

More powerful exact noninferiority and equivalence tests based on binary matched pairs

Chris J. Lloyd and Max V. Moldovan

Abstract

Assessing the therapeutic noninferiority or equivalence of one medical treatment compared to another is often based on the difference of response rates from a matched binary pairs design. This paper develops new exact unconditional tests for noninferiority and equivalence that are more powerful than available alternatives. There are three new elements presented in this paper. First we introduce the LR statistic as an alternative to the previously proposed score statistic of Nam (1997). Second, we eliminate the nuisance parameter by estimation followed by maximization as an alternative to the partial maximization of Berger and Boos (1994) or traditional full maximization. Third, for testing equivalence it is standard to combine two one-sided tests (TOST). We point out that even if the one-sided tests are exact and efficient, the TOST will be conservative and requires a further adjustment to remove this conservatism. Based on an extensive numerical study, we recommend tests based on the score statistic, the nuisance parameter being controlled by estimation followed by maximization.

1 Introduction

Testing for noninferiority plays an essential part in randomized clinical trials (RCT). The demonstration of the ability of a trial to distinguish between an effective treatment and a less effective or ineffective treatment, more formally known as assay sensitivity, is the first critical point in assessing the efficacy of a new treatment. Assay sensitivity can be confirmed by demonstration of superiority of a treatment to the concurrent placebo, and a two-arm double-blind placebo-controlled RCT is often suggested as an optimal choice for medical treatment evaluation as long as it is possible and justifiable in practice [1], [2]. When it is also required to compare the new treatment with a known effective alternative, an ideal RCT would be a three-arm trial that includes the new treatment, the active control and the concurrent placebo control. In this situation, the first step is to establish the assay sensitivity of the trial by demonstrating either that the active control is superior to placebo or, as recently argued by Koch and Röhmel [3], that the new treatment is superior to placebo. Only if assay sensitivity is confirmed, is it meaningful to proceed to the noninferiority test with the aim of showing that the new treatment is not substantially worse than the active control, demonstrated by retaining a certain fraction of its effect. The desired fraction is normally selected as the smallest

clinically important dominance of the active control over placebo and defined by the noninferiority margin δ . Even though the new treatment may be less effective than the active control, it may still be preferred according to several alternative criteria, such as better safety profile, lower toxicity, easier administration regime or lower cost.

Although the two-arm RCT with a placebo control and the three-arm RCT with concurrent placebo and active controls are the two most preferable designs, in practice it is not always possible to include a placebo arm, often due to ethical considerations [4], [5]. Design and analysis of RCT's which compare the new treatment with an active control without a concurrent placebo control is more problematic since it involves cross-trial inference under fairly strong assumptions, such as a well-known constancy assumption [6]. Another source of complexity in active control RCT's is the selection of an appropriate noninferiority margin δ [7], [8]. If the new treatment is shown to be noninferior to the active control, the appropriate noninferiority margin must not only ensure that the new treatment would be superior to placebo had it been included in the trial, but also guarantee that the new treatment preserves a certain clinically meaningful part of the effect of active control. D'Agostino et al. [1] give an excellent synthesis of academic and practical perspectives on these and several other relevant questions involved in active control RCT's.

In order to account for population heterogeneity, it is common to conduct matched-pairs design trials where subjects are either matched according to a specific covariate with further randomization to test and control groups, e.g. in intervention studies, or each assigned to receive the sequence of test and control in one of two possible orders, e.g. in cross-over design trials [9], [10]. In this paper, we assume a matched-pairs design active control RCT with established assay sensitivity and sensibly chosen noninferiority margin δ .

Let θ be the difference between the response probabilities of the treatment and the (active) control, where larger values of θ are favorable to the treatment. We are interested in testing the noninferiority hypothesis of the form

$$\mathcal{H}_0 : \theta \leq -\delta \quad vs. \quad \mathcal{H}_1 : \theta > -\delta. \quad (1)$$

The objective of our study is to present a more powerful exact unconditional test for noninferiority that is less conservative and more powerful than the traditional alternatives. The rest of the paper is structured as follows. In Section 2, we introduce the illustrative example used throughout the study together with relevant notation. In Section 3, standard approximate test statistics and P-values are described. Section 4 presents the range of transformations of approximate tests into exact form. In Section 5, we perform an extensive numerical comparison of the exact tests. Section 6 concludes the study.

