Testing general hypotheses under binomial sampling: The two sample case - asymptotic theory and exact procedures

A. Munk [1], G. Skipka, B. Stratmann,

Institute for Mathematical Stochastics, Georg-August University Göttingen, Germany

Abstract

A unified exact approach is given for testing hypotheses of the form $H_0 : \vartheta_1 \geq h(\vartheta_2)$ versus $H_1 : \vartheta_1 < h(\vartheta_2)$ where ϑ_1 and ϑ_2 are the failure rates of two independent groups and h is a monotone function. This includes the classical problem of testing equality of ϑ_1 and ϑ_2 as well as the problem of showing non-inferiority and superiority in clinical trials with respect to various commonly used measures, such as the difference, the relative risk or the odds ratio. Conditions on h are given in order to guarantee uniqueness of the restricted maximum likelihood estimator. Further, it is shown that the likelihood ratio test for H_0 versus H_1 follows asymptotically a $\frac{1}{2} + \frac{1}{2}F_{\chi_1^2}$-law, provided h is a smooth curve. However, it is found that this asymptotics often is not sufficiently reliable for small and moderate sample sizes, $n_1, n_2 \leq 100$, say, and an exact modification will be given. This test is compared to competitors from the literature and its superiority with respect to power will be shown numerically. Procedures will be illustrated by a clinical trial.

Key words: Non-inferiority, shifted null hypothesis, unconditional test, order-restricted inference, restricted maximum likelihood, exact test, ordered alternatives, equivalence testing, superiority

1 Introduction

A well known problem in statistical inference is the comparison of the effects of two treatments (one might be placebo) in a controlled clinical trial where

[1] Corresponding address: Institut für Mathematische Stochastik, Maschmühlenweg 8-10, D-37073 Göttingen, Germany, tel: +49-551-3913501, fax: +49-551-3913521, email: munk@math.uni-goettingen.de, url: www.stochastik.math.uni-goettingen.de

the primary endpoint is a dichotomous quantity, such as a success or failure rate. Recently, it became also of interest to show non-inferiority of a treatment group "1" with respect to a control group "2", i.e. to show that there is at most an irrelevant "difference" in failure rates of the new treatment. Closely related to this, in superiority trials the aim is to establish a relevant superiority of a new treatment compared to a standard, say. For an extensive discussion see Chan (1998); Chuang-Stein (2001); Dunnett and Tamhane (1997); Greco et al. (1996); Gustafsson et al. (1996); Moulton et al. (2001); Röhmel and Mansmann (1999b). The most common ways to deal with this problem is to test a proper hypotheses (to be described later on) or to base a decision on a confidence interval (see e.g. Newcombe (1998) for a survey). In this paper we will focus solely on testing methods. We mention, however, that in principle all procedures can be used to obtain confidence intervals by proper inversion (Casella and Berger, 2002, ch. 9.2). We refer to Chan (1998), Farrington and Manning (1990), Roebruck and Kühn (1995) or Skipka et al. (2004) for a survey on testing methods for the difference of the failure rates. However, there is a controversial discussion how to measure non-inferiority properly. In addition to the difference $\vartheta_1 - \vartheta_2 = \theta_{DI}$ various authors suggest the relative risk $\vartheta_1/\vartheta_2 = \theta_{RR}$ or the odds ratio $\vartheta_1(1-\vartheta_2)/(\vartheta_2(1-\vartheta_1)) = \theta_{OR}$. The ASSENT-2 trial (1999) compares two thrombolytic therapies with respect to 30 day mortality. Here θ_{DI} as well as θ_{RR} is evaluated.

Proper hypotheses associated with these quantities are of the form $H_0 : \theta \geq \theta_0$ where θ is a measure of non-inferiority (such as θ_{DI}, θ_{RR}, or θ_{OR}) and θ_0 is a positive quantity to be specified. Typical values are $\theta_0 = 0.1, 0.15, 0.2$ for the difference and $\theta_0 = 1.1, 1.2, 1.5$ for the relative risk or the odds ratio, say CPMP (2003, 1999); FDA (1992, 1998); ILAE (1998); InTIME-II (2000); Moliterno and Topol (2000). Phillips (2002) considered hypotheses with linear inequalities $\vartheta_1 \geq a + b\,\vartheta_2$ for fixed a and b and provided an asymptotic test (based on a standardized z-statistic with unpooled variance estimates). Here θ would correspond to $\theta_P = \vartheta_1 - a - b\vartheta_2$.

Recently, Röhmel and Mansmann (1999b) argued forcefully that even more general hypotheses are of interest. These authors consider various types of hypotheses which can be described as

$$H_0 \colon \vartheta_1 \geq h(\vartheta_2) \text{ versus } H_1 \colon \vartheta_1 < h(\vartheta_2)\,. \tag{1}$$

Here h is an increasing curve $h : [0,1] \to [0,1]$ which has to be specified in advance. This includes in particular the above mentioned quantities for

$$h_{DI}(\vartheta_2) = \vartheta_2 + \theta_0, \quad h_{RR}(\vartheta_2) = \vartheta_2\,\theta_0, \quad h_{OR}(\vartheta_2) = \frac{\theta_0}{\theta_0 + \vartheta_2^{-1} - 1} \tag{2}$$

or Phillips' (2002) hypotheses.

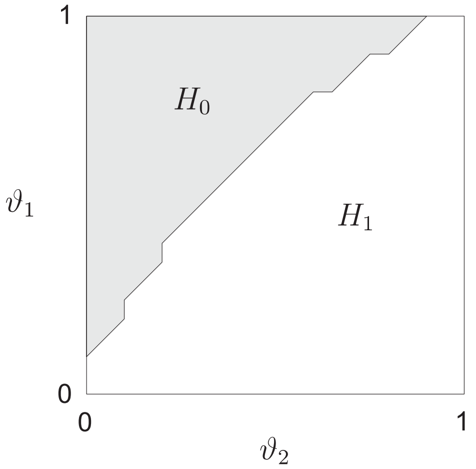

Fig. 1. Parameter space FDA

More general, h might take into account that different measures of discrepancies as well as different values of θ_0 have to be combined in one quantity, depending on the underlying response rate.

Based on recent guidelines of the FDA (1992) and CPMP (2003; 1999), Röhmel and Mansmann (1999b) (see also Bristol (1996)) considered such curves h. Some of them may even discontinuous, however, always being increasing.

As an example, the FDA (1992) requires that non-inferiority can be claimed if the two-sided 95% confidence interval around the difference in response rates must be within θ_0 with

$$\theta_0 = \begin{cases} 20\% & < 80\% \\ 15\% \quad \text{if} \quad \max\left\{\hat{\vartheta}_1, \hat{\vartheta}_2\right\} & \in [80\%, 90\%) \\ 10\% & \geq 90\% \end{cases} .$$

If this rule is applied by replacing the observed rates with the true rates and extrapolating the margins in a symmetric way for small rates (< 0.5) it results in the curve displayed in Figure 1. For a careful discussion and other examples we refer to (Röhmel and Mansmann, 1999b, p. 151-153).

