STATISTICS IN MEDICINE
Statist. Med. **17**, 891–908 (1998)

EQUIVALENCE TEST AND CONFIDENCE INTERVAL FOR THE DIFFERENCE IN PROPORTIONS FOR THE PAIRED-SAMPLE DESIGN

TOSHIRO TANGO*

Division of Theoretical Epidemiology, The Institute of Public Health, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

SUMMARY

This paper considers a model for the difference of two proportions in a paired or matched design of clinical trials, case-control studies and also sensitivity comparison studies of two laboratory tests. This model includes a parameter indicating both interpatient variability of response probabilities and their correlation. Under the proposed model, we derive a one-sided test for equivalence based upon the efficient score. Equivalence is defined here as not more than 100Δ per cent inferior. McNemar's test for significance is shown to be a special case of the proposed test. Further, a score-based confidence interval for the difference of two proportions is derived. One of the features of these methods is applicability to the 2×2 table with off-diagonal zero cells; all the McNemar type tests and confidence intervals published so far cannot apply to such data. A Monte Carlo simulation study shows that the proposed test has empirical significance levels closer to the nominal α-level than the other tests recently proposed and further that the proposed confidence interval has better empirical coverage probability than those of the four published methods. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

Recently, two groups of authors have proposed methods for testing equivalence in proportions arising from a paired-sample study design. Lu and Bean[1] considered a sensitivity comparison of two medical diagnostic laboratory tests. Morikawa and Yanagawa[2] discussed a comparison of two treatments in clinical trials. They have derived test statistics similar to McNemar's test statistic using Wald-type asymptotic standard errors.

Equivalence[3] is usually defined as not more than 100Δ per cent inferior, where $\Delta(>0)$ is a prespecified acceptable difference between two proportions. A naive question about the approaches above is whether a McNemar's type of test statistic comparing off-diagonal cells is the only valid approach for inference about equivalence. Motivated by this question, in this paper, we first formulate a reasonable model representing the structure of this kind of comparison study. This model includes parameters indicating interpatient variability of response probabilities and their correlation. Under the proposed model, we derive a test for equivalence of two proportions and also a confidence interval for their difference based upon the efficient score. Despite my initial question, the derived method is also shown to be of McNemar's type but to have better small sample properties. Further, one feature of this method is that we can make

* Correspondence to: Toshiro Tango, Division of Theoretical Epidemiology, The Institute of Public Health, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan. e-mail: tango@iph.go.jp

CCC 0277–6715/98/080891–18$17.50

Received January 1996
Accepted July 1997

reasonable inferences for data in a 2×2 table having zero frequencies in the off-diagonal cells; all methods published so far do not apply to such data.

Monte Carlo simulation studies are conducted to compare empirical significance levels of the proposed test with the two tests above and also to compare empirical coverage probabilities of the proposed method for confidence interval estimation with those of published methods, both unconditional and conditional. Finally, we illustrate our methods using data from a cross-over trial of soft contact lenses and an epidemiological study of sleeping difficulties in marijuana users.

Score-based methods for hypothesis testing and interval estimation have been widely proposed and used in many fields; Cox and Hinkley[4] provide theoretical justification for them. Vollset[5] reported favourable properties for the Wilson[6] score method for setting a confidence interval for a single proportion. Yanagawa *et al.*[7] have recently proposed Mantel–Haenszel type tests for testing equivalence for comparative parallel design clinical trials, based on the efficient score.

2. MODEL

Consider the comparison of a new and standard treatment (or diagnostic test) independently performed on the same patient (or matched-pairs of patients) and suppose we have n patients (or pairs). Further we assume that the probability of response to a treatment is a function of the individual patient's unobserved characteristics. Let $\theta_k (k = 1, \ldots, n)$ denote the kth patient's (matched-patients's common) characteristics which might be *multivariate*, though in this paper, we assume them to be univariate for simplicity. Further let N_k and S_k denote the dichotomous response random variable having values 1 (response) and 2 (non-response) of the new treatment and standard, respectively. Then, we have the following conditional probabilities:

$$\Pr\{N_k = 1 \mid \theta_k\} = p_{\mathrm{N}}(\theta_k) \tag{1}$$

$$\Pr\{S_k = 1 \mid \theta_k\} = p_{\mathrm{S}}(\theta_k). \tag{2}$$

In a matched case-control study, these two random variables N_k and S_k denote the incidence probability of some 'event' under study in the case and control groups, respectively. Given θ, N_k and S_k are mutually independent, then the 2×2 matrix $Q(\theta) = (q_{ij}(\theta))$, of response probabilities is shown as

$$Q(\theta) = \begin{pmatrix} p_{\mathrm{N}}(\theta)p_{\mathrm{S}}(\theta) & p_{\mathrm{N}}(\theta)(1 - p_{\mathrm{S}}(\theta)) \\ (1 - p_{\mathrm{N}}(\theta))p_{\mathrm{S}}(\theta) & (1 - p_{\mathrm{N}}(\theta))(1 - p_{\mathrm{S}}(\theta)) \end{pmatrix} \tag{3}$$

where

$$q_{ij}(\theta) = \Pr\{N = i, S = j \mid \theta\} = \Pr\{N = i \mid \theta\} \Pr\{S = j \mid \theta\}. \tag{4}$$

In general, we do not know the population distribution for θ. However, it can be assumed that θ_k's are mutually independent and identically distributed with unknown distribution

function F. Thus, we have

$$E_\theta\{p_N(\theta)\} = \pi_N \tag{5}$$

$$E_\theta\{p_S(\theta)\} = \pi_S \tag{6}$$

$$V_\theta\{p_N(\theta)\} = \sigma_N^2 \tag{7}$$

$$V_\theta\{p_S(\theta)\} = \sigma_S^2 \tag{8}$$

and

$$E_\theta\{p_S(\theta)p_N(\theta)\} = \pi_S\pi_N + \rho\sigma_S\sigma_N \tag{9}$$

where ρ is the correlation coefficient over patients. Therefore, the expected 2×2 response matrix $Q = (q_{ij}) = (E_\theta\{q_{ij}(\theta)\})$ is given as

$$Q = \begin{pmatrix} \pi_N\pi_S + \rho\sigma_N\sigma_S & \pi_N(1-\pi_S) - \rho\sigma_N\sigma_S \\ (1-\pi_N)\pi_S - \rho\sigma_N\sigma_S & (1-\pi_N)(1-\pi_S) + \rho\sigma_N\sigma_S \end{pmatrix}. \tag{10}$$

If the two treatments are truely identical, then, of course, $\rho = 1$ and the 2×2 response matrix will be

$$Q = \begin{pmatrix} \pi^2 + \sigma^2 & \pi(1-\pi) - \sigma^2 \\ \pi(1-\pi) - \sigma^2 & (1-\pi)^2 + \sigma^2 \end{pmatrix}. \tag{11}$$

Now consider the two extreme cases. When $\sigma = 0$, that is, the response probability is constant regardless of θ, then the matrix will be

$$Q = \begin{pmatrix} \pi^2 & \pi(1-\pi) \\ \pi(1-\pi) & (1-\pi)^2 \end{pmatrix}. \tag{12}$$

When $\sigma^2 = \pi(1-\pi)$, on the other hand, the matrix will be

$$Q = \begin{pmatrix} \pi & 0 \\ 0 & 1-\pi \end{pmatrix}. \tag{13}$$

The latter suggests the deterministic nature of treatment with threshold level θ_0. Namely, $p(\theta) = 1$ for $\theta > \theta_0$ and 0, otherwise. In this special case, $\pi = 1 - F(\theta_0)$.

