EQUIVALENCE TESTS FOR PAIR-MATCHED BINARY DATA

Morikawa, Toshihiko
Pharmaceutical Development Division, Takeda Chemical Industries, Ltd.,

Yanagawa, Takashi
Graduate School of Mathematics, Kyushu University

Endou, Akira
Faculty of Engineering, Science University of Tokyo

Yoshimura, Isao
Faculty of Engineering, Science University of Tokyo

https://doi.org/10.5109/13453

出版情報：Bulletin of informatics and cybernetics. 28 (1), pp.31-45, 1996-03. Research Association of Statistical Sciences
バージョン：
権利関係：

Bulletin of Informatics and Cybernetics, Vol. 28, No. 1, 1996

EQUIVALENCE TESTS FOR PAIR-MATCHED BINARY DATA

By

Toshihiko MORIKAWA*, Takashi YANAGAWA†, Akira ENDOU‡ and Isao YOSHIMURA‡

Abstract

Equivalence tests for binary data with pair-matched design in clinical trials are explored in this paper. Eight tests are considered altogether including those tests constructed by the GSK method, the tests which use estimated null variances among others. The Type I error rates and powers are investigated by a Monte Carlo study, and it is shown that the test associated with the log-risk difference has the highest power among those eight tests. The determination of the value of tolerance is also discussed.

Key Words : cross-over design; McNemar test; log-odds difference criterion; log-risk difference criterion; pair-matched prospective design; risk difference criterion.

1. Introduction

Equivalence tests for binary data with parallel design has been proposed (see, for example, Dunnett and Gent [3]; Makuch and Simon [8]; Blackwelder [2]; Hirotsu [6]; Farrington and Manning [4]; Yanagawa, Tango and Hiejima [10]). However, the one with pair-matched design seems not to be given yet and in some cases equivalence tests for parallel designs are applied without regard to the association among matched pairs, which may affect results. In this paper, we propose equivalence tests for pair-matched binary data, arising in the situation where two drugs, say, new drug T and standard drug C, are administered to the same person with the same condition in a homogeneous population and the effects of drugs are evaluated in binary response. Such data are obtained from, for example, a bilateral design, using both eyes or both ears, etc., a cross-over design, or other pair-matched prospective designs. Note that, in the application of matched pair methods, it is required that there is no bias between right and left sides in bilateral studies, and there is no residual effect or period effect in the cross-over studies. Also note that in the so-called clinical equivalence, it is sufficient to show that the response of T is not lower than that of C over a given value of tolerance,

* Pharmaceutical Development Division, Takeda Chemical Industries, Ltd., Osaka 541, Japan.

† Graduate School of Mathematics, Kyushu University 33, Fukuoka 812, Japan.

‡ Faculty of Engineering, Science University of Tokyo, Tokyo 162, Japan.

so-called "delta value". We consider the test statistics which are based on the GSK method (Grizzle, Starmer and Koch [5]) using the generalized least squares (GLS) for some function of multinomial distribution. The statistics utilize the empirical variance in denominators. We also consider the test statistics using the estimated null variance obtained by the maximum likelihood method, and several other statistics.

The problem and hypotheses are established in Section 2, and numerical illustration is given in Section 3 to show the value of the new tests of equivalence which will be developed in Section 4. Eight tests are developed altogether in this paper. The use of confidence intervals is briefly discussed in Section 5 in relation to the proposed tests. In Section 6 the Type I errors and powers of the proposed tests are compared by a Monte Carlo study. One of the most controversial points in the practice of the test of equivalence has been centered around the value of tolerance. The problem is studied in Section 7 and methods of determining the value of the tolerance is explored in the section. Section 8 summarizes our findings and gives suggestion for the use of the equivalence test in pair-matched designs.

2. Problem and Hypothesis

In a bilateral or cross-over trial, assume that two drugs T and C are allocated randomly to both sides or two treatment periods of n subjects, respectively. Let n_{ij} be the number of persons who responded as i for drug T and as j for control C, where i, j = 1, 2; 1 indicates a positive response (+) and 2 a negative one (-). The results of the trial may be summarized in Table 1, where $n_{1.}$, $n_{2.}$, $n_{.1}$ and $n_{.2}$ represent marginal frequencies, and n is the total sample size. Corresponding cell probabilities and associated marginal probabilities are shown in Table 2. We want to prove the equivalence of T to C based on the data in Table 1, in the sense that $\pi_{1.}$ is not lower than $\pi_{.1}$ over a predetermined value of tolerance, where marginal probabilities $\pi_{1.}$ and $\pi_{.1}$ represent the response rates of drugs T and C, respectively.

Table 1: A frequency table from pair-matched data

		C		
		+	-	total
T	+	n_{11}	n_{12}	$n_{1.}$
	-	n_{21}	n_{22}	$n_{2.}$
	total	$n_{.1}$	$n_{.2}$	n

Hereafter we use the word "risk" as a general terminology to represent a response probability. We consider three tolerance criteria according to the three types of scale to be evaluated, namely, original scale, log-risk scale or log-odds (or logit) scale. Corresponding hypotheses to be tested are as follows:

(a) *risk difference criterion*

Table 2: Cell probabilities corresponding to Table 1

		C		
		+	-	total
T	+	π_{11}	π_{12}	$\pi_{1.}$
	-	π_{21}	π_{22}	$\pi_{2.}$
	total	$\pi_{.1}$	$\pi_{.2}$	1

