Unconditional exact tests for the difference of binomial probabilities - contrasted and compared

Preprint - Final Version

G. Skipka [a,1], A. Munk [b], G. Freitag [a],

[a] *Department of Medical Informatics, Biometry and Epidemiology, Ruhr-University Bochum, Germany*

[b] *Institute for Mathematical Stochastics, Georg-August University Göttingen, Germany*

Abstract

Various exact tests for showing a difference between two treatments or the non-inferiority (therapeutic equivalence) based on the difference of two binomial proportions are compared. It is found that a frequently used test has to be applied with great caution due to its numerical instability. Furthermore, a test based on the score statistic can be recommended as a good compromise between a simple and powerful procedure. Finally, a likelihood ratio based exact test is introduced, which slightly outperforms all other tests from the literature with respect to power. The issue of sample size determination is briefly addressed. All methods are illustrated with help of an example where two antihelmintic agents are compared.

Key words: likelihood ratio test, shifted null hypotheses, unconditional exact p-values, therapeutic equivalence, non-inferiority trials, Fisher's exact test

1 INTRODUCTION

Testing the difference of two failure rates is a classical topic in statistics. Closely related to this is the assessment of non-inferiority, which has significantly gained in importance during the last years. This is due to the need of statistical methodology encountered with therapeutic equivalence trials, where

[1] Corresponding address: Institut für Mathematische Stochastik, Maschmühlenweg 8-10, D-37073 Göttingen, Germany, tel: +49-551-3913520, fax: +49-551-3913521, email: skipka@statlab.de, homepage: www.statlab.de/statistics_group

the aim is to show the non-inferiority of a new treatment as compared to standard one, instead of its superiority. In many trials, dichotomous endpoints are the primary quantities of interest, such as failure or success rates. For example, Rodary et al. (1989) used the rupture (of tumor) rate in patients with childhood nephroblastoma. In the ASSENT-2 trial (1999), the 30-days-mortality rate after acute myocardial infarction was the primary endpoint.

In an FDA draft-guidance (1998) it is recommended that the difference of failure rates $\delta = p_T - p_C$ of the treatment (T) and control (C) group should fall below the value of $\Delta_0 = 0.15$ in clinical trials for the development of antimicrobial drugs for urinary tract infections. A similar rule can be found in CPMP (2003). For further applications see e.g. CPMP (1999) or FDA (1992).

Hence, the testing problem

$$H_0 \colon \delta \geqslant \Delta_0 \quad \text{versus} \quad H_1 \colon \delta < \Delta_0 \tag{1}$$

results. If a suitable test allows for rejection of H_0, so-called *therapeutic equivalence* of T and C, or *non-inferiority* of the new treatment with respect to control is established.

Example 1 *In a randomized controlled clinical trial Chouela et al. (1999) assessed therapeutic equivalence of ivermectin (an antihelmintic agent) with respect to lindane (control) for the treatment of human scabies. The sample size was 43, of whom 19 patients received ivermectin. Chouela et al. (1999) defined the equivalence margin as 0.2 and argued that ivermectin is much simpler applicable than lindane. The statistical analysis was performed using Blackwelder's asymptotic test (Blackwelder, 1982) with* $\alpha = 0.05$*. The p-value was found to be 0.002, hence therapeutic equivalence of ivermectin and lindane was claimed. It is well known, however, that for this small sample size the actual level of Blackwelder's test could be twice as the nominal level* α*, depending on the unknown value of* δ*. Hence, this test should be used with great caution here. Therefore, this example is re-analyzed in Section 5 by means of various exact test procedures.*

In contrast to asymptotic tests (cf. Dunnett and Gent, 1977; Farrington and Manning, 1990; Roebruck and Kühn, 1995, for a survey), exact tests aim to control the type I error exactly, i.e. the actual level α^* of these tests should never exceed the pre-assigned nominal level α. For $\Delta_0 = 0$, the most prominent test is Fisher's test, which lacks, however, from practicability, because randomization is required to keep the nominal level exactly. If this randomization step is not performed, the power of this test becomes rather low as compared to other procedures (Boschloo, 1970; McDonald et al., 1977; Upton, 1982; D'Agostino et al., 1988). Furthermore, this test cannot be transferred to the testing problem (1) for $\Delta_0 > 0$, because it is based on the odds ratio $\rho = p_T(1-p_C)/(p_C(1-p_T))$ and hence only suitable for the hypotheses (1) if

$\Delta_0 = 0$ (in this case $\rho = 1$). Martín Andrés and Silva Mato (1994) have given a comprehensive survey about unconditional exact tests for (1) when $\Delta_0 = 0$.

In this paper we compare various exact tests for the problem (1). This includes the procedure suggested in the benchmark paper by Barnard (1947), the test by Chan (1998), and the π_{local}-test by Röhmel and Mansmann (1999b). Moreover, we present an exact version of the likelihood ratio test, which is in the spirit of Kang and Chen (2000). It is shown that this test slightly outperforms the afore mentioned competitors with respect to power. Furthermore, we observed serious numerical difficulties with Barnard's test which makes it hardly applicable in practice.

The paper is organized as follows. In the next section we describe more precisely the general methodology for all exact tests mentioned above. In particular, we will find that the computational effort for Barnard's test can be considerable. Furthermore, this test causes numerical difficulties due to the computation of extremely small probabilities for determining the region of rejection or its corresponding p-values. Because of the inductive construction of the rejection region this test is not capable to correct small numerical errors in subsequent calculations. This implies that the critical region of the Barnard test (or an associated p-value) cannot be computed fully automatically and has to be controlled visually by the statistician.

In Section 2.3 and 2.4 the π_{local}-test and a test by Chan (1998) are introduced. In Section 2.5 we present the exact version of the likelihood ratio test. The computational effort is much less than for Barnard's test and it is numerically more stable.

In Section 3 all tests are compared numerically with respect to power, size and computational time. To our knowledge such a comparison has not been published so far. It is found that the actual level of the exact LR-test gets closest to the nominal level in most cases. This test, Chan's test and Barnard's test are comparable with respect to power, with a slight tendency in favor of the LR-test. We mention, however, that this test may suffer from a theoretical gap. This test cannot be shown to fulfill a convexity condition introduced by Barnard (1947), although we found numerical evidence that this holds.

In Section 4 the issue of sample size determination (in order to control the type II error) is briefly addressed. A more comprehensive discussion of this issue is postponed to a separate publication. Finally, in Section 5 the performance of all methods is discussed and illustrated by means of the above example. SAS code for all tests can be obtained from the authors on request.

