Research Article

Statistics in Medicine

Received 12 June 2008, Accepted 13 August 2009 Published online 12 October 2009 in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3738

Confidence intervals for a difference between proportions based on paired data

Man-Lai Tang,[a*†] Man-Ho Ling,[a] Leevan Ling[a] and Guoliang Tian[b]

We construct several explicit asymptotic two-sided confidence intervals (CIs) for the difference between two correlated proportions using the method of variance of estimates recovery (MOVER). The basic idea is to recover variance estimates required for the proportion difference from the confidence limits for single proportions. The CI estimators for a single proportion, which are incorporated with the MOVER, include the Agresti–Coull, the Wilson, and the Jeffreys CIs. Our simulation results show that the MOVER-type CIs based on the continuity corrected Φ coefficient and the Tango score CI perform satisfactory in small sample designs and spare data structures. We illustrate the proposed CIs with several real examples. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: Agresti–Coull interval; Jeffreys interval; method of variance estimates recovery; paired binary data; Tango score interval; Wilson score interval

1. Introduction

Paired binary data arise in many statistical contexts such as crossover trials, equivalence trials, matched case-control studies, and pre- and post-test comparative studies. For instance, Miyanaga [1] conducted a crossover clinical trial that compared a chemical (hydrogen peroxide) disinfection system SA806 with a thermal disinfection system for soft contact lenses. Forty-four patients were randomized to one of two treatment sequences and the results are reported in Table I. In this trial, we are interested in the equivalence of the two disinfection systems. In this case, equivalence can be demonstrated by comparing the $100(1-\alpha)$ per cent confidence limits on the difference between the two effective rates with the equivalence limits, say $(-\Delta_0, \Delta_0)$ with Δ_0 being some small positive clinical acceptable threshold. That is, if the $100(1-\alpha)$ per cent confidence interval (CI) for the difference between the two effective rates is entirely lying inside the interval $(-\Delta_0, \Delta_0)$, one could safely conclude the equivalence between the two disinfection systems. Obviously, the construction of CI for the proportion difference from correlated binary data plays an important role here.

A variety of approaches have been proposed for the construction of CIs for differences between proportions based on paired data. A conceptually simple approach is to adapt an interval designed for unpaired proportions to take account of the correlation between the two classifications. Newcombe [2] showed that a simple, effective CI for the difference between independent proportions may be calculated by the so-called *square-and-add approach*, which calculates an interval for the difference in proportions from separate intervals for the two proportions. The square-and-add algorithm preserves boundary respecting as well as closed-form properties, which makes it particularly suitable for applications involving proportions. In his comparative evaluation of 11 methods, a square-and-add interval based on Wilson [3] score intervals for the two proportions performed as well as the score interval [4] and several other methods evaluated.

Based on this study, Newcombe [5] also compared 10 CIs for the difference between binomial proportions based on paired binary data, including methods derived from intervals for unpaired data but corrected for correlation. An interval derived from the square-and-add Wilson interval by incorporating a Φ-type measure of association performed as well as any other method evaluated. At the same time Tango [6] developed a score-based CI for the difference between two proportions for paired-sample designs. Newcombe [7] subsequently showed that the Tango score CI consistently slightly outperformed other CIs for difference between

[a] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China
[b] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, People's Republic of China
[*] Correspondence to: Man-Lai Tang, Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China.
[†] E-mail: mltang@math.hkbu.edu.hk

Contract/grant sponsor: Research Grant Council of the Hong Kong Special Administrative Region; contract/grant numbers: HKBU261007, HKBU261508

Table I. Clinical assessment of treatment in crossover trial of disinfection systems for soft contact lenses in Example 5.1

	Thermal disinfection system	
Hydrogen peroxide system	Effective	Ineffective
Effective	43	0
Ineffective	1	0

two correlated proportions. Although the score CI is preferred among the existing asymptotic CIs [7, 8], it is not computationally friendly in the sense that iterative methods are required to calculate the confidence limits [6], whereas the interval based on squaring and adding performs nearly as well and is of closed form.

