Journal of Biopharmaceutical Statistics, 15: 889–901, 2005

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400500265561

Taylor & Francis
Taylor & Francis Group

TESTS FOR EQUIVALENCE BASED ON ODDS RATIO FOR MATCHED-PAIR DESIGN

Jen-pei Liu
Division of Biometry, Department of Agronomy, National Taiwan University, Taipei, Taiwan and Division of Biostatistics and Bioinformatics, National Health Research Institutes, Taipei, Taiwan

Hsin-yi Fan and Mi-Chia Ma
Department of Statistics, National Cheng-kung University, Tainan, Taiwan

Currently, methods for evaluation of equivalence under a matched-pair design use either difference in proportions or relative risk as measures of risk association. However, these measures of association are only for cross-sectional studies or prospective investigations, such as clinical trials and they cannot be applied to retrospective research such as case-control studies. As a result, under a matched-pair design, we propose the use of the conditional odds ratio for assessment of equivalence in both prospective and retrospective research. We suggest the use of the asymptotic confidence interval of the conditional odds ratio for evaluation of equivalence. In addition, a score test based on the restricted maximum likelihood estimator (RMLE) is derived to test the hypothesis of equivalence under a matched-pair design. On the other hand, a sample size formula is also provided. A simulation study was conducted to empirically investigate the size and power of the proposed procedures. Simulation results show that the score test not only adequately controls the Type I error but it can also provide sufficient power. A numerical example illustrates the proposed methods.

Key Words: Equivalence; Match-paired design; δ-method; Odds ratio; Sample size; Score test.

1. INTRODUCTION

To increase the efficiency of treatment comparisons, a matched-pair design is often used in both prospective and retrospective clinical research. For example, to reduce the number of subjects exposed to the new unproved treatments, a matched-pair design is frequently used in prospective clinical trials. If the new treatment can provide advantages, such as a better safety profile, reduction of cost, or easy administration, assessment of the equivalence or noninferiority of the new treatment in efficacy to a standard treatment is a more appropriate objective of the study. In other words, the goal of the trial is to show that the efficacy of the new treatment is no worse than that of the standard within a prespecified clinically meaningful limit. Correlated binary endpoints are one of the most common responses obtained

Received February 2005; Accepted May 2005
Address correspondence to Jen-pei Liu, Division of Biometry, Department of Agronomy, National Taiwan University, 1 Section 4, Roosevelt Road, Taipei 106, Taiwan; E-mail: jpliu@ntu.edu.tw

from the matched-paired design. Current methods for evaluation of equivalence or noninferiority based on correlated binary endpoints include those proposed by Liu et al. (2002) and Hsueh et al. (2001) for difference in proportions or those proposed by Tang et al. (2003) for the relative risk defined as the ratio of proportions.

However, difference in proportions and relative risk are often used as measures of risk association for comparing different treatments in prospective clinical trials, but they cannot be applied to assess equivalence or noninferiority of risk or disease outcomes in retrospective studies. On the other hand, odds ratio is the most widely used measure of risk association that can be used in both prospective and retrospective studies. In case-control studies using a matched-pair design, a nonsignificant odds ratio with the corresponding confidence interval including 1 does not prove that the risk of the exposed subjects is equal or smaller than that of the unexposed or control. Similarly, a significant odds ratio with the corresponding confidence interval not including 1 does not imply that the risk of the exposed subjects exceeds that of the control by a prespecified level. Therefore, in addition to prospective clinical trials, equivalence or noninferiority is also an appropriate and legitimate objective for assessment of risk or disease outcomes in retrospective studies.

For the matched-pair design, the observed marginal odds ratio does not provide an estimate of the population odds ratio. On the other hand, the conditional retrospective odds ratio equals the conditional prospective odds ratio (Lachin, 2000). As a result, the conditional odds ratio is selected as one of the risk measures for evaluation of equivalence and noninferiority in both prospective and retrospective studies. In Section 2, the conditional odds ratio is defined for prospective and retrospective studies, and noninferiority hypothesis is formulated in terms of the conditional odds ratio. In Section 3, the procedure for the noninferiority testing using the delta method for the conditional odds ratio is suggested. In addition, an asymptotic score test based on the restricted maximum likelihood estimator (RMLE) is derived for evaluation of noninferiority using the conditional odds ratio. Furthermore, an approximate sample size formula is provided. Simulation results are presented in Section 4. In Section 5, a numerical example is given to illustrate the proposed procedures. Discussion and final remarks are given in Section 6.