2 Illustrative example: comparing true negative rates of two diagnostics tests

To illustrate and motivate our ideas, we describe a study reported by Kao et al. [11]. One of the objectives of the initial study was to compare the diagnostic accuracy of two alternative procedures for detecting recurrent or residual nasopharyngeal carcinomas (NPC). Four months after being treated by radiotherapy, which is the main treatment against NPC without distant metastases, the true state of the subjects was determined by biopsy. Of the total of 36 subjects, 25 showed no presence of NPC. These known disease negative patients were subsequently tested by two diagnostic methods: Method 1 is computed tomography (CT) and method 2 is technetium-99m methoxyisobutylisonitrile (Tc-MIBI) single photon emission computed tomography (SPECT). We use Tc-MIBI SPECT, which is more available and practical than CT though with varying diagnostic accuracy, as a test method. The CT method serves as a control in testing for noninferiority of Tc-MIBI SPECT.

The results of the study are summarized in Table 1. A correct (i.e. negative) diagnosis is coded as 1 and incorrect as 0. Out of the 25 patients, 22 were correctly diagnosed by both methods and another patient was incorrectly classified by both methods. Out of 2 observations where the two methods disagreed, Tc-MIBI SPECT was correct on both occasions. Overall, the empirical true negative proportions are $(x_{11}+x_{01})/n = 24/25 = 96\%$ for Tc-MIBI SPECT and $(x_{11}+x_{10})/n = 22/25 = 88\%$ for CT, so we estimate that Tc-MIBI SPECT has an 8% better true negative fraction than CT. This is not necessarily sufficient to conclude noninferiority. A test for noninferiority of Tc-MIBI SPECT is a one-sided test against the lower margin $-\delta$ given by (1). To make our illustrations more objective, we assume that $\delta = 0.10$ and is a properly selected noninferiority margin.

Table 1: Results of a comparison of true negative rates of CT and Tc-MIBI SPECT diagnoses of NPC, reported in Kao et al. [11].

Test	<u>Control</u>		Total
	CT (-)	CT (+)	
Tc-MIBI SPECT(-)	22(x_{11})	2(x_{01})	24
Tc-MIBI SPECT(+)	0(x_{10})	1(x_{00})	1
Total	22	3	25(n)

3 Standard approximate tests

In general, we have n individuals who have two binary responses measured. Denote by X_{jk} the number of responses $jk \in \{11, 01, 10, 00\}$ with corresponding probabilities π_{jk} .

We are interested in the parameter $\theta = \pi_{01} - \pi_{10}$, which is the probability of response from method 2 (the test treatment) minus the probability of response for method 1 (the control).

Provided individuals within each arm respond independently, the joint distribution of $\{X_{jk}\}$ is multinomial but can be expressed as a product of three binomial factors, namely

$$B(t; n, \phi)B(x_{01}; t, \eta)B(x_{11}; n - t, \psi),$$

where $B(x; n, p)$ is the probability of a binomial with parameters (n, p) equaling x and $t = x_{01} + x_{10}$ is the number of so-called “discordant” pairs. The parameter $\phi = \pi_{01} + \pi_{10}$ is the probability of a discordant pair, while $\eta = (\theta + \phi)/(2\phi)$ is the probability of a response (01) in favor of the treatment, conditional on the response being discordant. The parameter $\psi = \pi_{11}/(\pi_{00} + \pi_{11})$ and it is commonly agreed on the basis of sufficiency and/or conditionality principles that the factor involving this parameter has no relevance to inference on (θ, ϕ) , see Lloyd [12] for a more detailed discussion. Denoting x_{01} by x , the data comprise $y = (x, t)$ and the likelihood is