Subsequently, the square-and-add approach has been used in a variety of applications. Donner and Zou [9] and Zou and Donner [10] have provided a theoretical justification for this approach, summarized in the acronym method of variance estimates recovery (MOVER) proposed by Zou [11]. The MOVER and square-and-add approaches are evidently identical, yet neither is the original use of the method. Zou *et al.* [12] pointed out that methods similar to MOVER have previously been introduced by Howe [13] for approximate CIs for the mean of the sum of two independent random variables, Graybill and Wang [14] for CIs on non-negative linear combinations of variances, and Lee *et al.* [15] for CIs for linear combinations of variance components. These earlier works provided the rationale behind MOVER.

The MOVER or square-and-add algorithm per se combines CIs based on separate samples. It has been adapted as described above for paired comparisons, but hitherto only using score intervals for the two proportions. This article develops and evaluates similar methods for constructing hybrid CIs for the difference between two correlated proportions derived from other intervals for the two proportions. The resulting intervals do not involve the computational complexity of the Tango score interval and perform well.

The remainder of this article is organized as follows. Section 2 first delineates the traditional paired-sample design together with the Tango score interval for the difference between two correlated proportions. We then discuss general hybrid CI construction for the difference between two correlated proportions based on MOVER approach. Three existing best CIs (namely the Agresti–Coull, the Wilson score, and the Jeffreys CIs) for a single proportion are reviewed. A simulation study is conducted to investigate the performance of various CI estimators with respect to expected coverage probability, expected confidence width (ECW), and mesial and distal noncoverage probabilities (MNCPs and DNCPs) in Section 3. Both the sample correlation coefficient and Φ coefficient are considered as the estimate of the required correlation parameter. In Section 4, we illustrate our methods with the aforementioned equivalence trial for contact lenses disinfection systems. A discussion is given in Section 5.

2. CIs for the difference of two correlated proportions

In this section, we present our problem under the context of equivalence trial for two diagnostic test procedures, which involves comparisons of the response rates between paired binary endpoints. Let X_{0j} and $X_{1j}, j=0,1$ be the dichotomous responses representing the diagnostic outcomes for the test and referenced procedures, respectively. Let x_{00}, x_{01}, x_{10}, and x_{11} be the observed frequencies for outcomes $(0,0),(0,1),(1,0)$, and $(1,1)$, respectively, with $x_{00}+x_{01}+x_{10}+x_{11}=n$. The four outcomes and response probabilities can be classified as the 2×2 table given in Table II.

In this article, the parameter of interest is $\Delta=\pi_{1+}-\pi_{+1}$ (equivalently, $\Delta=\pi_{10}-\pi_{01}$). In particular, we would like to construct simple but reliable CIs for Δ. For this purpose, Tango [6] derived a score-based CI for Δ. Briefly, Tango score confidence limits for Δ are the two real roots to the equation $x_{10}-x_{01}-n\ \Delta=\pm z_{\alpha/2}\sqrt{n(2q+\Delta(1-\Delta)}$, where $q=\{\sqrt{y^2-4Nx_{01}\Delta(1-\Delta)}-y\}/(2N)$ and $y=(2n-x_{10}+x_{01})\Delta-x_{10}-x_{01}$. Recent empirical studies recommended Tango score CI to be an excellent CI for difference between two correlated proportions with respect to coverage probability and ECW [7, 8]. Here, the construction of Tango score CI involves iterative procedures for solving the above equation. We describe below a simple CI construction procedure that hybridizes a $100(1-\alpha)$ per cent CI for π_{1+} and a $100(1-\alpha)$ per cent CI for π_{+1} to form a $100(1-\alpha)$ per cent CI for π_{1+} CI for $\Delta=\pi_{1+}-\pi_{+1}$.

2.1. *The method of variance estimates recovery* (*MOVER*)

We first describe the general concept behind the MOVER. We would like to point out that the method presented here has been recently adopted by Zou and his fellows in epidemiological studies [11], psychological studies [16], environmental studies [17], medical studies [10, 18], and binary outcome studies [19]. In general, we would like to construct a $100(1-\alpha)$ per cent CI, denoted as $(L,\ U)$, for $\theta_1-\theta_0$. Let $\hat{\theta}_1$ and $\hat{\theta}_0$ be two estimates of θ_1 and θ_0, respectively, and (l_1,u_1) and (l_0,u_0) denote two separate $100(1-\alpha)$ per cent CIs for θ_1 and θ_0, respectively.