2 UNCONDITIONAL EXACT TESTS

2.1 General methodology

In a clinical trial we will denote the outcome of a patient by X or Y which is 1 or 0 according to a failure or not. Furthermore, we denote by p_C and p_T the failure probability of the control and treatment group, respectively. Let n_C and n_T denote the sample size in each treatment group. Under these assumptions we observe two independent i.i.d. Bernoulli samples
$X_1, \ldots, X_{n_C} \sim B(1, p_C)$ and $Y_1, \ldots, Y_{n_T} \sim B(1, p_T)$, which have joint likelihood

$$P_{(x,y)}(p_C, p_T) = \binom{n_C}{x} {p_C}^x \left(1 - p_C\right)^{n_C - x} \binom{n_T}{y} {p_T}^y \left(1 - p_T\right)^{n_T - y}, \tag{2}$$

where $x = \sum_{i=1}^{n_C} x_i,\ y = \sum_{j=1}^{n_T} y_j,\ (x, y) \in \{0, \ldots, n_C\} \times \{0, \ldots, n_T\}$.

This probability depends on the true failure rates p_C and p_T. If the outcome (x, y) is represented on a grid of $(n_C + 1) \times (n_T + 1)$ points, the critical region (CR) of any test for testing (1) is defined by a subset of this grid. Therefore, the probability of type I error is given for $(p_C, p_T) \in H_0$ by

$$P\left((X, Y) \in CR \mid (p_C, p_T)\right) = \sum_{(x,y) \in CR} P_{(x,y)}(p_C, p_T)\,. \tag{3}$$

The principle of unconditional exact tests is to maximize the function (3) over H_0 in order to eliminate the nuisance parameters p_C and p_T. Thus, for a given critical region (CR), the actual level for the probability of type I error is determined by

$$\alpha^* = \max_{(p_C, p_T) \in H_0} P\left((X, Y) \in CR \mid (p_C, p_T)\right). \tag{4}$$

In general it is a difficult task to maximize over the entire null hypothesis, a triangle in the (p_C, p_T)-plane. Röhmel and Mansmann (1999b) have shown that if a certain condition holds for the critical region, the maximum is attained always at the boundary $p_T = p_C + \Delta_0$.

This condition, called "C" for convexity, has been introduced by Barnard (1947). We say that a critical region CR fulfills condition "C" if for any $(x, y) \in CR$ it holds that $(x + 1, y) \in CR$ and $(x, y - 1) \in CR$. The reasoning for "C" is that, if we reject (and hence conclude non-inferiority of T) for x failures in the control group, we should certainly reject for more than x failures, and if we reject for y failures in the test group we are supposed to do the same for less. We mention that the term convexity is somehow misleading, because

"C" does not always force the critical region to be convex in the sense that $x, y \in C \Rightarrow \lambda x + (1-\lambda)\, y \in C$, $\lambda \in (0,1)$. Perhaps *quadrant monotonicity* would be a more appropriate term.

If "C" is fulfilled, the critical region can be identified with its boundary, which has the advantage of simpler storage (always less than $\min\{n_C, n_T\}$ values) as well as a reduction of computational complexity, as it is carefully discussed in Röhmel and Mansmann (1999b, Sect. 7). We mention that condition "C" reduces the number of possible critical regions and hence the number of computations for finding the maximum in (4) from $O(2^{n_C+n_T})$ to $O\Big(\binom{n_C+n_T}{n_C}\Big)$ steps. This can be seen by observing that any critical region satisfying condition "C" can be identified with an isotonic path from $(0,0)$ to (n_C+1, n_T+1). Now an argument as in the ballot theorem (cf. Feller, 1968, p. 68) can be applied.

We mention that we found that all four tests under consideration satisfy condition "C". However, for the exact LR-test this is based on extensive numerical investigations, and we have no rigorous proof for this.

Remark 2 *If the rejection region of a test is of the form* $\big\{(x,y) : T(x,y) > c\big\}$ *for a real valued function* T*, condition "C" reads as a monotonicity condition for* T*, viz.*

$$T(x+1, y) \geq T(x, y) \quad \text{and} \quad T(x, y-1) \geq T(x, y)\,.$$

Hence, under condition "C" T *is isotonic in its first argument and antitonic in the second one.*

2.2 Barnard's test

Barnard (1947) has constructed an unconditional exact test for $H_0\colon \delta = 0$ versus $H_1\colon \delta \neq 0$. Röhmel and Mansmann (1999b) have recognized that the principle of constructing Barnard's test is directly transferable to the testing problem (1). The idea of calculating the critical region is to start with the outcome $(n_C, 0)$, that is the most extreme outcome with respect to condition "C". Then the critical region is extended iteratively. Potential next outcomes are the adjacent points $(n_C - 1, 0)$ and $(n_C, 1)$, which fulfill condition "C". The actual levels α^* (see (4)) for $CR = \{(n_C, 0), (n_C - 1, 0)\}$ and $CR = \{(n_C, 0), (n_C, 1)\}$ are compared. From these adjacent points the outcome is included into the critical region, which increases the actual level by the smallest amount. Now, the next adjacent points and their amount to the actual level are calculated to determine the next point to be included in the critical region. This procedure is continued as long as α^* remains smaller than the nominal level. Loosely speaking, the critical region is based on the principle to include as much as possible points under the constraint "C" and $\alpha^* \leq \alpha$, whith α^* in (4).

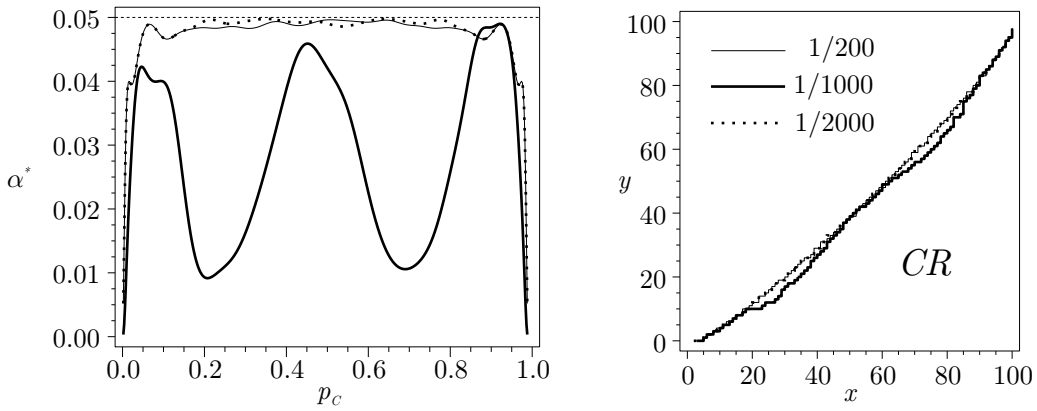

Fig. 1. Exact level as a function of p_C (left hand side) for calculated rejection regions (right hand side) using a grid width of 1/200 (thin line), 1/1000 (thick line) and 1/2000 (dotted line)