H_0: $\pi_{1.} \leq \pi_{.1} + \Delta \qquad (\Delta < 0)$

H_1: $\pi_{1.} > \pi_{.1} + \Delta$

(b) *log-risk difference criterion*

H_0: $\log \pi_{1.} \leq \log \pi_{.1} + \log \Gamma \qquad (\log \Gamma < 0)$

H_1: $\log \pi_{1.} > \log \pi_{.1} + \log \Gamma$

(c) *log-odds (or logit) difference criterion*

H_0: $\log[\pi_{1.}/(1 - \pi_{1.})] \leq \log[\pi_{.1}/(1 - \pi_{.1})] + \log \Psi \qquad (\log \Psi < 0)$

H_1: $\log[\pi_{1.}/(1 - \pi_{1.})] > \log[\pi_{.1}/(1 - \pi_{.1})] + \log \Psi.$

In order to treat the problems in a unified format, the logarithmic scales are used here for (b) and (c). Note that the hypotheses (b) and (c) may be represented in the original scale by:

(b') H_0:$\pi_{1.} \leq \Gamma \pi_{.1} \qquad (0 < \Gamma < 1)$

H_1:$\pi_{1.} > \Gamma \pi_{.1}$

and

(c') H_0: $[\pi_{1.}/(1 - \pi_{1.})] \leq \Psi[\pi_{.1}/(1 - \pi_{.1})] \qquad (0 < \Psi < 1)$

H_1: $[\pi_{1.}/(1 - \pi_{1.})] > \Psi[\pi_{.1}/(1 - \pi_{.1})]$,

which are often called the risk ratio (or relative risk) criterion and odds ratio criterion, respectively.

3. Numerical illustration

The equivalence tests that have been developed for parallel designs should not be applied for pair-matched designs. To illustrate it and also to show the value of the new tests which are developed in the next section we consider two numerical examples, and demonstrate the behavior of these tests. The tests considered for the parallel design are Z_{MSB}-test which was proposed in Makuch and Simon [8], and in Blackwelder [2], and Z_H-test, proposed in Hirotsu [6]. The new tests considered for the pair-matched design are Z_D-test for criterion (a), Z_R-test for criterion (b), and Z_L-test for criterion (c). The explicit formulae of these tests are given in the next section.

We first consider the data in Table 3. The cell frequencies of the table are generated by setting $p_{1.} = p_{.1} = 0.4$, cell odds ratio $\phi = 1$ (no association) and n = 100. The tolerance was selected as Δ = - 0.1, and Γ and Ψ in criteria (b) and (c) are determined so that $p_{.1} + \Delta = \Gamma p_{.1}$, and $(p_{.1} + \Delta)/(1 - p_{.1} - \Delta) = \Psi p_{.1}/(1 - p_{.1})$.

Applying the tests for a parallel design to the data in Table 3, we obtain

Z_{MSB} =1.443 ($p = 0.074$), and $Z_H = 1.450$ ($p = 0.074$).

For tests in pair-matched design, we obtain

$Z_D = 1.443$ ($p = 0.074$), $Z_R = 1.661$ ($p = 0.048$),

and $Z_L = 1.531$ ($p = 0.063$).

The results show that the behaviors of the Z_D, Z_{MSB}, and Z_H tests are similar. This would be reasonable since this is the case of no association in matched pairs, and furthermore all three tests are designed for the risk difference criterion. Note that the p-values of Z_R and Z_L-tests indicate that we might improve the powers of the tests of equivalence by considering criteria (b) and (c).

We next consider the data in Table 4. The data in the table are obtained by using the same sample size and the same marginal probabilities as Table 3 but the cell odds ratio $\phi \simeq 10$ (substantial association). Using the same values of tolerance as above, we obtain the following results:

$Z_D = 2.041$ ($p = 0.021$), $Z_R = 2.349$ ($p = 0.009$)

and Z_L= 2.165 ($p = 0.015$).

The values of Z_{MSB} and Z_H are the same as above since these tests use only marginal frequencies of a table. Generally, the responses of the new and standard drugs in a matched pair are positively associated, and this association is frequently substantial. If this is the case the above numerical values show that the new tests which will be constructed in the next section are the one that should be used in a pair-matched design. Inspecting the p-values of the new tests it is also indicated that criteria (b) and (c) could provide us more powerful tests of equivalence than the test for the the risk difference criterion.

Table 3: No association in a matched-pair ($\phi = 1$)

		C		
		+	-	total
T	+	16	24	40
	-	24	36	60
	total	40	60	100

Table 4: Substantial association in a matched-pair ($\phi \simeq 10$)

		C		
		+	-	total
T	+	28	12	40
	-	12	48	60
	total	40	60	100

4. Construction of tests

4.1. By GSK method

We first construct test statistics by the GSK (Grizzle, Starmer and Koch [5]) formulation. Define the vectors $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})'$ and $p = (p_{11}, p_{12}, p_{21}, p_{22})'$, where $p_{ij} = n_{ij}/n$ is a sample estimate of $E[p_{ij}] = \pi_{ij}$, for i, j=1, 2. Then the sample variance-covariance matrix of p is,

$$V_p = \frac{1}{n} \begin{pmatrix} p_{11}(1-p_{11}) & -p_{11}p_{12} & -p_{11}p_{21} & -p_{11}p_{22} \\ -p_{11}p_{12} & p_{12}(1-p_{12}) & -p_{12}p_{21} & -p_{12}p_{22} \\ -p_{11}p_{21} & -p_{12}p_{21} & p_{21}(1-p_{21}) & -p_{21}p_{22} \\ -p_{11}p_{22} & -p_{12}p_{22} & -p_{21}p_{22} & p_{22}(1-p_{22}) \end{pmatrix}$$

