ for the uniqueness of the RMLE $\hat{\vartheta}^*$ are given in Munk, Skipka & Stratmann (2005), which apply in the following as well.

We denote the two-sample RMLEs (for a boundary function $h$) as

$$\hat{\vartheta}^*_{n_i, n_j, x_i, x_j, h} := \arg\max_{\vartheta} \vartheta^{x_i} (1-\vartheta)^{n_i - x_i} (h(\vartheta))^{x_j} (1 - h(\vartheta))^{n_j - x_j} \ , \quad i \neq j \ . \tag{2.4}$$

Now, let us return to the three-sample case. Obviously, if $\hat{\vartheta} \in H^U$, then $\hat{\vartheta}^* = \hat{\vartheta}$, thus the LR statistic equals zero. Thus, let $\hat{\vartheta} \notin H^U$. The boundaries of the pairwise null spaces (cf. Figure 1) are denoted by

$$S_1 := \{\vartheta \in H^U | \vartheta_3 = h_1(\vartheta_1)\} \ , \quad S_2 := \{\vartheta \in H^U | \vartheta_3 = h_2(\vartheta_2)\} \ . \tag{2.5}$$

Let $\Theta_U^h = \{\vartheta \in S_1 | \vartheta_3 < h_2(\vartheta_2)\} \cup \{\vartheta \in S_2 | \vartheta_3 \leq h_1(\vartheta_1)\}$. Since

$$\max_{\vartheta \in H^U} L_x(\vartheta) = \max_{\vartheta \in \Theta_U^h} L_x(\vartheta) \leq \max_{S_1 \cup S_2} L_x(\vartheta) \leq \max_{\vartheta \in H^U} L_x(\vartheta) \ ,$$

where the above equal sign follows as in the proof of Theorem 1 in Munk, Skipka & Stratmann (2005), the maximum over $H^U$ is calculated by

$$\max_{\vartheta \in \Theta_U^h} L_x(\vartheta) = \max_{S_1 \cup S_2} L_x(\vartheta) = \max\{\max_{S_1} L_x(\vartheta), \max_{S_2} L_x(\vartheta)\} \ .$$

The parameter $\vartheta_2$ is unconstrained in $S_1$, and $\vartheta_1$ is unconstrained in $S_2$. Thus, the MLE constrained to $H^U$ is obtained from one of the two-sample RMLEs (cf. (2.4)),

$$\begin{aligned} \hat{\vartheta}^*_{S_1} &:= (\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1} \ , \ \hat{\vartheta}_2 \ , \ h_1(\hat{\vartheta}^*_{n_1,n_3,x_1,x_3,h_1}))^\top \ , \\ \hat{\vartheta}^*_{S_2} &:= (\hat{\vartheta}_1 \ , \ \hat{\vartheta}^*_{n_2,n_3,x_2,x_3,h_2} \ , \ h_2(\hat{\vartheta}^*_{n_2,n_3,x_2,x_3,h_2}))^\top \ . \end{aligned}$$

Therefore, the test statistic (2.2) is given by

$$T = 2[\ln L_x(\hat{\vartheta}) - \ln \max\{L_x(\hat{\vartheta}^*_{S_1}), L_x(\hat{\vartheta}^*_{S_2})\}] = \min\{T_1, T_2\} \ ,$$

where

$$T_i = 2[\ln L_x(\hat{\vartheta}) - \ln L_x(\hat{\vartheta}^*_{S_i})] \ , \ i = 1, 2 \ . \tag{2.6}$$

Hence, $T$ equals to the two-sample LR test statistic for the hypothesis $H^1 : \vartheta_3 \geq h_1(\vartheta_1)$ in case of $L_x(\hat{\vartheta}^*_{S_1}) \geq L_x(\hat{\vartheta}^*_{S_2})$, otherwise it equals the two-sample LR test statistic for the hypothesis $H^2 : \vartheta_3 \geq h_2(\vartheta_2)$. The following theorem guarantees that, asymptotically, $H^U$ is rejected at level $\alpha$ if $T$ is larger than the $(1-\alpha)$-quantile of the distribution of $\frac{1}{2} + \frac{1}{2}F_{\chi^2_1}$.

THEOREM 2. *Let $t > 0$. Then, under the conditions of Theorem 1 for the samples $X_{ij}$, $j = 1, \ldots, n_i$, $i = 1, 2, 3$, and $h_1, h_2$, respectively, for all $\vartheta \in \Theta_U^h$ it holds that*

$$P(T > t) \leq P(Z > t) + o(1) \ , \tag{2.7}$$

*where $Z$ is distributed as $\frac{1}{2} + \frac{1}{2}F_{\chi^2_1}$. Furthermore, for some $\vartheta \in \Theta_U^h$ we have strict equality in (2.7).*

### *2.2. The hypothesis $H^I$*

For the hypothesis $H^I$ the situation is rather different, since the LR test is not a combination of two pairwise comparisons. The boundary of the hypothesis $H^I$ in (1.2) is given by the union of the surfaces

$$\begin{aligned} K_1 &:= \{\vartheta \in [0,1]^3 \mid \vartheta_3 = h_1(\vartheta_1) \text{ and } \vartheta_3 \geq h_2(\vartheta_2)\} \ , \\ K_2 &:= \{\vartheta \in [0,1]^3 \mid \vartheta_3 = h_2(\vartheta_2) \text{ and } \vartheta_3 \geq h_1(\vartheta_1)\} \ . \end{aligned}$$

In contrast to the hypothesis $H^U$, the calculation of the LR statistic cannot be reduced to the two-sample case in general, since $K_1$ and $K_2$ are proper subsets of $S_1$ and $S_2$ (cf. (2.5)), respectively. Thus, the MLEs constrained to $S_1$ and $S_2$ are not included in $H^I$ for some outcomes. In that case the MLE constrained to $H^I$ is a projection onto the "edge" of $H^I$, i.e. on $K_3 := K_1 \cap K_2$ (cf. Figure 2).

THEOREM 3. *Let $\hat{\vartheta} \notin H^I$. With the notation of Section 2.1, the constrained MLE for the hypothesis $H^I$ is given by*

$$\hat{\vartheta}^* := \begin{cases} \hat{\vartheta}^*_{S_1} & \\ \hat{\vartheta}^*_{S_2} & \text{if} \\ \hat{\vartheta}^*_{K_3} & \end{cases} \begin{array}{l} \hat{\vartheta}^*_{S_1} \in H^I, \ (\hat{\vartheta}^*_{S_2} \notin H^I \ \vee \ (\hat{\vartheta}^*_{S_2} \in H^I, \ T_1 \leq T_2)) \\ \hat{\vartheta}^*_{S_2} \in H^I, \ (\hat{\vartheta}^*_{S_1} \notin H^I \ \vee \ (\hat{\vartheta}^*_{S_1} \in H^I, \ T_1 > T_2)) \\ \hat{\vartheta}^*_{S_1} \notin H^I, \ \hat{\vartheta}^*_{S_2} \notin H^I \end{array} \ ,$$

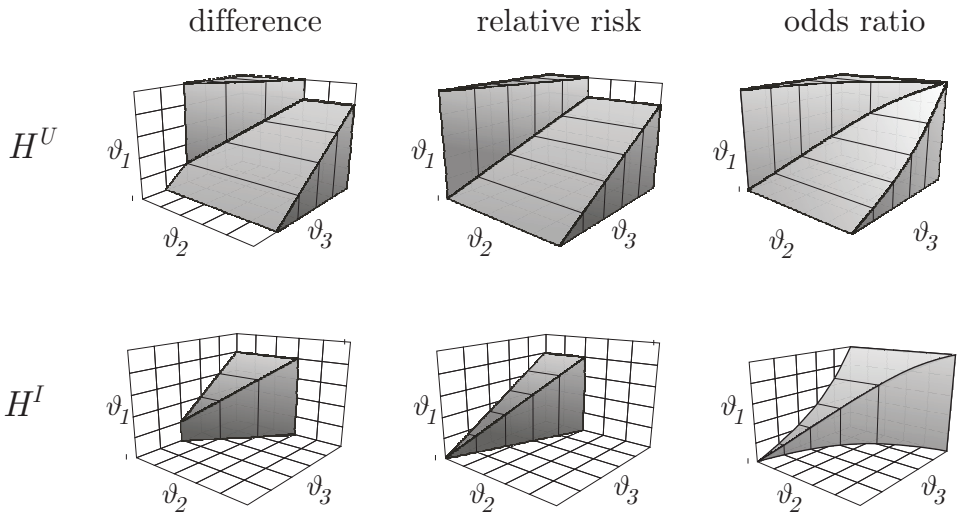


FIG 2. *Null spaces for $H^U$ and $H^I$ , respectively.*

*where* $\hat{\vartheta}^*_{K_3} := \arg\max_{K_3} L_x(\vartheta)$.