In fact, the proper choice of h is a subtle problem and will depend on many aspects, e.g. the clinical field of application. We will not pursue the issue of the most appropriate hypotheses further, instead we will present a general statistical methodology which allows to deal with any isotonic curve h. To this end in Section 2 the likelihood ratio test for (1) will be constructed and we will show that for smooth h the asymptotic distribution is $\frac{1}{2}+\frac{1}{2}F_{\chi_1^2}$, exactly as for the case where h is the identity (cf. Robertson and Wright (1981)).

The paper will be organised as follows. In Section 2 it is shown that uniqueness of the MLE depends heavily on the function h. Even for strictly increasing and smooth h uniqueness cannot be guaranteed, in general. Conditions on the "boundary function" h will be given which guarantee uniqueness of the MLE. This yields a unifying result for the uniqueness of the MLE for all measures of non-inferiority discussed so far. Further, it highlights an interesting difference between superiority and non-inferiority trials. In superiority trials often the null hypothesis will be convex, which immediately implies uniqueness of the MLE, whereas in non-inferiority trials convexity of H_0 is not the typical case. A simple counterexample is given, for which two solutions of the MLE exist.

Skipka et al. (2004) introduced an exact modification of the LR-test for the difference (boundary function h_{DI}), which is based on an idea of Storer and Kim (1990). In Section 3 we extend this unconditional exact approach to general boundary functions h. This even applies to nonsmooth curves h (cf. Figure 1). In Section 4 the exact LR-test is compared numerically with various competitors from the literature for the relative risk and the odds ratio, respectively. Our method is similar in spirit to Röhmel and Mansmann (1999). However, we propose the cumulative likelihood ratio function as a criterion to order the sample space. To this end additionally the restricted likelihood estimators have to be determined and will be used as initial estimators for the parameters of the likelihood function. It will be shown that this exact (modification of the) LR-test in general provides a larger power than its competitors for both specified curves h_{RR} and h_{OR}. This extends the results for the special case of a difference, h_{DI} by Skipka et al. (2004). The power enhancement can be quite substantial (up to 10%), which implies a reduction of the sample size up to 15%. It is demonstrated that in particular for the relative risk our methods perform very well.

In Section 5 a medical application is discussed and it is illustrated how the presented methods perform.

The paper closes by a summary section, where possible extensions are briefly discussed. SAS source code is available on request from the authors.

All proofs are postponed to Appendix A.

2 The likelihood ratio test for general hypotheses - asymptotic theory

Throughout the following let

$$X_1, \ldots, X_{n_1} \overset{i.i.d.}{\sim} B(1, \vartheta_1) \quad \text{and} \quad Y_1, \ldots, Y_{n_2} \overset{i.i.d.}{\sim} B(1, \vartheta_2)$$

two independent Bernoulli samples with failure rates ϑ_1 and ϑ_2, respectively. Hence the joint likelihood is given as

$$L(\vartheta_1, \vartheta_2) = \binom{n_1}{x} \vartheta_1^x \, (1 - \vartheta_1)^{n_1 - x} \binom{n_2}{y} \vartheta_2^y \, (1 - \vartheta_2)^{n_2 - y} \quad , \tag{3}$$

where $x = \sum_{i=1}^{n_1} x_i$ denotes the number of negative responses in treatment group 1 and $y = \sum_{j=1}^{n_2} y_j$ in control group 2, respectively. Our approach is based on the likelihood ratio test for hypotheses in (1), hence we are concerned with the MLE under the restriction H_0 in (1).

The next lemma ensures that the the constrained ML-estimator restricted to H_0 in (1) can be computed on the set, where $\vartheta_1 = h(\vartheta_2)$. Furthermore, we give conditions on h which guarantee the uniqueness of the MLE restricted to H_0. Surprisingly, we will find that this is not always the case and counterexamples will be given. Let us write $\vartheta = (\vartheta_1, \vartheta_2)$, $\Theta = [0,1]^2$. Let $\hat{\vartheta} = (\hat{\vartheta}_1, \hat{\vartheta}_2) = (\frac{x}{n_1}, \frac{y}{n_2})$, the unrestricted MLE.

Lemma 1 *Let $\Theta_0 = \{\vartheta : \ \vartheta_1 \geq h(\vartheta_2)\}$ and assume $X_1, \ldots, X_{n_1} \sim B(1, \vartheta_1)$ i.i.d. and independently $Y_1, \ldots, Y_{n_2} \sim B(1, \vartheta_2)$ i.i.d., where $n_1, n_2 \geq 1$. Let h be continuous and increasing, and not identically 1.*

a) Then, the MLE restricted to Θ_0 exists and is given as $\hat{\vartheta}^* = \hat{\vartheta}$ (the unrestricted MLE) if $\hat{\vartheta} \in \Theta_0$ and if $\hat{\vartheta} \notin \Theta_0$

$$\hat{\vartheta}^* = \left\{ \arg \max_{\{\vartheta : \vartheta_1 = h(\vartheta_2)\}} L(\vartheta) \right\} \subseteq \partial\Theta_0, \tag{4}$$

i.e. the MLE is attained on the boundary $\partial\Theta_0$ of Θ_0.

b) Let h twice differentiable; $h \in C^2[0,1]$. The restricted MLE $\hat{\vartheta}^*$ is unique if the following conditions (i) and (ii) or if the condition (iii) are satisfied on the set $\Theta_h = \{\vartheta_1 = h(\vartheta_2)\}$:

(i) $-(h')^2 + h \cdot h'' \leq 0$

(ii) $-(h')^2 - h'' + h \cdot h'' \leq 0$

(iii) h is convex.

If we apply Lemma 1 to the functions in (2) we find that the restricted MLE can always be computed on the set Θ_h for $h = h_{DI}, h_{RR}, h_{OR}$, respectively and that the MLE is unique. This follows for h_{DI} and h_{RR} by (iii), where for h_{OR} we observe that (i) reads as

$$-(h'(\vartheta_2))^2 + h(\vartheta_2) \cdot h''(\vartheta_2) = -\frac{\theta_0^2}{(1 + \vartheta_2(\theta_0 - 1))^3} \leq 0 \ ,$$

and (ii) as

$$-(h'(\vartheta_2))^2 - h''(\vartheta_2) + h(\vartheta_2) \cdot h''(\vartheta_2) = -\frac{\theta_0}{(1+\vartheta_2(\theta_0-1))^3} \leq 0 \ .$$