For some scalar function $f(p)$ of p, $f(p) = f(\pi) + \epsilon$, $\epsilon \sim N(0, S)$, asymptotically, where S is the asymptotic variance of $f(p)$ based on the delta method. Therefore the test of the null hypothesis $f(p) = c$ can be performed by using the test statistic

$$Z = (f(p) - c)/S^{1/2}, \tag{1}$$

and this statistic may be applied for testing $H_0 : f(\pi) \leq c$ against $H_1 : f(\pi) > c$ since $P(Z > a | f(\pi) < c) \leq P(Z > a | f(\pi) = c)$.

Statistic Z is distributed asymptotically as a standard normal under $f(\pi) = c$. If $Z \geq Z_\alpha$, we decide that T is clinically equivalent to C, where Z_α is an upper α point of a standard normal distribution. Note that the S is represented as $S = Q'V_pQ$ with $Q = \partial f(p)/\partial p$.

Now the criterion (a) utilizes a linear function $f(\pi) = A\pi$ with A = (0, 1, -1, 0), thus $Q = A$; The criterion (b) utilizes a log-linear function $f(\pi) = K \log\{A\pi\}$ with K = (1, -1) and

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

thus $Q = K\text{diag}\{Ap\}^{-1}A$, where the log operation applies to each element of the vector $A\pi$, i.e., $\log\{A\pi\} = \log(\pi_{1.}, \pi_{.1})' = (\log(\pi_{1.}), \log(\pi_{.1}))'$, and diag $\{Ap\}$ is the diagonal matrix with the elements of Ap as its diagonal elements; and the criterion (c) utilizes also a log-linear function $f(\pi) = K \log\{A\pi\}$ with K = (1, -1, -1, 1) and

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

Specifically we have

$$\begin{aligned} S \;=\; & [q_1^2 V_{11} + q_2^2 V_{22} + q_3^2 V_{33} + q_4^2 V_{44} + 2q_1q_2V_{12} + 2q_1q_3V_{13} + 2q_1q_4V_{14} \\ & + 2q_2q_3V_{23} + 2q_2q_4V_{24} + 2q_3q_4V_{34}], \end{aligned}$$

where V_{ij} is the (i,j) element of V_p and q_i is the i-th element of Q. Using these representations, the test statistics are explicitly given as follows:

For risk difference

$$Z_D = (p_{12} - p_{21} - \Delta)/S^{1/2},$$

where $q_1 = 0$, $q_2 = 1$, $q_3 = -1$, $q_4 = 0$.

For log-risk difference

$$Z_R = (\log p_{1.} - \log p_{.1} - \log \Gamma)/S^{1/2},$$

where $q_1 = p_{1.}^{-1} - p_{.1}^{-1}$, $q_2 = p_{1.}^{-1}$, $q_3 = -p_{.1}^{-1}$, $q_4 = 0$.

For log-odds (or logit) difference

$$Z_L = (\log p_{1.} - \log p_{.1} - \log p_{2.} + \log p_{.2} - \log \Psi)/S^{1/2},$$

where $q_1 = p_{1.}^{-1} - p_{.1}^{-1}$, $q_2 = p_{1.}^{-1} + p_{.2}^{-1}$, $q_3 = -p_{2.}^{-1} - p_{.1}^{-1}$, $q_4 = -p_{2.}^{-1} + p_{.2}^{-1}$.

Especially the test statistic Z_D, which is based on a risk difference, can be written as

$$Z_D = (n_{12} - n_{21} - n\Delta)/\{(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n\}^{1/2}.$$

If we set $\Delta = 0$, this reduces to

$$Z_D^0 = (n_{12} - n_{21})/[(n_{12} + n_{21}) - (n_{12} - n_{21})^2/n]^{1/2}.$$

Note that the McNemar's test statistic

$$Z_M = (n_{12} - n_{21})/(n_{12} + n_{21})^{1/2}$$

is constructed by taking into account the heterogeneity among individuals, whereas the homogeneity among individuals is assumed in this paper. We would emphasize that in randomized matched-pair prospective designs, we often come across the situations where this assumption is justifiable. $Z_D^0 > Z_M$ for $n_{12} \neq n_{21}$, but usually in practice $(n_{12} - n_{21})^2/n$ is relatively small and the difference of these statistics will not be substantial.

Also statistic Z_D can be represented as

$$Z_D = (p_{12} - p_{21} - \Delta)/[\{p_{1.}(1 - p_{1.}) + p_{.1}(1 - p_{.1}) - 2(\phi - 1)p_{12}p_{21}\}/n]^{1/2},$$

where $\phi = p_{11}p_{22}/p_{12}p_{21}$. Note that ϕ is an estimate of the population cell odds ratio $\Phi = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$. If we assume that the data are mutually independent in a matched pair, ϕ would be close to one. If we set $\phi = 1$, we have

$$Z_{MSB} = (p_{1.} - p_{.1} - \Delta)/[\{p_{1.}(1 - p_{1.}) + p_{.1}(1 - p_{.1})\}/n]^{1/2}.$$

This is the equivalence test statistic proposed by Makuch and Simon [8] or Blackwelder [2]. Thus Z_D-test can be considered as a matched-pair analog of Makuch, Simon, and Blackwelder-test in a parallel design. Pair-matched data in our situation are usually positively associated, i.e., $\Phi > 1$, thus the value of Z_D tends to be larger than that of Z_{MSB}. Furthermore the above representation of the statistic Z_D indicates that the power of the Z_D-test increases as the increase of the positive association.