We found, however, that a serious difficulty encountered with the practical use of Barnard's consists in the effective numerical computation of its rejection region. To this end the maximum α^* in (4) has to be determined for any possible extension of the critical region. Numerically, this can only be achieved by calculating α^* on a discrete grid of the interval $[0, 1-\Delta_0]$, the domain of p_C, say. The following algorithm was implemented for computing critical regions:

(1) The initial critical region consists of the most extreme possible outcome only: $CR_1 = \{(n_C, 0)\}$.
(2) The adjacent outcomes not violating condition C are $(n_C - 1, 0)$ and $(n_C, 1)$. α^* is computed for $CR = CR_1 \cup \{(n_C - 1, 0)\}$ and $CR = CR_1 \cup \{(n_C, 1)\}$, respectively. The maxima are determined by calculating the values $P((X,Y) \in CR \mid p_T = p_C + \Delta_0)$ iteratively for $p_C \in \{\epsilon, 2\epsilon, \ldots, 1 - \Delta_0 - 2\epsilon, 1 - \Delta_0 - \epsilon\}$ with, e.g., $\epsilon = 1/1000$. The critical region is extended by the outcome with the smaller resulting α^*. If α^* is (numerically) equal for both outcomes (e.g. if $n_T = n_C$) then $CR_2 = CR_1 \cup \{(n_C - 1, 0), (n_C, 1)\}$.
(3) Step 2 is iterated according to condition "C" until α^* exceeds the nominal level.
(4) Stop the iteration and choose the preceding critical region.

Note that in step 2 the selection of a possible point to be included in the critical region depends heavily on the grid width ϵ chosen to determine α^*. We found that this yields an intrinsically numerical difficulty, because the corresponding α^*'s to be compared are below the numerical precision of any standard software. Due to the iterative structure of the algorithm a wrong selection of a point in iteration step "i" will affect the entire subsequent construction and may lead to completely wrong rejection regions. This is in contrast to the subsequent algorithms in 2.3 - 2.5.

Figure 1 shows the exact levels (l.h.s.) as a function of the nuisance parameter

p_C (with $n_C = n_T = 100$ and $\Delta_0 = 0.01$) for rejection regions, which are calculated using three different grid widths (1/200, 1/1000 and 1/2000). The right hand figure displays the corresponding rejection regions. Observe that, for a width of 1/1000, the rejection region and its level curve differ tremendously from the regions with 1/200 and 1/2000, respectively.

We will illustrate this with the following numerical example. For the construction of the rejection region it is essential that the probabilities corresponding to the potential points are ordered correctly. These probabilities may be extremely small, in particular for larger sample sizes. E.g., for $n_T = n_C = 100$, $\Delta_0 = 0.01$, and the initial critical region $CR = \{(100, 0)\}$, the maximum is $\alpha^* \approx 8.34 * 10^{-62}$ (for $p_C = 0.495$). In contrast, for unconditional exact tests which use a test statistic T as an ordering criterion this is a minor problem (just rounding errors may cause difficulties). As a consequence, if T_i and T_j are wrongly ordered, this will not affect the following T_k, $k > i, j$. However, by constructing the critical region of Barnard's test the preceding sequence of the points in the rejection region essentially determines all subsequent points. To illustrate this, assume that the correct "Barnard ordering" is $(x_1, y_1), \ldots, (x_i, y_i), (x_{i+1}, y_{i+1})$, and

$$\max_{(p_C,p_T)\in H_0} P\left((X,Y) \in \{(x_1,y_1),\ldots,(x_i,y_i),(x_{i+1},y_{i+1})\} \mid (p_C,p_T)\right) \leq \alpha.$$

Assume further, that instead of (x_i, y_i) the outcome (x_i', y_i') is wrongly inserted into the critical region. Then it may happen that the "correct" outcomes (x_i, y_i) and/or (x_{i+1}, y_{i+1}) will not be included into the critical region, because

$$\max_{(p_C,p_T)\in H_0} P\left((X,Y) \in \{(x_1,y_1),...,(x_i',y_i'),(x_i,y_i)\} \mid (p_C,p_T)\right) > \alpha,$$

$$\max_{(p_C,p_T)\in H_0} P\left((X,Y) \in \{(x_1,y_1),...,(x_i',y_i'),(x_{i+1},y_{i+1})\} \mid (p_C,p_T)\right) > \alpha, \text{ or}$$
$$\max_{(p_C,p_T)\in H_0} P\left((X,Y) \in \{(x_1,y_1),...,(x_i',y_i'),(x_i,y_i),(x_{i+1},y_{i+1})\} \mid (p_C,p_T)\right) > \alpha.$$

Remark 3 *We mention that there is some divergent terminology in the literature. In various papers and software packages "Barnard's test" does not refer to the test introduced by Barnard (1947). E.g., the software product StatXact refers to "Barnard's test", but actually uses the unconditional exact test from Chan (see below), in case of the testing problem (1). Testimate advertises the "Barnard type exact test for non-inferiority using the Röhmel-Mansmann procedure", but this is in fact equal to the π_{local}-test (see below).*

2.3 The π_{local}-test

Röhmel and Mansmann (1999b) have suggested an additional unconditional exact test for the problem (1). Here probabilities $\pi_{min}(x, y)$ are calculated

for all possible outcomes (x, y) $(0 \leq x \leq n_C, 0 \leq y \leq n_T)$. These are the H_0-probabilities for outcomes (i, j) with $i \geq x$ and $j \leq y$:

$$\pi_{min}(x, y) = \max_{H_0} \sum_{i \geq x} \binom{n_C}{i} p_C{}^i (1 - p_C)^{n_C - i} \sum_{j \leq y} \binom{n_T}{j} p_T{}^j (1 - p_T)^{n_T - j}.$$