From my experience with the analysis of clinical laboratory data where we can easily make repeated measurements or diagnostic tests by splitting one *blood sample* into two, it seems to me that well-recognized diagnostic tests might have the latter characteristic, that is, σ^2 can be nearly equal to $\pi(1-\pi)$. Therefore, the equivalence in the sensitivities between two diagnostic tests can be reduced to the equivalence problem for the π's only. It should be noted here that, as Lu and Bean[1] have already pointed out, equivalence for diagnostic tests should focus on both sensitivity and specificity, but the same principles can apply to examine the latter comparison using a different study population.

As to the response probability to treatments in clinical trials, on the other hand, variability σ^2 will probably lie between 0 and $\pi(1-\pi)$ depending on both the treatment under study and the patients entered into the trial. Therefore, if we could apply the same treatment to the same patient (or matched-pairs of patients), we could estimate not only π but also σ:

$$\hat{\pi} = \frac{2a + b + c}{2n}, \quad \hat{\sigma}^2 = \frac{4ad - (b+c)^2}{4n^2} \tag{14}$$

where a, b, c and d are observed frequencies defined in (15) in the next section. This implies that in equivalence testing in clinical trials, we have to evaluate not only π but also σ^2. Since σ^2 can be at most $\pi(1-\pi)$, it only characterizes variability in the context of a given value of π. Therefore, the ideal procedure is (i) apply an equivalence test for π and then (ii) under the equivalence of π, conduct an equivalence test for σ^2. However, I admit that it will usually be impractical to repeat the same treatment in the same patient and so we can not obtain such measures of variability in practice. Therefore we are obliged to estimate π only in clinical trials.

Nevertheless, the proposed modelling is useful to consider the structure of the problem under study and also can give some insight. For example, Lu and Bean[1] suggested that the primary evidence of equivalence in sensitivity is the probability of discordance and thus the smaller the probability the more likely the two sensitivities will be equivalent. This view is based on the property that the overall degree of discordance $q_{12}+q_{21}$ is the upper bound for the degree of marginal heterogeneity $q_{12}-q_{21}=\pi_{\mathrm{N}}-\pi_{\mathrm{S}}$. However, this *view* is not always correct. As shown above, whether the probability of discordance is small or large depends on the *unknown* size of variability σ^2.

3. SCORE TEST FOR EQUIVALENCE TESTING

Consider a random sample from the multinomial distribution defined in (10), then we have

$$Q_{\mathrm{DATA}}=\frac{1}{n}\begin{pmatrix} a & b \\ c & d \end{pmatrix}. \tag{15}$$

Then, by letting

$$\phi=\rho\sigma_{\mathrm{N}}\sigma_{\mathrm{S}}$$

the log-likelihood for this sample can be written

$$L(\pi_{\mathrm{N}},\pi_{\mathrm{S}},\phi)=a\log(q_{11})+b\log(q_{12})+c\log(q_{21})+d\log(q_{22})+\text{constant}. \tag{16}$$

An equivalence hypothesis will be formulated as

$$H_0:\pi_{\mathrm{N}}=\pi_{\mathrm{S}}-\Delta,\quad H_1:\pi_{\mathrm{N}}>\pi_{\mathrm{S}}-\Delta \tag{17}$$

where $\Delta(>0)$ is a prespecified acceptable difference in two proportions. Let

$$\beta=\pi_{\mathrm{N}}-(\pi_{\mathrm{S}}-\Delta) \tag{18}$$

then the above hypothesis is equivalent to the following:

$$H_0:\beta=0,\quad H_1:\beta>0 \tag{19}$$

where the expected 2×2 response matrix $Q=(q_{ij})$ is given as

$$Q=\begin{pmatrix} (\beta+\pi_{\mathrm{S}}-\Delta)\pi_{\mathrm{S}}+\phi & (\beta+\pi_{\mathrm{S}}-\Delta)(1-\pi_{\mathrm{S}})-\phi \\ (1-\beta-\pi_{\mathrm{S}}+\Delta)\pi_{\mathrm{S}}-\phi & (1-\beta-\pi_{\mathrm{S}}+\Delta)(1-\pi_{\mathrm{S}})+\phi \end{pmatrix}. \tag{20}$$

Therefore, the score test can be applied to the log-likelihood function $L(\beta, \pi_S, \phi)$ and its test statistic is given by

$$Z = \left[\left. \frac{\partial L}{\partial \beta} \right|_{\pi_S = \hat{\pi}_S, \phi = \hat{\phi}, \beta = 0} \right] \sqrt{\{(\hat{I}^{-1})_{33}|_{\pi_S = \hat{\pi}_S, \phi = \hat{\phi}, \beta = 0}\}}$$
$$= \left[\frac{a\hat{\pi}_S}{\hat{q}_{11}} + \frac{b(1-\hat{\pi}_S)}{\hat{q}_{12}} - \frac{c\hat{\pi}_S}{\hat{q}_{21}} - \frac{d(1-\hat{\pi}_S)}{\hat{q}_{22}} \right] \sqrt{\left(\frac{\hat{q}_{12} + \hat{q}_{21} - \Delta^2}{n} \right)} \tag{21}$$

where $(\hat{\pi}_S, \hat{\phi})$ is the maximum likelihood estimator under the null hypothesis $\beta = 0$, which is usually the unique solution to the following equations:

$$\frac{\partial L}{\partial \pi_S} = \frac{\partial L}{\partial \phi} = 0. \tag{22}$$

Further $(\hat{I}^{-1})_{33}$ indicates the (3, 3)th element of the inverse Fisher information matrix evaluated at maximum likelihood estimators and is algebraically simplified to $(\hat{q}_{12} + \hat{q}_{21} - \Delta^2)/n$. The details of the Fisher information matrix are given in Appendix I. The test statistic Z is known to have asymptotically a standard normal distribution under H_0. In terms of π_S and ϕ, equations (22) can be solved iteratively by the scoring method as follows:

$$\begin{bmatrix} \hat{\pi}_S \\ \hat{\phi} \end{bmatrix}_k = \begin{bmatrix} \hat{\pi}_S \\ \hat{\phi} \end{bmatrix}_{k-1} + \begin{bmatrix} \hat{I}_{11} & \hat{I}_{12} \\ \hat{I}_{21} & \hat{I}_{22} \end{bmatrix}_{k-1}^{-1} \begin{bmatrix} \dfrac{\hat{\partial} L}{\partial \pi_S} \\ \dfrac{\hat{\partial} L}{\partial \phi} \end{bmatrix}_{k-1} \tag{23}$$

where the possible ranges of parameters are restricted by the conditions that $0 \leqslant q_{ij} < 1$, that is, the range for the π_S is

$$0 \leqslant \Delta \leqslant \pi_S < 1$$

and the range for ϕ is

$$-\pi_S(\pi_S - \Delta) \leqslant \phi \leqslant (1 - \pi_S)(\pi_S - \Delta) \quad \text{for} \quad \Delta \leqslant \pi_S \leqslant (1 + \Delta)/2$$
$$-(1 - \pi_S)(1 - \pi_S + \Delta) \leqslant \phi \leqslant (1 - \pi_S)(\pi_S - \Delta) \quad \text{for} \quad (1 + \Delta)/2 \leqslant \pi_S < 1.$$