4.2. Statistics with estimated null variances in the denominators

The statistics obtained by the GSK method utilize empirical variances in the denominators. Alternatively, we may use estimated null variances. For a parallel design Hirotsu [6] employed this idea and proposed the following test statistic;

$$Z_H = (p_{1.} - p_{.1} - \Delta)/[\{p_{1.}^*(1 - p_{1.}^*) + p_{.1}^*(1 - p_{.1}^*)\}/n]^{1/2},$$

where the $p_{1.}^*$ and $p_{.1}^*$ are the restricted maximum likelihood estimators (MLEs) with

the restriction $p^*_{1.} - p^*_{.1} = \Delta$.

Now we develop corresponding tests in a pair-matched design. The variance of $f(p) - c$ is approximated by $S^* = Q^* V_{p^*} Q^{*\prime}$, where V_{p^*} and Q^* are the same quantities as V_p and Q except p is replaced by p^*, where p^* is the restricted MLE of π obtained under the null hypothesis $f(\pi) = c$. Using the Lagrangean multiplier, the MLE, p^*, is obtained by maximizing

$$\ell = n_{11} \log \pi_{11} + n_{12} \log \pi_{12} + n_{21} \log \pi_{21} + n_{22} \log \pi_{22}$$
$$-\lambda(\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} - 1) - \eta(f(\pi) - c).$$

The test statistics are given by $Z^* = (f(p) - c)/(S^*)^{1/2}$. These statistics follow the standard normal distribution under the null hypothesis asymptotically. In particular, the statistics for the risk difference and log-risk difference are explicitly given as follows:

(a) *For risk difference*

$$Z^*_D = (p_{12} - p_{21} - \Delta)/[\{(p^*_{12} + p^*_{21}) - \Delta^2\}/n]^{1/2},$$

where p^*_{21} is given as a proper solution of a quadratic equation $Ax^2 - Bx + C = 0$ with $A = 2n$, $B = [(n_{12} + n_{21}) + \Delta(n_{12} - n_{21}) - 2n\Delta]$, $C = -n_{21}\Delta(1 - \Delta)$, and the other MLEs are obtained as $p^*_{12} = p^*_{21} + \Delta$, $p^*_{11} = n_{11}/\lambda$, $p^*_{22} = n_{22}/\lambda$, and $\lambda = (n_{12}/p^*_{12} + n_{21}/p^*_{21})/2$.

(b) *For log-risk difference*

$Z^*_R = (\log p_{1.} - \log p_{.1} - \log \Gamma)/(S^*)^{1/2}$, where p^*_{12} and p^*_{21} are given by $p^*_{12} = Dp^*_{21}$ and $p^*_{21} = E/F$,

with

$$A = \Gamma n_{21}, \quad B = n_{1.} - \Gamma^2 n_{.1}, \quad C = -\Gamma n_{12},$$
$$D = [-B + (B^2 - 4AC)^{1/2}]/2A, \quad E = \Gamma n_{12} + D n_{21},$$
$$F = (1 + \Gamma) D n_{22} - (\Gamma n_{12} + D n_{21})(\Gamma D - 1)/(1 - \Gamma),$$

and the other MLEs are given by

$$p^*_{11} = -\{(D - \Gamma)/(1 - \Gamma)\} p^*_{21}, \quad p^*_{22} = 1 + \{(\Gamma D - 1)/(1 - \Gamma)\} p^*_{21}.$$

We do not consider the test for the log-odds difference criterion in this paper since it requires an iterative solution. It might be interesting to apply the Z^*_D and Z^*_R tests to the data in Tables 3 and 4. The values of Z^*_D are obtained from Tables 3 and 4, respectively, by 1.442 (p=0.075) and 1.968 (p=0.025). The MLEs of cell probabilities are $p^*_{11} = 0.16$, $p^*_{12} = 0.20$, $p^*_{21} = 0.30$, and $p^*_{22} = 0.35$, for Table 3; and $p^*_{11} = 0.27$, $p^*_{12} = 0.08$, $p^*_{21} = 0.18$, and $p^*_{22} = 0.46$ for Table 4. The corresponding values of Z^*_R are obtained, respectively, as 1.606 (p=0.054) and 2.105 (p=0.018), and the MLEs of cell probabilities are $p^*_{11} = 0.15$, $p^*_{12} = 0.19$, $p^*_{21} = 0.30$, $p^*_{22} = 0.36$ for Table 3; and $p^*_{11} = 0.25$, $p^*_{12} = 0.08$, $p^*_{21} = 0.19$, $p^*_{22} = 0.48$ for Table 4. As anticipated, these p-values are slightly larger than those p-values given in section 3. It will be shown in the next section by simulation that these tests provide closer levels to the nominal levels than the tests by GSK method, in particular, when the size of the sample is small.