The set of all possible outcomes (x, y) is sorted in ascending order by $\pi_{min}(x, y)$, which is denoted as $S = \left(S_1, \ldots, S_{(n_C+1)\cdot(n_T+1)}\right)$. Now define

$$\alpha_k^* = \alpha^* \left(\bigcup_{l=1}^{k} S_l \right), \tag{5}$$

which denotes the maximal actual level of the rejection region $\bigcup_{l=1}^{k} S_l$ of the "k" smallest values in S with respect to the ordering induced by π_{min}. Finally, the critical region CR is defined by

$$\arg\max_k \{\alpha_k^* \leq \alpha\}. \tag{6}$$

2.4 Chan's test

In an approach recommended by Chan (1998), the test statistic of Farrington and Manning (1990) is used to construct an unconditional exact test. Farrington and Manning introduced an asymptotic method based on the normal approximation of the observed difference of the failure rates, where the variance is estimated with help of an ML-estimator restricted to $p_T = p_C + \Delta_0$. An explicit solution of the maximum likelihood equation can be given as

$$\begin{aligned} \widetilde{p}_T &= 2\frac{\sqrt{r^2 - 3s}}{3} \cos\left[\frac{1}{3} \arccos\left(-\frac{\frac{2r^3}{27} - \frac{rs}{3} + t}{2\left(\frac{\sqrt{r^2-3s}}{3}\right)^3} \right) + \frac{4}{3}\pi \right] - \frac{r}{3}, \\ \widetilde{p}_C &= \widetilde{p}_T - \Delta_0, \end{aligned} \tag{7}$$

where

$$\begin{aligned} r &= -\frac{n_T(1 + \hat{p}_T + 2\Delta_0) + n_C(1 + \hat{p}_C + \Delta_0)}{n_T + n_C}, \\ s &= \frac{n_T(\hat{p}_T + 2\Delta_0\hat{p}_T + \Delta_0 + \Delta_0^2) + n_C(\hat{p}_C + \Delta_0)}{n_T + n_C}, \\ t &= \frac{-n_T\hat{p}_T\Delta_0(1 + \Delta_0)}{n_T + n_C}. \end{aligned}$$

The critical region of Chan's test is constructed in the same way as in the π_{local} approach. However, Farrington and Manning's test statistic,

$$\frac{\frac{y}{n_T} - \frac{x}{n_C} - \Delta_0}{\sqrt{\frac{\widetilde{p}_T(1-\widetilde{p}_T)}{n_T}\frac{\widetilde{p}_C(1-\widetilde{p}_C)}{n_C}}},$$

is used as the ordering criterion, instead of $\pi_{min}(x, y)$.

Röhmel and Mansmann (1999a) (letter to the editor) have remarked that searching for the maximum at the boundary of H_0 might be not correct here, because Chan did not prove that his ordering criterion fulfills the condition "C"- in contrast to Barnard's test and the π_{local}-test. However, Chan (author's reply) has given a heuristic argument that the condition is satisfied, which was finally proved by Martín Andrés and Herranz Tejedor (2003).

2.5 The exact likelihood ratio test

The likelihood ratio statistic for the testing problem (1) is given by

$$\lambda = \lambda(x, y, \Delta_0) = \frac{\sup\limits_{H_0} P_{x,y}(p_C, p_T)}{\sup\limits_{H_0 \cup H_1} P_{x,y}(p_C, p_T)} \tag{8}$$

$$= \begin{cases} 1 & \text{if } \hat{p}_T \geq \hat{p}_C + \Delta_0 \\ \dfrac{\widetilde{p}_C{}^x \left(1-\widetilde{p}_C\right)^{n_C - x} \widetilde{p}_T{}^y \left(1-\widetilde{p}_T\right)^{n_T - y}}{\hat{p}_C{}^x \left(1-\hat{p}_C\right)^{n_C - x} \hat{p}_T{}^y \left(1-\hat{p}_T\right)^{n_T - y}} & \text{if } \hat{p}_T < \hat{p}_C + \Delta_0 \end{cases}, \tag{9}$$

with $\widetilde{p}_C$ and $\widetilde{p}_T$ as in (7).

It can be shown that the asymptotic distribution of $-2\ln(\lambda)$ is a $\frac{1}{2} + \frac{1}{2}\chi^2_1$ -law (work in preparation). However, the finite sample distribution of $-2\ln(\lambda)$ for sample sizes smaller than 100, say, differs significantly from the asymptotic law and depends on the particular values of p_C and p_T, where $p_T - p_C = \Delta_0$. Therefore, we extend an idea of Storer and Kim (1990) and suggest to construct a test by estimating the values of p_C and p_T restricted to $\delta = \Delta_0$ and using these values as estimators for the parameters p_C and p_T of the exact distribution of λ. To this end for every possible pair of x and y the LR-statistic λ and the probability of observing this outcome are computed. The probabilities are calculated by inserting restricted ML-estimates $(\widetilde{p}_C, \widetilde{p}_T)$ for the failure rates into $P_{(x,y)}(p_C, p_T)$ (see equation (7)). With that, estimated p-values p^*

are calculated for every outcome (x, y),

$$p^*(x,y) = \sum_{(a,b):\lambda(a,b,\Delta_0)\leq\lambda(x,y,\Delta_0)} P_{(a,b)}\left(\widetilde{p}_C, \widetilde{p}_T\right). \tag{10}$$

In a second step these p-values are used as an ordering criterion. The critical region for an unconditional exact version of the LR-statistic is constructed in the same way as described for the π_{local}-test. The estimated p-values $p^*(x,y)$ are used to sort all possible outcomes in ascending order. Now consider (5), where, again, S_j denotes the outcome with the "j" smallest estimated p-value based on (10). Finally, the points to be included into the critical region CR are determined as in (6).

Hence, our method differs from the above mentioned tests by basing the decision on which points to be included in CR on a cumulative likelihood ratio criterion.

Remark 4 *To our knowledge it is not possible to prove Barnard's condition "C" (mentioned in Section 2.1) for the exact LR-test. Therefore, the actual level α^* has to be determined by maximizing over the entire null space, in principle. Nevertheless we found numerically that it is feasible to restrict the calculation of the maximum to the boundary of the null space for all parameter settings in Section 3. This was feasible, since we checked condition "C" after sorting the outcomes for every parameter setting without finding any violation.*

We mention that the LR-statistics without estimating the unknown parameters p_C and p_T in $P(p_C, p_T)$ leads to a rather conservative test and cannot be recommended. Hence, the cumulative LR function in (10) has been used.