However, it is easily understood that we do not have to obtain these estimators directly for testing purposes and we need only a maximum likelihood estimator $\hat{q}_{21}$. In fact, as described in Appendix II, the test statistic Z can be simply expressed as

$$Z(b, c; n, \Delta) = Z = \frac{b - c + n\Delta}{\sqrt{\{n(2\hat{q}_{21} - \Delta(\Delta + 1))\}}} \tag{24}$$

where the estimator $\hat{q}_{21}$ is the larger root of the quadratic equation $Ax^2 + Bx + C = 0$, that is

$$\hat{q}_{21} = \frac{\sqrt{(B^2 - 4AC)} - B}{2A} \tag{25}$$

where

$$A = 2n, \quad B = -b - c - (2n - b + c)\Delta, \quad C = c\Delta(\Delta + 1) \tag{26}$$

and $\hat{q}_{12} = \hat{q}_{21} - \Delta$, $\hat{q}_{11} = \dfrac{a}{a+d}(1 - \hat{q}_{12} - \hat{q}_{21})$ and $\hat{q}_{22} = \dfrac{d}{a+d}(1 - \hat{q}_{12} - \hat{q}_{21})$.

It should be noted that if an observed sample has one or more cells with zero frequencies then the log-likelihood (16) tends to take the maximum value on the *boundary* of the parameter space and thus equations (22) are no longer a necessary condition. However, as is shown in Appendix II, the estimator $\hat{q}_{21}$ defined in (25) still provides the correct maximum likelihood estimator, which is summarized as follows. If

$$b = 0 \quad \text{and} \quad c < \frac{2n\Delta}{1+\Delta} \, (\Delta > 0) \tag{27}$$

then the maximum likelihood occurs at $\hat{Q}$ with $\hat{q}_{12} = 0$ and $\hat{q}_{21} = \Delta$ which is a boundary point of the parameter space and equations (22) are no longer satisfied. However, $\hat{q}_{21}$ defined in (25) also equals to Δ.

It is well known that if the true parameter value is on the boundary of the parameter space then the regular asymptotic property of the likelihood ratio test breaks down whilst the score test may be applied as usual (for example, see Chant[8] and Self and Liang[9]). However, it seems to be not well known whether the score-based method still works if the maximum likelihood estimator is on the boundary of the parameter space. Therefore, as a way of checking the validity of Z defined by (24) in this boundary situation, we shall calculate $Z(0, c; n, \Delta)$ and compare it with the predicted value. First, we can easily prove the following inequality:

$$|Z(b, c; n, \Delta)| > |Z(b+1, c+1; n, \Delta)| \quad \text{for} \quad b > 0 \tag{28}$$

where $Z(b, c; n, \Delta)$ with $b > 0$ is free of the boundary problem. The prediction considered here is an extrapolation of this trend, that is, to *predict* $Z(0, c; n, \Delta)$ by applying a low-order polynomial regression to a series of Z's

$$\{Z(k, k+c; n, \Delta), k = 1, 2, \ldots, K\}$$

for some K. We shall use a cubic polynomial regression and $K = 10$. The precision for this sort of prediction scheme is examined by predicting $Z(b, c; n, \Delta)$ for $b > 0$ that is free from the boundary problem, and, as a result, cubic polynomials with $K = 10$ have been confirmed good enough. The results are shown in Table I for data with $n = 30$, 50 and 80 and $c = 0$, 1 and 2, indicating that the predicted Z's are surprisingly consistent with Z's. In general, as n become large, the predicted Z's are shown to be very close to Z's. Therefore, we conclude that the proposed test works even if the maximum likelihood estimator lies on the boundary of the parameter space.

As special cases, consider the following:

1. Zero off-diagonal cells, $b = c = 0$. In this case, $\hat{q}_{21} = \Delta$ and thus the test statistic Z reduces simply to

$$Z(0, 0; n, \Delta) = \sqrt{\left(\frac{n\Delta}{1-\Delta} \right)}. \tag{29}$$

   That is to say, we can declare 'clinically equivalent' if

$$n > n_{\min} = \frac{1-\Delta}{\Delta} Z_{\alpha}^2. \tag{30}$$

   For example, when $\Delta = 0{\cdot}1$ and $\alpha = 0{\cdot}05$, we have $n_{\min} = 9 \times 1{\cdot}645^2 = 24{\cdot}35$.
2. $\Delta = 0$ and $b + c > 0$. In this case, $\hat{q}_{21} = (b+c)/2n$ and thus this test coincides with the well-known McNemar test for significance testing, namely,

$$Z(b, c; n, \Delta = 0) = (b-c)/\sqrt{(b+c)}. \tag{31}$$

Table I. Comparison of the proposed Z and λ and their predicted values via cubic polynomial regression models for the case $b = 0$ and $\Delta = 0{\cdot}1$, where $\hat{q}_{12} = 0{\cdot}0$ and $\hat{q}_{21} = 0{\cdot}1$, maximum likelihood estimates on a boundary point of the parameter space

n	c	Test statistic ($\Delta = 0{\cdot}1$)		90% confidence interval	
		Z	predicted Z	λ_{low}, λ_{up}	predicted λ_{low}, λ_{up}
30	0	1·83	1·81	(− 0·083, 0·083)	(− 0·086, 0·086)
	1	1·22	1·18	(− 0·136, 0·052)	(− 0·138, −)
	2	0·61	0·57	(− 0·183, 0·022)	(− 0·186, −)
50	0	2·36	2·37	(− 0·051, 0·051)	(− 0·054, 0·054)
	1	1·89	1·88	(− 0·085, 0·032)	(− 0·086, −)
	2	1·41	1·40	(− 0·114, 0·013)	(− 0·114, −)
80	0	2·98	2·99	(− 0·033, 0·033)	(− 0·034, −0·034)
	1	2·61	2·61	(− 0·054, 0·021)	(− 0·054, −)
	2	2·24	2·24	(− 0·073, 0·009)	(− 0·073, −)

When $c > 0$, predicted λ_{up} is not shown here for $c > 0$ since $\hat{\lambda}_{up}$ is free from the boundary problem

4. CONFIDENCE INTERVAL

Testing clinical equivalence with an acceptable difference Δ at one-sided significance level α is equivalent to judging whether the lower limit of the $1 - 2\alpha$ level confidence interval is greater than $-\Delta$. The score-based approximate confidence limits for the difference of two proportions

$$\lambda = \pi_N - \pi_S = q_{12} - q_{21} \tag{32}$$

are the two solutions to the equation

$$Z(b, c; n, -\lambda) = \pm Z_\alpha \tag{33}$$

where the plus and minus signs indicate the lower $\lambda_{low}(b, c; n, 1 - 2\alpha)$ and the upper limit $\lambda_{up}(b, c; n, 1 - 2\alpha)$, respectively, and Z_α is the upper α percentile of the standard Normal distribution. These two limits can be easily found by the secant method (see, for example, Gart and Nam[11]).