It is shown in Theorem 3 that the calculation of the MLE can be reduced to the two-sample case if $\hat{\vartheta}^*_{S_1} \in H^I$ or $\hat{\vartheta}^*_{S_2} \in H^I$. Otherwise, $L(\vartheta)$ has to be maximized under the constraint $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$. In order to find the arg max in this situation, one has to compute the zeros of the function

$$F(\vartheta_3) = \frac{x_3}{\vartheta_3} - \frac{n_3 - x_3}{1-\vartheta_3} + \frac{h_1^{-1'}(\vartheta_3)(x_1 - n_1 h_1^{-1}(\vartheta_3))}{h_1^{-1}(\vartheta_3)(1-h_1^{-1}(\vartheta_3))} + \frac{h_2^{-1'}(\vartheta_3)(x_2 - n_2 h_2^{-1}(\vartheta_3))}{h_2^{-1}(\vartheta_3)(1-h_2^{-1}(\vartheta_3))} \ ,$$

which are often the roots of a multi-degree-polynomial. For $h^{DI}$ it is a 5-degree-polynomial, for $h^{RR}$ and $h^{OR}$ we get a 3-degree-polynomial. In general, this can be determined numerically by Newton's method. Note that simultaneous testing of one-sided superiority in the two-sample comparisons is contained as a special case ($h_i(\vartheta) \equiv \vartheta$, $i = 1, 2$), in this case the solution is the overall rate $\vartheta^*_3 = \sum x_i / \sum n_i$. Note further, that for the resulting test any local maximum in $K_1 \cup K_2$ can be used as an MLE.

The LR statistic $T$ is calculated as in (2.2), with $\hat{\vartheta}^*$ given by Theorem 3. The asymptotic distribution of $T$ for $H^I$ is rather complicated and is given in Theorem A.1 (see Appendix). In contrast to the two-sample case (cf. Theorem 1), Theorem A.1 shows that the asymptotic distribution of the LR depends on the parameters $\vartheta_i$ and the functions $h_i$ for $\vartheta \in H^I$, hence asymptotically the LR test statistic is not free of nuisance parameters and the particular choice of the boundary functions.

To investigate the magnitude of the dependence of the probability $P(T > t)$ in Theorem A.1 on the nuisance parameter $\vartheta_1$, a numerical study is performed for several parameter configurations. Figure 3 shows the asymptotic probability $P(T > 3.84)$ for the difference, the relative risk, and the odds ratio. The critical value 3.84 is chosen such that it produces asymptotic probabilities near 0.05. However, note that in order to obtain a level $\alpha$-test the critical value has to be chosen, s.t. $\max_{\vartheta \in H^I} P_\vartheta (T > c) \leq \alpha$. In a comprehensive numerical study (not displayed) we found that this leads to a rather conservative test, which we do not recommend in practice. In Section 4.2 we suggest a modification which overcomes this drawback.

However, the merits of this section are twofold: first we have shown that the LR principle leads to a well known IU method for $H^U$, and to an asymptotically valid test for $H^I$. This result forms the basis for two modifications of the LR test - an exact version for small and an asymptotic modification for larger sample sizes. This will be discussed in detail in Section

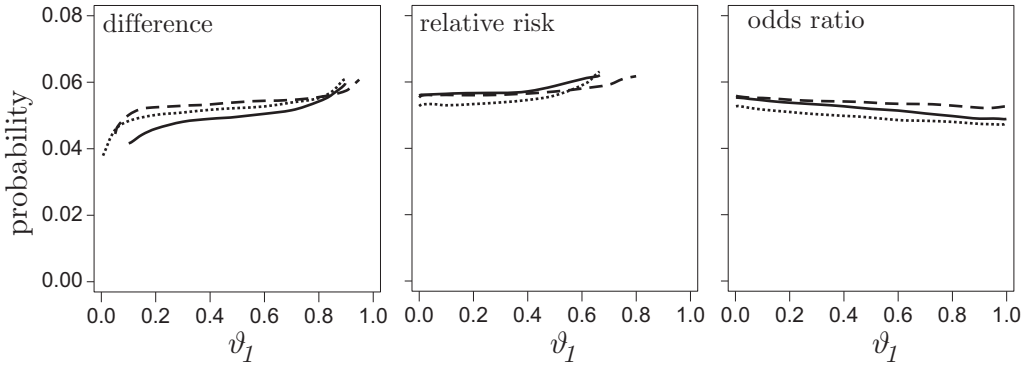


FIG 3. *The asymptotic probability* $P(T > 3.84)$ *as a function of the rate* $\vartheta_1$ *for several parameter configurations of* $\theta_1, \theta_2, c_{n1}, c_{n2}$ *with* $c_{ni} = \frac{n_i}{n_3}$, $i = 1, 2$ *(solid line:* $0.1, 0.2, 1, 1$ *for the difference,* $1.5, 2, 1, 0.5$ *for the relative risk and the odds ratio; dotted line:* $0.1, 0.1, 1, 0.5$ *for the difference,* $1.5, 1.5, 1, 1$ *for the relative risk and the odds ratio; dashed line:* $0.05, 0.1, 0.5, 0.5$ *for the difference,* $1.25, 1.5, 0.5, 0.5$ *for the relative risk and the odds ratio) and for hypothesis* $H^I$ *using the difference, the relative risk and the odds ratio.*

4. Moreover, in Section 5 we will show that these tests are more powerful than various competitors and practically feasible.

## 3. Pairwise comparisons

We have seen in Section 2.1 that the LR test for hypotheses of type $H^U$ will automatically lead to performing both two-sample comparisons. In contrast, the LR test for hypotheses of type $H^I$ does not yield immediate information on the pairwise comparisons, which could be criticized from a practical point of view. However, it is actually possible to always perform these pairwise comparisons in addition to the overall test, while keeping the nominal level as will be discussed in the following.

In fact, our argument does not only apply to the asymptotic test from Section 2.2, rather it applies to any global test for $H^I$, including those tests suggested in Section 4.

Assume now that an overall level-$\alpha$ test for $H^I$ in (1.2) has lead to a rejection of the null hypothesis. Then it is possible, in a second step, to perform two two-sample tests for $H^1$ and $H^2$ from (1.4), respectively, each at level $\alpha$. Note that the overall level $\alpha$ is not exceeded. This is due to the closed testing principle (Marcus, Peritz & Gabriel 1976), since the set $\mathbf{H} := \{H^I, H^1, H^2\}$ of hypotheses is closed under intersection, and since a hypothesis $H \in \mathbf{H}$ is tested (at level $\alpha$) only if each $\tilde{H} \subseteq H$ has been tested and rejected (at level $\alpha$).

Thus, if the LR test from Section 2.2 rejects $H^I$ at level $\alpha$, we may additionally perform the pairwise comparisons and conclude which one was "successful", each at an error rate $\alpha$. An alternative procedure would be to use directly a multiple test procedure based on the two-sample comparisons. The most prominent single step method to adjust the global level $\alpha$ for the hypothesis $H^I$ is the *Bonferroni adjustment*. Here $\alpha$ is evenly divided to each pairwise comparison, and $H^I$ is rejected if $p_{(1)} < \frac{\alpha}{2}$, where $p_{(1)}$, $p_{(2)}$ are the smaller and the larger of both p-values of the two-sample tests, respectively. Holm (1979) suggested a step down procedure improving the Bonferroni adjustment. However, for $H^I$ and the case of three samples, both procedures coincide.

Note that the possibility to perform each of the pairwise LR tests for $H^1$ and $H^2$ at level $\alpha$, *given the overall three-sample LR test for* $H^I$ *was successful*, leads automatically to a larger power for the pairwise comparisons if compared to the Bonferroni-adjusted two-sample LR tests for $H^i$, $i = 1, 2$. However, even *unconditionally*, we find in the subsequent numerical analysis (cf. Section 5) that the LR test for $H^I$ has larger power than the Bonferroni adjustment in most cases.