(A) $\hat{\theta}_1$ *and* $\hat{\theta}_0$ *are uncorrelated*

In this case, a plausible choice for (L,U) is $L=(\hat{\theta}_1-\hat{\theta}_0)-z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_1)+\mathrm{Var}(\hat{\theta}_0)}$ and $U=(\hat{\theta}_1-\hat{\theta}_0)+z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_1)+\mathrm{Var}(\hat{\theta}_0)}$. According to the duality between two-sided hypothesis testing and CI construction, the lower (i.e. L) and upper (i.e. U)

Table II. Frequency and probability table for equivalence trial for comparing two test procedures

	Diagnosis for reference procedure		
Diagnosis for test procedure	1 (+)	0 (−)	Total
1 (+)	x_{11} (π_{11})	x_{10} (π_{10})	n_{1+} (π_{1+})
0 (−)	x_{01} (π_{01})	x_{00} (π_{00})	n_{0+} (π_{0+})
Total	n_{+1} (π_{+1})	n_{+0} (π_{+0})	n

confidence limits can be regarded as the minimum and maximum parameter values that, asymptotically, satisfy

$$\frac{[(\hat{\theta}_1-\hat{\theta}_0)-L]^2}{\text{Var}(\hat{\theta}_1)+\text{Var}(\hat{\theta}_0)}=z_{\alpha/2}^2$$

and

$$\frac{[U-(\hat{\theta}_1-\hat{\theta}_0)]^2}{\text{Var}(\hat{\theta}_1)+\text{Var}(\hat{\theta}_0)}=z_{\alpha/2}^2$$

respectively. It is worth noting that (l_1, u_1) and (l_0, u_0) contain the plausible values of θ_1 and θ_0, respectively. Among these plausible values for θ_1 and θ_0, the values closest to the minimum L and maximum U are, respectively, l_1-u_0 and u_1-l_0 in spirit of the score-type CI [9]. According to the central limit theorem, the variance estimates can now be recovered from $\theta_1=l_1$ as $\widehat{\text{Var}}(\hat{\theta}_1)=(\hat{\theta}_1-l_1)^2/z_{\alpha/2}^2$ and from $\theta_0=u_0$ as $\widehat{\text{Var}}(\hat{\theta}_0)=(u_0-\hat{\theta}_0)^2/z_{\alpha/2}^2$ for setting L. As a result,

$$L=\hat{\theta}_1-\hat{\theta}_0-\sqrt{(\hat{\theta}_1-l_1)^2+(u_0-\hat{\theta}_0)^2}$$

Similarly, we have

$$U=\hat{\theta}_1-\hat{\theta}_0+\sqrt{(u_1-\hat{\theta}_1)^2+(\hat{\theta}_0-l_0)^2}$$

(B) *$\hat{\theta}_1$ and $\hat{\theta}_0$ are correlated*
Using the similar arguments given in the independence case, we can easily obtain

$$L=\hat{\theta}_1-\hat{\theta}_0-\sqrt{(\hat{\theta}_1-l_1)^2+(u_0-\hat{\theta}_0)^2-2\widehat{\text{Corr}}(\hat{\theta}_1,\hat{\theta}_0)(\hat{\theta}_1-l_1)(u_0-\hat{\theta}_0)}$$

and

$$U=\hat{\theta}_1-\hat{\theta}_0+\sqrt{(u_1-\hat{\theta}_1)^2+(\hat{\theta}_0-l_0)^2-2\widehat{\text{Corr}}(\hat{\theta}_1,\hat{\theta}_0)(u_1-\hat{\theta}_1)(\hat{\theta}_0-l_0)}$$

To construct CI for $\Delta=\pi_{1+}-\pi_{+1}$, we can simply set $\theta_1=\pi_{1+}$ and $\theta_0=\pi_{+1}$. Since n_{1+}/n and n_{+1}/n are unbiased estimates for π_{1+} and π_{+1}, respectively, it is sensible to set $\hat{\theta}_1=n_{+1}/n$ and $\hat{\theta}_0=n_{+1}/n$. Obviously, $\hat{\theta}_1$ and $\hat{\theta}_0$ are correlated in the present setting and the correlation can be estimated by