$$L(\theta, \phi; y, n) \propto \phi^t (1 - \phi)^{n-t} \eta^x (1 - \eta)^{t-x} \propto (1 - \phi)^{n-t} (\theta + \phi)^x (\phi - \theta)^{t-x} \quad (2)$$

defined over the parameter space $\Omega := \{(\theta, \phi) : 0 \leq |\theta| \leq \phi \leq 1\}$. The maximum likelihood (ML) estimate of θ is $\hat{\theta} = (x_{01} - x_{10})/n = (2x - t)/n$ and with variance σ^2/n , where $\sigma^2 = (\phi - \theta^2)$. The profile ML estimate $\hat{\phi}_\theta$ of ϕ when θ is assumed to be known is given by the larger solution of the quadratic equation

$$\phi^2 - \phi(\hat{\phi} - \hat{\theta}\theta) + \hat{\theta}\theta - (1 - \hat{\phi})\theta^2 = 0,$$

which is necessarily within the interval $[|\theta|, 1]$. The unrestricted ML estimate of ϕ is $\hat{\phi} = t/n$, which is equivalent to the profile estimate when $\theta = 0$.

Standard likelihood theory leads to Wald-type test statistics $\sqrt{n}(\hat{\theta} - \theta)/\hat{\sigma}$ with approximate standard normal distribution. Using the restricted ML estimator $\hat{\phi}_\theta - \theta^2$ of σ^2 under the null hypothesis, leads to the statistic

$$R(y; \theta) = \sqrt{n}(\hat{\theta} - \theta)/\sqrt{\hat{\phi}_\theta - \theta^2}$$

suggest by Nam [13] and Tango [14], who independently showed that this is identical to Rao’s score statistic. Other estimators of σ in the Wald statistic have been suggested by Lu and Bean [15] and Hsueh et al. [9] but are not considered in our study for reasons described in Hsueh et al. [9] and Liu et al. [16]. All Wald-type statistics reduce to the McNemar’s [17] statistic when $\theta = 0$.

An obvious alternative to $R(y; \theta)$ that has not been previously proposed is the signed root likelihood ratio (LR) statistic that has the form

$$L(y; \theta) = \text{sign}(\hat{\theta} - \theta)\text{LR}(y, \theta)^{1/2},$$

where $\text{LR}(y, \theta) = 2 \log(L(\hat{\theta}, \hat{\phi})/L(\theta, \hat{\phi}_\theta))$ with the conventions that both $0 \log 0$ and $\text{sign}(0)$ equal zero. Both the score and LR statistics are asymptotically normal when n is large and when ϕ is not near the boundary of $[|\theta|, 1]$.

Let $S(y; \theta)$ denote a generic test statistic for testing θ , either the score or LR statistics in our case. For testing noninferiority large positive values of $S(y; \delta)$ lead to rejection of the null hypothesis and an approximate P-value based on the approximating normal distribution is $P(y; \delta) = 1 - \Phi(S(y; \delta))$.

4 Exact tests

All our tests can be expressed in terms of rejecting the null if a P-value $P(y; \delta)$ is smaller than the chosen significance level α . The exact P-value

$$\pi(y; \delta, \phi) := \Pr(P(Y; \delta) \leq P(y; \delta); \delta, \phi) \quad (3)$$

depends on $\phi \in (|\delta|, 1)$. This probability is calculated using the binomial distribution of T with parameters (n, ϕ) and the conditional binomial distribution of X given T with parameters (t, η) , as indicated in Section 3. There are three main methods of controlling for the effect of ϕ . Each method is based on consideration of $\pi(y; \delta, \phi)$ as a function of ϕ , which we call the *significance profile*. This theory is identical for both one- and two-sided tests and the three P-values defined below are guaranteed to satisfy the characteristic property of a valid P-value, namely

$$\sup_{\phi} \Pr(P(Y; \delta) \leq \alpha; \delta, \phi) \leq \alpha. \quad (4)$$

Full Maximization: M P-values. Bickel and Doksum [18], p168 define the P-value to be the supremum of the significance profile. One can think of this as a transformation

$$\mathcal{M} : P(y; \delta) \rightarrow P^*(y; \delta) := \sup_{\phi} \{\pi(y; \delta, \phi), \phi \in (|\delta|, 1)\}. \quad (5)$$

It can be shown that $P^*(Y; \delta)$ is as small as possible amongst valid P-values that are non-decreasing functions of the original statistic $P(Y; \delta)$. It also satisfies (4) with the inequality replaced by a corresponding equality, see for instance Röhmel and Mansmann [19] for strong theorems to this effect.