Note further, that it may happen that the overall test leads to rejection of $H^I$, indicating that one of the subhypotheses is not valid, albeit both pairwise tests do not lead to a significant rejection. Simulation studies show that this occurs very rarely, but it cannot be excluded, in general.

This effect is well known from other situations, e.g. within an analysis of variance accompanied by multiple testing strategies. We will discuss the impact of this phenomenon in the specific context of a three sample noninferiority trial. Here, we are faced with the situation that some noninferiority effect between the three groups is significant, however, it is not possible to assess at level $\alpha$, which one. Numerical experiments show that this situation only occurs, if the observed rates $\hat{\vartheta}$ fall close to the boundary $K_1 \cap K_2$.

There are situations where this does not cause a serious problem, e.g. when the primary goal of the trial is to show that a new treatment is as effective as one of two standards (cf. Section 1). In this case rejection of the global hypothesis reveals the new treatment as effective as one of the standards, and it might be of contiguous interest, to which one. In contrast, if the goal is to investigate whether one of two new treatments is as effective as a standard, for a regulatory agency and the sponsor it will be important to know which one. Receiving a significant result for the global test and two non significant results for the pairwise tests here, is of very limited use, and finally will leave the drug authority in a difficult situation.

To avoid these difficulties in general, we recommend to plan a clinical trial such that a power, of 0.8, say, for both subhypotheses is guaranteed, respectively. Note that in this case, the overall power of the test is at least 0.8 as well. Note further, that in contrast to directly applying pairwise comparisons to $H^I$ (each at level $\alpha/2$), one does not need to adjust the level $\alpha$ (as shown above), resulting in a larger power and reduction of sample sizes. In summary, we have seen that it is always advisable to perform the overall test in a preliminary step. If this test rejects and the (rare) event happens, that none of the pairwise tests leads to rejection, further sampling is required.

## 4. Modification of the LR test

### *4.1. Exact modification*

Exact tests for general hypotheses in the two-sample case were first introduced in two seminal papers by Barnard (1945, 1947). It has been shown, however, that Barnard's original test bears intrinsic numerical difficulties due to its specific iterative way to construct the region of rejection (Skipka, Munk & Freitag 2004). During the last two decades various other exact methods were suggested. Most of them were developed for $H\colon \vartheta_1 = \vartheta_2$ (Boschloo 1970; Upton 1982; D'Agostino, Chase & Belanger 1988), or for specific choices of $h$ in the hypothesis $H : \vartheta_1 = h(\vartheta_2)$ (see e.g. Chan 1998). Finally, Röhmel & Mansmann (1999) presented a general exact method for arbitrary hypotheses $H : \vartheta_1 \geq h(\vartheta_2)$, based on ideas of Barnard (1947).

The methodology of unconditional exact approaches for two samples is directly transferable to more than two samples (in the following described for three samples). In our context this reads as follows. The actual level $\alpha^*$ for a statistical test which specifies the critical region, i.e. the subset $CR$ of the sample space $\mathcal{S} = (0, \ldots, n_1) \times (0, \ldots, n_2) \times (0, \ldots, n_3)$ for which the null hypothesis $H^I$ is rejected, is calculated by

$$P(X \in CR \mid \vartheta) = \sum_{x \in CR} L_x(\vartheta) \ , \tag{4.1}$$

where $L_x(\vartheta)$ is given in (2.1) and $X = (X_1, X_2, X_3)^\top$. A commonly used approach is to eliminate the unknown parameter $\vartheta$ by maximizing the function (4.1) over $H^I$, yielding

$$\alpha^* = \alpha^*(CR) = \max_{H^I} P(X \in CR \mid \vartheta) . \tag{4.2}$$

Hence, an exact test fulfills $\alpha^* \leq \alpha$.

The performance of an exact test will mainly be determined by a criterion function which allows for a proper ordering of the sample to determine which values are included in the critical region, since the computational complexity to compare all possible $2^{n_1+n_2+n_3+3}$ subsets of the sample space is too large, in general. Various approaches for the two-sample case have been investigated in Skipka, Munk & Freitag (2004), and it has been shown that the cumulative likelihood function outperforms other methods suggested in the literature with respect to power. This will be extended in the following to the case of three samples.

In a first step, based on an idea of Storer & Kim (1990), the exact distribution of the LR statistic is estimated by inserting the constrained MLE $\hat{\vartheta}^*$ into (2.1). With that, p-values can be estimated for any outcome $x = (x_1, x_2, x_3)^\top \in \mathcal{S}$ by calculating

$$p^*(x) = \sum_{a:\, T(a) \geq T(x)} L_a(\hat{\vartheta}^*) \,, \tag{4.3}$$

where $a = (a_1, a_2, a_3)^\top$ and $T(a)$ is the likelihood ratio given in (2.2). These are the exact p-values under the assumption that $\hat{\vartheta}^*$ (the MLE constrained to $H^I$) is the true parameter.

In a second step these estimated p-values $p^*(x)$ are used to sort all possible outcomes $x \in \mathcal{S}$ in ascending order. Thus, we obtain a vector

$$S = \left( x^{(1)}, \ldots, x^{((n_1+1)\cdot(n_2+1)\cdot(n_3+1))} \right),$$

with the corresponding increasing values $p^*\left(x^{(i)}\right) =: p_i^*$. Now define

$$\alpha_i^* = \alpha^* \left( \bigcup_{j=1}^{i} \left\{ x^{(j)} \right\} \right), \tag{4.4}$$

which denotes the maximal actual level of the rejection region $CR_i = \bigcup_{j=1}^{i} x^{(j)}$ of the "$i$ smallest" values in $S$ with respect to the ordering induced by $p^*$. Finally, the critical region is defined by

$$CR = CR_k, \quad k = \arg\max_i \{\alpha_i^* \leq \alpha\} \,.$$

In Section 5, this unconditional exact modification of the LR test for $H^I$ - denoted by *exact LR test* in the following - will be compared to pairwise two-sample tests. Note that this test does not share the specific numerical problems of Barnard's test (cf. Skipka, Munk & Freitag 2004).

### *4.2. Quasi exact modification*

For larger sample sizes (50 per group, say) the computation of the exact LR test is rather time consuming. Hence, for these situations we now will suggest a modification of this test, which is based on the asymptotic distribution in Theorem A.1. Since the asymptotic null distribution of the LR statistic depends on $\vartheta$ (cf. Section 2.2), we estimate the null distribution by inserting that $\vartheta$ which is most likely under the null hypothesis. This test will be called a quasi exact test in the sequel. The following algorithm describes this approach in detail. As above, let $x = (x_1, x_2, x_3)$ be the observed outcomes and $h_1$, $h_2$ specified.

1. Calculate from $x$ the MLE $\hat{\vartheta}^*$ constrained to $H^I$ (denoted as $\hat{\vartheta}^*$ in Theorem 3).
2. Compute a large number of 3 binomial samples with parameter $\hat{\vartheta}^*$ numerically by simulations (in our investigation we used 100,000 repetitions).
3. Calculate the LR statistic $T$ for each sample.
4. Calculate the $(1-\alpha)$-quantile $t$ from the sample of $T$s.
5. Reject $H^I$ in case of $T(x) > t$.

Although this test is not exact, we will see that it keeps its nominal level quite accurately. In the next section this is investigated in detail by a numerical comparison of level and power with two commonly used asymptotic methods. Our main finding is that the improvement in power is considerable where at the same time the nominal level is not more exceeded.

## 5. Level and power comparisons for tests of $H^I$

The LR principle for the hypothesis $H^U$ leads to a combination of the two-sample LR tests, where no level adjustment is necessary. Therefore, no further investigations are carried out for these hypotheses and we refer to the investigations for the two-sample case in Skipka, Munk & Freitag (2004) and Munk, Skipka & Stratmann (2005). However, as seen in Section 2.2, for the hypothesis $H^I$ the LR test cannot be reduced to the two-sample case. Therefore, the LR test for $H^I$ (exact and asymptotic) will be compared with Bonferroni-adjusted test procedures based on exact two-sample tests (other than the LR test) proposed in the literature.

### *5.1. Exact tests*

Munk, Skipka & Stratmann (2005) investigated exact two-sample LR tests for general hypotheses. Based on these results, the best competitors proposed in the literature are chosen: Chan's test (Chan 1998) is an unconditional approach for $h^{DI}$ and $h^{RR}$ based on Farrington and Manning's $z$ statistic (Farrington & Manning 1990). For $h^{OR}$, Fisher's exact unconditional test can be applied, which is based on the generalized hypergeometric distribution. We refer to Munk, Skipka & Stratmann (2005) for a detailed description of these competitors.