2. CONDITIONAL ODDS RATIO AND NONINFERIORITY HYPOTHESIS

For a cross-sectional or prospective study under the 1:1 match, each member of samples' matched sets consists of an exposed or treated subject and an unexposed or control subject, and both exposed and control subjects are uniquely matched with respect to a set of covariates. A clinical trial for evaluation of equivalence between two diagnostic tests provides a typical example of the prospective matched-pair design. The subjects in such a trial serve as his or her own control and receive both the new (treated) and the standard (control) diagnostic tests. The primary endpoints for evaluation of diagnostic tests are usually binary responses such as sensitivity, specificity, or diagnostic agreement. It follows that each subject can be classified according to the values of the binary responses from the two diagnostic tests. Therefore, the results of N subjects can be presented in the 2×2 table as given in Table 1. On the other hand, for a matched-pair retrospective study, the sampling

Table 1 Data structure 2×2 table for a prospective matched-pair study

	Unexposed (control)		
Exposed (treatment)	Response (1)	Nonresponse (2)	Total
Response (1)	a (p_{11})	b (p_{12})	$a+b$ (p_{1+})
Nonresponse (2)	c (p_{21})	d (p_{22})	$c+d$ (p_{2+})
Total	$a+c$ (p_{+1})	$b+d$ (p_{+2})	N (1)

procedure is performed retrospectively in time. The selection of a case with the disease and a set of covariates and the selection of a control without the disease matching with the same set of covariates were performed first. Whether the case and control had been previously exposed to the risk factor of interest or treatment was then determined later. The results for the binary outcomes from a retrospective matched-pair study can be summarized in Table 2. In Tables 1 and 2, a, b, c, and d are observed numbers of pairs with outcomes $(1, 1)$, $(1, 2)$, $(2, 1)$, and $(2, 2)$, respectively, where a and d are referred to the concordant pairs, and b and c are called discordant pairs. In addition, let $p_{11}(q_{11})$, $p_{12}(q_{12})$, $p_{21}(q_{21})$, and $p_{22}(q_{22})$ be corresponding probabilities of the pairs.

The traditional unconditional observed marginal odds ratio from either prospective or retrospective matched-pair design does not provide an estimate of the unconditional population odds ratio for the general unselected population (Lachin, 2000). However, for the prospective matched-pair design, the matched exposed and unexposed pair members are sampled independently from their respective population. Conditioned on the values of the covariates, pair members are independent. Under the assumption of a constant conditional marginal odds ratio for all values of the matching covariate, the conditional marginal odds ratio can be expressed in terms of the ratio of the population averaged discordant probabilities as Lachin (2000)

$$\delta_{\text{pros}} = p_{12}/p_{21}, \quad \text{for prospective matched-pair design.}$$

Similarly, the conditional marginal odds ratio for the retrospective matched-pair design can be defined as

$$\delta_{\text{retro}} = q_{12}/q_{21}, \quad \text{for retrospective matched-pair design.}$$

Table 2 Data structure 2×2 table for a retrospective matched-pair study

	Nondiseased		
Disease	Exposed (1)	Unexposed (2)	Total
Exposed (1)	a (q_{11})	b (q_{12})	$a+b$ (q_{1+})
Unexposed (2)	c (q_{21})	d (q_{22})	$c+d$ (q_{2+})
Total	$a+c$ (q_{+1})	$b+d$ (q_{+2})	N (1)

Lachin (2000) further showed that the conditional retrospective odds ratio equals to the conditional prospective odds ratio; hence, they are referred to as the conditional odds ratio for the matched-pair design (i.e., $\delta = \delta_{\text{pros}} = \delta_{\text{pros}}$). As a result, under the prospective or retrospective matched-pair designs, one should use the conditional odds ratio because the traditional unconditional odds ratio is not an estimate of the unconditional population odds ratio. Because a common conditional odds ratio is used for evaluation of equivalence, for the sake of convenience, the setting in Table 1 for prospective matched-pair studies is used to illustrate the proposed methods. The results obtained from the prospective matched-pair design can be equally applied to the retrospective matched-pair studies. If the objective of a prospective clinical trial is to verify whether the efficacy of the new treatment is no worse than that of the standard treatment within a prespecified clinically meaningful margin δ_0, $0 < \delta_0 < 1$, the hypothesis of equivalence can be formulated as

$$H_0 : \delta \leq \delta_0 \quad \text{vs.} \quad H_a : \delta > \delta_0. \tag{1}$$