Explicit formulae for the MLE's in case of h_{DI} and h_{RR} are given by Miettinen and Nurminen (1985) and Farrington and Manning (1990). For h_{OR} the restricted MLE is found to be

$$\begin{aligned}\hat{\vartheta}_2^* =& \frac{1}{2n_2\,(\theta_0-1)}\Bigg[\theta_0(x+y-n_1) - x - y - n_2 \\ &+ \sqrt{(x+y+n_2-x\theta_0-y\theta_0+n_1\theta_0)^2 + 4\,(x+y)\,n_2\,(\theta_0-1)}\Bigg], \\ \hat{\vartheta}_1^* =& \left[1+\theta_0^{-1}(\hat{\vartheta}_2^{*^{-1}}-1)\right]^{-1} \ .\end{aligned}$$

It is interesting to note that condition (iii) is in general not satisfied by most hypotheses for non-inferiority (see (Röhmel and Mansmann, 1999b, fig. 1d-1f)). However, in superiority trials these hypotheses are more important, because here H_0 will be a convex set in many cases. This makes a subtle distinction between non-inferiority and superiority trials: The restricted MLE in the latter case will be typically a projection onto a convex set (the null hypothesis), and hence unique, in non-inferiority trials often the alternative is a convex set, hence uniqueness has to be checked carefully, e.g. by means of Lemma 1b) (i),(ii). Observe finally that (i) together with (ii) guarantee that the likelihood function is strictly concave on Θ_h which allows a quick computation by the use of any standard maximization routine.

Remark 2 *As mentioned above uniqueness of the MLE (albeit always located on the set* $\{\vartheta : \vartheta_1 = h(\vartheta_2)\}$*) is not valid for every increasing function h. In fact various global maxima can occur for certain outcomes* (x,y) *and hypotheses* H_0*. A simple class of counterexamples is as follows. Let* $n_1 = n_2 = n$ *and* $\frac{x}{n} = 1 - \frac{y}{n}$ *and consider the case where* $(\frac{x}{n}, \frac{y}{n}) \in H_1$*, i.e. where the restricted MLE does not equal the unrestricted one. Let us define h as*

$$h(\vartheta_2) = \begin{cases} \frac{1}{\gamma}\ \vartheta_2 & \textit{for } \vartheta_2 \leq \frac{\gamma}{1+\gamma} \\ \gamma\ \vartheta_2 + 1 - \gamma & \textit{elsewhere} \end{cases} \tag{5}$$

for some constant $0 < \gamma < 1$ *(cf. Figure 2, here* $\gamma = 0.33$*). Observe, that h is piecewise linear and symmetric (as well as L) w.r.t.* $D = \{(\vartheta_1, \vartheta_2) : \vartheta_1 = 1 - \vartheta_2\}$*.*

Now, it can be shown (a detailed proof can be obtained from the authors on request) that for any $\gamma \in (0,1)$ *there are exactly two solutions of the MLE (denoted as* $\hat{\vartheta}_A^*$ *and* $\hat{\vartheta}_B^*$ *in Figure 2), which are symmetrical w.r.t. D, located on each of the two branches A and B of h in (5), respectively.*

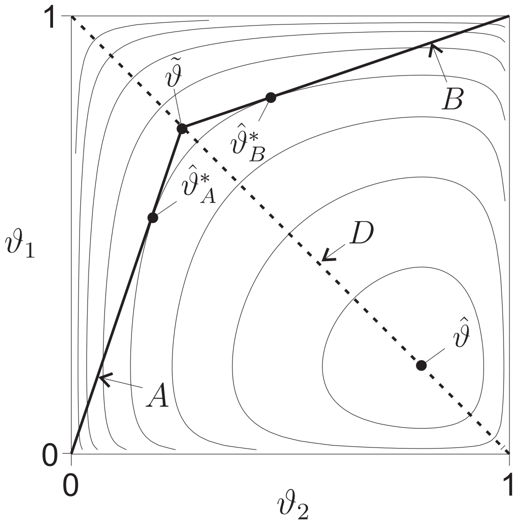

Fig. 2. Two solutions of the restricted MLE, where $n_1 = n_2 = n$ and $x = n - y$, for the hypotheses in (5). Here the contour plot of the two sample binomial likelihood shows the existence of two MLE's $\hat{\vartheta}^*_A$ and $\hat{\vartheta}^*_B$ on each branch A and B of the boundary of the hypothesis H_0, respectively

In Figure 2 the contourlines of the likelihood are displayed for this particular case and $\gamma = 0.33$. *The solutions* $\hat{\vartheta}^*_A, \hat{\vartheta}^*_B$ *are such that the hypothesis function* h *is tangent to the likelihood. From this it can also be drawn that various other hypotheses may even lead to more than two solutions of the MLE.*

Remark 3 *Lemma 1a) is related to a theorem of Röhmel and Mansmann (1999b p. 161) who showed that for fixed* $(\overline{x}, \overline{y}) := \hat{\vartheta}$ *the supremum over* $\vartheta \in \Theta_0$ *of* $\sum_{T(x,y) \leq T(\overline{x},\overline{y})} L(\vartheta)$ *is attained at the boundary of* Θ_0*, provided the statistic* T *satisfies a convexity condition "C" introduced by Barnard (1947). It states that for any* $(x, y) \in CR$ *(the critical region of a test) holds that* $(x, y+1) \in CR$ *and* $(x-1, y) \in CR$*. Note, however, that our result is different and the proof relies essentially on the uniqueness of the unrestricted MLE.*

The inherit of Lemma 1 is twofold. First, it allows to restrict the parameter space to a one dimensional curve Θ_h *for numerical computation of the restricted MLE. Second, this will be the key property for the derivation of the asymptotic distribution of the likelihood ratio statistic*

$$\lambda = \lambda(x, y) = \frac{\sup_{\vartheta \in \Theta_0} L(\vartheta)}{\sup_{\vartheta \in \Theta} L(\vartheta)}. \tag{6}$$

Theorem 4 *Let* h *be increasing,* $h : [0, 1] \to [0, 1]$ *and* $h \in C^{(1)}[0, 1]$*. Then, under the assumption of Lemma 1a) we have for* $\vartheta_1 = h(\vartheta_2)$ *and for any solution* $\hat{\vartheta}^*$

$$-2 \ln \lambda \xrightarrow{\mathcal{D}} U \sim \tfrac{1}{2} + \tfrac{1}{2} F_{\chi^2_1},$$

as $\min\{n_1, n_2\} \to \infty$*, s.t.* $\frac{n_1}{n_2} \to c \in (0, \infty)$*, where* $F_{\chi^2_1}$ *denotes the c.d.f. of the square of a standard normal random variable.*

Table 1
The actual probability (times 100) for $-2\ln(\lambda) > \left(\frac{1}{2} + \frac{1}{2}F_{\chi_1^2}\right)_{0.95}$.

		exact level		
		difference	relative risk	odds ratio
n_1, n_2	ϑ_2	$\theta_0 = 0.1$	$\theta_0 = 1.5$	$\theta_0 = 1.5$
$10, 10$	0.1	8.93	5.69	6.15
$10, 25$		10.22	9.46	10.25
$25, 25$		5.33	5.59	6.17
$25, 10$		5.27	6.27	7.01
$50, 50$		5.45	6.32	4.4
$50, 100$		5.19	5.22	6.01
$100, 100$		5.22	5.24	4.52
$100, 50$		5.37	4.86	4.31
$500, 500$		5.05	4.97	4.98
$10, 10$	0.4	5.95	4.76	5.72
$10, 25$		5.63	5.45	5.63
$25, 25$		4.46	5.26	4.36
$25, 10$		5.00	5.47	5.00
$50, 50$		4.51	5.16	4.33
$50, 100$		4.84	4.97	4.61
$100, 100$		5.05	4.86	5.05
$100, 50$		4.74	4.91	4.57
$500, 500$		5.18	4.91	5.04

Nevertheless, the approximation using the asymptotics of Theorem 4 is not always sufficient for small and moderate sample sizes.