The idea to estimate the unknown nuisance parameters themselves in order to improve the accuracy dates back to Storer and Kim (1990) for the classic null hypothesis $H_0 : p_C = p_T$. They calculate approximate p-values by using a standardized Z statistic with the pooled variance estimate for the nuisance parameter. This idea was carried on to shifted hypotheses like (1) by Kang and Chen (2000), who suggested an approximate unconditional test, which does not keep the nominal level exactly. Note that the exact likelihood ratio test keeps its level exactly.

Remark 5 *The exact LR-test as well as Chan's test and the π_{local}-test are numerically stable, since no grid search is needed for* sorting *the outcomes. This is in contrast to Barnard's test (cf. Sect. 2.2). Note that for the afore mentioned tests also a numerical maximization step is required in order to compute α_i^* in (5). This is, however, numerically feasible.*

2.6 Computational time

Finally, we briefly comment on the computational time in order to compute the critical region and a p-value, respectively. As an example, on a PC using a Pentium IV processor we have found that Chan's rejection region requires about 10 seconds to be computed when $n_1 = n_2 = 70$, $\Delta_0 = 0.15$ and $\alpha = 0.05$. The π_{local}-test requires about 30 seconds for the same setting, whereas the exact LR-test takes about 50 seconds, which is due to the additional effort to compute the estimated p-values accordingly. For all these tests a grid width of 1/1000 was chosen to determine the maximum level in (5). Computation of Barnard's test is much more time consuming. It requires about 8 minutes for a grid width of 1/1000. Increasing the grid width increases the computing time linearly, of course. Hence, for a grid width of 1/2000 about 16 minutes are required. To reduce the computational effort, Martín Andrés and Silva Mato (1994) have introduced a modification of Barnard's test for $\Delta_0 = 0$, where the restricted ML-estimators are used to approximate α^* in (4). Note that this modification will not solve the numerical problems of Barnard's test, as described in Sect. 2.2.

3 POWER INVESTIGATIONS - NUMERICAL STUDY

All tests under investigation aim at keeping the nominal level exactly. However, we will see that they differ with respect to the actual level and power. To investigate the exact levels (see (3)), we have calculated these as a function of the nuisance parameter p_C on the boundary of H_0, i.e. when $p_T - p_C = \Delta_0$. Generally, for the exact LR-test, Chan's test and the π_{local}-test we used a grid width of 1/1000. In Figure 2 the actual level of all four tests is displayed for two parameter constellations, $n_C = n_T = 100$, $\Delta_0 = 0.1$, and $n_C = n_T = 50$, $\Delta_0 = 0.15$. The Barnard test was computed using a grid width of 1/1000.

We find that the actual levels of Barnard's test and the exact LR-test are above the levels of Chan's and the π_{local}-test, uniformly over the entire null hypothesis. In particular, when p_C is close to zero or to $1 - \Delta_0$, this becomes a quite drastic difference. The differences between Barnard's and the exact LR-test are surprisingly small. Nevertheless, both tests outperform each other in different regions of the parameter space. These figures are somewhat typical. We have computed a broad scenario of various other settings, including unequal sample sizes and other values of Δ_0:

- *Boundary of hypothesis:* We chose $\Delta_0 \in \{0.05, 0.1, 0.15, 0.2, 0.25\}$.

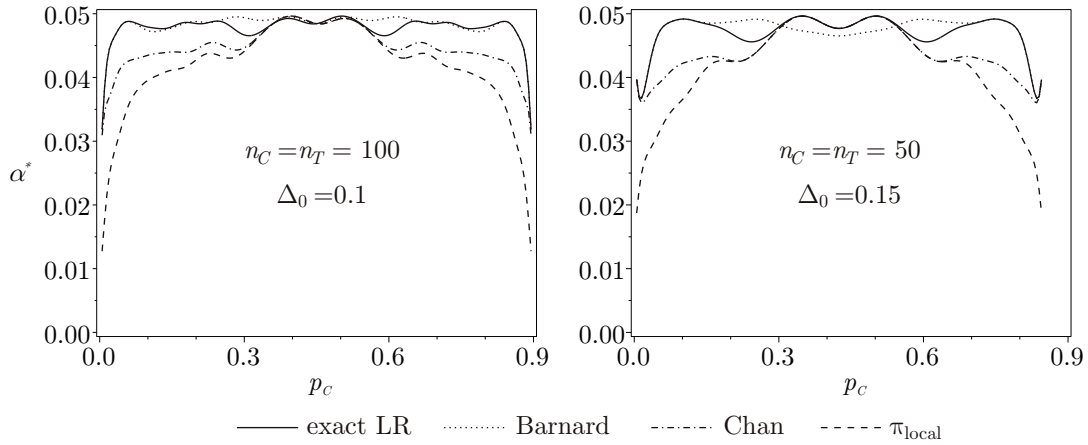

Fig. 2. Exact level of all four tests as a function of p_C for two different parameter constellations

- *Sample size:* We chose balanced sample sizes $n_T = n_C \in \{20, 25, 30, 35, 40, 50, 60, 80, 100\}$ and unbalanced settings $(n_T, n_C) \in \{(30, 20), (40, 20), (50, 25), (60, 30), (60, 40), (80, 40), (80, 50), (80, 60), (100, 50), (100, 60), (100, 80)\}$.
- *Nuisance parameter:* We chose $p_C \in \{0.1, 0.2, 0.3, 0.5, 0.8, 0.9\}$.

This gives 600 different parameter configurations. Configurations were omitted in case of non-feasible settings (i.e. $p_C \geq 1 - \Delta_0$). We have chosen the parameter $p_T \leq p_C$ for every configuration 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 Δ_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 Δ_0 some parameter constellations result in a power larger than 0.9 for all tests. These cases were omitted, too. Finally, 407 parameter constellations were extracted. The resulting values of the power function were calculated exactly for all tests under investigation by computing the exact binomial probabilities (2) for all $(x, y) \in CR$.

Observe finally, that for Barnard's test, for the computation of the critical region we have determined that grid, where the test with the best performance results (grid width 1/1000). For other settings we found quite different grid widthes. This is a very cumbersome proceeding and cannot be done automatically.