Hypothesis (1) can also be used to test whether the risk of a certain factor exceeds a predetermined level. In this case, margin δ_0 can be set greater than 1. For the goal of verifying whether the risk of a certain factor does not exceed a prespecified level, the hypothesis can be expressed as

$$H_0 : \delta \geq \delta_0 \quad \text{vs.} \quad H_a : \delta < \delta_0. \tag{2}$$

3. TEST STATISTICS AND SAMPLE SIZE

The conditional odds ratio δ can be consistently estimated by the sample proportions $\hat{p}_{12} = b/N$ and $\hat{p}_{21} = c/N$ for probabilities p_{12} and p_{21}, respectively (i.e., $\hat{\delta} = b/c$). It can be shown (Lachin, 2000) that asymptotically, $\ln(\hat{\delta}) = \ln(b/c)$ follows as a normal distribution with mean $\ln(\delta)$ and variance σ_1^2, where

$$\sigma_1^2 = V[\ln(\hat{\delta})] \approx \frac{1}{N}\left(\frac{1}{p_{12}} + \frac{1}{p_{21}}\right), \tag{3}$$

and ln represents the nature logarithm.

Because the cell proportions provide consistent estimates of cell probabilities, the variance can be estimated consistently as

$$\hat{\sigma}_1^2 = \widehat{V}[\ln(\hat{\delta})] = \frac{1}{N}\left(\frac{1}{\hat{p}_{12}} + \frac{1}{\hat{p}_{21}}\right) = \frac{1}{b} + \frac{1}{c}.$$

Hence, by Slutsky's theorem (Serfling, 1980) asymptotically, the statistic

$$Z_D = \frac{\ln(\hat{\delta}) - \ln(\delta_0)}{\sqrt{\frac{1}{b} + \frac{1}{c}}}. \tag{4}$$

follows the standard normal distribution.

Because nature logarithm is a monotonic function that transforms odds ratio from a range of $(0, \infty)$ into $(-\infty, \infty)$, the noninferiority hypothesis in Eq. (1) can be reformulated in terms of the log conditional odds ratio as

$$H_0 : \ln(\delta) \leq \ln(\delta_0) \quad \text{vs.} \quad H_a : \ln(\delta) > \ln(\delta_0). \tag{5}$$

It follows that H_0 in Eq. (4) is rejected, and the noninferiority of the new treatment to the standard is concluded at the α significance level if $Z_D > z_\alpha$, where z_α is the αth upper percentile of a standard normal variable. Equivalently, H_0 in Eq. (1) is rejected at the α significance level if the lower limit $(1 - 2\alpha)100\%$ confidence interval for $\ln(\delta)$ is greater than $\ln(\delta_0)$, i.e.,

$$\ln(\hat{\delta}) - z_\alpha \hat{\sigma}_1 > \ln(\delta_0). \tag{6}$$

The asymptotic variance of the delta method in Eq. (3) does not take the equivalence margin into consideration and may have impact on the performance of the procedure. To resolve this issue, a score test based on RMLE is proposed below. Because the observed frequencies $(b, c, N - b - c)$ follow a trinomial distribution with probabilities $(p_{12}, p_{21}, p_{11} + p_{22})$, therefore, the likelihood function of the data is given as

$$f(p_{12}, p_{21}) = \frac{N!}{b!c!(N-b-c)!} p_{12}^b p_{21}^c (1 - p_{12} - p_{21})^{N-b-c}.$$

When $\delta = p_{12}/p_{21}$, the log likelihood can be rewritten as

$$L(\delta, p_{21}) = b[\ln(\delta)] + (b + c)[\ln(p_{21})] + (N - b - c)[\ln(1 - \delta p_{21} - p_{21})] + \text{constant}$$

where δ is the parameter of interest and p_{21} is a nuisance parameter.

Therefore, the corresponding score functions can be shown as

$$U_\delta[\delta, p_{21}] = \frac{\partial L(\delta, p_{21})}{\partial \delta} = \frac{b}{\delta} + \frac{(N - b - c)(-p_{21})}{1 - \delta p_{21} - p_{21}}$$

and

$$U_{p_{21}}[\delta, p_{21}] = \frac{\partial L(\delta, p_{21})}{\partial p_{21}} = \frac{b + c}{p_{21}} + \frac{(N - b - c)(-\delta - 1)}{1 - \delta p_{21} - p_{21}}.$$

It follows that the score statistic for noninferiority hypothesis (1) when $\delta = \delta_0$ is given as

$$Z_S = \frac{U_\delta[\delta_0, \tilde{p}_{21}]}{\sqrt{Var(\delta_0, \tilde{p}_{21})}} = \frac{b - \delta_0 c}{\delta_0(\delta_0 + 1)} \sqrt{\frac{\delta_0(\delta_0 + 1)}{N\tilde{p}_{21}}}$$

where $\tilde{p}_{21}$ is the RMLE of p_{21} given at the boundary $\delta = \delta_0$ and is equal to $(b + c)/[N(\delta_0 + 1)]$.