In Appendix III, the relationship between $\hat{q}_{21}$ and $\hat{\lambda}$ is described to show a solution in the boundary situation. The results are summarized as follows:

1. $b = 0$ and $c > 0$: $\hat{Q}$ that attains a lower limit lies on the boundary with $\hat{q}_{12} = 0$ of the parameter space. $\hat{Q}$ that attains an upper limit lies on the boundary with $\hat{q}_{12} = 0$ if $Z_\alpha^2 \leqslant c(n - c)/2n$ and on an interior point, otherwise.
2. $b > 0$ and $c = 0$: $\hat{Q}$ that attains an upper limit lies on the boundary with $\hat{q}_{21} = 0$. $\hat{Q}$ that attains a lower limit lies on the boundary with $\hat{q}_{21} = 0$ if $Z_\alpha^2 \leqslant b(n - b)/2n$ and on an interior point, otherwise.
3. $b = c = 0$: Both $\hat{Q}$ that attains a lower limit and $\hat{Q}$ that attains an upper limit lie on the boundary:

Using a similar method as the prediction for Z's, we shall examine the validity of the confidence interval based on a boundary value of $\hat{Q}$ by applying a cubic polynomial regression to predict one or both limits. For example, we predicted the limits of the 90 per cent confidence interval, for

the data with $b = 0$, $c = 0, 1, 2$ and $n = 30, 50, 80$ using a series:

$$\{\hat{\lambda}_{\text{low}}(k, c + k; n, 0{\cdot}90), k = 1, 2, \ldots, 10\}$$

for the lower limit, and

$$\{\hat{\lambda}_{\text{up}}(k, c + k; n, 0{\cdot}90), k = 1, 2, \ldots, 10\}$$

for the upper limit. The results are also shown in Table I, which also indicates very good consistency. It should be noted that when $b = 0$ and $c = 1, 2$, each of the $\hat{Q}$'s attaining an upper limit is an interior point of the parameter space and thus the corresponding predicted limit is unnecessary and not shown there.

Here also, let us consider the special case of zero off-diagonal cells, $b = c = 0$. From (29), we easily have

$$[\hat{\lambda}_{\text{low}}(0, 0; n, 1 - 2\alpha), \hat{\lambda}_{\text{up}}(0, 0; n, 1 - 2\alpha)] = \left[-\frac{Z_\alpha^2}{n + Z_\alpha^2}, \frac{Z_\alpha^2}{n + Z_\alpha^2} \right]. \tag{34}$$

It should be noted that when $n = 0$, that is, we have no information, this interval becomes a quite reasonable interval $[-1, 1]$.

5. SIMULATION

5.1. Equivalence Test

So far, three tests have been proposed for this purpose. Lu and Bean[1] proposed an unconditional test based on McNemar's test as

$$Z_{\text{LB}} = \frac{b - c + n\Delta}{\sqrt{(b + c - n\Delta^2)}} \tag{35}$$

for the problem of testing one-sided equivalence in the sensitivities of screening tests. They also proposed a conditional version

$$Z_{\text{CLB}} = \sqrt{\left(\frac{b + c}{bc}\right)\left(\frac{b - c + n\Delta}{2}\right)}. \tag{36}$$

Morikawa and Yanagawa[2] proposed, on the other hand, a test statistic similar to Z_{LB} for the problem of clinical equivalence of drugs:

$$Z_{\text{MY}} = \frac{b - c + n\Delta}{\sqrt{\{b + c - (b - c)^2/n\}}}. \tag{37}$$

This statistic is based on the asymptotic Normality of some function of the multinomial distribution.[10] When $\Delta = 0$, the three tests above are not equivalent and only Z_{LB} is equivalent to McNemar's Z statistic. To investigate the small-sample distribution of the proposed and two others tests, Z_{LB} and Z_{MY}, under the null hypothesis, some Monte Carlo simulations were performed for $\Delta = 0{\cdot}1$ $\pi_{\text{S}} = 0{\cdot}5, 0{\cdot}8$, $\phi = 0, 0{\cdot}1, 0{\cdot}14, 0{\cdot}15$ and $0{\cdot}20$ where applicable. Lu and Bean's conditional test Z_{CLB} was excluded from this comparison since the test cannot apply to data with $b = 0$ or $c = 0$. Samples of multinomial proportions of size $n(= 30, 50, 80)$ whose parameters are defined as the matrix Q in equation (20) with $\beta = 0$, were randomly generated. For

Table II. Empirical significance level of the proposed score test with $\Delta = 0{\cdot}1$ based on 10,000 trials

π_S	n	ϕ	q_{12}	Test statistic		
				Z	Z_{LB}	Z_{MY}
0·5	30	0·0	0·2	4·8	5·7	5·7
		0·1	0·1	5·0	6·3	6·3
		0·15	0·05	4·2	7·7	7·7
		0·20	0·0	4·4	18·7	18·7
	50	0·0	0·2	5·2	5·5	5·5
		0·1	0·1	5·2	6·2	6·2
		0·15	0·05	4·3	6·4	6·4
		0·20	0·0	3·7	11·2	11·2
	80	0·0	0·2	5·0	5·0	5·1
		0·1	0·1	5·2	5·4	5·9
		0·15	0·05	4·4	5·9	6·5
		020	0·0	3·6	8·7	8·7
0·8	30	0·0	0·14	5·1	5·8	5·8
		0·1	0·04	4·2	9·2	9·2
		0·14	0·0	4·1	18·6	18·6
	50	0·0	0·14	5·1	6·2	6·2
		0·1	0·04	4·8	7·2	7·2
		0·14	0·0	4·0	11·9	11·9
	80	0·0	0·14	5·3	5·4	5·6
		0·1	0·04	4·9	6·7	7·4
		0·14	0·0	3·8	8·6	8·6

each simulated sample of proportions each of the three test statistics was computed and compared with the 95th quantile of the standard Normal distribution.

Table II presents empirical significance levels based on 10,000 replications. It appears that the empirical significance levels for the proposed Z are generally closer to the nominal α-level than those for the other two tests. Especially as n becomes small and ϕ becomes large, the empirical significance levels for Z_{LB} and Z_{MY} tend to inflate whereas that of Z tends to decrease gradually. When $q_{12} = 0$, the inflation is clear for Z_{LB} and Z_{MY}.