It is found that in general the power differences between the exact LR-test and its competitors are small. Nevertheless, the power of the exact LR-test tends to be larger. In order to illustrate this, the differences of the exact LR-test's power and the power of its competitors are displayed in Figure 3. From this figure we can draw several conclusions. First, the exact LR-test outperforms the π_{local}-test, and this difference can be quite substantial. For a fixed type II error we found numerically that this leads to differences in the required sample size up to 25%. Barnard's test and the LR-test outperform

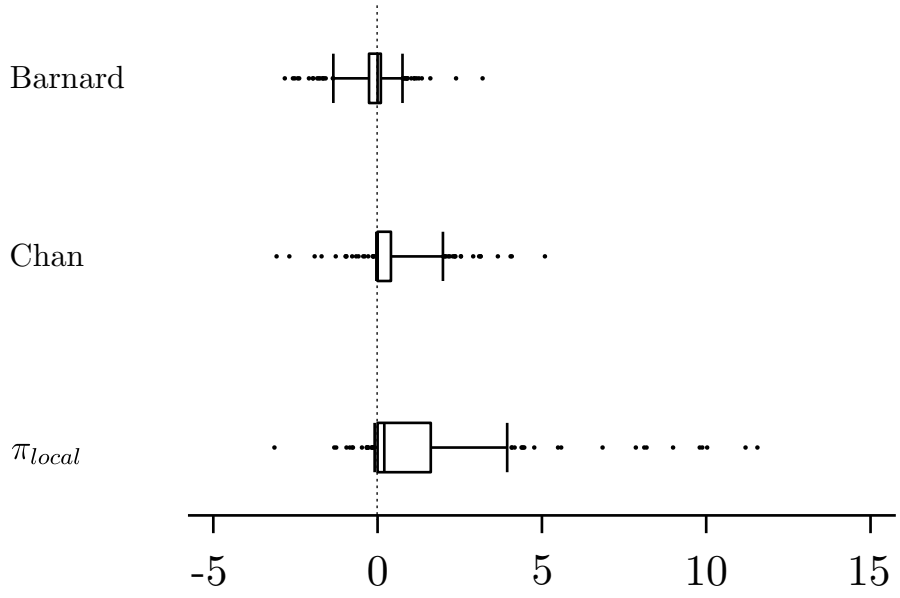

Fig. 3. Boxplot (whiskers are the 5% and 95% quantiles) for the power differences (times 100) between the exact LR test and its competitors.

each other for different parameter constellations, whereas the LR-test slightly outperforms Chan's test in some cases, but in most cases no differences were found.

The most extreme power differences and the corresponding parameter constellations are displayed in Table 1. More precisely, those values are displayed where the power of the exact LR-test differs by more than 2% from Barnard's test, more than 3% from Chan's test, or by more than 5% from the π_{local}-test.

4 SAMPLE SIZE DETERMINATION

In this section we briefly discuss various issues encountered with the sample size determination when planning a non-inferiority trial in order to control the type II error. One might think at a first glance that this will be in general achieved for equal sample sizes, $n_T = n_C$. This is, however, not true, and we will provide in a forthcoming paper tables for optimal allocations of the sample size, when the total number of observations $n = n_C + n_T$ is kept fixed.

In order to illustrate this effect, in Figure 4 for each of the tests the allocations of sample sizes where $40 \leq n_T, n_C \leq 80$ are displayed, specifying $\Delta_0 = 0.15$, $\alpha = 0.05$ and $p_C = 0.1$. The black dots indicate an allocation of sample sizes for which the test results in a power larger than 80%. For the white dots the power is less than 80%. From this figure the following conclusions can be

Table 1
Exact power (times 100) for different values of the parameters, where $\alpha = 0.05$.

n_T	n_C	Δ_0	p_C	p_T	$Barnard$	$Chan$	π_{local}	exact LR
20	20	0.05	0.01	0.2	86.2	86.2	75	86.2
20	20	0.15	0.02	0.1	85.1	85.1	77	85.1
25	25	0.05	0.02	0.2	85.4	85.4	75.4	85.4
25	25	0.1	0.02	0.1	80.3	80.3	72.1	80.3
25	25	0.2	0.07	0.1	81.8	81.7	74	81.8
30	20	0.1	0.01	0.1	88.7	88.7	79.7	88.7
30	20	0.15	0.08	0.2	84.9	82.3	82.1	82.3
30	20	0.15	0.14	0.3	84	81.6	81.6	81.6
30	20	0.2	0.08	0.1	82.1	82.1	72.9	84.5
30	20	0.2	0.19	0.3	80.3	82.4	82.5	79.4
30	20	0.2	0.08	0.1	82.1	82.1	72.9	84.5
35	35	0.05	0.01	0.1	80.2	80.2	74.6	81.4
35	35	0.15	0.07	0.1	77.9	77	71.3	81.1
40	40	0.05	0.01	0.1	89	89	79.2	89
50	25	0.15	0.08	0.1	80.4	80.4	74.9	80.4
50	50	0.1	0.06	0.1	80.2	77	75.8	80.2
50	50	0.15	0.09	0.1	82.5	80	76.9	82.5
60	30	0.05	0.01	0.1	91.2	86.5	86.1	88.4
60	30	0.05	0.06	0.2	85.2	84.3	82.6	82.6
60	30	0.05	0.59	0.8	80	76.8	80.2	80.4
60	30	0.05	0.73	0.9	80.3	75.2	77.5	80.3
60	60	0.1	0.06	0.1	85.8	82.7	81.8	85.8
60	60	0.2	0.46	0.5	84	81.9	81.9	81.9
80	60	0.05	0.04	0.1	82.1	81.6	78.9	79.7
100	60	0.05	0.66	0.8	83.2	80	82.7	83.2
100	60	0.05	0.8	0.9	80.7	77.3	77.6	81.3

drawn.

(1) For all tests, the choice of equal sample sizes (displayed on the diagonal $n_C = n_T$) is not optimal in the sense that the total sample size can be