Partial Maximization: B P-values. Berger and Boos [20] suggested the P-value

$$\mathcal{B} : P(y; \delta) \rightarrow P_{\gamma}(y; \delta) := \sup_{\phi} \{\pi(y; \delta, \phi) : \phi \in C_{\gamma}\} + \gamma, \quad (6)$$

where C_{γ} is a $100(1 - \gamma)\%$ confidence interval for ϕ under the null. This methodology has been applied in several recent papers, see for instance Berger and Sidik [21] and Sidik [22]. Proponents of this approach acknowledge that dependence of results on the choice of γ can be extreme, notwithstanding the general recommendation by Berger and Sidik that γ be small. It is essential, therefore, that the confidence interval error γ be strictly controlled. Sidik [22] suggested using a Clopper-Pearson interval for ϕ intersected with the null interval $(|\delta|, 1)$. Clearly, B P-values depend on the choice of γ , while no such subjective choices are required for M P-values.

Estimation Followed by Maximization: E+M P-values. An alternative and much older approach is to replace the nuisance parameter by its estimate under the null. This \mathcal{E} -step produces the statistic

$$\mathcal{E} : P(y; \delta) \rightarrow \hat{P}(y; \delta) := \pi(y; \delta, \hat{\phi}_\delta). \quad (7)$$

While $\hat{P}(Y; \delta)$ is not a valid P-value, it is valid after a further \mathcal{M} -step given by (5). This involves computing the significance profile of the statistic $\hat{P}(Y; \delta)$, which is different from that of $P(Y, \delta)$ and is typically much better behaved. Estimation followed by maximisation has been successfully applied to the likelihood ratio statistic for testing the difference between two independent proportions (see [23], [24] and [25]).

Both $R(y; \theta)$ and $L(y; \theta)$ are non-decreasing in x and non-increasing in t . This property is known as Barnard convexity, following Barnard [26] in the context of independent binary trials (see [19] and [22]). Apart from being logically essential, the monotonicity property is important for two reasons. Firstly, it ensures that the maximum probability of $\{P(Y) \leq P(y)\}$ over the null hypothesis $\theta \leq -\delta$ is attained at the value $\theta = -\delta$. Secondly, the tail probability in (3) can be much more efficiently computed using the cumulative binomial approach. Berger and Sidik [21] show that, for matched pairs, monotonicity is maintained by their \mathcal{B} -step. It can be demonstrated ([27]) that monotonicity is also preserved by the \mathcal{M} -step and the \mathcal{E} -step.

In terms of computational effort, the \mathcal{B} -step requires slightly less computation than the \mathcal{M} -step since the supremum search is over a restricted range. The \mathcal{E} -step is much easier computationally because it involves evaluation of the significance profile at a single point. A rank ordering from least to most computationally demanding is $\mathcal{E} < \mathcal{B} < \mathcal{M} < \mathcal{E} + \mathcal{M}$. For large enough sample sizes $n > 200$, E+M P-values will become computationally demanding though in our study no P-value takes more than a few seconds to compute.

Table 2 gives all P-values discussed above for the illustrative example. Recall that the estimates of the true negative rates were 96% and 88% for Tc-MIBI SPECT and CT respectively, leading to the estimated difference of $\hat{\theta} = 8\%$. We defined a hypothetical noninferiority margin as 10%. The noninferiority test (1) is testing for the alternative that $\theta > -10\%$.

Table 2: Various P-values for data set $(x, t, n; \delta) = (2, 2, 25; 0.10)$. The estimate is $\hat{\theta} = 0.08$. The \mathcal{B} -step is performed with $\gamma = 0.001$.