It should be noted that (i) simulated samples with $b = c = 0$ are replaced by $b = c = 1$ (with n unchanged) for the two test statistics Z_{LB} and Z_{MY}, and (ii) samples with the square root of negative values of Z_{LB} or Z_{MY} are excluded from this comparison since these data cannot be used with these two tests. These samples occurred mainly when $\phi = 0{\cdot}1 \sim 0{\cdot}2$ in which the off-diagonal cell probability q_{12} tends to be small. Therefore the results of this simulation are biased in favour of Z_{LB} and Z_{MY}.

It should also be noted that Lu and Bean[1] gave their formulae for use in hypothesis testing only in passing, as an adjunct to establishing sample size requirements and they are not claiming good properties for these formulae.

Table III shows empirical powers for the case when $\pi_N = \pi_S$ based on 10,000 replications with parameter values similar to those in Table III. The difference in powers between Z and Z_{LB} or Z_{MY} seems to be due to the difference in empirical significance levels.

Table III. Empirical power of the score test with $\Delta = 0{\cdot}1$ for the case $\pi_N = \pi_S$ based on 10,000 trials

π_S	n	ϕ	q_{12}	Test statistic		
				Z	Z_{LB}	Z_{MY}
0·5	30	0·0	0·2	18·4	20·9	20·9
		0·1	0·1	25·1	27·3	27·3
		0·15	0·05	29·6	36·9	36·9
	50	0·0	0·2	26·1	26·8	26·8
		0·1	0·1	34·9	38·8	38·8
		0·15	0·05	44·0	49·8	49·8
	80	0·0	0·2	35·4	35·4	35·9
		0·1	0·1	49·1	49·5	51·4
		0·15	0·05	60·7	64·0	66·1
0·8	30	0·0	0·14	24·5	26·5	26·5
		0·1	0·04	33·7	53·3	53·3
	50	0·0	0·14	33·6	37·5	37·5
		0·1	0·04	58·1	67·4	67·4
	80	0·0	0·14	48·1	48·4	50·0
		0·1	0·04	77·2	83·5	84·2

5.2. Confidence Interval

In this section, we evaluate the coverage probability of the proposed confidence interval by comparing several existing methods, including both unconditional and conditional ones. The most well-known unconditional textbook formula (for example, Altman[12]) is

$$W: \frac{b-c}{n} \pm \frac{Z_\alpha}{n}\sqrt{\left\{b+c-\frac{(b-c)^2}{n}\right\}}. \tag{38}$$

Recently, Vollset[5] has compared thirteen methods for computing binomial confidence intervals in the one-sample problem and recommended (a) continuity corrected score intervals mainly because of simplicity and good coverage probabilities. Further, he recommended (b) exact intervals and (c) Mid-p exact intervals. Therefore, we use these three in our comparison as conditional methods. To obtain conditional confidence intervals, we have only to apply these three methods for the proportion $\hat{p} = b/(b+c)$ where $b+c$ is fixed:

$$\frac{b}{n} - \frac{c}{n} = \frac{b+c}{n}\left\{2\frac{b}{b+c} - 1\right\} = \frac{b+c}{n}\{2\hat{p}-1\}.$$

A $(1-2\alpha)$ level confidence continuity corrected score interval (SCC) for p is given by

$$\mathrm{SCC}: \frac{(b \pm 0{\cdot}5) + \dfrac{Z_\alpha^2}{2} \pm Z_\alpha\sqrt{\left\{(b \pm 0{\cdot}5) - \dfrac{(b \pm 0{\cdot}5)^2}{b+c} + \dfrac{Z_\alpha^2}{4}\right\}}}{b+c+Z_\alpha^2}. \tag{39}$$

Exact confidence intervals for p, called Clopper–Pearson intervals, are calculated from the cumulative binomial distribution and the lower and upper limits are solutions to the following polynomial equations:

$$\mathrm{EX}_{\mathrm{lower}} = \left\{ p : \sum_{i=0}^{b-1} \binom{b+c}{i} p^i (1-p)^{b+c-i} = 1-\alpha \right\} \tag{40}$$

$$\mathrm{EX}_{\mathrm{upper}} = \left\{ p : \sum_{i=0}^{b} \binom{b+c}{i} p^i (1-p)^{b+c-i} = \alpha \right\}. \tag{41}$$

Needless to say, these limits can also be obtained by using percentiles of the F distribution. We used these equations mainly because we can easily use the secant method to find the roots. Mid-p exact limits for p are found from similar equations with half the probability assigned to the observed outcome:

$$\mathrm{MID}_{\mathrm{lower}} = \left\{ p : \frac{1}{2}\binom{b+c}{b} p^b (1-p)^c + \sum_{i=0}^{b-1} \binom{b+c}{i} p^i (1-p)^{b+c-i} = 1-\alpha \right\} \tag{42}$$

and

$$\mathrm{MID}_{\mathrm{upper}} = \left\{ p : \frac{1}{2}\binom{b+c}{b} p^b (1-p)^c + \sum_{i=0}^{b} \binom{b+c}{i} p^i (1-p)^{b+c-i} = \alpha \right\}. \tag{43}$$

For the conditional methods, 0 and 1 are utilized as lower and upper limits when $b = 0$ and $c = 0$, respectively. Using a similar method as the evaluation of equivalence testing procedures, simulated data with $b = c = 0$ are replaced by $b = c = 1$ (with n unchanged) for the conditional methods due to their inapplicability to these data. Therefore the results of this simulation also have biases in favour of the conditional approaches.

Table IV shows empirical coverage probabilities of the above five methods for computing 95 per cent confidence intervals based on 1000 replications under the hypothesis $\pi_{\mathrm{N}} = \pi_{\mathrm{S}} - \Delta$ where $\pi_{\mathrm{S}} = 0{\cdot}8$, $\phi = 0{\cdot}0, 0{\cdot}1$ and $0{\cdot}14$ and $\Delta = 0{\cdot}0$ and $0{\cdot}1$. It is seen that the proposed confidence interval performs very well except when $q_{12} = 0$ where we have *conservative* confidence intervals for small n. The unconditional simple method also performs well with relatively large sample sizes and large discordant probability but not so well for other cases. On the other hand, conditional methods are shown to be no good. Especially when q_{12} is small, their empirical coverage probabilities seem to depend strongly on the parameter values and sample sizes. When $\phi = 0{\cdot}1$ and $\Delta = 0{\cdot}1$, the coverage probabilities extend down to 80 per cent. On the other hand, when $\phi = 0{\cdot}1$ and $\Delta = 0$, they go up to 99·9 per cent. These phenomena are probably the result of *discreteness* as clearly seen in Figures 1–3 of Vollset's paper.[5] Further, when $q_{12} = 0$, the empirical coverage probabilities of conditional methods are around 34–40 per cent, indicating that we cannot recommend conditional methods when the discordant cell frequencies are small. We have also carried out the same simulation study for the case $\pi_{\mathrm{S}} = 0{\cdot}5, 0{\cdot}6, 0{\cdot}7$ and $0{\cdot}9$ but we have not shown the results here since the resultant performances are similar to those shown in Table IV and no new features have been observed.