The exact procedures are investigated for various sample sizes (up to 50 per group). Note that the computation time (approximately 5 minutes for a sample size of 25 per group and 20 minutes for 50 per group with a Pentium IV, 3 GHz, SAS V8) increases rapidly for larger sample sizes. In comparison, the asymptotic test described at the end of Section 4.2 is computationally much more feasible. The power of all tests has been calculated always exactly. In all simulation studies a broad scenario of parameter settings $(\theta_1, \theta_2, n_1, n_2, n_3, \vartheta_1)$ is considered for the distance measures difference, relative risk and odds ratio:

- *Equivalence margins*:
  $(\theta_1, \theta_2) \in \{(0.15, 0.15), (0.15, 0.2), (0.15, 0.25), (0.2, 0.2), (0.2, 0.25), (0.25, 0.25)\}$ for $h^{DI}$ and $(\theta_1, \theta_2) \in \{(1.5, 1.5), (1.5, 2), (1.5, 2.5), (2, 2), (2, 2.5), (2.5, 2.5)\}$ for $h^{RR}$ and $h^{OR}$.
- *Sample size*: Balanced sample sizes $n_1 = n_2 = n_3 \in \{20, 25, 30, 40, 50\}$ and unbalanced sample sizes $(n_1, n_2, n_3) \in \{(20, 20, 40), (25, 25, 50), (40, 40, 20), (50, 50, 25)\}$.
- *Nuisance parameter*: $\vartheta_1 \in \{0.1, 0.2, 0.3, 0.5, 0.8, 0.9\}$ .
- *Distances between the groups*: $\vartheta_1 = \vartheta_2 \leq \vartheta_3$ and $\vartheta_1 \geq \vartheta_2 = \vartheta_3$ .

Overall, 648 parameter configurations were investigated for each distance measure, respectively. Configurations are omitted in case of non-feasible settings (e.g. $\vartheta_1 \geq 1 - \theta_1$ for $h^{DI}$, $\vartheta_1 \geq 1/\theta_1$ for $h^{RR}$). The parameters $\vartheta_1, \vartheta_2, \vartheta_3$ are chosen such that, if possible, the resulting power is larger than 0.8, at least for one of the tests compared. Here two settings are investigated: either the parameter $\vartheta_3$ is chosen equal to or smaller than $\vartheta_1 = \vartheta_2$, or the parameter $\vartheta_1$ is chosen equal to or greater than $\vartheta_2 = \vartheta_3$. In a second step, parameter configurations are omitted for which all tests achieve a power larger than 0.9. Finally, 261 parameter configurations remain for the difference, 85 parameter configurations remain for the relative risk, and 260 parameter configurations remain for the odds ratio.

Figure 4 represents the power of the exact LR test (vertical axes) and its competitors (horizontal axes) for the three distance measures $h^{DI}$, $h^{RR}$ and $h^{OR}$, respectively. The calculations show that the power of the exact LR test compared to its pairwise competitors is larger in general. For nearly each parameter configuration and distance measure the LR test outperforms the pairwise procedures using Bonferroni's adjustment. Figure 5 gives Boxplots

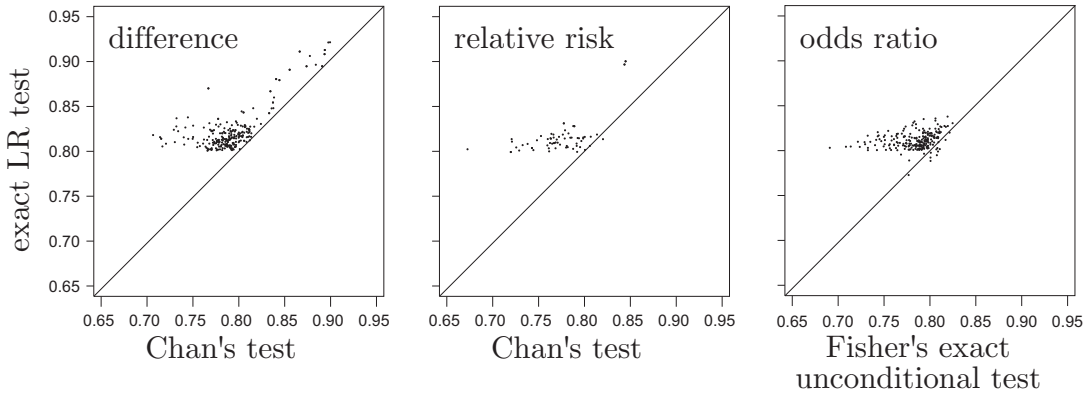


FIG 4. *The power of the exact 3-sample LR test (vertical axis) in comparison to the pairwise 2-sample tests with Bonferroni's adjustment (horizontal axis) for several parameter configurations and for hypothesis $H^I$ using the difference, relative risk, and odds ratio, respectively.*

(results of all distance measures combined) of the power differences between the exact LR test and its competitors with Bonferroni's adjustment.

As a conclusion, the calculations show that in most cases the exact LR test for $H^I$ improves the power compared to the Bonferroni adjusted pairwise procedures, and this improvement can be quite substantial. We mention that similar pictures are obtained if the exact LR test for $H^I$ is compared with the Bonferroni adjusted procedure using the exact two-sample LR tests instead of the two-sample tests used in Figures 4 and 5 (results not shown).

Thus, the sample size can be significantly reduced when applying the exact LR test for $H^I$ instead of the Bonferroni adjusted procedures. This will be illustrated by the following example. Let $h = h^{OR}$, $(\theta_1, \theta_2) = (2, 2)$ and $(\vartheta_1, \vartheta_2, \vartheta_3) = (0.8, 0.8, 0.6)$. Then a sample size of 25 per group is required to give a power of 0.8 when using the Bonferroni adjusted procedure (Fisher's exact unconditional test). Applying the exact LR test for $H^I$, a sample size of 20 per group yields a power of 0.8, i.e. the sample size can be reduced by 20%.

### *5.2. The quasi exact test*

The quasi exact LR test - described at the end of Section 4.2 - is numerically investigated by simulations (100,000 simulation runs in each scenario) for sample sizes between 50 and 500 per group. Level and power of this test are compared to the pairwise asymptotic two-sample tests based on Bonferroni's adjustment of commonly used score tests. To this end, for $h^{DI}$ and $h^{RR}$ Farrington and Manning's test (Farrington & Manning 1990) is chosen, because among asymptotic tests this test has been revealed in a variety of papers as a benchmark w.r.t. power, albeit sometimes slightly liberal (cf. e.g. Munk, Skipka & Stratmann 2005). For $h^{OR}$ a test based on the standardized log odds ratio is applied, which is the commonly used approach. We mention that a survey on these and various other asymptotic approaches can be found in Chen, Tsong & Hang (2000).

Table 1 shows the simulated level and power for the three distance measures difference, relative risk and odds ratio, respectively. Different parameter settings are implemented, analogously to the investigations mentioned above. Note, that the simulated levels are quite accurate for all approaches. Overall, it can be seen that the quasi exact LR test is superior to its competitors with respect to level and power in most cases.