In Table 1 the actual exact levels are drawn for different parameter constellations when using the 95% quantile $\left(\frac{1}{2} + \frac{1}{2}F_{\chi_1^2}\right)_{0.95}$ of the asymptotic distribution as critical value for a level 5% test.

From Table 1 it can be seen that the nominal level is exceeded up to twice for small sample sizes. As a very rough rule of thumb we state that the asymptotic test can be recommended if $n_1, n_2 \geq 100$, say. Of course, this will depend on the underlying (unknown) response rates. Note, as $\vartheta_1, \vartheta_2 \to 0$, Theorem 4 does not hold anymore, instead a Poisson limit is valid. In the next section, an exact modification of the asymptotic LR-test is presented, i.e. a test which keeps its nominal level exactly.

3 Exact modification

Exact tests for general hypotheses (1) were first introduced in two seminal papers by Barnard (1945, 1947). In Skipka et al. (2004) it is shown, however, that Barnard's original test bears intrinsically numerical difficulties due to its specific iterative way to construct the region of rejection. During the last two decades various other exact methods were suggested, most of them were developed for $H : \vartheta_1 = \vartheta_2$ (Boschloo (1970); McDonald et al. (1977); Upton (1982); D'Agostino et al. (1988); Berger and Boos (1994)) or for specific hypotheses in (1) (see e.g. Martín Andrés and Silva Mato (1994); Chan (1998)). Finally Röhmel and Mansmann (1999b) presented a general exact method for arbitrary hypotheses in (1), based on ideas of Barnard (1947). For details see the next section.

Our general strategy is as follows. To guarantee that the LR-test keeps its nominal level α, a modification of the LR statistic λ in (6) is applied, which is described by Skipka et al. (2004) for the boundary function h_{DI}. This approach will be transfered in the following to arbitrary boundary functions h. In a first step, based on an idea of Storer and Kim (1990), the exact distribution of the LR statistic is estimated by inserting the restricted ML estimates $(\vartheta_2^*, \vartheta_1^*)$ in (3). With that, p-values can be estimated for any outcome (x, y). In a second step these estimated p-values $p^*(x, y)$ are used to sort all possible outcomes in ascending order. This ordering defines the unconditional exact test, similary as for the case of h_{DI} (see (Skipka et al., 2004, Ch. 2) for details).

Remark 5
a) *Obviously, it is computationally more feasible to calculate the maximum on the boundary of H_0, if possible. Röhmel and Mansmann (1999a) have shown that the maximum is attained always at the boundary $\vartheta_1 = h(\vartheta_2)$, if the test fulfills Barnard's convexity condition "C". We were not able to prove that condition "C" holds for the modified LR-test (denoted as exact LR-test in the following). Therefore, the actual level α^* has to be determined by maximizing over the entire null space, in principle. Nevertheless, note that it is still feasible to restrict the calculation of the maximum to the boundary of the null space in Section 4. To this end we simply check numerically condition "C" after sorting the outcomes for every parameter setting. We mention that in all numerical examples investigated so far we never found a violation of condition "C". Hence, for a given testing problem, we recommend to check condition "C" numerically. If this is satisfied, numerical maximization can be performed on the boundary. If not, maximization over the entire null space has*

to be performed.

b) *It would be tempting to base the ordering of the sample space directly on the likelihood ratio statistic $\lambda(x, y)$, instead of the cumulative likelihood function. We found, however, that this approach does lead to a test with rather low power compared to the present approach and, hence, cannot be recommended in practice.*

c) *Berger and Boos (1994) introduced an exact method where the actual level is determined by maximization over a confidence region of the unknown parameters instead of maximization over the entire parameter space. We applied this approach to the exact unconditional tests - mentioned in the following section, however, we found no improvement. This is in accordance with extensive investigations done by Chan and Zhang (1999) who argued that "the search of nuisance parameter over a restricted domain does not offer benefits ... as the tail probability often peaks in the middle of the domain of the nuisance parameter". Agresti and Min (2001) came to the same conclusion.*

4 Power investigation

The exact LR-test is compared with various commonly used exact approaches. The competitors will be briefly introduced in the sequel.

- *Chan's test*: Chan (1998) has suggested an unconditional exact approach for h_{DI} and h_{RR}, where the standardized z statistic with ML variance estimates, restricted to $\vartheta_1 = h(\vartheta_2)$, is used as an ordering criterion. This test statistic was originally introduced by Farrington and Manning (1990) (see this reference for explicit formulae). Chan's test is constructed in a similar way as the exact LR-test, but uses Farrington & Manning's test statistic as the ordering criterion.
- π_{local}*-test*: Röhmel and Mansmann (1999a) have constructed an unconditional exact test, denoted as π_{local}, by using the "smallest possible p-value according to condition C" stated in sect. 2,

  $$\pi_{min}(x, y) = \max_{\vartheta_1 = h(\vartheta_2)} P(X \leq x, Y \geq y | (\vartheta_1, \vartheta_2)),$$

  as the ordering criterion. This test is applicable for all specifications of a monotone curve h.
- *Fisher's exact test (unconditional exact adaption)*: If $m := x + y$ is fixed, the conditional exact test (see e.g. Gart (1971)) can be adapted for h_{OR} by using the conditional exact p-values as an ordering criterion to construct an unconditional exact version in the same way as before. Therefore this adaption, denoted as Fisher's exact unconditional test in the following, is

the generalization of the test introduced by McDonald et al. (1977) for the classical hypothesis with $\vartheta_1 = h(\vartheta_2) = \vartheta_2$.