Table IV. Comparison of five methods for computing 95 per cent confidence intervals via their coverage probabilities under the null hypothesis $\pi_N = \pi_S - \Delta$ where $\pi_S = 0{\cdot}8$, based on 1000 trials for each simulation

ϕ	Δ	q_{12}	n	Unconditional methods		Conditional methods		
				W	Proposed	SCC	Exact	Mid-p
0·0	0·1	0·14	30	92·8	95·4	97·7	97·8	96·1
			50	93·8	95·1	97·0	96·7	94·9
			80	94·4	94·7	96·2	96·0	94·5
	0·0	0·16	30	93·3	95·3	97·6	97·6	95·6
			50	94·5	95·0	98·2	97·3	95·1
			80	94·6	94·8	97·2	97·2	95·9
0·1	0·1	0·04	30	91·7	96·5	81·2	81·2	80·0
			50	92·9	96·0	87·3	87·3	86·3
			80	94·1	94·6	91·7	91·3	88·6
	0·0	0·06	30	92·5	95·6	99·9	99·9	99·6
			50	94·0	95·6	98·4	98·4	97·5
			80	94·3	95·0	97·4	97·6	95·6
0·14	0·1	0·0	30	80·0	98·4	36·5	36·5	34·5
			50	87·7	98·0	37·5	37·0	36·5
			80	91·5	94·3	39·4	39·0	37·2

6. EXAMPLES

6.1. Cross-over Clinical Trials on Soft Contact Lenses

First, let us consider cross-over clinical trials in which patients are randomized to one of two treatment sequences AB or BA. Under the assumption that there is no *carry-over effects* and no *period effects*, dichotomous data for the efficacy are summarized in the same form of 2×2 table as (15).

Miyanaga[13] conducted cross-over clinical trials comparing a chemical (hydrogen peroxide) disinfection system SA806 with a thermal disinfection system for soft contact lenses. It seems well recognized that appropriate conditions for carrying out cross-over trials are likely to be met in this field. 44 patients were randomized to one of two treatment sequences and the results are summarized in Table V. In this trial, we are interested in the equivalence of two disinfection methods. In this trial, the Fisher exact test was applied to test for equivalence; the one-tailed p-value was $p = \binom{1}{0}\frac{1}{2} = 0{\cdot}5$ indicating clear *non-significance* and so it was concluded that two methods are *equivalent*. However, this sort of inference is not acceptable. So, let us apply tests for equivalence. With the small off-diagonal cell frequency, however, we cannot apply Z_{LB} and Z_{MY} as is shown in Table II. Instead we apply our test. We have

$$Z = 1{\cdot}709 > Z_{0{\cdot}05} = 1{\cdot}645 \text{ (one-tailed } p\text{-value} = 0{\cdot}044)$$

and the 90 per cent confidence lower limit is

$$\lambda_{low} = -0{\cdot}096 > -\Delta = -0{\cdot}1.$$

Based on these results, it will be concluded that the two methods are equivalent at the 5 per cent significance level. To examine the empirical significance level of Z for this kind of data with

Table V. Clinical assessment of treatments in cross-over trials of disinfection systems for soft contact lenses

		Thermal disinfection		
		Effective	Ineffective	Total
Hydrogen peroxide	Effective	43	0	43
	Ineffective	1	0	1
	Total	44	0	44

Table VI. Empirical significance level of the four tests with $\Delta = 0{\cdot}1$ for such data with $n = 44$ and small off-diagonal cell frequencies as shown in Table II, based on 10,000 trials

q_{12}	π_S	ϕ	Test statistic		
			Z	Z_{LB}	Z_{MY}
0·0	0·95	0·0425	5·4	17·2	17·2
	0·90	0·08	5·7	16·8	16·8
0·02	0·95	0·0225	4·3	9·6	9·6
	0·90	0·06	4·6	10·0	10·0

$n = 44$ and small off-diagonal cell frequencies, we performed a simulation study (the same as that in Section 5) for each of the combinations of parameter values shown in Table VI. The resultant empirical significance levels for Z are shown to be around 4·3–5·7 per cent, reasonably close to the nominal 5 per cent level.

6.2. Epidemiological Study

Here we consider the data analysed by Karacan *et al.*[14] and also by Altman.[12] Karacan *et al.* compared a group of 32 marijuana users with 32 matched controls with respect to their sleeping difficulties. Data are reproduced in Table VII. In this example, we are interested in the statistical significance of the difference of the proportions experiencing sleeping difficulties, not in their equivalence. Therefore, by letting $\Delta = 0$, the test statistic Z coincides with the McNemar test and we have

$$Z = \frac{b - c}{\sqrt{(b + c)}} = \frac{9 - 3}{\sqrt{(9 + 3)}} = 1{\cdot}73$$

which gives two-tailed $p = 0{\cdot}08$, indicating weak evidence that marijuana users experience fewer sleeping difficulties than controls. The unconditional simple Wald type 95 per cent confidence interval for the difference in the proportions is $-0{\cdot}014$ to 0·389. The proposed score-based confidence interval is $-0{\cdot}027$ to 0·390.

It should be noted that the 95 per cent confidence interval calculated in Altman's textbook (page 237) is $-0{\cdot}03$ to 0·41. This calculation is wrong since the standard error of the difference in proportions is falsely calculated as $\frac{1}{32}\sqrt{(3 + 9 + 6^2/32)}$. The correct standard error is $\frac{1}{32}\sqrt{(3 + 9 - 6^2/32)}$.

Table VII. Numbers with (+) or without (−) sleeping difficulties among marijuana users and matched controls (Karacan *et al.*[14])

		Marijuana group		
		+	−	Total
Control group	+	4	9	13
	−	3	16	19
	Total	7	25	32

7. DISCUSSION

Regarding the model proposed in Section 2, as an alternative, but more restricted one, we might consider the following *mixed-effects* model:

$$\Pr\{N_k = 1 \mid \theta_k\} = \theta_k + \lambda$$

and

$$\Pr\{S_k = 1 \mid \theta_k\} = \theta_k$$

where $\{\theta_k\}, k = 1, 2, \ldots, n$ is assumed to be a sequence of unobserved independent and identically distributed random variables which have unknown distribution F with mean π and variance σ^2. In this model, λ indicates the common difference in proportions but the variability σ^2 is assumed to be the same for the two treatments and the correlation coefficient is 1·0. Even this simpler model can lead to the same score-based test statistic and its associated confidence interval procedure since the proposed statistical inference procedure starts from the parameterization $\phi = \rho\sigma_N\sigma_S$ as in equation (16). Liang and Zeger[15] considered a similar model to propose a new estimator of the common odds ratio in a matched case-control study, in which the link function used is logit.

One of the characteristics of the proposed procedure was shown to be applicability to tables with zero off-diagonal cells since other published methods do not apply to such data. We have carried out simulation study to evaluate the performance of the proposed procedures. The proposed test for equivalence has been shown to have empirical significant levels closer to a nominal α-level compared with the other two tests. The evaluation of confidence intervals has focused on the coverage probabilities since it is very important to confirm the designed $1 - \alpha$ coverage at least. The proposed confidence interval has been shown to perform well in general, however, it has been shown to be *conservative* when one of the off-diagonal cell probabilities is zero. On the other hand, all the conditional approaches including the exact one have been shown to be inadequate when one or both discordant probabilities are small.