$$\widehat{\text{Corr}}(\hat{\theta}_1,\hat{\theta}_0)=\frac{\hat{\pi}_{11}(1-\hat{\pi}_{11})-\hat{\pi}_{11}\hat{\pi}_{10}-\hat{\pi}_{11}\hat{\pi}_{01}-\hat{\pi}_{10}\hat{\pi}_{01}}{\sqrt{[\hat{\pi}_{11}(1-\hat{\pi}_{11})+\hat{\pi}_{10}(1-\hat{\pi}_{10})-2\hat{\pi}_{11}\hat{\pi}_{10})(\hat{\pi}_{11}(1-\hat{\pi}_{11})+\hat{\pi}_{01}(1-\hat{\pi}_{01})-2\hat{\pi}_{11}\hat{\pi}_{01}]}}$$

where $\hat{\pi}_{11}=x_{11}/n$, $\hat{\pi}_{10}=x_{10}/n$, and $\hat{\pi}_{01}=x_{01}/n$. When x_{1+} or x_{+1} equals to 0 or n, the denominator of $\widehat{\text{Corr}}(\hat{\theta}_1,\hat{\theta}_0)$ is equal to 0 and the correlation estimate is then undefined. To overcome this issue, we simply set x_{01} (or x_{10}) to be 0.5 when x_{01} (or x_{10}) is equal to 0. Alternatively, Newcombe [2] suggested the following Φ coefficient with continuity correction as an estimator for the correlation

$$\hat{\Phi}=\frac{x_{11}x_{00}-x_{01}x_{10}}{\sqrt{x_{1+}+x_{0+}+x_{+1}+x_{+0}}}$$

Here, $\hat{\Phi}=0$ if the denominator is 0 and the numerator of $\hat{\Phi}$ is replaced by $\max(x_{11}x_{00}-x_{01}x_{10}-n/2, 0)$ if $x_{11}x_{00}-x_{01}x_{10}-n/2>0$.

2.2. *Reliable CIs for single binomial proportion*

By constructing CI for $\Delta=\pi_{1+}-\pi_{+1}$ based on MOVER, one needs two separate CIs for $\theta_1=\pi_{1+}$ and $\theta_0=\pi_{+1}$. On the other hand, it is noted that $n_{1+}\sim\text{Bin}(n,\theta_1)$ and $n_{+1}\sim\text{Bin}(n,\theta_0)$. Brown *et al.* [20] presented a thorough literature review on CI estimation for a single binomial proportion. Noticing that the standard Wald interval for a binomial proportion could behave poorly even for large sample sizes, they recommended the Wilson [3] and the equal-tailed Jeffreys (i.e. Bayesian with Jeffreys non-informative prior) intervals for small sample sizes (i.e. $\leqslant 40$). For large sample sizes (i.e. $\geqslant 40$), they reported that the Wilson, the Jeffreys, and the Agresti–Coull [21] intervals are all comparable and for simplicity, they recommended the Agresti–Coull interval [22]. In general, let $Y \sim \text{Bin}(n,\theta)$ and $\hat{\theta}=Y/n$. We briefly review the three CIs for θ as follows:

(i) *The Wilson score interval*
According to the central limit theorem, it is noted that $n^{1/2}(\hat{\theta}-\theta)/\sqrt{\theta(1-\theta)}$ converges in distribution to the standard normal distribution. Thus,

$$\Pr(-z_{\alpha/2}\leqslant n^{1/2}(\hat{\theta}-\theta)/\sqrt{\theta(1-\theta)}\leqslant z_{\alpha/2})=\Pr(n(\hat{\theta}-\theta)^2/[\theta(1-\theta)]\leqslant z^2_{\alpha/2})=1-\alpha$$

The $100(1-\alpha)$ per cent Wilson CI is defined as (l,u), with l and u being the smaller and larger roots to the quadratic equation $n(\hat{\theta}-\theta)^2/[\theta(1-\theta)]=z^2_{\alpha/2}$ for θ. After some algebra, it can be shown that

$$l=\tilde{\theta}-\frac{z_{\alpha/2}}{\tilde{n}}\sqrt{n\hat{\theta}(1-\hat{\theta})+\frac{z^2_{\alpha/2}}{4}} \quad \text{and} \quad u=\tilde{\theta}+\frac{z_{\alpha/2}}{\tilde{n}}\sqrt{n\hat{\theta}(1-\hat{\theta})+\frac{z^2_{\alpha/2}}{4}}$$

where $\tilde{\theta}=(Y+0.5z^2_{\alpha/2})/(n+z^2_{\alpha/2})$ and $\tilde{n}=n+z^2_{\alpha/2}$.