Remark 6 *The comparison with Barnard's test (1947) (more precise: with Röhmel & Mansmann's (1999a) adaption of Barnard's test for the hypothesis (1)) is omitted because the investigations of Skipka et al. (2004) have shown that this test is hardly applicable in practice due to intrinsic numerical difficulties.*

All tests under investigation are exact methods, i.e. they all keep the nominal level exactly. We refer to (Skipka et al., 2004, Ch. 3) for a comparison of these exact approaches for the boundary function h_{DI}. In the following, these tests are compared numerically for the two distance measures relative risk and odds ratio w.r.t. power for a broad scenario of parameter settings $(\theta_0, n_1, n_2, \vartheta_2)$:

- *Boundary of hypothesis:* We choose $\theta_0 \in \{1.1, 1.25, 1.5, 2, 2.5\}$ for h_{RR} and h_{OR}.
- *Sample size:* We choose balanced sample sizes $n_1 = n_2 \in \{20, 25, 30, 35, 40, 50, 60, 80, 100\}$ and unbalanced sample sizes $(n_1, n_2) \in \{(30, 20), (40, 20), (50, 25), (60, 30), (60, 40), (80, 40), (80, 50), (80, 60), (100, 50), (100, 60), (100, 80)\}$.
- *Nuisance parameter:* We choose $\vartheta_2 \in \{0.1, 0.2, 0.3, 0.5, 0.8, 0.9\}$.

This gives 600 different parameter configurations for every function h. Configurations regarding the the relative risk are omitted in case of non feasible settings (i.e. $\vartheta_2 \geq 1/\theta_0$ for h_{RR}). We have chosen the parameter $\theta_{RR} \leq 1$ and $\theta_{OR} \leq 1$ such that the resulting power is larger than 0.8, at least for one of the tests compared. Of course, for small sample sizes and small θ_0 there exist parameter constellations, for which no test achieves a power larger than 0.8. On the other hand, for large sample sizes and large θ_0 some parameter constellations result in a power larger than 0.9 for all tests. These cases are omitted, too. Finally, for h_{RR} 330 parameter constellations were extracted and for h_{OR} 522. The resulting values of the power function are calculated exactly for all tests under investigation by computing the exact binomial probabilities (3) for all $(x, y) \in CR$.

The Figures 3 and 4 show the power of the exact LR-test (vertical axes) and its competitors (horizontal axes) for the two distance measures h_{RR} and h_{OR}, respectively

It is found that in general the power differences between the exact LR-test and its competitors are small. But for both distance measures the power of the exact LR-test tends to be larger. In some cases the power enhancement is up to 0.1, whereas the inferiority is much smaller, if present at all. In order to illustrate this, the differences of the exact LR-test's power and the power of its competitors are displayed in Figure 5. The LR-test performs better in most

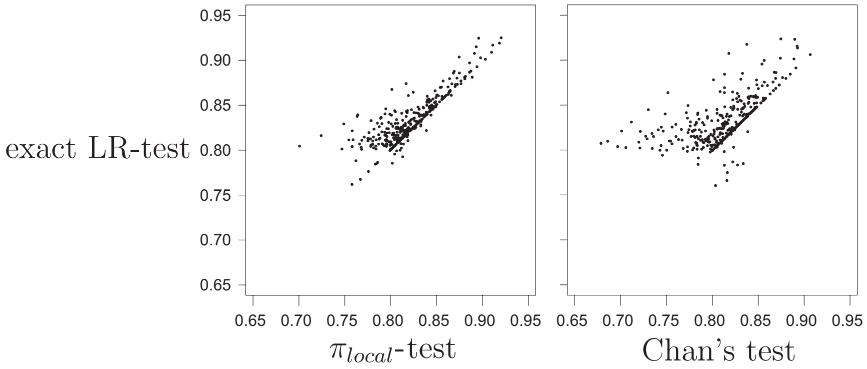

Fig. 3. The power of LR_{exact} (vertical axis) in comparison to π_{min} and Chan (horizontal axis) for several parameter constellations with $h_{RR}(\vartheta_2) = \vartheta_2\ \theta$

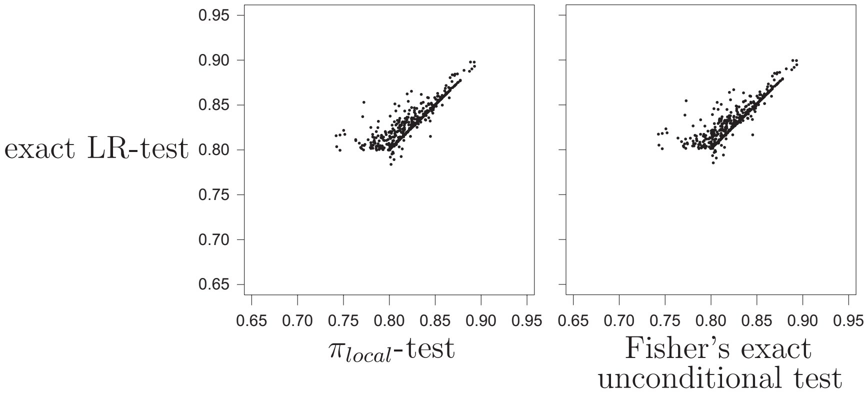

Fig. 4. The power of LR_{exact} (vertical axis) in comparison to π_{min}, and Fisher's exact unconditional (horizontal axis) for several parameter constellations with $h_{OR}(\vartheta_2) = \frac{\theta}{\theta+\vartheta_2^{-1}-1}$

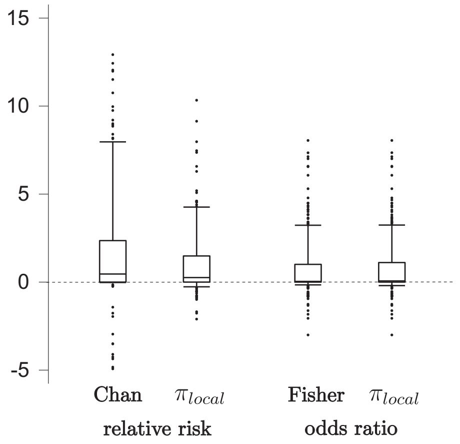

Fig. 5. Boxplot (whiskers are the 5% and 95% quantiles) for the power differences (times 100) between the exact LR-test and its competitors for the two distance measures relative risk and odds ratio.

cases of parameter constellations, even if the median power enhancement is always near zero.

The most extreme power differences and its parameter constellations are displayed in the Tables B.1 and B.2 in Appendix B. We have displayed those values separately, where the power of the exact LR-test differs from the largest power of its competitors by more than 0.015 (for h_{RR}) and 0.03 (for h_{OR}), respectively.

Finally, we briefly comment on the computational time in order to compute the critical regions of all tests. We always have found that Chan's test is the fastest method. The other methods are more time consuming: Fisher's exact unconditional test requires about 1.5 times of the computational time of Chan's test, the π_{min}-test 4 times and the exact LR-test about 6 times. In summary, however, all tests considered are computationally feasible and numerically stable.