	P	E	M	B	E+M
Score	0.0127	0.0075	0.0174	0.0109	0.0085
LR	0.0023	0.0073	0.0077	0.0087	0.0077

We used $\gamma = 0.001$ for the \mathcal{B} -step. For the LR test, the B P-value is 0.001 larger than the M P-value. This happens when the maximum of the profile occurs inside

C_γ and yet the γ -penalty is still imposed. For the Score test, the B P-value smaller than the M P-value, though in this case the E+M P-value is even smaller than any alternatives. A detailed numerical comparison of the competing methods is presented in the next section.

5 Numerical study

We have two basic test statistics (Score and LR) and three exact transformations (\mathcal{M} , \mathcal{B} and $\mathcal{E} + \mathcal{M}$), making six P-values in all. All our comparisons are based on a full enumeration of the competing P-values for all $(n + 1)(n + 2)/2$ possible data sets. Our first comparison is based on the fact that all six P-values are valid P-values and we would like them to be as small as possible, especially under the alternative. A straightforward way to measure this is by the simple average across the sample space. To concentrate on the interesting part of the sample space, we calculate the average conditional on at least one of the six P-values being less than 0.10. This comparison does not involve a nominal size α . The second comparison is based on the rejection probability

$$\beta(\theta, \phi) = \Pr(P(Y) \leq \alpha; \theta, \phi),$$

where the P-value $P(Y)$ depends on δ . The achieved size of the test is measured by $\sup_\phi \beta(-\delta, \phi)$ with $\phi \in (|\delta|, 1)$ and is guaranteed to be less than α . As a third comparison, we will look at the power at the most interesting alternative parameter value, namely $\theta = 0$, i.e. we look at the power curve $\beta(0, \phi)$ where $\phi \in (0, 1)$. This power profile typically has high values, often close of equal to 1, when ϕ is close to 0 and a much lower minimum value when ϕ is close to 1. We will give a typical power profile plot below, but will end up using the mean value of corresponding power functions.

To give an indication of computation times, the computation of all 1326 possible score-based E+M P-values for $n = 50$ takes approximately 110 seconds using a Pentium-IV 2.00 GHz processor. Computational time for individual P-values varies from sample point to sample point depending on the size and shape of the rejection region.

Table 3 shows that the E+M P-values are in each case notably smaller than M P-values on average and also smaller than the B P-values, but to a lesser extent. These averages give equal weight to each sample point. We have also computed averages using the joint binomial weights, as suggested by Lloyd and Moldovan [28] in the confidence limits context, and obtained very similar results. There is simply no question that the E+M P-values are collectively smaller than the alternatives. There is no practical difference between E+M P-values based on the score and LR statistics, using this crude mean measure.

Table 4 lists mean powers of the tests at the alternative $\theta = 0$. It is evident that both \mathcal{B} and $\mathcal{E} + \mathcal{M}$ based tests achieve higher power than \mathcal{M} tests, that $\mathcal{E} + \mathcal{M}$ tests are slightly more powerful than \mathcal{B} tests and that there is almost no practical power difference between score and LR-based tests, with a slight advantage to score-based

Table 3: Comparison of six exact P-values. Tabulated figures are average P-values conditional on at least one of the six being less than 0.10.

n	δ	M		B		E+M	
		Score	LR	Score	LR	Score	LR
20	0.05	0.019	0.024	0.018	0.021	0.017	0.017
25		0.017	0.033	0.017	0.020	0.016	0.016
30		0.015	0.033	0.015	0.018	0.014	0.014
40		0.013	0.025	0.013	0.014	0.012	0.012
50		0.011	0.018	0.011	0.012	0.010	0.010
75		0.009	0.012	0.009	0.009	0.008	0.008
100		0.007	0.010	0.008	0.008	0.007	0.007
20	0.10	0.018	0.032	0.018	0.022	0.017	0.017
25		0.016	0.024	0.016	0.018	0.015	0.015
30		0.013	0.018	0.013	0.015	0.012	0.012
40		0.012	0.015	0.012	0.013	0.011	0.011
50		0.010	0.013	0.010	0.011	0.009	0.009
75		0.008	0.009	0.008	0.008	0.007	0.007
100		0.007	0.007	0.007	0.007	0.006	0.006

tests. We have also computed the achieved size of the tests, given by $\sup_{\phi} \beta(\delta, \phi)$ with $\phi \in (\delta, 1)$ and found that $\mathcal{E} + \mathcal{M}$ tests achieve size slightly closer to nominal than its competitors. The entire set of results is available from the authors by request.