4.3. Other test statistics

Using the MLEs in the previous subsection, we easily obtain the likelihood ratio (LR) chi-square

$$\chi^2_{LR} = 2(L_{max} - L_{max}(p*)),$$

where L_{max} is the unrestricted maximum log-likelihood and $L_{max}(p*)$ is the restricted maximum log-likelihood with restriction $p^*_{1\cdot} = p^*_{\cdot 1} + \Delta$ or $p^*_{1\cdot} = \Gamma p^*_{\cdot 1}$, depending on the criteria employed. Setting

$$\begin{aligned} Z_{LR} &= \chi_{LR} \quad \text{if } p_{1\cdot} \geq p_{\cdot 1} + \Delta, \\ &= -\chi_{LR} \quad \text{otherwise,} \end{aligned}$$

we may obtain one-sided LR test. We included these statistics in the Monte Carlo study in the next section, but the behaviors of these statistics are quite similar to the corresponding statistics with the empirical variance in the denominators and we omit to describe the results.

If we apply the GSK approach directly to the risk ratio criterion we may obtain

$$Z_{RR} = (p_{1\cdot} - \Gamma p_{\cdot 1})/S^{1/2},$$

where S is given in (1) with $q_1 = 1-\Gamma$, $q_2 = 1$,$q_3 = -\Gamma$, and $q_4 = 0$. We also included this statistic in our Monte Carlo study but its behavior is quite similar to the Z_R-test and we omit to describe the results. Similarly, we may construct the test statistic by applying the GSK method to criterion (c'). Since the expression of the statistic is involved it is not included in our simulation.

5. Confidence intervals

Confidence intervals are utilized frequently to show the equivalence of two drugs. We may immediately obtain a two-sided (1-α) confidence interval of $f(\pi)$ as

$$|f(p) - f(\pi)|/S_{1/2} \leq Z_\alpha \quad \text{or} \quad f_L,\, f_U = \{f(p) \pm Z_\alpha S^{1/2}\}.$$

It would be reasonable to decide that T is clinically equivalent to C if and only if the lower limit $f_L \geq c$. Now this formulation is equivalent to the equivalence test based on (1), and the behaviors of the confidence intervals which are constructed by selecting the function f according to the risk difference, log-risk difference, or log-odds difference criterion, and also using the empirical variance or the null variance in the denominators, are the same as those of the corresponding test statistics. Especially the confidence interval using the empirical variance for the risk difference is described in some text books, e.g., Agresti [1] or Sakuma [9].

6. Monte Carlo Study

A Monte Carlo study was conducted by SAS ver.6.08 to investigate the Type I error rates and powers for the eight tests, $Z_D, Z_D{}^*, Z_R, Z_R{}^*, Z_L, \chi^2_{LR}$ (original scale), χ^2_{LR} (log-risk scale) and Z_{RR} (risk ratio). For the reason mentioned above we only show the results of the $Z_D, Z_D{}^*, Z_R, Z_R{}^*$ and Z_L. The response rate of the standard drug C was selected as $\pi_{\cdot 1} = 0.2(0.2)0.8$, and that of the new drug T was $\pi_{1\cdot} = \pi_{\cdot 1} + \delta$, for each $\pi_{\cdot 1}$, where the true difference δ was selected as δ = -0.1(0.05)0.1. The cell odds ratio was Φ = 1, 2, 4, 8 and 10. For each combination of the values of $\pi_{\cdot 1}$, δ and Φ, the cell probability vector π was calculated, and from each π, 10000 multinomial samples of size

n were generated using the SAS RANTBL function. The tolerance value Δ was always set to - 0.1, and the corresponding values of Γ and Ψ were generated so that $\pi_{.1} + \Delta = \Gamma\pi_{.1}$ and $(\pi_{.1} + \Delta)/(1 - \pi_{.1} - \Delta) = \Psi\pi_{.1}/(1 - \pi_{.1})$, respectively. Each data was tested using the above eight tests at the significant level $\alpha = 0.05$. Although the simulations were performed for sample sizes n = 30, 50 and 100, only the results for n = 30, 100 (and Φ = 1, 10) are shown in this paper.

Table 5 (for n=30) and Table 6 (for n=100) show the simulated Type I error rates and powers of the five test statistics, Z_D, Z_D^*, Z_R, Z_R^*, and Z_L, for Φ = 1, 10 and $\pi_{.1}$ = 0.2(0.2)0.8. In the tables, the rows corresponding to the true difference δ = -0.1 list the Type I error rates and the other rows list the powers, because we set Δ = -0.1.

Table 6 shows that if the sample size is as large as or larger than n = 100, all five tests control the Type I error approximately to the nominal level. But if one look into it carefully, although not substantial, the Type I errors of the tests tend to be slightly larger than the nominal level, especially in the case of Φ = 10. It seems that this tendency is stronger in the lower or the higher response rates. Table 5 shows that when the sample size is small, i.e., n=30, these findings are amplified in the tests that use the empirical variance, i.e., Z_D, Z_R and Z_L, but the table shows that those tests using the estimated null variance, i.e., Z_D^* and Z_R^*, remarkably improve the overestimation.

Concerning the power, Tables 5 and 6 show that

(i) the power of Z_R is higher than those of Z_D and Z_L, and also the power of Z_R^* is higher than that of Z_D^*,

(ii) when the sample size is small the powers of Z_D and Z_R are higher than those of Z_D^* and Z_R^*, respectively; the difference is large especially when Φ = 10, whereas when the sample size is large those powers are similar,

(iii) when $\pi_{.1}$ is small ($\pi_{.1}$ = 0.2), the power of Z_D is lower than those of Z_R and Z_L,

(iv) when $\pi_{.1}$ is intermediate ($\pi_{.1}$ = 0.4 or 0.6), the powers of the five tests are relatively similar,

(v) when $\pi_{.1}$ is large ($\pi_{.1}$ = 0.8), the powers of Z_D and Z_R are higher than that of Z_L.