5 Example

In a multicenter randomized double-blind study in Heliobacter pylori-positive patients, Dammann et al. (2000) compared the eradication rate of two pantoprazole-based triple therapies of different length. One group (PCM-7) received a combination of pantoprazole, clarithromycin and metronidazole during the first 7 days, followed by 7 days with placebo tablets. The other group (PCM-14) received the same combination of drugs for 14 days. An equivalence margin for the odds ratio of $\theta_0 = 0.33$ was specified. It results in an eradication rate of 89/121 for PCM-7 and 92/123 for PCM-14 referring to the intention-to-treat (ITT) population. In our notation (referring to failure rates) this results in $\hat{\vartheta}_1 = 32/121$ and $\hat{\vartheta}_2 = 31/123$. The tests, described above, show all the non-inferiority of PCM-7 over PCM-14 for the odds ratio with $\theta_0 = 3.03 \approx 1/0.33$ (p-values: 0.00021 with exact LR, 0.00025 with π_{local} and Fisher's exact unconditional test, respectively). The exact LR-test gives a slightly smaller p-value than its competitors.

The corresponding test-based upper 95%-confidence limit for the odds ratio (the smallest θ_0, for which the p-value is smaller than 0.05) are 1.76 (exact LR) and 1.74 (π_{local} and Fisher's exact unconditional). Interestingly, the confidence limit based on the exact LR-test is slightly larger, albeit in general this test was seen to be more powerful as π_{local} and Fisher's exact unconditional version.

This finding relies on the fact that the p-value as a function of the margin θ_0 is not monotone which is well known for many exact procedures (cf. Röhmel (2004)).

Dammann et al. (2000) calculated a lower 95% confidence limit (for the eradication rates) of 0.579, based on the Mantel-Haenszel-test - presumably stratified for centers (it is not exactly described in their paper). In our notation it results in $1/0.579 \approx 1.73$, which is similar to our findings, altogether.

6 Conclusions

The asymptotics for the likelihood ratio test for general hypotheses of non-inferiority or superiority was derived as a $\frac{1}{2} + \frac{1}{2} F_{\chi_1^2}$-law, independently of the function h. It is found that for small sample sizes this often is not sufficiently reliable and an exact modification of the LR-test was proposed. Nevertheless, for sample sizes larger than 100, say, the asymptotic test yields quite accurate results. This will depend on the underlying values ϑ_1, ϑ_2, of course. In an extensive numerical study it is found that the (exact) LR-test outperforms various competitors for hypotheses concerning the odds ratio and the relative risk. This is in accordance with the results from Skipka et al. (2004) for the difference. Furthermore, we have given general conditions on the boundary function h of the hypotheses to yield a unique ML estimator. This conditions are easy to check and are satisfied by current approaches.

We have not addressed sample size issues for the exact LR-test in order to control the type II error, which is , in general, an important task and will be postponed to a separate paper.

Acknowledgements

This research is supported by the Deutsche Forschungsgemeinschaft grant TR 471/1. We are indepted to A. Martin Andrés and S. Senn for helpful discussions. The helpful comments of a referee are grateful acknowledged.

A Proofs

Proof of Lemma 1:
a) Assume that $\hat{\vartheta} \notin \Theta_0$. Θ_0 is compact, L is continuous, and hence the supremum of L is attained in Θ_0, denoted as $\hat{\vartheta}^*$. In order to prove that the maximum is attained on the set $\Theta_h := \{\vartheta : \vartheta_1 = h(\vartheta_2)\}$ assume that $\hat{\vartheta}^* \notin \Theta_h$. If $\vartheta_1 \in \{0,1\}$ or if $\vartheta_2 \in \{0,1\}$ then $L(\vartheta) = 0$. But for all $\vartheta \in \overset{\circ}{\Theta}_0$, where $\overset{\circ}{\Theta}_0 = \Theta_0 \backslash \partial\Theta_0$, $L(\vartheta)$ is strictly positive. Therefore $\hat{\vartheta}^* \in \overset{\circ}{\Theta}_0$, which is not empty, because $h \not\equiv 1$.

$\overset{\circ}{\Theta}_0$ is open, hence there exists U_0 open in $[0,1]^2$; s.t. $\hat{\vartheta}^*$ is a local maximum in $U_0 \subseteq \overset{\circ}{\Theta}_0$. However, $L \in C^{(1)}(\Theta)$ and $\frac{\partial L}{\partial \vartheta}$ has at most a single zero in $\overset{\circ}{\Theta} = \Theta \backslash \partial\Theta$ because for any $x, y \in \{0, \ldots, n_1\} \times \{0, \ldots, n_2\}$ we have

$$\begin{aligned}
\frac{\partial}{\partial \vartheta_1} L\left(\vartheta_1, \vartheta_2\right) = 0 \quad &\Leftrightarrow \quad \vartheta_2 \in \{0,1\} \vee \vartheta_1 \in \left\{0, 1, \frac{x}{n_1}\right\} \\
\frac{\partial}{\partial \vartheta_2} L\left(\vartheta_1, \vartheta_2\right) = 0 \quad &\Leftrightarrow \quad \vartheta_1 \in \{0,1\} \vee \vartheta_2 \in \left\{0, 1, \frac{y}{n_2}\right\}.
\end{aligned}$$

Hence, any local maximum in $\overset{\circ}{\Theta}_0$ is also global, i.e. $\hat{\vartheta}^* = \hat{\vartheta}$, the unrestricted MLE. However, we have assumed that $\hat{\vartheta} \notin \Theta_0$, which gives a contradiction.

b) Define the function $\Psi(\vartheta_2) = \ell(h(\vartheta_2), \vartheta_2)$ where

$$\ell(\vartheta_1, \vartheta_2) := x \log \vartheta_1 + (n_1 - x) \log(1 - \vartheta_1) + y \log \vartheta_2 + (n_2 - y) \log(1 - \vartheta_2)$$

denotes the log-likelihood, $\ell = \log L$ (omitting the constant term $\log \binom{n_1}{x} + \log \binom{n_2}{y}$). We have

$$\begin{aligned}
\Psi''(\vartheta_2) = &-\frac{y}{\vartheta_2^2} - \frac{n_2 - y}{(1-\vartheta_2)^2} - \frac{x}{h^2(\vartheta_2)}(h'(\vartheta_2))^2 + \frac{x}{h(\vartheta_2)} h''(\vartheta_2) \\
&- \frac{n_1 - x}{(1 - h(\vartheta_2))^2}(h'(\vartheta_2))^2 - \frac{n_1 - x}{1 - h(\vartheta_2)} h''(\vartheta_2) \\
= &-\frac{y}{\vartheta_2^2} - \frac{n_2 - y}{(1-\vartheta_2)^2} + \frac{x}{h^2(\vartheta_2)}\Big(- (h'(\vartheta_2))^2 + h(\vartheta_2) \cdot h''(\vartheta_2)\Big) \\
&+ \frac{n_1 - x}{(1 - h(\vartheta_2))^2}\Big(- (h'(\vartheta_2))^2 - (1 - h(\vartheta_2)) h''(\vartheta_2)\Big) .
\end{aligned}$$

Now, if (i) and (ii) are fulfilled, $\Psi''(\vartheta_2) < 0$ and hence Ψ is strictly concave on the set Θ_h.