It can be seen from Table 4 that in some cases the powers of the competing tests are identical and that this pattern is itself erratic. For instance, with $\delta = 0.10$, we see that for $n = 30, 50$ the powers are identical for all six tests, while for $n = 75$ there is a distinct advantage for the $\mathcal{E} + \mathcal{M}$ tests. It is worth clarifying the reasons for this inconsistent behavior, as well as the computation of the summary mean powers.

Figure 1 describes score-based P-values with $n = 50$ and $\delta = 0.10$. The left panel plots the E+M P-values against the M P-values, and it is apparent that there is a tendency for E+M P-values to be smaller, which is consistent with the results in Table 3. For a test of nominal size 5%, the two tests happen to be identical. They reject the null for precisely the same data sets. This is more an accident of the choice of α and the sample size n , and for a nominal size of $\alpha = 0.10$, the two tests disagree, with the E+M test rejecting the null more often than the M test. The right panel plots the power $\beta(0, \phi)$ against ϕ for the $\mathcal{E} + \mathcal{M}$ test (solid) and the \mathcal{M} test (dotted), with $\alpha = 0.10$. We are using the mean value of these plots to summarize the power. For this specific example, the mean powers are 41.8% and 44.9%, a difference of 3.1% in favor of the $\mathcal{E} + \mathcal{M}$ test. This approach to power evaluation is different from Liu et al. [16] and Sidik [22] who consider several distinct values of the nuisance parameter.

Table 4: Comparison of six nominal 5% tests. Entries are mean of power profile $\beta(\theta, \phi)$ over the interval $\phi \in [0, 1]$ at the alternative $\theta = 0$.

n	(δ, θ)	M		B		E+M	
		Score	LR	Score	LR	Score	LR
25	(0.05, 0)	0.099	0.070	0.099	0.072	0.099	0.099
30		0.098	0.066	0.105	0.076	0.105	0.105
40		0.159	0.108	0.159	0.127	0.159	0.159
50		0.152	0.086	0.164	0.129	0.178	0.169
75		0.255	0.264	0.264	0.264	0.271	0.271
100		0.309	0.284	0.318	0.310	0.326	0.326
25	(0.10, 0)	0.195	0.120	0.216	0.136	0.216	0.198
30		0.308	0.308	0.308	0.308	0.308	0.308
40		0.372	0.337	0.372	0.347	0.372	0.372
50		0.462	0.462	0.462	0.462	0.462	0.462
75		0.550	0.493	0.566	0.545	0.577	0.577
100		0.663	0.621	0.663	0.651	0.665	0.665

6 Discussion

In the presented study, we compared three types of exact unconditional tests for non-inferiority in the binary matched pairs context. On the basis of the numerical results presented above, we recommend that noninferiority tests be based on the E+M P-value from the score statistic. The power gains compared to the partially maximized P-values of Sidik [22] are modest but distinct. Moreover, our proposed P-values do not require a subjective choice of γ . At the same time, our tests demand the computational effort very similar to the known alternatives and we provide software for their implementation.