(ii) *The Agresti–Coull interval*
The Agresti–Coull interval is an adjusted standard Wald interval with a different estimate of θ, namely the center of the Wilson score interval $\tilde{\theta}=(Y+0.5z^2_{\alpha/2})/(n+z^2_{\alpha/2})$. Hence, the $100(1-\alpha)$ per cent Agresti–Coull CI is defined as (l,u) with l and u being

$$l=\tilde{\theta}-z_{\alpha/2}\sqrt{\tilde{\theta}(1-\tilde{\theta})/\tilde{n}} \quad \text{and} \quad u=\tilde{\theta}+z_{\alpha/2}\sqrt{\tilde{\theta}(1-\tilde{\theta})/\tilde{n}}$$

(iii) *The Jeffreys interval*
It is noted that β distributions are the standard conjugate priors for binomial distributions. Further suppose that $\theta\sim\beta(\gamma_1,\gamma_2)$. Then, the posterior distribution of θ is $\beta(Y+\gamma_1,n-Y+\gamma_2)$. Hence, the $100(1-\alpha)$ per cent equal-tailed Bayesian interval is given by $[\beta(\alpha/2;\ Y+\gamma_1,n-Y+\gamma_2),\beta(1-\alpha/2;Y+\gamma_1,n-Y+\gamma_2)]$, where $\beta(\alpha;\ \lambda_1,\lambda_1)$ denotes the α quantile of a $\beta(\lambda_1,\lambda_1)$ distribution. The well-known non-informative Jeffreys prior is $\beta(\frac{1}{2},\frac{1}{2})$. Therefore, the $100(1-\alpha)$ per cent Jeffreys CI is defined as (l,u) with l and u being

$$l=\frac{2Y+1}{2Y+1+(2[n-Y]+1)F_{\alpha/2}(2[n-Y]+1,2Y+1)}$$

and

$$u=\frac{2Y+1}{2Y+1+(2[n-Y]+1)F_{1-\alpha/2}(2[n-Y]+1,2Y+1)}$$

where $F_\gamma(v_1,v_2)$ is the γ quantile from the F-distribution with (v_1,v_2) degrees of freedom. It is worth noting that the well-known Clopper–Pearson 'exact' confidence limits take a very similar form to Bayesian confidence limits, whether with a Jeffreys $\beta(\frac{1}{2},\frac{1}{2})$, uniform $\beta(1,\ 1)$, or any other conjugate beta prior. However, the Clopper–Pearson lower and upper limits use different parameter values, resulting in a wider interval. The Jeffreys interval is always contained within the Clopper–Pearson CI, accordingly they have very similar location properties but quite different coverage; while the 'exact' interval ensures a minimum coverage of $1-\alpha$, the Jeffreys interval aligns the mean coverage with this nominal level.

To construct CI for $\Delta=\pi_{1+}-\pi_{+1}$ based on MOVER, we can simply set $\theta=\pi_{1+}$ with $Y=n_{1+}$ for π_{1+}, and $\theta=\pi_{+1}$ with $Y=n_{+1}$ for π_{+1} in the confidence limits (i.e. l's and u's) defined in (i), (ii), and (iii).

3. Evaluation

In this section, we investigate the performance of various CIs with respect to their exact coverage probabilities (ECPs), ECWs, and DNCPs and MNCPs. Coverage and directional noncoverage are evaluated following the paradigm set out in Newcombe [2, 5, 23]. First and the most importantly, a CI is good if it is able to guarantee its CPs close to the pre-specified coverage level. Given the CPs are well controlled, one would prefer those CIs that generally yield shorter ECWs. Finally, directional noncoverage may be characterized here as mesial- and distal-noncoverage, indicating that the CI is wholly above or below the true value of $\Delta=\pi_{1+}-\pi_{+1}>0$, respectively. For $-1<\Delta<0$, they would be interchanged. Here, Newcombe [24] suggested a ratio measure MNCP/(MNCP+DNCP) for it ranges on the interval $[0,1]$ and can effectively separate the function of assessing location from assessment of overall coverage. We classify this ratio measure as *satisfactory* if it is between 0.4 and 0.6, the interval is too *mesially* located if it is below 0.4, and too *distally* located if it is above 0.6.