In order to prove (iii) observe that ℓ is strictly concave and hence the maximum on Θ_0, which is convex because h is convex, is unique. □

Proof of Theorem 4: First, note that by means of Lemma 1 we have

$$\lambda = \begin{cases} 1 & \text{if } \hat{\vartheta} \in \Theta_0 \\ \dfrac{L(\hat{\vartheta}^*)}{L(\hat{\vartheta}\,)} & \text{if } \hat{\vartheta} \notin \Theta_0 \end{cases} . \tag{A.1}$$

Furthermore, for $t \geq 0$

$$\begin{aligned} P\left(-2\ln\lambda \leq t\right) = & P\left(\{-2\ln\lambda \leq t\} \cap \left\{\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2)\right\}\right) \\ & + P\left(\{-2\ln\lambda \leq t\} \cap \left\{\hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right\}\right) =: I + II\,. \end{aligned}$$

By means of (A.1) we have $\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2) \Leftrightarrow \lambda = 1 \Leftrightarrow -2\ln\lambda = 0$ and hence if $\vartheta_1 = h\left(\vartheta_2\right)$

$$I = P\left(\hat{\vartheta}_1 \geq h(\hat{\vartheta}_2)\right) = P\left(\hat{\vartheta}_1 - \vartheta_1 \geq h(\hat{\vartheta}_2) - h\left(\vartheta_2\right)\right) \overset{n_2,n_1\to\infty}{\longrightarrow} P\left(Z_1 \geq Z_2\right),$$

where Z_1 and Z_2 are independent normal random variables with mean zero and variance $\tau_1 = \vartheta_1\left(1-\vartheta_1\right)$ and $\tau_2 = c\left(h'\left(\vartheta_2\right)\right)^2 \vartheta_2\left(1-\vartheta_2\right)$, respectively. Observe, that $P\left(Z_1 \geq Z_2\right) = \frac{1}{2}$ always, even if $h'\left(\vartheta_2\right) = 0$. Now

$$\begin{aligned} II = & P\left(-2\ln\lambda \leq t | \hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) P\left(\hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) \\ = & \frac{1}{2} P\left(-2\ln\lambda \leq t | \hat{\vartheta}_1 < h(\hat{\vartheta}_2)\right) + o\left(1\right) \\ = & \frac{1}{2} P\left(-2\ln\lambda \leq t | -2\ln\lambda > 0\right) + o\left(1\right) \\ = & \frac{1}{2} P\left(\chi_1^2 \leq t\right) + o\left(1\right), \end{aligned}$$

where in the last step we have applied a modification of a theorem of Pruscha (2000, p. 253). In order to apply this theorem note, that $-2\ln\lambda > 0$ ensures that $\hat{\vartheta}^* \in \Theta_h$ and $\hat{\vartheta} \in [0,1]^2 \backslash \Theta_0$. □

B Tables

References

Agresti, A., Min, Y., 2001. On small-sample confidence intervals for parameters in discrete distributions. Biometrics 57, 963–971.

ASSENT, 1999. Single-bolus tenecteplase compared with front-loaded alteplase in acute myocardial infarction: The assent-2 double-blind randomised trial. Assessment of the safety and efficacy of a new thrombolytic investigators. The Lancet 354 (9180), 716–722.

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

Barnard, G. A., 1947. Significance tests for 2x2 tables. Biometrika 34, 123–138.

Table B.1
Relative risk: The exact power γ (times 100) for parameter constellations, which gives the most extreme power differences between the exact LR test and the best of it's competitors ($|\gamma_{exact\,LR} - \max\{\gamma_{competitors}\}| > 0.015$)

θ_0	n_1	n_2	ϑ_2	ϑ_1	exact LR	Chan	π_{local}
1.1	60	30	0.3	0.09	84	78.4	81.2
1.5	60	40	0.2	0.07	82.8	78.1	80
2.5	100	50	0.1	0.04	82.3	74.1	79.5
2.5	50	25	0.2	0.09	81.1	75.3	78.4
1.1	100	50	0.9	0.81	84.1	74.9	81.8
1.1	80	40	0.3	0.12	81.6	76.1	79.4
1.1	60	60	0.8	0.64	87.6	85.7	85.7
2.5	30	30	0.3	0.24	81.4	79.5	79.6
2.5	80	40	0.1	0.025	82.4	76.6	80.6
1.1	100	100	0.9	0.855	84.7	82.3	83
1.1	100	60	0.9	0.81	88.7	86.9	86.9
1.1	60	40	0.9	0.765	86.5	82.2	84.8
1.1	60	30	0.9	0.72	88.5	81.3	86.8
1.1	30	20	0.9	0.675	80	74.3	78.4
1.25	60	60	0.5	0.35	85.5	83.2	84
1.5	30	30	0.3	0.12	78.5	78.5	80.2
1.25	25	25	0.5	0.225	82.1	83.9	83.9
1.1	80	80	0.1	0.015	80.4	82.3	78.4
2	30	30	0.3	0.18	79.2	78.5	81.3
2	80	60	0.1	0.035	78.4	81.3	78.3
2	25	25	0.2	0.04	78.7	82.2	76.2
1.5	40	40	0.2	0.05	81.2	84.7	81.1
2	40	20	0.2	0.04	77.6	81.7	77.6
2	80	40	0.1	0.02	76.1	80.4	75.8
2.5	30	20	0.2	0.05	78.4	82.8	78.4
1.1	80	60	0.1	0.01	78.6	83.4	78.6
1.5	60	60	0.1	0.015	76.7	81.6	76.7

Table B.2
Odds ratio: The exact power γ (times 100) for parameter constellations, which gives the most extreme power differences between the exact LR test and the best of it's competitors ($|\gamma_{exact\,LR} - \max\{\gamma_{competitors}\}| > 0.03$)