TABLE 1

*The simulated power (level) (times 100) of the quasi exact LR test and its competitors ($\tilde{\theta}_1, \tilde{\theta}_2$ as the true differences, relative risks, or odds ratios, respectively).*

| $n_1$ | $n_2$ | $n_3$ | $\theta_1$ | $\theta_2$ | $\vartheta_1$ | $\tilde{\theta}_1$=$\tilde{\theta}_2$ | LR test | score test |
|---|---|---|---|---|---|---|---|---|
| | | | | | difference | | | |
| 50 | 50 | 100 | 0.1 | 0.15 | 0.2 | -0.02 | 84.9 (5.0) | 84.7 (5.0) |
| 75 | 75 | 75 | 0.1 | 0.15 | 0.3 | -0.05 | 85.3 (5.0) | 83.6 (4.7) |
| 75 | 75 | 150 | 0.1 | 0.15 | 0.25 | 0 | 80.5 (4.9) | 79.2 (5.0) |
| 100 | 100 | 100 | 0.1 | 0.15 | 0.2 | 0 | 80.9 (5.1) | 77.8 (4.2) |
| 100 | 100 | 200 | 0.1 | 0.15 | 0.3 | 0 | 85.7 (4.9) | 84.5 (4.8) |
| 200 | 200 | 200 | 0.1 | 0.1 | 0.25 | 0 | 80.7 (5.0) | 78.1 (4.3) |
| 200 | 200 | 400 | 0.1 | 0.1 | 0.4 | 0 | 86.4 (4.9) | 84.3 (4.8) |
| 400 | 400 | 400 | 0.05 | 0.1 | 0.5 | 0 | 83.8 (4.9) | 82.9 (4.7) |
| 500 | 500 | 500 | 0.05 | 0.1 | 0.5 | 0 | 90.6 (5.3) | 90.0 (4.7) |
| | | | | | relative risk | | | |
| 50 | 50 | 100 | 1.5 | 1.75 | 0.4 | 1 | 82.4 (4.8) | 79.8 (5.2) |
| 75 | 75 | 75 | 1.5 | 1.75 | 0.35 | 1 | 80.9 (4.9) | 79.1 (5.2) |
| 75 | 75 | 150 | 1.5 | 1.75 | 0.3 | 1 | 81.8 (5.0) | 80.6 (5.2) |
| 100 | 100 | 100 | 1.5 | 1.5 | 0.2 | 0.75 | 81.5 (5.0) | 80.3 (4.9) |
| 100 | 100 | 200 | 1.5 | 1.5 | 0.3 | 1 | 79.4 (5.0) | 77.1 (5.2) |
| 200 | 200 | 200 | 1.25 | 1.5 | 0.3 | 1 | 80.0 (5.1) | 78.9 (4.9) |
| 200 | 200 | 400 | 1.25 | 1.5 | 0.25 | 1 | 80.7 (4.8) | 79.8 (5.1) |
| 400 | 400 | 400 | 1.25 | 1.25 | 0.35 | 1 | 83.6 (4.9) | 80.7 (4.6) |
| 500 | 500 | 500 | 1.25 | 1.25 | 0.3 | 1 | 82.8 (4.8) | 80.7 (4.6) |
| | | | | | odds ratio | | | |
| 50 | 50 | 100 | 1.5 | 1.75 | 0.4 | 0.75 | 80.2 (4.7) | 77.6 (4.9) |
| 75 | 75 | 75 | 1.5 | 1.75 | 0.5 | 0.8 | 77.8 (4.9) | 75.3 (4.5) |
| 75 | 75 | 150 | 1.5 | 1.75 | 0.3 | 0.8 | 83.0 (4.9) | 80.6 (5.0) |
| 100 | 100 | 100 | 1.5 | 1.5 | 0.4 | 0.75 | 83.9 (5.3) | 82.2 (4.8) |
| 100 | 100 | 200 | 1.5 | 1.5 | 0.45 | 0.85 | 84.5 (4.9) | 82.4 (4.9) |
| 200 | 200 | 200 | 1.25 | 1.5 | 0.3 | 0.85 | 79.8 (5.0) | 78.2 (4.6) |
| 200 | 200 | 400 | 1.25 | 1.5 | 0.25 | 0.9 | 79.5 (5.0) | 77.6 (4.9) |
| 400 | 400 | 400 | 1.25 | 1.25 | 0.35 | 0.9 | 79.0 (5.1) | 75.5 (4.6) |
| 500 | 500 | 500 | 1.25 | 1.25 | 0.2 | 0.85 | 84.3 (4.9) | 81.8 (4.5) |

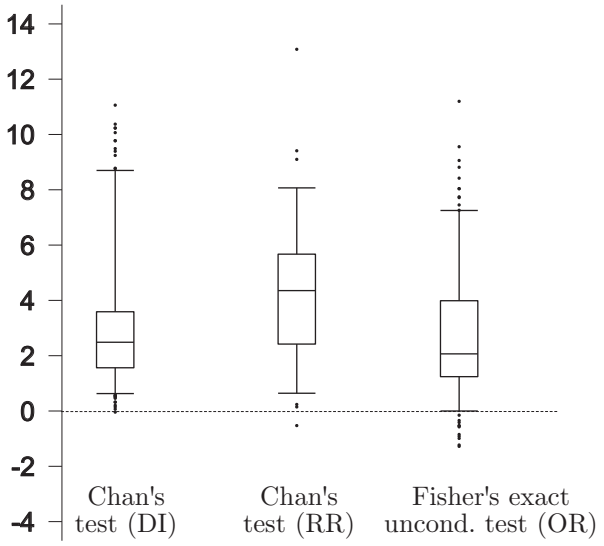


FIG 5. *Boxplot (whiskers are the 5% and 95% quantiles) for the power differences (times 100) between the exact LR test and the multiple comparison procedures using Bonferroni's adjustment for the three distance measures difference (DI), relative risk (RR) and odds ratio (OR).*

## 6. Example

In a randomized double-blind comparison in patients with cancer, Hesketh *et al.* (1996) assess the efficacy of antiemetic agents in preventing cisplatin-induced nausea and vomiting. The trial was performed to show noninferiority of dolasetron mesylate at doses of 1.8 mg/kg ($E_1$) and 2.4 mg/kg ($E_2$), respectively, over the standard ondansetron ($C$) at its approved dose of 32 mg. The primary analysis was done by comparing the failure rates of $E_1$ and $E_2$, respectively, with $C$. Patients having emetic episodes or receiving rescue medication during 24 hours were classified as failures. For both comparisons the equivalence margin for the odds ratio was specified as 2. It is not clearly described by Hesketh *et al.* (1996) whether it was the goal to show noninferiority of both doses of dolasetron compared to ondansetron, or to show that at least one of the doses of dolasetron is non-inferior to ondansetron.

The resulting failure rates were similar in the three groups: 110/198 (56%) in $E_1$, 123/205 (60%) in $E_2$, and 118/206 (57%) in $C$. Comparing $E_1$ versus $C$ and $E_2$ versus $C$, the authors calculated an odds ratio (upper 97.5% confidence limit) of 0.97 (1.47) and 1.16 (1.75), respectively. They concluded that dolasetron (1.8 or 2.4 mg/kg) has comparable efficacy to ondansetron, since the upper confidence limits were smaller than 2 (without specifying any level adjustment; actually, there was no significance level stated at all).

If we apply the asymptotic LR test for the odds ratio for $H^U$ (which equals the pairwise comparisons with level $\alpha$, respectively), i.e. for showing that *both* treatments $E_1, E_2$ are non-inferior to $C$, we obtain p-values of 0.00007 and 0.0019 comparing $E_1$ versus $C$ and $E_2$ versus $C$, respectively, i.e. we can reject the null hypothesis $H^U$ at level $\alpha = 0.05$. The same p-values result for the asymptotic score tests. For comparison with the results in Hesketh *et al.* (1996), we can determine test-based upper 97.5% confidence limits by calculating the hypotheses boundaries for which the respective exact two-sample LR tests do not reject the null hypotheses at level 2.5%. This results in boundaries 1.38 for $E_1$ versus $C$ and 1.66 for $E_2$ versus $C$, which are even a bit smaller than the boundaries given by Hesketh *et al.* (1996). Even if their boundaries - calculated with adjustment for covariates - are not directly comparable to our boundaries, this indicates how powerful the LR test is.

If it is of interest to show noninferiority of *at least one* of both doses of dolasetron compared to ondansetron, we can apply the LR test for $H^I$. To embed this setting into hypothesis $H^I$,

we have to regard success rates instead of failure rates. Thus, let $\vartheta_1$, $\vartheta_2$, and $\vartheta_3$ denote the true *success rates* for $E_1$, $E_2$, and $C$, respectively. For this example we obtain $T = 15.9$ as the value of test statistic (2.2). Applying the quasi exact test described in Section 4.2, we get $\hat{\vartheta}^* = (0.37, 0.37, 0.54)$ as the constrained MLE and $t = 3.81$ as the 95%-quantile. The approximated p-value is about 0.00009, hence we can reject $H^I$ at level $\alpha = 0.05$. Thus, we can now proceed with the two-sample comparisons, as was discussed in Section 3. Since these can be performed each at level $\alpha$, we can immediately use the results obtained above when considering the null hypothesis $H^U$.

## 7. Discussion

We mention that the LR test for three armed trials can be extended to other settings. More general, if the observations come from an exponential family, similar Theorems as 1 and 2 can be proved. Besides of the binomial model discussed so far the most common assumption is normality. Here, various simplifications are possible.