The vast majority of research has concentrated on one-sided tests of noninferiority, with the usually implicit recommendation that equivalence be tested using the procedure based on combining two one-sided tests (TOST) with the same target size, see Theorem 1, in Berger and Hsu [29]. There appears to be no recognition in the literature that this approach is unnecessarily conservative and that the conservatism can be easily reduced by applying maximisation to P-values obtained from the TOST. Under some circumstances (Theorem 2, [29]), the size of the TOST will achieve the upper bound α so that combining efficient one-sided tests automatically leads to an efficient equivalence test. However, the conditions for this do not hold for matched pairs nor for most discrete data models. Even beginning with an efficient one-sided test, the corresponding TOST will be conservative. Results of applying the $\mathcal{E} + \mathcal{M}$ methodology to two-sided tests of equivalence are available from the authors in a working paper [30]. The conclusion of that research is to recommend the TOST based on estimated score P-values, maximized with respect to the nuisance parameter.

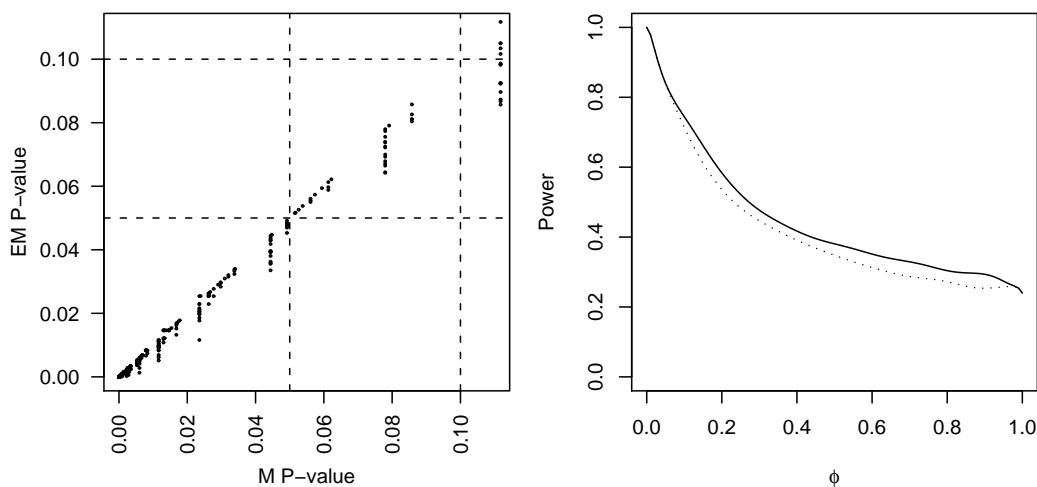


Figure 1: Power of score-based \mathcal{M} and $\mathcal{E} + \mathcal{M}$ tests with $n = 50$. *Left.* E+M P-values versus M P-values for $\delta = 0.10$. *Right.* Power profiles for $\alpha = 0.10$.

It has been pointed out by a referee that applying the \mathcal{M} -step to the Berger-Boos P-value can potentially produce P-values which may combine the attractive logical features of B P-values with exactness. Indeed, it may turn out that B P-values with quite large values of γ , appropriately maximized by the \mathcal{M} step, have good statistical properties. However, there are computational challenges in computing these P-values since the partially maximized P-values must be computed for all possible data sets before computing the significance profile and we leave this question as a topic for future research.

References

- [1] D’Agostino RB, Massaro JM, Sullivan LM. Non-inferiority trials: Design concepts and issues – the encounters of academic consultants in statistics. *Statistics in Medicine* 2003; **22**:169–186.
- [2] Gomberg-Maitland M, Frison L, Halperin JL. Active-control clinical trials to establish equivalence or noninferiority: Methodological and statistical concepts linked to quality. *American Heart Journal* 2003; **146**:398–403.
- [3] Koch K, Röhmel J. Hypothesis testing in the ”Golden Standard” design for proving the efficacy of an experimental treatment relative to placebo and a reference. *Journal of Biopharmaceutical Statistics* 2004; **14**:315–325.
- [4] Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments - Part 1: Ethical and scientific issues. *Annals of Internal Medicine* 2000; **133**:455–463.