Hence, Z_S reduces to

$$Z_S = \frac{b - \delta_0 c}{\sqrt{\delta_0(b+c)}}. \tag{7}$$

When $\delta_0 = 1$, Z_S reduces to the well-known McNemar's test statistic for testing no association. The Z_S statistic has an asymptotic standard normal distribution under the null hypothesis. It follows that H_0 in Eq. (1) is rejected and the noninferiority of the new treatment to the standard is concluded at the α significance level if $Z_S > z_\alpha$.

For determination of sample size for evaluation of noninferiority, we consider the following simple hypothesis

$$H_0 : \delta = \delta_0 \quad \text{vs.} \quad H_a : \delta = \delta_1, \tag{8}$$

where $\delta_0 < \delta_1$.

However, the hypothesis in Eq. (8) can be reformulated in terms of discordant probabilities as

$$H_0 : p_{12} - \delta_0 p_{21} = 0 \quad \text{vs.} \quad H_a : p_{12} - \delta_1 p_{21} = 0. \tag{9}$$

Hence, the difference of the parameters under the null and alternative hypotheses is given as $\eta = p_{12} - \delta_0 p_{21} - (p_{12} - \delta_1 p_{21}) = (\delta_1 - \delta_0)p_{21}$. Under the null hypothesis, the variance of the test statistic is given as

$$\begin{aligned} \sigma_0^2 &= V[\hat{p}_{12} - \delta_0 \hat{p}_{21} \mid p_{12} = \delta_0 p_{21}] \\ &= \frac{p_{21}\delta_0(\delta_0 + 1)}{N}, \end{aligned}$$

and the variance of the test statistic under the alternative hypothesis is given as

$$\begin{aligned} \sigma_{a1}^2 &= V[\hat{p}_{12} - \delta_0 \hat{p}_{21} \mid p_{12} = \delta_1 p_{21}] \\ &= \frac{p_{21}(\delta_1 + \delta_0^2) - [\delta_1 p_{21} - \delta_0 p_{21}]^2}{N}. \end{aligned}$$

It follows that the sample size required for power $1 - \beta$ at the α nominal level is given as

$$N = \left[\frac{z_\alpha \sqrt{\bar{p}_{21}\delta_0(\delta_0 + 1)} + z_\beta \sqrt{p_{21}(\delta_1 + \delta_0^2) - [p_{21}(\delta_1 - \delta_0)]^2}}{(\delta_1 - \delta_0)p_{21}} \right]^2 \tag{10}$$

where $\bar{p}_{21}$ is the asymptotic limit of $\tilde{p}_{21}$ for large N under $\delta = \delta_1$, and $\bar{p}_{21}$ is equal to $p_{21}[(\delta_1 + 1)/(\delta_0 + 1)]$.

4. SIMULATION STUDY

A simulation study was conducted to investigate and compare the performance of the empirical size and power between the delta method and

score test for evaluation of the noninferiority hypothesis. Fortran 90 and IMSL STAT/LIBRARY Fortran subroutines were used in the simulation study. To investigate the impact of sample size, the magnitude of occurrence rate in the control treatment or unexposed members, and the difference in the conditional odds ratios under the null and alternative hypotheses, we chose $N = 50$, 100, 200, and 500; p_{21} from 0.01 to 0.30 by an increment of 0.05, and $\delta_0 = 0.8$, 0.9 with corresponding $\delta_1 = 1.25$, 1.11. For each of the 36 combinations, 10,000 random samples were independently generated from trinomial distribution. If zero counts of b and c were generated, a value of 0.5 is added to these cell frequencies. For a 5% nominal significance level, a simulation study with 10,000 random samples implies that 95% of empirical size evaluated at the equivalence limits will be within 0.04782 and 0.05218.