Since these Monte Carlo experiments have been based on a relatively small number of sets of parameter values and sample sizes, the conclusions derived here may not be representative. However, as I have examined typical sets of parameter values and sample sizes, I expect that drastically different conclusions would not be derived for the parameter values not examined here, although we need a further simulation study for more detailed comparisons.

In this paper, we have considered tests for the class of hypotheses that have $\Delta \geqslant 0$, which includes McNemar's test for significance and a test for equivalence. We can also generalize the test to cope with the hypothesis with $\Delta < 0$. The latter alternative indicates that the true difference is greater than some *medically significant* difference $(-\Delta)$ which is not zero.

APPENDIX I: CALCULATION OF PARTIAL DERIVATIVES

For simplicity, let

$$\xi = 1 - (\pi_N + \pi_S) = 1 - 2\pi_S - \beta + \Delta.$$

Then, the three scores are given by:

$$\frac{\partial L}{\partial \phi} = \left[\frac{a}{q_{11}} - \frac{b}{q_{12}} - \frac{c}{q_{21}} + \frac{d}{q_{22}} \right]$$

$$\frac{\partial L}{\partial \beta} = \pi_S \frac{\partial L}{\partial \phi} + \frac{b}{q_{12}} - \frac{d}{q_{22}}$$

$$\frac{\partial L}{\partial \pi_S} = (1 - \xi) \frac{\partial L}{\partial \phi} + \frac{b}{q_{12}} + \frac{c}{q_{21}} - \frac{2d}{q_{22}}$$

where $\{q_{ij}\}$'s are defined by (20). The ijth element of the Fisher information matrix I are given by the following:

$$I_{11} = E\left[-\frac{\partial^2 L}{\partial \pi_S^2} \right] = n\left[\frac{(1-\xi)^2}{q_{11}} + \frac{\xi^2}{q_{12}} + \frac{\xi^2}{q_{21}} + \frac{(1+\xi)^2}{q_{22}} \right]$$

$$I_{12} = E\left[-\frac{\partial^2 L}{\partial \pi_S \partial \phi} \right] = n\left[\frac{1-\xi}{q_{11}} - \frac{\xi}{q_{12}} - \frac{\xi}{q_{21}} - \frac{(1+\xi)}{q_{22}} \right]$$

$$I_{13} = E\left[-\frac{\partial^2 L}{\partial \pi_S \partial \beta} \right] = n\left[\frac{(1-\xi)\pi_S}{q_{11}} + \frac{\xi(1-\pi_S)}{q_{12}} - \frac{\xi \pi_S}{q_{21}} + \frac{(1+\xi)(1-\pi_S)}{q_{22}} \right]$$

$$I_{22} = E\left[-\frac{\partial^2 L}{\partial \phi^2} \right] = n\left[\frac{1}{q_{11}} + \frac{1}{q_{12}} + \frac{1}{q_{21}} + \frac{1}{q_{22}} \right]$$

$$I_{23} = E\left[-\frac{\partial^2 L}{\partial \phi \partial \beta} \right] = n\left[\frac{\pi_S}{q_{11}} + \frac{\pi_S - 1}{q_{12}} + \frac{\pi_S}{q_{21}} + \frac{\pi_S - 1}{q_{22}} \right]$$

$$I_{33} = E\left[-\frac{\partial^2 L}{\partial \beta^2} \right] = n\left[\frac{\pi_S^2}{q_{11}} + \frac{(\pi_S - 1)^2}{q_{12}} + \frac{\pi_S^2}{q_{21}} + \frac{(\pi_S - 1)^2}{q_{22}} \right].$$

Needless to say, we have $I_{ij} = I_{ji}$.

APPENDIX II: DERIVATION OF EQUATIONS (24)–(26)

The simultaneous equations (22) yield

$$\frac{a}{q_{11}} + \frac{d}{q_{22}} = \frac{b}{q_{12}} + \frac{c}{q_{21}} = 2\frac{d}{q_{22}}.$$

Then we have

$$\frac{\partial L}{\partial \beta} = \frac{1}{2}\left(\frac{b}{q_{21} - \Delta} - \frac{c}{q_{21}} \right).$$

Further, by noting that

$$\frac{a}{q_{11}} = \frac{d}{q_{22}} = \frac{a+d}{q_{11}+q_{22}} = \frac{n-b-c}{1-q_{12}-q_{21}}$$

it is shown that q_{21} is a solution of the equation

$$\frac{b}{q_{21}-\Delta} + \frac{c}{q_{21}} = -2\frac{n-b-c}{2q_{21}-\Delta-1} \tag{44}$$

which reduces to the quadratic equations described in Section 3:

$$f(x) = 2nx^2 - (b+c+(2n-b+c)\Delta)x + c\Delta(\Delta+1) = 0.$$

So does the following equation:

$$\frac{b}{q_{21}-\Delta} - \frac{c}{q_{21}} = 2\frac{b-c+n\Delta}{2q_{21}-\Delta(\Delta+1)}.$$

Therefore the score statistic Z is given by (24).

Next, we describe why $\hat{q}_{21}$ should be the larger root of $f(x) = 0$ as defined in (25). We consider here not only the case $\Delta > 0$ but also $\Delta < 0$ since the result below can be applied to the problem of confidence limits (Appendix III).

Consider the following four cases.

1. $b > 0$ and $c > 0$. Since $0 < \hat{q}_{21} = x < 1$ and $0 < \hat{q}_{12} = x - \Delta < 1$, the appropriate root of $f(x) = 0$ must satisfy

$$\Delta < x < 1 \quad \text{for } \Delta > 0$$

$$0 < x < 1 + \Delta \quad \text{for } \Delta < 0.$$

   When $\Delta > 0$, we have

$$f(0) = c\Delta(\Delta+1) > 0, \quad f(1) = (1-\Delta)(2n-b-c(1+\Delta)) > 0, \text{ and } f(\Delta) = -b\Delta(1-\Delta) < 0,$$

   which indicates the larger root satisfies $\Delta < x < 1$. When $\Delta < 0$, we have

$$f(0) < 0, \text{ and } f(1+\Delta) = (1+\Delta)(2n-c-b(1-\Delta)) > 0$$

   and thus the larger root satisfies $0 < x < 1 + \Delta$. Therefore, when $b > 0$ and $c > 0$, regardless of the sign of Δ, $\hat{q}_{21}$ is the larger root.
2. $b = 0$ and $c > 0$. In this case, we have two roots:

$$x_1 = \Delta, \text{ and } x_2 = \frac{c(1+\Delta)}{2n}.$$

   However, equation (44) has a single root x_2. The root x_1 is a boundary point of the parameter space since $\hat{q}_{12} = 0$. This means the following:

(a) Case $\Delta > 0$: the condition that the simultaneous equations (22) have the unique root x_2 is $0 \leqslant \hat{q}_{12} = x_2 - x_1$, that is

$$c \geqslant \frac{2n\Delta}{1+\Delta}. \tag{45}$$

If $x_2 < x_1$, then the simultaneous equations (22) are no longer a necessary condition for maximality and the log-likelihood (16) attains its maximum at $\hat{q}_{21} = x_1 = \Delta$ on a boundary point.