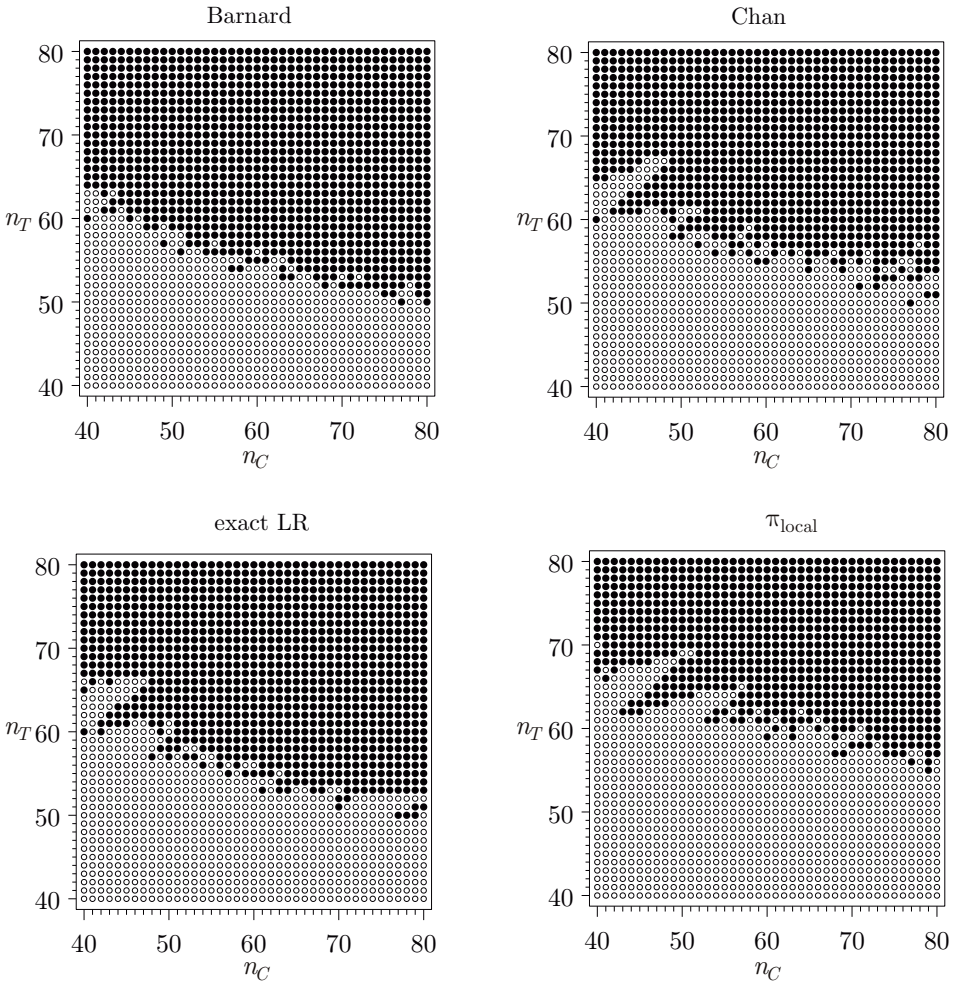

Fig. 4. Visualization of sample sizes (n_T, n_C), for which the power is larger (black dots) or less (white dots) than 0.8, specifying $\Delta_0 = 0.15$, $\alpha = 0.05$ and $p_C = 0.1$

reduced for a different allocation.

In fact, Figure 4 shows that the total sample size n can be reduced by overweighting the treatment group (i.e. $n_T/n_C > 1$). This was found for various other values of p_C. Except for the π_{local}-test 56 patients per group are needed for balanced allocation in order to achieve a power of 80%. For the π_{local}-test 62 patients are needed, respectively. Minimizing the total number of observations where the power of 80% is kept fixed gives in this particular case for the Barnard, the exact LR and Chan's test the same result, $(n_T, n_C) = (60, 40)$. Hence, the total sample size can be reduced by 12, i.e. by about 10% of the total sample size using a balanced design. For the π_{local}-test we even obtain a reduction by an amount of 18 observations.

(2) As described in Skipka and Trampisch (2001) and Finner and Strassburger (2001), exact tests do not have a monotone increasing power function, in general. In particular, Figure 4 shows that for all tests there are pairs of (n_T, n_C), for which the power is larger than for $(n_T + 1, n_C)$ or

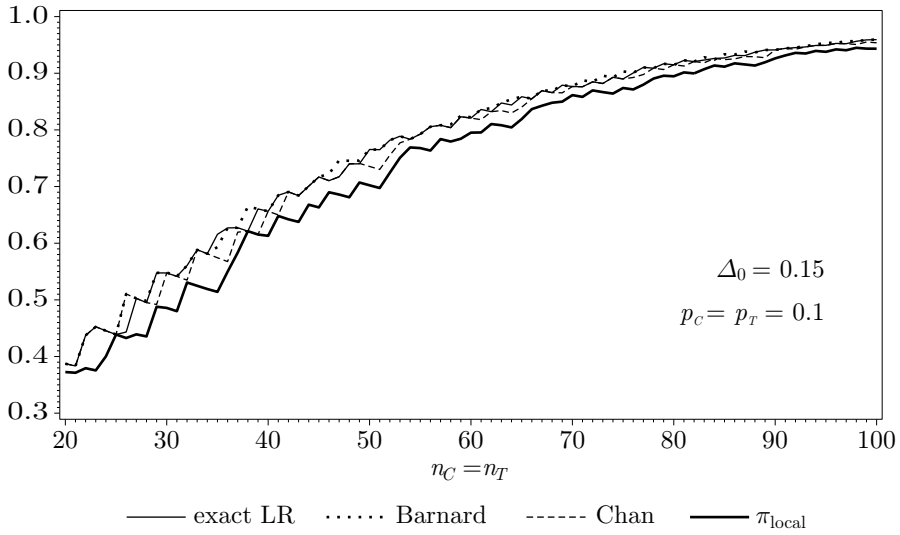

Fig. 5. Exact power of all four tests as a function of the sample size (balanced)

(n_T, n_C+1). This is found for all tests under investigation. Figure 5 shows the exact power as a function of $n_T = n_C$.

Due to the lack of monotonicity of the power function of these tests it is computationally extremely intensive to determine the optimal allocation of sample size. A way out of this problem might consist in asymptotic considerations. However, we will not pursue this topic here and leave it as a challenging task for further research.

5 EXAMPLE

In this section we re-analyze the randomized controlled trial from Chouela et al. (1999) for the treatment of human scabies (cf. Example 1). We draw from Chouela et al. (1999) that 29 days after the treatments were administered, 18 of 19 patients treated with ivermectin (5.3% failure rate) and 23 of 24 patients who received lindane (4.2% failure rate) were healed from their scabies. Chouela et al. (1999) used Blackwelder's test (Blackwelder, 1982) with $\Delta_0 = 0.2$, which results in a p-value of 0.002. The nominal level was $\alpha = 0.05$.