Table 3 presents the empirical Type I error for the noninferiority test. When p_{21} is smaller than 0.1 or sample size is either 50 or 100, the empirical sizes of both delta method and score test can fall below 0.04782. Therefore, when the incidence

Table 3 Empirical size (%) by simulation

		Margin = 0.8		Margin = 0.9	
p_{21}	Sample size	Score test	Delta method	Score test	Delta method
0.01	50	0.42	0.01	0.04	0.00
	100	1.79	0.09	0.60	0.12
	200	3.82	0.96	2.72	0.91
	500	4.95	3.71	4.62	3.31
0.05	50	4.61	1.87	3.42	1.35
	100	4.95	3.80	5.34	3.95
	200	5.00	5.00	4.87	4.68
	500	5.19	4.42	5.13	4.86
0.10	50	5.36	4.18	5.19	3.82
	100	4.63	4.29	4.84	4.71
	200	5.66	5.12	5.42	5.20
	500	5.07	5.07	4.51	4.44
0.15	50	5.13	4.73	5.06	4.72
	100	5.05	5.00	4.99	4.71
	200	5.20	4.61	5.05	4.98
	500	5.40	5.32	4.98	4.89
0.20	50	4.96	4.36	5.01	4.90
	100	5.59	5.09	5.06	4.80
	200	5.47	5.06	5.35	5.31
	500	5.17	4.16	4.42	4.42
0.25	50	4.32	4.21	4.89	4.44
	100	5.77	4.77	5.10	4.91
	200	4.84	4.84	4.82	4.67
	500	5.21	5.21	5.04	5.04
0.30	50	5.29	5.28	5.09	4.78
	100	5.00	4.57	5.08	5.01
	200	4.91	4.79	4.72	4.72
	500	5.03	5.02	5.34	5.34

rate of the control treatment or unexposed member of the matched pairs is below 0.1 or sample size is small, both methods tend to be conservative. In addition, under these combinations, the empirical size of both methods is smaller with more stringent limit of $\delta_0 = 0.9$ than that of a more liberal limit $\delta_0 = 0.8$. For example, when $\delta_0 = 0.9$, the empirical size of the delta method can be as small as 0 for $N = 50$ and $p_{21} = 0.01$. Therefore, the delta method is very conservative when $p_{21} \leq 0.05$ and sample size is fewer than 100. On the other hand, when $p_{21} \geq 0.10$, almost all empirical sizes of both methods are between 0.04782 and 0.05218. Therefore, both methods can adequately control the size under the 5% nominal level when sample size is at least 100 and $p_{21} \geq 0.10$, although the delta method is conservative when sample size is small and incidence rate of the control treatment is low.

Empirical powers obtained from the simulation study are presented in Table 4, which shows the empirical powers of both methods are increasing function of sample size, the incidence rate p_{21}, and difference between δ_1 and δ_0. When p_{21} is less than 0.10 and sample size is fewer than 100, the score test provides a better

Table 4 Empirical power (%) by simulation for $\delta_1 = 1.25$

		$\delta_0 = 0.8$ and $\delta_1 = 1.25$		$\delta_0 = 0.9$ and $\delta_1 = 1.11$	
p_{21}	Sample size	Score test	Delta method	Score test	Delta method
0.01	50	1.60	0.09	0.21	0.00
	100	5.04	0.63	1.20	0.26
	200	10.15	4.66	5.21	2.09
	500	18.12	16.40	9.99	8.26
0.05	50	12.36	7.34	6.75	3.53
	100	18.54	16.57	9.96	8.07
	200	26.76	26.03	11.56	11.02
	500	50.26	48.26	19.51	18.98
0.10	50	18.81	17.04	9.98	8.16
	100	26.38	25.73	12.69	12.14
	200	45.86	41.98	17.05	16.45
	500	76.58	76.41	28.66	28.55
0.15	50	22.99	21.42	10.54	10.16
	100	38.24	37.44	14.96	14.20
	200	58.77	56.35	21.90	21.50
	500	89.29	89.20	36.29	36.24
0.20	50	26.73	26.28	11.91	11.50
	100	45.80	42.39	17.32	16.69
	200	68.38	68.37	25.75	25.51
	500	95.22	95.22	47.31	47.31
0.25	50	37.72	32.70	13.41	12.45
	100	50.49	48.78	19.48	19.05
	200	76.40	76.30	28.11	28.04
	500	98.26	98.24	53.53	53.53
0.30	50	38.16	37.77	14.94	14.32
	100	60.38	57.04	20.97	20.53
	200	83.07	82.87	33.36	33.30
	500	99.41	99.39	59.55	59.55

Table 5 Sample size, empirical size (%), and power (%)