Let n represent the total sample size. All these measures are then examined for small (e.g. $10\leqslant n\leqslant 20$) and moderate (e.g. $30\leqslant n\leqslant 50$) sample sizes via simulation studies. For the marginal probability $\pi_{+1}=\pi_{11}+\pi_{01}$, we consider the following three zones: (p1) the rare response group with $0.05\leqslant\pi_{+1}\leqslant 0.1$; (p2) the moderate response group with $0.4\leqslant\pi_{+1}\leqslant 0.6$; and (p3) the high response group with $0.8\leqslant\pi_{+1}\leqslant 0.95$. For proportion difference Δ, we consider the following three different zones: (d1) the no difference group with $\Delta=0$; (d2) the small difference group with $0.01\leqslant\Delta\leqslant 0.05$; and (d3) the moderate difference group with $0.1\leqslant\Delta\leqslant 0.2$. Here, $\pi_{1+}=\pi_{11}+\pi_{10}$ can be determined by $\pi_{+1}+\Delta$. To introduce dependence/correlation between the paired binary outcomes, we assume that the bivariate binary observations are coming from a bivariate distribution with the correlation coefficient defined by

$$\rho=(\pi_{11}-\pi_{1+}\pi_{+1})/[\pi_{1+}(1-\pi_{1+})\pi_{+1}(1-\pi_{+1})]^{1/2}$$

Hence, given π_{1+},π_{+1} and ρ, we have

$$\pi_{11}=\pi_{1+}\pi_{+1}+\rho[\pi_{1+}(1-\pi_{1+})\pi_{+1}(1-\pi_{+1})]^{1/2}$$

For correlation ρ, we consider the following four zones: (r1) the slightly negative correlation group with $-0.1\leqslant\rho\leqslant 0$; (r2) independence group with $\rho=0$; (r3) slightly positive correlation group with $0\leqslant\rho\leqslant 0.2$; and (r4) moderately positive correlation group with $0.4\leqslant\rho\leqslant 0.6$.

Our simulation is conducted as follows. We first randomly generate 10 000 vectors (π_{+1},Δ,ρ) for each of the 36 zones of $\{(\text{p1}),(\text{p2}),(\text{p3})\}\times\{(\text{d1}),(\text{d2}),(\text{d3})\}\times\{(\text{r1}),(\text{r2}),(\text{r3}),(\text{r4})\}$. For each generated vector (π_{+1},Δ,ρ), we then randomly generate a sample size n from each sample size group (i.e. $10\leqslant n\leqslant 20$ or $30\leqslant n\leqslant 50$). For each sample size n generated from each of the 36 zones, the ECP, ECW, MNCP, and DNCP based on the seven 95 per cent CIs discussed in Section 2 (i.e. the Tango score CI (T), the MOVER-Agresti–Coull CI based on the sample correlation coefficient (A), the MOVER-Wilson CI based on the sample correlation coefficient (W), the MOVER-Jeffreys CI based on the sample correlation coefficient (J), the MOVER-Agresti–Coull CI based on the sample Φ coefficient (N_{A}), the MOVER-Wilson CI based on the sample Φ coefficient (N_{W}) as proposed by Newcombe [5, Method 10], and the MOVER-Jeffreys CI based on the sample Φ coefficient (N_{J})) can be computed as follows. For ECP, it is defined as

$$\text{ECP}=\sum_{x_{11}=0}^{n}\sum_{x_{10}=0}^{n-x_{11}}\sum_{x_{01}=0}^{n-x_{11}-x_{10}} I(L\leqslant\Delta\leqslant U)\cdot f(n,x_{11},x_{10},x_{01})$$

where $[L,U]$ is any CI based on the seven methods, and

$$f(n,x_{11},x_{10},x_{01})=\frac{n!}{x_{11}!x_{10}!x_{01}!(n-x_{11}-x_{10}-x_{01})!}\pi_{11}^{x_{11}}\pi_{10}^{x_{10}}\pi_{01}^{x_{01}}(1-\pi_{1+}-\pi_{01})^{n-x_{11}-x_{10}-x_{01}}$$

The ECW is defined as

$$\sum_{x_{11}=0}^{n}\sum_{x_{10}=0}^{n-x_{11}}\sum_{x_{01}=0}^{n-x_{11}-x_{10}}(U-L)\cdot f(n,x_{11},x_{10},x_{01})$$