In Figure 6 the exact levels of the tests of Blackwelder, Barnard, and Chan, and of the exact likelihood ratio test are displayed for the situation in Chouela et al. (1999). Figure 6 shows that the actual level of Blackwelder's test is heavily exceeded for small and large values of p_C (up to twice of the nominal level), hence this test is not appropriate here. For example, if we equate the observed failure rate 0.042 with the exact one, the actual level of Blackwelder's test is to be expected as 0.09. Barnard's test has a maximum actual level of 0.05, the exact LR-test has a maximum actual level of 0.049. Finally, the p-value for the

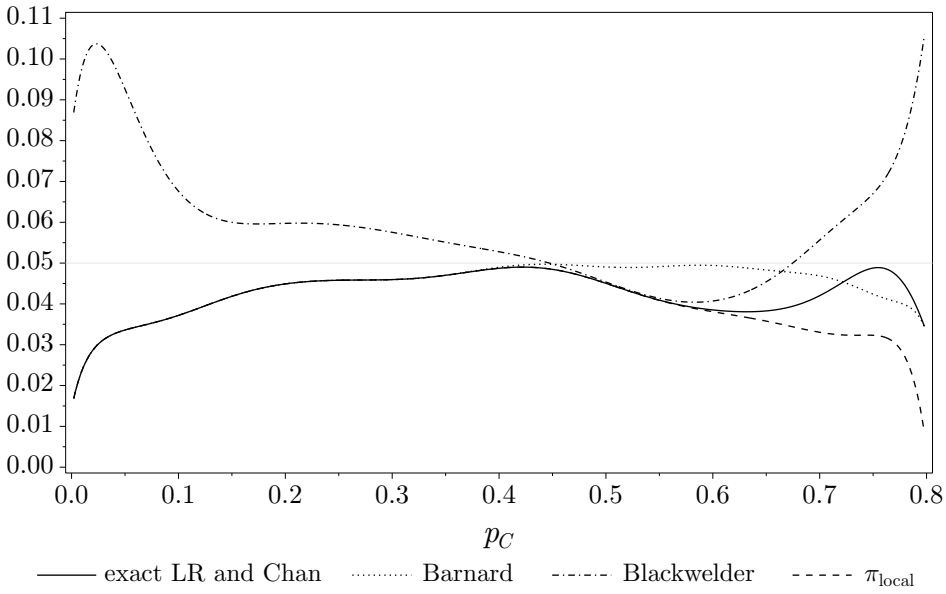

Fig. 6. Exact level as a function of p_C for the parameter constellation of the example (sec. 5)

data in Chouela et al. (1999) of Barnard's test is 0.0092, which is slightly larger than the p-value of the exact LR-test (p-value = 0.0087). The p-values for the π_{local}-test (0.0152) and Chan's test (0.0172) are somewhat larger. In summary all, of these tests show a significant therapeutic equivalence of ivermectin and lindane.

Remark 6 *Therapeutic equivalence would had been demonstrated too for* $\Delta_0 = 0.15$*, which gives the following p-values: Barnard 0.0305, exact LR 0.0309, Chan 0.04, and* π_{local} *0.0434. Even if the equivalence margin is chosen smaller,* $\Delta_0 = 0.13$*, say, the exact LR-test and Barnard's test give a significant (*$\alpha = 0.05$*) result (p-value =* 0.0493*, respectively). However, Chan's test (p-value =* 0.0544*) and the* π_{local}*-test (p-value =* 0.0677*) do not claim equivalence for* $\Delta_0 = 0.13$*.*

6 CONCLUSION

In this paper we have compared four exact tests, Barnard's test, Chan's test, π_{local}-test, and an exact version of the likelihood ratio test, which is similar in spirit to the approximate unconditional test of Kang and Chen (2000) and relies on an idea by Storer and Kim (1990). We found that the exact LR-test and Barnard's test are comparable with respect to power, and that the exact LR-test slightly outperforms Chan's test. Barnard's test suffers from the fact that its computation depends sensitively on the chosen grid width of the probabilities p_C in order to compute the actual level. Hence, this test has

to be used with great caution and cannot be applied fully automatically in practice. In contrast, the exact LR-test is computationally much more feasible and stable.

Overall, Chan's test seems to be a good compromise between a simple and efficient procedure, whereas the test by Röhmel and Mansmann is inferior to all competitors with respect to power. An advantage of this test as well as of the LR-test, however, consists in its generality, because it can easily be transferred to other measures of non-inferiority, such as the odds ratio.

The tests proposed by Chan and Röhmel & Mansmann are simpler and faster to compute than the LR-test. The ratio of computation time is approximately 1:3:5 (Chan : π_{local} : exact LR).

In contrast to asymptotic tests, the exact computation of required sample size is very time consuming due to the non-monotonicity of the power of all tests. This will be investigated in detail elsewhere.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft DFG grant TR 471/1. We are indebted to A. Martín Andrés, S. Lange, J. Röhmel and H.-J. Trampisch for helpful comments and discussions. Various comments of two referees have led to a much clearer presentation of this paper.

References

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

Blackwelder, W. C., 1982. "Proving the null hypothesis" in clinical trials. Controlled Clinical Trials 3 (4), 345–353.

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.

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.

Chouela, E. N., Abeldano, A. M., Pellerano, G., Forgia, M. L., Papale, R. M., Garsd, A., Balian, M. C., Battista, V., Poggio, N., 1999. Equivalent therapeutic efficacy and safety of ivermectin and lindane in the treatment of human scabies. Archives of Dermatology 135 (6), 651–655.

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.

Dunnett, C. W., Gent, M., 1977. Significance testing to establish equivalence between treatments, with special reference to data in the form of 2x2 tables. Biometrics 33 (4), 593–602.

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.

Feller, W., 1968. An Introduction to Probability Theory and its Applications. Wiley & Sons.

Finner, H., Strassburger, K., 2001. $ump(u)$-tests for a binomial parameter: A paradox. Biometrical Journal 43, 667–675.

Kang, S. H., Chen, J. J., 2000. An approximate unconditional test of non-inferiority between two proportions. Statistics in Medicine 19 (16), 2089–2100.

Martín Andrés, A., Herranz Tejedor, I., 2003. Exact unconditional non-classical tests on the difference of two proportions. Computational Statistics and Data Analysis (in press).

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.

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.

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

Roebruck, P., Kühn, A., 1995. Comparison of tests and sample size formu-

lae for proving therapeutic equivalence based on the difference of binomial probabilities. Statistics in Medicine 14, 1583–1594.

Skipka, G., Trampisch, H., 2001. Unconditional exact tests for comparing two independent proportions. In: Kunert, J., Trenkler, G. (Eds.), Festschrift in Honour of Siegfried Schach: Mathematical Statistics with Applications in Biometry. Eul publishers, pp. 189–196.

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.