				Score test		Delta method	
δ_0	δ_1	p_{21}	Sample size	Size	Power	Size	Power
0.8	1.25	0.01	5586	4.93	79.46	4.86	79.28
		0.05	1116	4.54	80.74	4.48	80.50
		0.10	557	4.82	80.61	4.74	80.40
		0.15	371	4.66	80.54	4.58	80.39
		0.20	278	5.13	80.68	5.05	80.59
		0.25	222	5.05	80.12	5.00	80.10
		0.30	185	4.94	80.00	4.83	79.96
0.9	1.11	0.01	26455	4.51	80.22	4.50	80.18
		0.05	5290	5.29	80.44	5.28	80.38
		0.10	2644	5.10	80.64	5.09	80.61
		0.15	1762	5.12	79.88	5.12	79.85
		0.20	1321	5.20	79.40	5.20	79.36
		0.25	1057	4.74	79.92	4.73	79.89
		0.30	880	4.85	79.71	4.83	79.69

empirical power than the delta method. However, when p_{21} is greater than 0.05 and sample size is at least 100, although the power of the score test is still numerically larger than that of the delta method, the powers of the two methods are in fact indistinguishable. On the other hand, under the same combination of sample size and p_{21}, the power for detecting a smaller difference between $\delta_1 = 1.11$ and $\delta_0 = 0.9$ is considerably less than a wider difference between $\delta_1 = 1.25$ and $\delta_0 = 0.8$.

Table 5 provides the required sample sizes calculated by Eq. (10) to achieve 80% power at the 5% nominal level for different combinations of p_{21}, δ_1, and δ_0.

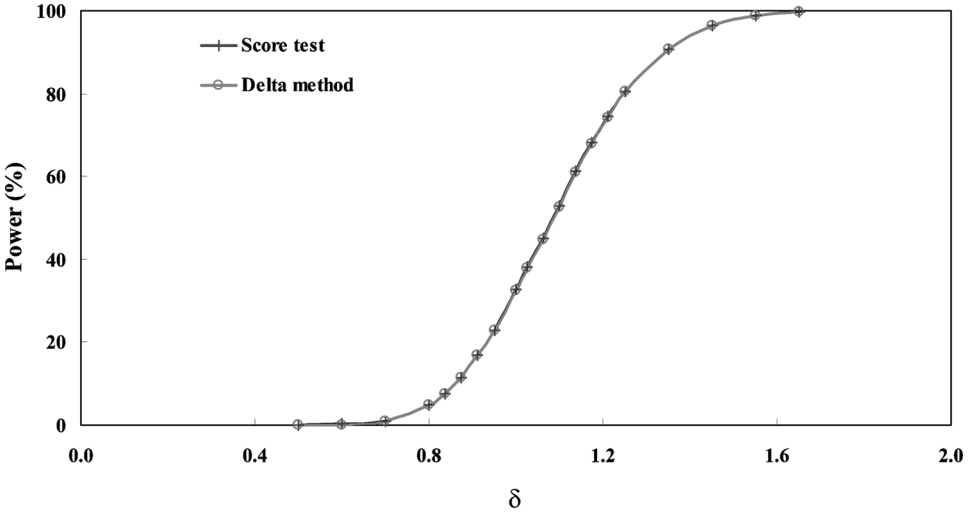

Figure 1 Empirical power curve of the score test and delta method for $N = 557$ and $p_{21} = 0.1$.