Summarizing these findings in the Monte Carlo study we may conclude that (a) when the sample size is small, the Z_R^*-test is recommended, (b) when the sample size is large the Z_R-test is suggested, and (c) when the sample size is large and response rate is intermediate or high, test Z_D is recommended since the test is simple.

7. Determination of the value of tolerance

One of the most controversial points in the practice of the test of equivalence is about the value of tolerance. The risk difference criterion is conventionally employed with Δ=-0.10 in testing equivalence in parallel designs. It seems that this value has been empirically accepted as a reasonable value of tolerance in many cases where the response rate of the standard drug is in the range between 0.2 and 0.8 (or more safely

Table 5: Type I Errors and Powers for n=30

Φ	δ	$\pi_{.1}$	Z_D	Z_D^*	Z_R	Z_R^*	Z_L
1	-0.10	0.2	0.066	0.048	0.054	0.034	0.056
		0.4	0.055	0.045	0.052	0.043	0.051
		0.6	0.059	0.049	0.055	0.051	0.057
		0.8	0.054	0.049	0.056	0.050	0.050
	-0.05	0.2	0.140	0.123	0.181	0.140	0.172
		0.4	0.122	0.104	0.126	0.112	0.117
		0.6	0.115	0.095	0.116	0.109	0.107
		0.8	0.123	0.112	0.134	0.116	0.102
	0.00	0.2	0.260	0.240	0.388	0.345	0.357
		0.4	0.209	0.183	0.235	0.218	0.206
		0.6	0.214	0.186	0.223	0.212	0.200
		0.8	0.261	0.240	0.285	0.254	0.206
	0.05	0.2	0.424	0.395	0.612	0.575	0.563
		0.4	0.337	0.302	0.389	0.370	0.337
		0.6	0.348	0.312	0.376	0.365	0.318
		0.8	0.467	0.438	0.517	0.463	0.358
	0.10	0.2	0.567	0.536	0.777	0.752	0.728
		0.4	0.480	0.440	0.554	0.536	0.481
		0.6	0.499	0.459	0.542	0.532	0.454
		0.8	0.701	0.662	0.763	0.699	0.546
10	-0.10	0.2	0.086	0.039	0.090	0.059	0.080
		0.4	0.067	0.047	0.066	0.050	0.063
		0.6	0.065	0.046	0.069	0.050	0.063
		0.8	0.071	0.044	0.079	0.054	0.068
	-0.05	0.2	0.197	0.119	0.286	0.230	0.254
		0.4	0.174	0.135	0.196	0.159	0.174
		0.6	0.164	0.131	0.185	0.150	0.157
		0.8	0.191	0.127	0.212	0.163	0.162
	0.00	0.2	0.372	0.275	0.566	0.504	0.515
		0.4	0.316	0.268	0.374	0.327	0.324
		0.6	0.319	0.266	0.363	0.306	0.303
		0.8	0.376	0.282	0.419	0.347	0.301
	0.05	0.2	0.582	0.490	0.792	0.749	0.745
		0.4	0.522	0.471	0.614	0.567	0.538
		0.6	0.534	0.480	0.600	0.540	0.508
		0.8	0.616	0.497	0.680	0.597	0.466
	0.10	0.2	0.764	0.697	0.921	0.893	0.892
		0.4	0.704	0.664	0.794	0.762	0.721
		0.6	0.717	0.668	0.787	0.747	0.673
		0.8	0.836	0.740	0.887	0.830	0.651

Table 6: Type I Errors and Powers for n=100

Φ	δ	$\pi_{.1}$	Z_D	Z_D^*	Z_R	Z_R^*	Z_L
1	-0.10	0.2	0.052	0.048	0.053	0.058	0.052
		0.4	0.051	0.051	0.052	0.052	0.050
		0.6	0.049	0.049	0.052	0.051	0.050
		0.8	0.052	0.051	0.056	0.053	0.049
	-0.05	0.2	0.252	0.243	0.375	0.393	0.345
		0.4	0.183	0.183	0.218	0.219	0.197
		0.6	0.171	0.171	0.194	0.193	0.168
		0.8	0.219	0.213	0.243	0.233	0.180
	0.00	0.2	0.548	0.537	0.770	0.786	0.726
		0.4	0.410	0.410	0.500	0.502	0.449
		0.6	0.420	0.419	0.475	0.473	0.404
		0.8	0.557	0.547	0.605	0.592	0.455
	0.05	0.2	0.817	0.814	0.963	0.966	0.940
		0.4	0.696	0.696	0.795	0.797	0.731
		0.6	0.708	0.708	0.775	0.773	0.683
		0.8	0.877	0.870	0.912	0.904	0.766
	0.10	0.2	0.949	0.948	0.997	0.997	0.992
		0.4	0.883	0.883	0.944	0.945	0.906
		0.6	0.910	0.910	0.949	0.949	0.890
		0.8	0.989	0.988	0.994	0.993	0.952
10	-0.10	0.2	0.063	0.050	0.062	0.063	0.060
		0.4	0.053	0.048	0.055	0.051	0.052
		0.6	0.055	0.052	0.061	0.054	0.054
		0.8	0.054	0.048	0.059	0.048	0.054
	-0.05	0.2	0.354	0.323	0.527	0.531	0.493
		0.4	0.276	0.266	0.339	0.322	0.299
		0.6	0.267	0.256	0.309	0.292	0.260
		0.8	0.324	0.301	0.357	0.324	0.269
	0.00	0.2	0.766	0.745	0.930	0.932	0.911
		0.4	0.670	0.660	0.769	0.757	0.711
		0.6	0.669	0.660	0.730	0.714	0.649
		0.8	0.762	0.741	0.806	0.776	0.647
	0.05	0.2	0.960	0.957	0.997	0.997	0.995
		0.4	0.924	0.919	0.966	0.964	0.942
		0.6	0.928	0.923	0.959	0.954	0.910
		0.8	0.977	0.973	0.987	0.982	0.921
	0.10	0.2	0.998	0.998	1.000	1.000	1.000
		0.4	0.993	0.992	0.999	0.999	0.996
		0.6	0.995	0.994	0.999	0.999	0.990
		0.8	1.000	1.000	1.000	1.000	0.995