θ_0	n_1	n_2	ϑ_2	ϑ_1	exact LR	Fisher's exact uncond.	π_{local}
1.25	100	50	0.1	0.011	85.4	77.3	77.3
2.5	40	40	0.1	0.016	81.6	74.3	74.3
1.5	80	50	0.1	0.016	82.2	75.1	75.1
2.5	50	50	0.1	0.027	81.7	74.7	74.7
1.1	35	35	0.2	0.024	83.8	77.2	77.2
2.5	20	20	0.2	0.036	81.8	75.2	75.2
1.25	30	30	0.2	0.024	80.4	74.3	74.3
2	80	80	0.1	0.043	80	74.7	74.7
1.1	60	60	0.2	0.059	81.2	76.4	76.4
2.5	80	40	0.1	0.022	82.6	78.1	78.1
2	50	50	0.1	0.011	85.2	80.7	80.7
1.5	100	60	0.1	0.022	83.2	78.9	78.9
1.25	35	35	0.2	0.024	86.3	82	82
1.1	60	30	0.2	0.024	86.6	82.5	82.5
2.5	100	60	0.1	0.048	80.8	76.8	76.8
2	80	50	0.1	0.027	81.1	77.2	76.4
2	25	25	0.2	0.024	84.9	81.1	81.1
1.1	100	80	0.1	0.016	83.8	80.2	80.2
1.25	40	40	0.2	0.036	84.4	80.8	80.8
2	100	100	0.1	0.053	81.2	77.8	77.8
1.1	40	40	0.2	0.036	80.5	77.1	77.1
1.25	80	60	0.1	0.011	85.5	82.1	82.1
1.25	60	60	0.2	0.07	80.3	77	77
2.5	60	60	0.1	0.037	80.7	77.4	77.4
1.1	50	50	0.2	0.048	81.5	78.3	78.3
2.5	100	80	0.1	0.058	82.5	79.3	78.8
1.25	100	60	0.1	0.016	80.2	77.1	77.1
2	80	50	0.2	0.121	80.5	77.5	77.3

Berger, R. L., Boos, D. D., 1994. P values maximized over a confidence set for the nuisance parameter. Journal of the American Statistical Association 89, 1012–1016.

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

Bristol, D. R., 1996. Determining equivalence and the impact of sample size in anti-infective studies: A point to consider. Journal of Biopharmaceutical Statistics 6 (3), 319–326.

Casella, G., Berger, R. L., 2002. Statistical inference, 2nd Edition. Duxbury, USA.

Chan, I. S. F., 1998. Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies. Statistics in Medicine 17 (12), 1403–1413.

Chan, I. S. F., Zhang, Z., 1999. Test-based exact confidence intervals for the difference of two binomial proportions. Biometrics 55 (4), 1202–1209.

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

CPMP, 1999. Committee for Proprietary Medicinal Products. Note for guidance on clinical evaluation of new vaccines.

CPMP, 2003. Committee for Proprietary Medicinal Products. Note for guidance on evaluation of medicinal products indicated for treatment of bacterial infections.

D'Agostino, R. B., Chase, W., Belanger, A., 1988. The appropriateness of some common procedures for testing the equality of two independent binomial populations. The American Statistician 42 (3), 198–202.

Dammann, H. G., Folsch, U. R., Hahn, E. G., von Kleist, D. H., Klor, H. U., Kirchner, T., Strobel, S., Kist, M., 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 (1), 41–51.

Dunnett, C. W., Tamhane, A. C., 1997. Multiple testing to establish superiority/equivalence of a new treatment compared with k standard treatments. Statistics in Medicine 16 (21), 2489–2506.

Farrington, C. P., Manning, G., 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 (12), 1447–1454.

FDA, 1992. Points to consider. Clinical development and labeling of anti-infective drug products.

FDA, 1998. Guidance for industry. Complicated urinary tract infections and pyelonephritis - developing antimicrobial drugs. Draft guidance.

Gart, J. J., 1971. The comparison of proportions: A review of significance tests, confidence intervals and adjustments for stratification. Review of the International Statistical Institute 39 (2), 148–169.

Greco, D., Salmaso, S., Mastrantonio, P., Giuliano, M., Tozzi, A. E., Anemona, A., degli Atti, M. L. C., Giammanco, A., Panei, P., Blackwelder, W. C., Klein, D. L., Wassilak, S. G., 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 (6), 341–348.

Gustafsson, L., Hallander, H. O., Olin, P., Reizenstein, E., Storsaeter, J., 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 (6), 349–355.

ILAE, 1998. Considerations on designing clinical trials to evaluate the place of new antiepileptic drugs in the treatment of newly diagnosed and chronic patients with epilepsy. Report of the ilae commission on antiepileptic drugs. Epilepsia 39 (7), 799–803.

InTIME-II, 2000. Intravenous npa for the treatment of infarcting myocardium early; InTIME-II, a double-blind comparison of single-bolus lanoteplase vs accelerated alteplase for the treatment of patients with acute myocardial infarction. European Heart Journal 21, 2005–2013.

Martín Andrés, A., Silva Mato, A., 1994. Choosing the optimal unconditioned test for comparing two independent proportions. Computational Statistics and Data Analysis 17, 555–574.

McDonald, L. L., Davis, B. M., Milliken, G. A., 1977. A nonrandomized unconditional test for comparing two proportions in 2x2 contingency tables. Technometrics 19 (2), 145–157.

Miettinen, O., Nurminen, M., 1985. Comparative analysis of two rates. Statistics in Medicine 4 (2), 213–226.

Moliterno, D. J., Topol, E. J., 2000. A direct comparison of tirofiban and abciximab during percutaneous coronary revascularization and stent placement: Rationale and design of the target study. American Heart Journal 140 (5), 722–726.

Moulton, L. H., O'Brien, K. L., Kohberger, R., Chang, I., Reid, R., Weatherholtz, R., Hackell, J. G., Siber, G. R., Santosham, M., 2001. Design of a group-randomized streptococcus pneumoniae vaccine trial. Controlled Clinical Trials 22 (4), 438–452.

Newcombe, R. G., 1998. Interval estimation for the difference between independent proportions: Comparison of eleven methods. Statistics in Medicine 17, 873–890.

Phillips, K. F., 2002. A new test of non-inferiority for anti-infective trials. Statistics in Medicine (in press).

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

Röhmel, J., 2004. Problems with existing procedures to calculate exact unconditional p-values for non-inferiority/superiority and confidence intervals for two binomials and how to resolve them. Submitted .

Röhmel, J., Mansmann, U., 1999a. Letter to the Editor: Exact tests of equivalence and efficacy with a non-zero lower bound for comparative studies by I. S. F. Chan, Statistics in Medicine, 17, 1403-1413 (1998). Statistics in

Medicine 18 (13), 1734–1737.

Röhmel, J., Mansmann, U., 1999b. Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority. Biometrical Journal 41 (2), 149–170.

Robertson, T., Wright, F. T., 1981. Likelihood ratio tests for and against a stochastic ordering between multinomial populations. The Annals of Statistics 9, 1248–1257.

Roebruck, P., Kühn, A., 1995. Comparison of tests and sample size formulae for proving therapeutic equivalence based on the difference of binomial probabilities. Statistics in Medicine 14, 1583–1594.

Skipka, G., Munk, A., Freitag, G., 2004. Unconditional exact tests for the difference of binomial probabilities - contrasted and compared. Computational Statistics and Data Analysis In press.

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

Upton, G. J. G., 1982. A comparison of alternative tests for the 2x2 comparative trial. Journal of the Royal Statistical Society Series A 145 (Part 1), 86–105.