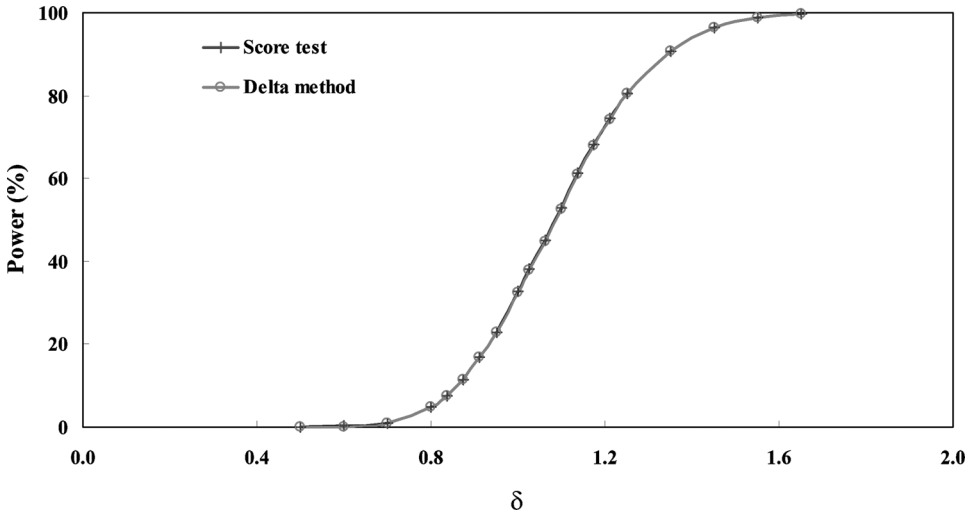

Figure 2 Empirical power curve of the score test and delta method for $N = 222$ and $p_{21} = 0.25$.

In addition, the empirical size and power obtained by simulation with 10,000 random samples were also given in Table 5. As shown in Table 5, sample size is a decreasing function of p_{21} and the difference between δ_1 and δ_0. In addition, there is a huge variation in sample sizes. The sample size ranges from 185 for the combination of $p_{21} = 0.3$, $\delta_1 = 1.25$, and $\delta_0 = 0.80$ to 26455 for the combination of $p_{21} = 0.01$, $\delta_1 = 1.11$, and $\delta_0 = 0.9$. All empirical sizes of both methods for all sample size are between 0.04783 and 0.05218, which indicates both delta method and score test can adequately control the 5% nominal level. For a simulation with 10,000 replicates, the power provided by the sample size is significantly smaller than 80% if the observed empirical power is less than 79.22%. All empirical powers for all sample sizes by both methods are above 79.22%. In addition, the empirical powers of all sample sizes by both methods are also under 81.00%. These results show that Eq. (10) can adequately provide a sample size with sufficient but no excessive power. The power curves for $N = 557$ and $p_{21} = 0.1$ with margin $\delta_0 = 0.80$ and for $N = 222$ and $p_{21} = 0.25$ with margin $\delta_0 = 0.80$ are given in Figs. 1 and 2 respectively. From Figs. 1 and 2, the power functions of both methods are increasing function of the conditional odds ratio. The size at margin $\delta_0 = 0.80$ is about 0.05. Because the sample size of 222 and 557 can provide 80% for their respective combinations, the power curves of the score test and delta methods are overlapped on top of each other and indistinguishable as shown in Figs. 1 and 2.

5. NUMERICAL EXAMPLE

Lachin (2000) presents the results of a retrospective matched case-control study, originally given in Mack et al. (1976), to examine the association between the use of conjugate estrogens and occurrence of endometrial cancer. The data are

Table 6 The 2×2 table of endometrial cancer and use of conjugated estrogen

	No endometrial cancer		
Endometrial cancer	Conjugate estrogen (%)	No conjugate estrogen (%)	Total (%)
Conjugate estrogen	18 (28.57)	33 (52.38)	51 (80.95)
No conjugate estrogen	6 (9.52)	6 (9.52)	12 (19.04)
Total	24 (38.10)	39 (61.90)	63 (100.0)

Source: Lachin (2000).

given in Table 6. The case and control for each of 63 pairs were matched within 1 year of their ages. Suppose that in addition to investigating the existence of association between the use of conjugated estrogen and occurrence of endometrial cancer, we also want to verify that the odds of endometrial cancer among the users of conjugated estrogen is at least twice as high as that among the nonusers. It follows that δ_0 in hypothesis (1) is 2. Readers can verify that the point estimate of the conditional odds ratio is 5.5 and its logarithm is 1.7047 with an estimated asymptotic variance of 0.1970. Therefore, Z_D is 2.2793, which is greater that $z_{0.05} = 1.645$. In addition, the lower limit of the 90% confidence interval for δ is 2.65, which is greater than the equivalence limit of 2. As a result, at the 5% significance level, both test statistic Z_D and the confidence limit approach of the delta method reach the same conclusion that the odds of endometrial cancer among the users of conjugated estrogen is at least twice as high as that of the nonusers.

On the other hand, the observed statistic of the score test Z_D is 2.378, which is also greater than 1.645. Therefore, both the score test and the delta method provide the same conclusion at the 5% significance level. However, if δ_0 in hypothesis (1) is 3, readers can verify that at the 5% significance level, score test and delta method cannot conclude that the odds of endometrial cancer among the users of conjugated estrogen is at least 3 times as high as that of the nonusers.