In this case we observe normally distributed samples $X_{ij} \sim N(\mu_i, \sigma^2)$ $(j = 1, \ldots, n_i;$ $i = 1, 2, 3)$. Most commonly, the hypotheses corresponding to (1) and (2) are formulated in terms of mean differences, e.g. Hypericum Depression Trial Study Group (2002),

$$H^U : \mu_3 - \mu_1 \geq \delta_1 \text{ or } \mu_3 - \mu_2 \geq \delta_2 \tag{7.1}$$

and

$$H^I : \mu_3 - \mu_1 \geq \delta_1 \text{ and } \mu_3 - \mu_2 \geq \delta_2, \tag{7.2}$$

where $\delta_1$, $\delta_2$ are threshold values specified in advance. In this particular case, without changing the testing problem, $\delta_1$ and $\delta_2$ can be set to zero when adding $\delta_1$ to $x_{1j}$ and $\delta_2$ to $x_{2j}$. Of course, when other distance measures are specified, shifting of the margins may no longer be possible (e.g. for the standardized difference). Pigeot *et al.* (2003) or Tang & Tang (2004) investigated trials where the noninferiority margin $\delta_1$ for a new treatment compared to a standard treatment is specified as a fraction of the true difference between the standard treatment and placebo. This setting can also be treated with help of the LR test - both asymptotically and exact. We will present this case in more detail in a further publication.

For one-sided hypotheses and more than two groups the statistical theory is based on methods of order restricted statistical inference which was extensively developed since the early 1950s. Barlow, Bartholomew, Bremner & Brunk (1972) have summarized much of the early work. For $k$ independent groups with normally distributed data and means $\mu = (\mu_1, \ldots, \mu_k)$, Robertson, Wright & Dykstra (1988) consider hypotheses of the type

$$H_0 : \mu \text{ is isotonic with respect to } \preceq \quad \text{vs.} \quad \neg H_0 \ ,$$

where $\preceq$ is a partial ordering of $\mu$. Robertson, Wright & Dykstra (1988) developed the LR test for different partial orderings. Applying their formulae to the hypothesis (7.1) it can be easily shown that the LR test is equivalent to the IUT, if for the IUT the two-sample variance estimates are replaced by the pooled three-sample variance estimates. This leads to an improvement over the pairwise testing when using the two-sample pooled standard deviation, due to the larger number of $n_1 + n_2 + n_3 - 3$ degrees of freedom. The hypothesis (7.2) is a particular case of a simple tree hypothesis for $k > 2$ groups. The formulae for the LR test in this case can be found in Robertson, Wright & Dykstra (1988).

To summarize, we have shown that in a three armed noninferiority or superiority trial in the binomial setting the LR principle leads to two different tests. For null hypotheses which are the union of two sub-hypotheses of the type $\vartheta_3 \geq h_1(\vartheta_1)$ and $\vartheta_3 \geq h_2(\vartheta_2)$, the intersection union test results asymptotically, whereas for testing the intersection of these sub-hypotheses a rather complicated asymptotic test results. An exact modification of this test yields numerically feasible solutions for small sample sizes. This unifies various approaches suggested in the literature for particular choices of $h_1$, $h_2$, and it leads to tests

which outperform Bonferroni-adjusted tests in terms of power. In addition, if the test for an intersection hypothesis leads to a rejection, then the pairwise comparisons can be performed in a second step, where no level adjustment is required.

There are several methodological aspects which remain to be investigated in the future. Issues related to the construction of test-based confidence intervals and to the determination of necessary sample sizes for the exact tests will be addressed in forthcoming publications. Further, we will deal with the problem of showing the retention of the fraction of the control effect in the three-sample design.

## Appendix A: Proof of Theorem 2.

First note that

$$P(T > t) = P(T_1 > t \ , \ L_x(\hat{\vartheta}^*_{S_1}) \geq L_x(\hat{\vartheta}^*_{S_2})) + P(T_2 > t \ , \ L_x(\hat{\vartheta}^*_{S_1}) < L_x(\hat{\vartheta}^*_{S_2})) \ .$$

In case of $\vartheta_3 = h_1(\vartheta_1), \vartheta_3 < h_2(\vartheta_2)$ ,

$$P(L_x(\hat{\vartheta}^*_{S_1}) \geq L_x(\hat{\vartheta}^*_{S_2})) = 1 + o(1) \ ,$$

and thus,

$$P(T > t) = P(T_1 > t) + o(1) \ .$$

Analogously, in case of $\vartheta_3 = h_2(\vartheta_2), \vartheta_3 < h_1(\vartheta_1)$ ,

$$P(T > t) = P(T_2 > t) + o(1) \ .$$

Hence, if $\vartheta$ is not located on the edge $S_1 \cap S_2 = \{\vartheta : \vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)\}$, the test statistic follows the same distribution $(\frac{1}{2} + \frac{1}{2}F_{\chi^2_1})$ as in the two-sample case (cf. Theorem 1).

If $\vartheta$ is located on the edge $S_1 \cap S_2$,

$$P(T > t) = P(T_1 > t \ , \ T_2 > t) \leq P(T_1 > t) \ ,$$

i.e. $P(T > t) \leq \alpha$ for $t = (\frac{1}{2} + \frac{1}{2}F_{\chi^2_1})_{1-\alpha}$.

## Appendix B: Proof of Theorem 3.

A similar argument as in Munk, Skipka & Stratmann (2005, Lemma 1a) yields that the maximum of $L_x(\vartheta)$ over $H^I$ is attained in $\Theta^h_I = K_1 \cup K_2$.

If $\hat{\vartheta}^*_{S_1} \notin H^I$, we have that $\arg\max_{K_1} L(\vartheta) \in K_3$, since $L_x(\vartheta)$ is isotonic in $\vartheta_2$ ($\vartheta_2 < \hat{\vartheta}_2$) for fixed parameters $\vartheta_3 = h_1(\vartheta_1)$. Analogously, $\arg\max_{K_2} L_x(\vartheta) \in K_3$ holds for $\hat{\vartheta}^*_{S_2} \notin H^I$. It follows that $\hat{\vartheta}^* = \hat{\vartheta}^*_{K_3}$.

If $\hat{\vartheta}^*_{S_1} \in H^I$ and $\hat{\vartheta}^*_{S_2} \notin H^I$, it follows that $\arg\max_{K_2} L_x(\vartheta) \in S_1$, since $K_3 \subset S_1$. If $\hat{\vartheta}^*_{S_1} \in H^I$ and $\hat{\vartheta}^*_{S_2} \in H^I$, $\max_{\vartheta \in H^I} L_x(\vartheta) = L_x(\hat{\vartheta}^*_{S_1})$ for $T_1 \leq T_2$. This proves the case $\hat{\vartheta}^* = \hat{\vartheta}^*_{S_1}$, and by symmetry the case $\hat{\vartheta}^* = \hat{\vartheta}^*_{S_2}$, also.

## Appendix C: Asymptotic null distribution of the LR for $H_I$

THEOREM A.1. *Let $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$ and $X = (X_1, X_2, X_3)^\top$ a 3-dimensional normally distributed random vector with zero mean and covariance matrix $\Sigma^{-1}$, where*

$$\Sigma := diag\left(\frac{1}{\vartheta_1(1-\vartheta_1)} \ , \ \frac{1}{\vartheta_2(1-\vartheta_2)} \ , \ \frac{1}{\vartheta_3(1-\vartheta_3)}\right) \ .$$

*Let further, for* $i = 1, 2$,

$$\begin{aligned}
\Sigma_i &:= diag\left(\frac{1}{\vartheta_i(1-\vartheta_i)}\ ,\ \frac{1}{\vartheta_3(1-\vartheta_3)}\right)\ ,\\
C_i &:= (\sqrt{c_i}\ ,\ h_i'(\vartheta_i))^\top\ ,\\
\Sigma_i^* &:= \frac{c_i}{\vartheta_i(1-\vartheta_i)} + \frac{h_i'(\vartheta_i)^2}{h_i(\vartheta_i)(1-h_i(\vartheta_i))}\ ,
\end{aligned}$$