(b) Case $\Delta < 0$: x_2 is always the unique root of (22) since $0 < \hat{q}_{12} < 1$.

In this case also, $\hat{q}_{21}$ is the larger root of $f(x) = 0$.

3. $c = 0$ and $b > 0$. In this case, we have two roots:

$$x_1 = 0, \text{ and } x_2 = \Delta + \frac{b(1-\Delta)}{2n}$$

but equation (44) has a single root x_2.

(a) Case $\Delta > 0$: x_2 is always the unique root of (22) since $0 = x_1 < x_2 < 1$ and $0 < \hat{q}_{12} < 1$.

(b) Case $\Delta < 0$: we always have $0 < \hat{q}_{12} < 1$. Then the condition that equations (22) have the unique root should be $0 \leqslant x_2 < 1$, that is

$$b \geqslant -\frac{2n\Delta}{1-\Delta}.$$

If $x_2 < 0 = x_1$ then equations (22) have no roots and the log-likelihood (16) takes its maximum at $\hat{q}_{21} = x_1 = 0$.

Here also, $\hat{q}_{21}$ is shown to be the larger root of $f(x) = 0$.

4. $b = c = 0$. We have two roots $x_1 = 0$ and $x_2 = \Delta$. In this case, equations (22) have no roots, but the log-likelihood (16) is maximized at $\hat{q}_{21} = x_2 = \Delta$ if $\Delta > 0$, and at $x_1 = 0$, otherwise. In this case also, the larger root is $\hat{q}_{21}$.

In summary, regardless of the sign of Δ and of the values of (b, c), $\hat{q}_{21}$ is always shown to be the larger root of $f(x) = 0$; hence the positive root value of the square root is taken in equation (25).

APPENDIX III: RELATIONSHIP BETWEEN $\hat{\lambda}$ AND $\hat{q}_{21}$

From the results of Appendix II, by letting $\lambda = -\Delta$, we have the following results:

1. $b > 0$ and $c > 0$. Each of the two confidence limits can be calculated by using $\hat{q}_{21}$ which is the unique root of equations (22).
2. $b = 0$ and $c > 0$. Using (45), we have the following relationship:

$$\hat{q}_{21} = \begin{cases} -\hat{\lambda} & \text{(boundary point)} \quad \text{if } \hat{\lambda} \leqslant -c/(2n-c) \\ c(1-\hat{\lambda})/2n & \text{(interior point)} \quad \text{otherwise.} \end{cases}$$

It should be noted that if $\hat{q}_{21} = -\hat{\lambda}$ then $\hat{q}_{12} = 0$. If $\lambda \geqslant -c/(2n-c)$ then $Z(0, c; n, -\lambda) < 0$, indicating that the lower limit is always based on the boundary value $\hat{q}_{21,\text{low}} = -\hat{\lambda}_{\text{low}}$. The upper limit requires $c + n\lambda > 0$, which means

$$\hat{q}_{21,\text{up}} = \begin{cases} -\hat{\lambda}_{\text{up}} & \text{if } Z_\alpha^2 \leqslant c(n-c)/2n \\ c(1-\hat{\lambda}_{\text{up}})/2n & \text{otherwise.} \end{cases}$$

3. $b > 0$ and $c = 0$. In a similar manner, $\hat{\lambda}_{\mathrm{up}}$ is always based on the boundary value of $\hat{q}_{21,\mathrm{up}} = 0$. The lower limit $\hat{\lambda}_{\mathrm{low}}$ is based on $\hat{q}_{21,\mathrm{low}}$ defined as

$$\hat{q}_{21,\mathrm{low}} = \begin{cases} 0 & \text{if } Z_\alpha^2 \leqslant b(n-b)/2n \\ -\hat{\lambda}_{\mathrm{low}} + b(1+\hat{\lambda}_{\mathrm{low}})/2n & \text{otherwise.} \end{cases}$$

4. $b = 0$ and $c = 0$. Clearly, we have that $\hat{q}_{21,\mathrm{low}} = -\hat{\lambda}_{\mathrm{low}}$ and $\hat{q}_{21,\mathrm{up}} = 0$.

ACKNOWLEDGEMENTS

The authors thanks two anonymous referees for invaluable comments on an earlier draft of the paper that led to substantial improvements.

REFERENCES

1. 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**, 1831–1839 (1995).
2. Morikawa, T. and Yanagawa, T. 'Taiounoaru 2chi data ni taisuru doutousei kentei. (Equivalence testing for paired dichotomous data)', *Proceedings of Annual Conference of Biometric Society of Japan*, 123–126 (1995) (in Japanese).
3. Machin, D. and Campbell, M. J. *Statistical Tables for the Design of Clinical Trials*, Blackwell Scientific Publications, Oxford, 1987.
4. Cox, D. R. and Hinkley, D. V. *Theoretical Statistics*, Chapman and Hall, London, 1974.
5. Vollset, S. E. 'Confidence intervals for a binomial proportion', *Statistics in Medicine*, **12**, 809–824 (1993).
6. Wilson, E. B. 'Probable inference, the law of succession, and statistical inference', *Journal of the American Statistical Association*, **22**, 209–212 (1927).
7. Yanagawa, T., Tango, T. and Hiejima, Y. 'Mantel-Haenszel type tests for testing equivalence or more than equivalence in comparative clinical trials', *Biometrics*, **50**, 859–864 (1994).
8. Chant, D. 'On the asymptotic tests of composite hypotheses in nonstandard conditions', *Biometrika*, **61**, 291–298 (1974).
9. Self, S. G. and Liang, K. Y. 'Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions', *Journal of the American Statistical Association*, **82**, 605–610 (1987).
10. Grizzle, J. E., Starmer, C. F. and Koch, G. G. 'Analysis of categorical data by linear models', *Biometrics*, **25**, 489–503 (1969).
11. Gart, J. J. and Nam, J. 'Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness', *Biometrics*, **44**, 323–338 (1988).
12. Altman, D. G. *Practical Statistics for Medical Research*, Chapman and Hall, London, 1991.
13. Miyanaga, Y. 'Clinical evaluation of the hydrogen peroxide SCL disinfection system (SCL-D)', *Japanese Journal of Soft Contact Lenses*, **36**, 163–173 (1994) (in Japanese).
14. Karacan, I., Fernandez, S. A. and Coggins, W. S. 'Sleep electrocephalographic-electrooculographic characteristics of chronic marijuana users: part 1', *New York Academy of Science*, **282**, 348–374 (1976).
15. Liang, K. Y. and Zeger, S. L. 'On the use of concordant pairs in matched case-control studies', *Biometrics*, **44**, 1145–1156 (1988).