- [5] Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments - Part 2: Practical issues and specific cases. *Annals of Internal Medicine* 2000; **133**:464–470.
- [6] Wang S-J, James Hung HM, Yi Tsong. Utility and pitfalls of some statistical methods in active controlled clinical trials. *Controlled Clinical Trials* 2002; **23**:15–28.
- [7] Wiens B. Choosing an equivalence limit for non-inferiority or equivalence studies. *Controlled Clinical Trials* 2002; **23**:2–14.
- [8] James Hung HM, Wang S-J, O’Neill R. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. *Biometrical Journal* 2005; **47**:28–36.
- [9] Hsueh H-M, Liu J-P, Chen JJ. Unconditional exact tests for equivalence or non-inferiority for paired binary endpoints. *Biometrics* 2001; **57**:478–483.
- [10] Tang M-L, Tang N-S, Carey VJ. Confidence interval for rate ratio in a 2×2 table with structural zero: An application in assessing false-negative rate ratio when combining two diagnostic tests. *Biometrics* 2004; **60**:550–555.
- [11] Kao CH, Shiau YC, Shen YY, Yen RF. Detection of recurrent or persistent nasopharyngeal carcinomas after radiotherapy with technetium-99m methoxyisobutylisonitrile single photon emission computed tomography and computed tomography: Comparison with 18-fluoro-2-deoxyglucose positron emission tomography. *Cancer* 2002; **94**:1981–1986.
- [12] Lloyd, CJ. A new exact and more powerful unconditional test of no treatment effect from binary matched pairs. *Biometrics* 2007; in press.
- [13] Nam J. Establishing equivalence of two treatments and sample size requirements in matched pairs designs. *Biometrics* 1997; **53**:1422–1430.
- [14] Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 1998; **17**:891–908.
- [15] Lu Y, Bean JA. On the sample size for one-sided equivalence of sensitivities based upon McNemar’s test. *Statistics in Medicine* 1995; **14**:1831–1839.
- [16] Liu J-P, Hsueh H-M, Hsieh E, Chen JJ. Test for equivalence or non-inferiority for paired binary data. *Statistics in Medicine* 2002; **21**:231–245.
- [17] McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* 1947; **17**:153–157.
- [18] Bickel PJ, Doksum KA. *Mathematical Statistics*. Oakland: Holden-Day, 1977.

- [19] Röhmel J, Mansmann U. Unconditional non-asymptotic one-sided tests for independent binomial proportions when interest lies in showing non-inferiority or superiority. *Biometrical Journal* 1999; **41**:149–170.
- [20] Berger RL, Boos DD. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**:1012–1016.
- [21] Berger RL, Sidik K. Exact unconditional tests for a 2×2 matched-pairs design. *Statistical Methods in Medical Research* 2003; **12**:91–108.
- [22] Sidik K. Exact unconditional tests for testing non-inferiority in matched-pairs design. *Statistics in Medicine* 2003; **22**:265–278.
- [23] Skipka G, Munk A, Freitag G. Unconditional exact tests for the difference of binomial probabilities - contrasted and compared. *Computational Statistics and Data Analysis* 2004; **47**:757–773.
- [24] Munk A, Skipka G, Stratmann B. Testing general hypotheses under binomial sampling: The two sample case - asymptotic theory and exact procedures. *Computational Statistics and Data Analysis* 2005; **49**:723–739.
- [25] Munk A, Milke M, Munk A, Freitag G. Testing noninferiority in three-armed clinical trials based on the likelihood ratio statistics. *Canadian Journal of Statistics* 2006; **35**:413–431.
- [26] Barnard GA. Significance tests for 2×2 tables. *Biometrika* 1947; **34**:123–138.
- [27] Lloyd CJ. A New Exact and More Powerful Unconditional Test of no Treatment Effect from Binary Matched Pairs. *Biometrics* 2008; **64**
- [28] Lloyd CJ, Moldovan MV. Exact one-sided confidence limits for the difference between two correlated proportions. *Statistics in Medicine* 2007; **26**:3369–3384.
- [29] Berger RL, Hsu JC. Bio-equivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science* 1996; **11**:283–319.
- [30] Lloyd CJ, Moldovan MV. More powerful exact noninferiority and equivalence tests based on binary matched pairs. Working paper 2007-03: www.mbs.edu