0.3 and 0.7). The tests for the log-risk difference and the log-odds difference are seldom employed in the parallel designs since it is not easy to determine the values of tolerance for these criteria. The determination of the value of tolerance in pair-matched designs is more complicated than the case of parallel designs since the power of the test depends on the association in a pair. Ideally the value of the tolerance should be determined from the clinical point of view in the stage of planning of a trial, in consideration of sample sizes and powers of the test. We consider the statistic Z_D for a matched-pair design and Z_{MSB} for a parallel design in this section and explore statistical methods of the determination of the value of tolerance in a pair-matched design. Our basic idea is to decide the value of tolerance, or sample size, of the test Z_D in a pair-matched design so that Z_D has the equal power as Z_{MSB} in a parallel design.

Now it is easy to show that the Z_D is approximately distributed as N(μ_1,1) under H_1, where

$$\mu_1 = n^{1/2}(\pi_{1.} - \pi_{.1} - \Delta_{MP})/[(\pi_{12} + \pi_{21}) - (\pi_{12} - \pi_{21})^2]^{1/2},$$

and Δ_{MP} is the tolerance parameter in a pair-matched design, thus denoting by F the distribution function of the standard normal distribution the power of the Z_D-test is approximated by

$$P_{MP} = 1 - F(Z_\alpha - \mu_1).$$

Similarly the power of the Z_{MSB}-test is approximated by

$$P_P = 1 - F(Z_\alpha - \mu_2),$$

where

$$\mu_2 = n^{1/2}(\pi_{1.} - \pi_{.1} - \Delta_P)/[\pi_{1.}(1 - \pi_{1.}) + \pi_{.1}(1 - \pi_{.1})]^{1/2},$$

and Δ_P is the tolerance parameter in a parallel design.

A method of the determination is to set $\Delta_{MP} = \Delta_P$, using the conventional value of Δ_P; that is, to decide Δ_{MP}= -0.1. The determination elevates the power of Z_D, and could violate the base of the rule of thumb which has been empirically accepted in a parallel design. Thus this determination must be accompanied with the selection of the sample size at the initiation of the study. The sample size of the Z_D-test, say n_{MP}, which provides the same power as the Z_{MSB}-test with the sample size n_P is given by

$$n_{MP} = n_P[\{(\pi_{12} + \pi_{21}) - (\pi_{12} - \pi_{21})^2\}/\{\pi_{1.}(1 - \pi_{1.}) + \pi_{.1}(1 - \pi_{.1})\}].$$

We have $n_{MP} \leq n_P$ for $\Phi \geq 1$, and $n_{MP} = n_P$ if and only if Φ=1. Using the relationship of n_{MP} and n_P, and also using the conventional value of Δ_P employed in the parallel design, one may choose the sample size n_{MP} for a matched-pair design. For illustration the values of n_{MP}/n_P are listed in the last column of Table 7 for selected values of Φ, $\pi_{1.}$, $\pi_{.1}$, and δ ($= \pi_{1.} - \pi_{.1}$).

However in many cases in practice a test of equivalence is required after the study is designed; that is, after the sample size has been decided. If this is the case, we suggest to select the value of Δ_{MP} so that the powers of the two tests are identical, by supposing that n_{MP}=n_P and that the value of Δ_P is known, i.e., -0.1. Specifically the value of Δ_{MP} is given by

$$\Delta_{MP} = (\pi_{1.} - \pi_{.1}) + \{\Delta_P - (\pi_{1.} - \pi_{.1})\}[\{(\pi_{12} + \pi_{21}) - (\pi_{12} - \pi_{21})^2\}/\{\pi_{1.}(1 - \pi_{1.}) + \pi_{.1}(1 - \pi_{.1})\}]^{1/2}$$

It follows that $\Delta_{MP} \leq \Delta_P$ for $\Phi \geq 1$, and $\Delta_{MP} = \Delta_P$ when $\Delta_P = (\pi_{1.} - \pi_{.1})$ or $\Phi = 1$.

For illustration numerical values of Δ_{MP} are given in the second last column of Table 7 for selected values of $\pi_{1.}$, $\pi_{.1}$ and Φ, when Δ_P = -0.1.