Suppose that the occurrence rate of endometrial cancer among the nonusers of conjugated estrogen is 10% and we want to know the required sample size for 80% power to test the hypothesis that the odds of endometrial cancer among the users of conjugated estrogen is at least twice as high as that of the nonusers at the 5% nominal level. In addition, δ_1 in the alternative hypothesis is assumed to be 5.5. It follows that $p_{21} = 0.1$, $\delta_0 = 2$, $\delta_1 = 5.5$, $z_{0.05} = 1.645$, and $z_{0.20} = 0.842$. Substitution of these figures into Eq. (10) yields a sample size of 57. If δ_1 decrease to 5, the required sample size increases to 73.

6. DISCUSSION

Current methods for risk assessment under a matched-pair design are to investigate the relationship of outcome with a certain risk factor through the following hypothesis for association based on the conditional odds ratio:

$$H_0 : \delta = 1 \quad \text{vs.} \quad H_a : \delta \neq 1. \tag{11}$$

However, the above hypothesis can only perform a qualitative evaluation of whether there is association between the risk of interest and exposure to the potential risk factor. Therefore, results for testing hypothesis (11) cannot provide a quantitative assessment of the magnitude of risk. Furthermore, with respect to hypothesis (11), a large sample size can always declare statistical significance for any minute difference without any clinical or practical meaning. On the other hand, failure to reject the null hypothesis (11) does not prove that there is no risk associated with the factor under investigation. Therefore, the concept of noninferiority testing currently used in prospective clinical trials can be extended to a quantitative evaluation of the magnitude of the risk. If the objective of the study is to examine whether the risk of the exposed subjects exceeds a prespecified level, then hypothesis (1) is the appropriate hypothesis. On the other hand, if the goal of the research is to verify that the factor under study possesses no risk, then one needs to prove that the conditional odds ratio does not exceed a predetermined margin by hypothesis (2).

Determination of noninferiority or equivalence limits is the most critical and yet the most difficult task for any noninferiority or equivalence tests. For example, for assessing whether the risk of exposure to a certain risk factor exceeds a prespecified level, if the affected population is large and the outcome can be catastrophic, the noninferiority margin should be very tight. However, noninferiority or equivalence limits should be determined jointly by clinicians, epidemiologists, and statisticians. Furthermore, determination of noninferiority or equivalence limits should consider many factors, such as the usage of treatment, the efficacy of the standard treatment, the background risk of the unexposed control, the seriousness of the outcome, the size of the targeted population, and feasibility of the required sample size.

For the matched-pair design, methods have been proposed to test equivalence and noninferiority based on correlated binary endpoints. These methods use either difference in proportions or relative risk, which can only be applied to prospective or cross-sectional studies and not to retrospective studies. To overcome this issue, we suggest the use of the conditional odds ratio and propose the delta method and a score test for evaluation of noninferiority hypothesis based on the conditional odds ratio. Our proposed methods can be applied to both prospective and retrospective studies. In addition, we also present a formula for determination of sample size. Simulation shows that both delta method and the score test perform satisfactorily. However, the methods we proposed are the large-sample methods. Therefore, when sample size is small or the incidence rate of unexposed subjects is very low, the proposed delta method and score test will tend to be conservative. The exact method for noninferiority testing based on the conditional odds ratio for the small sample size or a lower incidence requires further research.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their careful, thoughtful, and thorough review and comments, which greatly improve the content and presentation of our work. This work is partially supported by the Taiwan National Science Grants: NSC 92-2118-M-006-001 and NSC 93-2118-M-006-002.

REFERENCES

Hsueh, H. M., Liu, J. P., Chen, J. J. (2001). Unconditional exact tests for equivalence or noninferiority for paired binary endpoints. *Biometrics* 57:478–483.

Lachin, J. M. (2000). *Biostatistical Methods: The Assessment of Relative Risks*. New York: Wiley, pp. 175–190.

Liu, J. P., Hsueh, H. M., Hsieh, E., Chen, J. J. (2002). Tests for equivalence or noninferiority for paired binary data. *Stat. Med.* 21:231–245.

Mack, T. M., Pike, M. C., Henderson, B. E., Pfeffer, R. I., Gerkins, V. R., Arthus, B. S., Brown, S. E. (1976). Estrogens and endometrial cancer in a retirement community. *N. Eng. J. Med.* 294:1262–1267.

Serfling, R. J. (1980). *Approximation Theorems of Mathematics Statistics*. New York: Wiley.

Tang, N. S., Tang, M. L., Chan, I. S. F. (2003). On tests of equivalence via non-unity relative risk for matched-pair design. *Stat. Med.* 22:1217–1233.