*and*

$$\begin{aligned}
C &:= (\sqrt{c_1}\ ,\ \sqrt{c_2}[h_2^{-1}(h_1(\vartheta_1))]'\ ,\ h_1'(\vartheta_1))^\top\ ,\\
\Sigma^* &:= \frac{c_1}{\vartheta_1(1-\vartheta_1)} + \frac{c_2([h_2^{-1}(h_1(\vartheta_1))]')^2}{h_2^{-1}(h_1(\vartheta_1))[1-h_2^{-1}(h_1(\vartheta_1))]} + \frac{h_1'(\vartheta_1)^2}{h_1(\vartheta_1)(1-h_1(\vartheta_1))}\ .
\end{aligned}$$

*Then, as* $\min_{i=1,2,3}\{n_i\} \to \infty$, *such that* $c_{ni} := \frac{n_i}{n_3} \to c_i \in (0,\infty)$ $(i = 1, 2)$, *it holds for* $t > 0$ *that* $P(T > t) \to p_1(t) + p_2(t) + p_3(t)$, *where*

$$\begin{aligned}
p_1(t) &:= P\Bigg((X_1, X_3)A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} > t\ \cap\ [X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}}X_1 \cup X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}}X_2]\\
&\quad \cap\ B_1X_1 \geq \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_2}}X_2 \cap \Bigg[B_2X_2 < \frac{(h_2^{-1}[h_1(\vartheta_2)])'}{\sqrt{c_1}}X_1\\
&\quad \cup\ \ \{B_2X_2 \geq \frac{(h_2^{-1}[h_1(\vartheta_2)])'}{\sqrt{c_1}}X_1 \cap (X_1, X_3)A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} \leq (X_2, X_3)A_2\begin{pmatrix}X_2\\X_3\end{pmatrix}\}\Bigg]\Bigg)\ ,
\end{aligned}$$

$$\begin{aligned}
p_2(t) &:= P\Bigg((X_2, X_3)A_2\begin{pmatrix}X_2\\X_3\end{pmatrix} > t\ \cap\ [X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}}X_2 \cup X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}}X_1]\\
&\quad \cap\ B_2X_2 \geq \frac{(h_2^{-1}[h_1(\vartheta_2)])'}{\sqrt{c_1}}X_1 \cap \Bigg[B_1X_1 < \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_2}}X_2\\
&\quad \cup\ \ \{B_1X_1 \geq \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_2}}X_2 \cap (X_1, X_3)A_1\begin{pmatrix}X_1\\X_3\end{pmatrix} > (X_2, X_3)A_2\begin{pmatrix}X_2\\X_3\end{pmatrix}\}\Bigg]\Bigg)\ ,
\end{aligned}$$

$$\begin{aligned}
p_3(t) &:= P\Bigg(X^\top AX > t\ \cap\ [X_3 < \frac{h_1'(\vartheta_1)}{\sqrt{c_1}}X_1 \cup X_3 < \frac{h_2'(\vartheta_2)}{\sqrt{c_2}}X_2]\\
&\quad \cap\ \ B_1X_1 < \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_2}}X_2 \cap B_2X_2 < \frac{(h_2^{-1}[h_1(\vartheta_2)])'}{\sqrt{c_1}}X_1\Bigg)\ ,
\end{aligned}$$

*with (letting* $(x)_1$ *denote the first component of the vector* $x$*)*

$$\begin{aligned}
A &= \Sigma - \Sigma C{\Sigma^*}^{-1}C^\top\Sigma\ ,\\
A_i &= \Sigma_i - \Sigma_i C_i\Sigma_i^{*^{-1}}C_i^\top\Sigma_i\ ,\quad i = 1, 2\ ,\\
B_i &= (\Sigma_i^{*^{-1}}C_i^\top\Sigma_i)_1\ ,\quad i = 1, 2\ .
\end{aligned}$$

*Proof Theorem A.1.* If $\hat\vartheta^* = \hat\vartheta^*_{S_i}$ $(i = 1, 2)$, i.e. the MLE constrained to $H^I$ is in $S_i$, the test statistic $T_i$ is given by (2.6). Since for $\hat\vartheta^* = \hat\vartheta^*_{K_3}$ the MLE is calculated under the constraint $K_3$, the test statistic is given by $T_3 := 2[\ln L(\hat\vartheta) - \ln L(\hat\vartheta^*_{K_3})]$.

With Theorem 3 it holds for $t > 0$ that

$$\begin{aligned}
P(T > t) = P(&T_1 > t \cap \hat\vartheta \notin H^I \cap \hat\vartheta^*_{S_1} \in H^I \cap [\hat\vartheta^*_{S_2} \notin H^I \cup \{\hat\vartheta^*_{S_2} \in H^I \cap T_1 \leq T_2\}])\\
&+P(T_2 > t \cap \hat\vartheta \notin H^I \cap \hat\vartheta^*_{S_2} \in H^I \cap [\hat\vartheta^*_{S_1} \notin H^I \cup \{\hat\vartheta^*_{S_1} \in H^I \cap T_1 > T_2\}])\\
&+P(T_3 > t \cap \hat\vartheta \notin H^I \cap \hat\vartheta^*_{S_1} \notin H^I \cap \hat\vartheta^*_{S_2} \notin H^I)\ .
\end{aligned}$$

From Pruscha (2000, Th. 4.3, p. 253) it follows that $T_i$ $(i = 1, 2, 3)$ is asymptotically equivalent to $(\hat{X}_i, \hat{X}_3)\, A_i\, (\hat{X}_i, \hat{X}_3)^\top$ if $\vartheta_3 = h_i(\vartheta_i)$ for $i = 1, 2$, and to $\hat{X}^\top A\ \hat{X}$ if $\vartheta_3 = h_1(\vartheta_1) = h_2(\vartheta_2)$ for $i = 3$, where

$$\hat{X} \;=\; (\hat{X}_1, \hat{X}_2, \hat{X}_3)^\top \;=\; (\sqrt{n_j}(\hat{\vartheta}_j - \vartheta_j))_{j=1,2,3}\ .$$

Note that $\hat{\vartheta} \in H^I$ is equivalent to $\hat{\vartheta}_3 \geq h_1(\hat{\vartheta}_1) \ \cap\ \hat{\vartheta}_3 \geq h_2(\hat{\vartheta}_2)$ and hence to $\sqrt{n_3}(\hat{\vartheta}_3 - \vartheta_3) \geq \sqrt{\frac{n_1}{c_{n1}}}(h_1(\hat{\vartheta}_1) - h_1(\vartheta_1)) \ \cap\ \sqrt{n_3}(\hat{\vartheta}_3 - \vartheta_3) \geq \sqrt{\frac{n_2}{c_{n2}}}(h_2(\hat{\vartheta}_2) - h_2(\vartheta_2))$. Since we have $h_i(\hat{\vartheta}_i) - h_i(\vartheta_i) = h_i'(\vartheta_i)(\hat{\vartheta}_i - \vartheta_i) + o_p(|\hat{\vartheta}_i - \vartheta_i|)$ as $\min_{i=1,2,3}\{n_i\} \to \infty$, it holds that

$$\sqrt{n_3}[\hat{\vartheta}_3 - h_i(\hat{\vartheta}_i)] \;-\; [\hat{X}_3 - \frac{h_i'(\vartheta_i)}{\sqrt{c_{ni}}}\hat{X}_i] \;\xrightarrow{P}\; 0\ .$$

Now $\hat{\vartheta}^*_{S_1} \in H^I$ is equivalent to $h_1(\hat{\vartheta}^*_{S_1,1}) \geq h_2(\hat{\vartheta}_2)$ and hence to $\sqrt{n_3}(\hat{\vartheta}^*_{S_1,1} - \vartheta_1) \geq \sqrt{\frac{n_2}{c_{n2}}}[h_1^{-1}(h_2(\hat{\vartheta}_2)) - h_1^{-1}(h_2(\vartheta_2))]$, where $\hat{\vartheta}^*_{S_1,1}$ is the first component of $\hat{\vartheta}^*_{S_1}$ (note that $h_1^{-1}(h_2(\vartheta_2)) = \vartheta_1$). An application of Pruscha (2000, Corollary 4.1, p. 249) gives that this is asymptotically equivalent to the condition $B_1\hat{X}_1 \geq \frac{(h_1^{-1}[h_2(\vartheta_2)])'}{\sqrt{c_{n2}}}\hat{X}_2$. The proof for $\hat{\vartheta}^*_{S_2} \in H^I$ is carried out analogously.

From Pruscha (2000, Th. 3.4, p. 194) it follows that $\hat{X} \xrightarrow{\mathcal{D}} N_3(0, \Sigma^{-1}(\vartheta))$. Slutsky's theorem finishes the proof.

## Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft DFG grant TR 471/1. We are indebted to S. Senn, S. Lange, J. Röhmel, and H.J. Trampisch for helpful comments and discussions. Various comments of a referee are gratefully acknowledged, which lead to an improved version of this manuscript.

## References

R. E. Barlow, D. J. Bartholomew, J. M. Bremner & H. D. Brunk (1972). *Statistical Inference under Order Restrictions*. John Wiley & Sons, London.

G. A. Barnard (1945). A new test for 2x2 tables. Nature, 156, 177.

G. A. Barnard (1947). Significance tests for 2x2 tables. *Biometrika*, 34, 123-138.

R. L. Berger (1997). Likelihood ratio tests and intersection-union. In S. Panchapakesan and N. Balakrishnan, editors, *Advances in Statistical Decision Theory and Applications*, Chapter 15, pages 225-237. Birkhäuser.

W. C. Blackwelder (1982). "Proving the null hypothesis" in clinical trials. *Controlled Clinical Trials*, 3, 345-353.

R. D. Boschloo (1970). Raised conditional level of significance for the 2x2 table when testing the equality of two probabilities. *Statistica Neerlandica*, 24, 1-35.

I. S. F. Chan (1998). Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. *Statistics in Medicine*, 17, 1403-1413.

J. J. Chen, Y. Tsong & S.-H. Kang (2000). Tests for equivalence or noninferiority between two proportions. *Drug Information Journal*, 34, 569-578.

E. N. Chouela, A. M. Abeldano, G. Pellerano, M. La Forgia, R. M. Papale, A. Garsd, M. C. Balian, V. Battista & N. Poggio (1999). Equivalent therapeutic efficacy and safety of ivermectin and lindane in the treatment of human scabies. *Archives of Dermatology*, 135, 651-655.

C. Chuang-Stein (2001). Testing for superiority or inferiority after concluding equivalence? *Drug Information Journal*, 35, 141-143.

Committee for Proprietary Medicinal Products (1998a). Note for guidance on clinical investigation of medicinal products in the treatment of Parkinsons disease. CPMP/EWP/563/95.

Committee for Proprietary Medicinal Products (1998b). Note for guidance on the clinical investigation of medicinal products in the treatment of schizophrenia. CPMP/EWP/563/95.

Committee for Proprietary Medicinal Products (2001). Note for guidance on clinical evaluation of medicinal products for the treatment and prevention of bipolar disorder. CPMP/EWP/567/98.

Committee for Proprietary Medicinal Products (2002a). Note for guidance on clinical investigation of medicinal products for treatment of nociceptive pain. CPMP/EWP/612/00.

Committee for Proprietary Medicinal Products (2002b). Note for guidance on clinical investigation of medicinal products in the treatment of depression. CPMP/EWP/518/97, Rev. 1.

R. B. D'Agostino, W. Chase & A. Belanger (1988). The appropriateness of some common procedures for testing the equality of two independent binomial populations. *The American Statistician*, 42, 198-202.

H. G. Dammann, U. R. Folsch, E. G. Hahn, D. H. von Kleist, H. U. Klor, T. Kirchner, S. Strobel & M. Kist (2000). Eradication of h. pylori with pantoprazole, clarithromycin, and metronidazole in duodenal ulcer patients: A head-to-head comparison between two regimens of different duration. *Helicobacter*, 5, 41-51.

C. Diehm, H. J. Trampisch, S. Lange & C. Schmidt (1996). Comparison of leg compression stocking and oral horse-chestnut seed extract therapy in patients with chronic venous insufficiency. *Lancet*, 347, 292-294.

C. W. Dunnett & M. Gent (1977). Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. *Biometrics*, 33, 593-602.

C. W. Dunnett & A. C. Tamhane (1997). Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. *Statistics in Medicine*, 16, 2489-2506.

C. P. Farrington & G. Manning (1990). Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk. *Statistics in Medicine*, 9, 1447-1454.

D. Greco, S. Salmaso, P. Mastrantonio, M. Giuliano, A. E. Tozzi, A. Anemona, M. L. Ciofi degli Atti, A. Giammanco, P. Panei, W. C. Blackwelder, D. L. Klein & S. G. Wassilak (1996). A controlled trial of two acellular vaccines and one whole-cell vaccine against pertussis. Progetto Pertosse Working Group. *New England Journal of Medicine*, 334, 341-348.

L. Gustafsson, H. O. Hallander, P. Olin, E. Reizenstein & J. Storsaeter (1996). A controlled trial of a two-component acellular, a five-component acellular, and a whole-cell pertussis vaccine. *New England Journal of Medicine*, 334, 349-355.

P. Hesketh, R. Navari, T. Grote, R. Gralla, J. Hainsworth, M. Kris, L. Anthony, A. Khojasteh, E. Tapazoglou, C. Benedict & W. Hahne (1996). Double-blind, randomized comparison of the antiemetic efficacy of intravenous dolasetron mesylate and intravenous ondansetron in the prevention of acute cisplatin-induced emesis in patients with cancer. Dolasetron comparative chemotherapy-induced emesis prevention group. *Journal of Clinical Oncology*, 14, 2242-2249.

S. Holm (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65-70.

Hypericum Depression Trial Study Group (2002). Effect of hypericum perforatum (St John's wort) in major depressive disorder: a randomized controlled trial. *Journal of the American Medical Association*, 287, 1807-1814.

S. Lange (2003). *Äquivalenzbereiche in klinischen Therapiestudien*. 2nd thesis, Ruhr-University Bochum, Germany.

S. Lange & G. Freitag (2005). Choice of delta: requirements and reality - results of a systematic review. *Biometrical Journal*, 47, 12-27.

R. Marcus, E. Peritz & K. R. Gabriel (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63, 655-660.

A. Martín Andrés & I. Herranz Tejedor (2004). Exact unconditional non-classical tests on the difference of two proportions. *Computational Statistics and Data Analysis*, 45, 373-388.

O. Miettinen & M. Nurminen (1985). Comparative analysis of two rates. *Statistics in Medicine*, 4, 213-226.

A. Munk, G. Skipka & B. Stratmann (2005). Testing general hypotheses under binomial sampling: The two sample case - asymptotic theory and exact procedures. *Comp. Stat. Data Analysis*,

49, 723-739.

I. Pigeot, J. Schäfer, J. Röhmel & D. Hauschke (2003). Assessing non-inferiority of a new treatment in a three-arm clinical trial including a placebo. *Statistics in Medicine*, 22, 883-899.

H. Pruscha (2000). *Vorlesungen über Mathematische Statistik*. B. G. Teubner, Stuttgart.

J. Röhmel & U. Mansmann (1999). Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. *Biometrical Journal*, 41, 149-170.

T. Robertson, F. T. Wright & R. L. Dykstra (1988). *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester.

C. Rodary, C. Com-Nougue & M. F. Tournade (1989). How to establish equivalence between treatments: A one-sided clinical trial in paediatric oncology. *Statistics in Medicine*, 8, 593-598.

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

G. Skipka, A. Munk & G. Freitag (2004). Unconditional exact tests for the assessment of non-inferiority for the difference of binomial probabilities - contrasted and compared. *Computational Statistics and Data Analysis*, 47, 757-773.

B. E. Storer & C. Kim (1990). Exact properties of some exact test statistics for comparing two binomial proportions. *Journal of the American Statistical Association*, 85, 146-155.

M.-L. Tang & N.-S. Tang (2004). Tests of noninferiority via rate difference for three-arm clinical trials with placebo. *Journal of Biopharmaceutical Statistics*, 14, 337-347.

U. Tebbe, R. Michels, J. Adgey, J. Boland, A. Caspi, B. Charbonnier, J. Windeler, H. Barth, R. Groves, G. R. Hopkins, W. Fenell, A. Betriu, M. Ruda & J. Mlczoch (1998). Randomized, doubleblind study comparing saruplase with streptokinase therapy in acute myocardial infarction: The compass equivalence trial. *Journal of the American College of Cardiology*, 31, 487-493.

G. J. G. Upton (1982). A comparison of alternative tests for the 2x2 comparative trial. *Journal of the Royal Statistical Society Series A*, 145, 86-105.

B. L. Wiens & B. Iglewicz (1999). On testing equivalence of three populations. *Journal of Biopharmaceutical Statistics*, 9(3), 465-483.