Table 7: Tolerance and Sample Size Ratio to give the Same Power as the Parallel Design

Φ	$\pi_{.1}$	δ	$\pi_{1.}$	Δ_{MP}	N_{MP}/N_P
2	0.2	0.00	0.20	0.094	0.877
		0.05	0.25	0.090	0.870
		0.10	0.30	0.086	0.866
	0.4	0.00	0.40	0.091	0.833
		0.05	0.45	0.087	0.832
		0.10	0.50	0.082	0.832
	0.6	0.00	0.60	0.091	0.833
		0.05	0.65	0.087	0.837
		0.10	0.70	0.084	0.843
	0.8	0.00	0.80	0.094	0.877
		0.05	0.85	0.091	0.889
		0.10	0.90	0.091	0.908
10	0.2	0.00	0.20	0.075	0.556
		0.05	0.25	0.061	0.549
		0.10	0.30	0.049	0.556
	0.4	0.00	0.40	0.070	0.487
		0.05	0.45	0.055	0.489
		0.10	0.50	0.041	0.498
	0.6	0.00	0.60	0.070	0.487
		0.05	0.65	0.056	0.496
		0.10	0.70	0.044	0.515
	0.8	0.00	0.80	0.075	0.556
		0.05	0.85	0.065	0.584
		0.10	0.90	0.061	0.648

In assessing the sample size and the value of Δ_{MP} in a matched-pair design, we need prior information on the cell probabilities. If $\pi_{1.}$, $\pi_{.1}$ and Φ are assessed, π_{11} is obtained by

$$\pi_{11} = [B - (B^2 - 4AC)^{1/2}]/(2A),$$

where $A = (\Phi - 1)$, $B = [(\Phi - 1)(\pi_{1.} + \pi_{.1}) + 1]$ and $C = \Phi\pi_{1.}\pi_{.1}$, provided $\Phi \neq 1$. If $\Phi = 1$, then $\pi_{11} = \pi_{1.}\pi_{.1}$. Once π_{11} is obtained, other cell probabilities are obtained easily and Δ_{MP} or n_{MP} is finally obtained by the above formulae.

When $\pi_{1.} = \pi_{.1}$, the above formulae reduce to the simple formulae

$$n_{MP} = n_P[\pi_{12}/\{\pi_{1.}(1 - \pi_{1.})\}], \quad \text{and} \quad \Delta_{MP} = \Delta_P[\pi_{12}/\{\pi_{1.}(1 - \pi_{1.})\}]^{1/2}.$$

These formulae require no information on δ, and we might be able to use these formulae if the information is not available.

8. Discussion

We developed eight tests for the test of equivalence in a matched pair design in this paper, and studied their behavior in a simulation study. We first excluded three tests among them since their behavior is quite similar to one of the other five tests. Of the remaining five tests it is shown that Z_D^* and Z_R^*-tests which use the estimated null variance are more faithful to the nominal α level than the other tests for small sample sizes (n=30), and their powers are similar to the Z_D and Z_R-tests, respectively, for large sample sizes (n=100). The Z_R^*-test has higher power than the Z_D^*-test. Thus the use of the Z_R^* test is recommended. Because of its relative simplicity and of its highest power among the tests considered one could suggest the use of Z_R-test when the sample size is large (n=100).

A crucial point that should be taken into account in practice is the determination of the value of the tolerance. We developed the methods of the determination in this paper for the risk difference criterion with empirical variance. This approach using the empirical variance can apply to two other criteria, log-risk difference and log-odds difference, whereas in the case of using null variance, the approach cannot be applied directly.

After this paper was submitted for publication, we noticed the publication of the paper by Lu and Bean [7] in the latest issue of the Statistics in Medicine. They investigated the sample sizes for the equivalence tests for binary pair-matched data. Their test statistics are different from those eight statistics discussed in this paper. We would like to compare their statistics and ours in a follow up paper.

References

[1] AGRESTI, A.: *Categorical Data Analysis*, John Wiley & Sons, New York, (1990), 347-365.

[2] BLACKWELDER,W.C.: *"Proving the Null Hypothesis" in Clinical Trials*, Controlled Clinical Trials, **3** (1982), 345-353.

[3] DUNNETT, C.W. and GENT, M.: *Significance Testing to Establish Equivalence Between Treatments, with Special Reference to Data in the Form of 2 × 2 Tables*, Biometrics, **33** (1977), 593-602.

[4] FARRINGTON, C.P. and MANNING, G.: *Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk or Non-Unity Relative Risk*, Statistics in Medicine, **9** (1990), 1447-1454.

[5] GRIZZLE, J.E., STARMER, C.F. and KOCH, G.G.: *Analysis of Categorical Data by Linear Models*, Biometrics, **25** (1969), 489-505.

[6] HIROTSU, C.: *Some Statistical Problems in Clinical Trials (1) - Test for the Equivalence of Two Drugs*, Clinical Evaluation, **14** (1986), 467-475, in Japanese.

[7] LU, Y. and BEAN, J.A.: *On the Sample Size for One-Sided Equivalence of Sensitivities Based upon McNemar's Test*, Statistics in Medicine, **14** (1995), 1831-1839.

[8] MAKUCH, R. and SIMON, R.: *Sample Size Requirements for Evaluating a Conservative Therapy*, Cancer Treatment reports, **62** (1978), 1037-1040.

[9] SAKUMA, A.: *Seminar Text for Biostatistics in Clinical Trials*, JUSE, Tokyo, (1994). In Japanese.

[10] YANAGAWA, T., TANGO, T. and HIEJIMA, Y.: *Mantel-Haenszel-Type Tests for Testing Equivalence or More Than Equivalence in Comparative Clinical Trials*, Biometrics, **50** (1994), 859-864. (Correction: Biometrics, **51** (1995), 392).

Received October 20, 1995
Revised March 29, 1996