Finally, the MNCP and DNCP are defined as

$$\text{MNCP}=\sum_{x_{11}=0}^{n}\sum_{x_{10}=0}^{n-x_{11}}\sum_{x_{01}=0}^{n-x_{11}-x_{10}} I(\Delta<L)\cdot f(n,x_{11},x_{10},x_{01})$$

and

$$\text{DNCP}=\sum_{x_{11}=0}^{n}\sum_{x_{10}=0}^{n-x_{11}}\sum_{x_{01}=0}^{n-x_{11}-x_{10}} I(\Delta>U)\cdot f(n,x_{11},x_{10},x_{01})$$

4. Results of evaluation

A CI with a shorter ECW and its ECP close to the pre-specified confidence level (i.e. 95 per cent in this study) is assessed to be a better CI. The boxplots of ECPs, ECWs, and the ratios of MNCPs and NCPs stratified by different factors are reported in Figures 1–6. It should be noted that each box in the figures contains the middle 50 per cent of the data. The upper edge of the box indicates the 75th percentile of the data set, and the lower edge indicates the 25th percentile. The whiskers indicate the minimum and maximum data values, unless outliers are present in which case the whiskers extend to a maximum of 1.5 times the interquartile range. The points outside the ends of the whiskers are outliers.

(1) *General observation*: To have a better investigation of the ECP performance, we also report in Table III the proportions of ECPs lying above 96 per cent, between 94 and 96 per cent, or below 94 per cent. A CI will be classified as conservative, good, or deflated if its ECP is lying above 96 per cent, between 94 and 96 per cent, or below 94 per cent, respectively. According to the results, none of the CIs under consideration are deflated as their proportions of ECPs lying below 94 per cent are very small. However, all CIs are slightly conservative in small sample designs as their proportions of ECPs lying above 96 per cent are all close to 70 per cent. When sample size gets larger, all CIs become less conservative. In terms of ECWs, it is found that CIs based on (J), (N_{W}), and (N_{J}) are shorter than (T) generally. For the ratios of MNCPs and NCPs, the medians are generally lying within

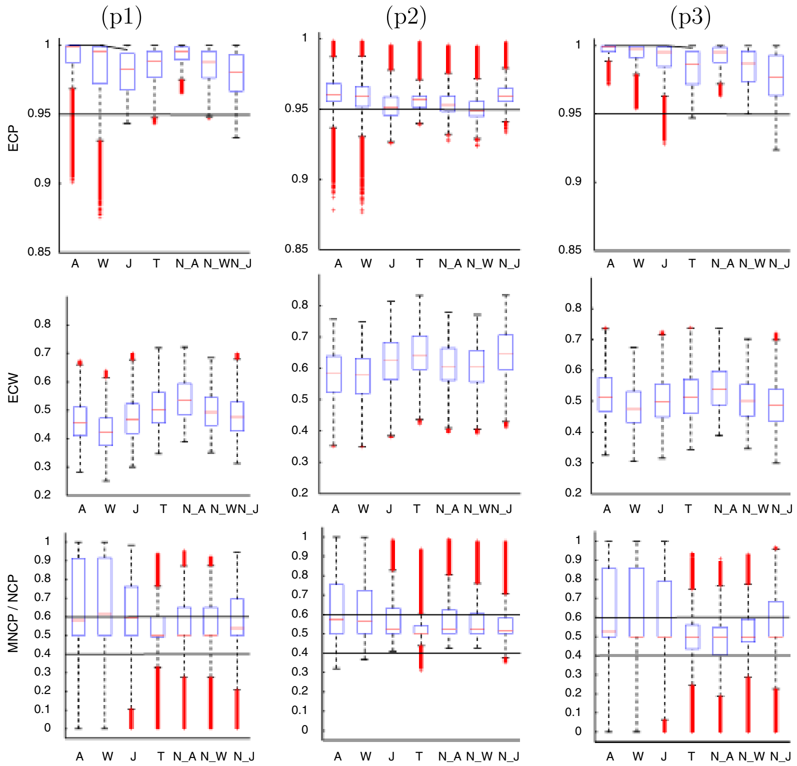

Figure 1. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of π_{+1} for small sample designs (i.e. $10\leqslant n\leqslant 20$): (p1) $0.05\leqslant\pi_{+1}\leqslant 0.1$; (p2) $0.4\leqslant\pi_{+1}\leqslant 0.6$; and (p3) $0.8\leqslant\pi_{+1}\leqslant 0.95$.

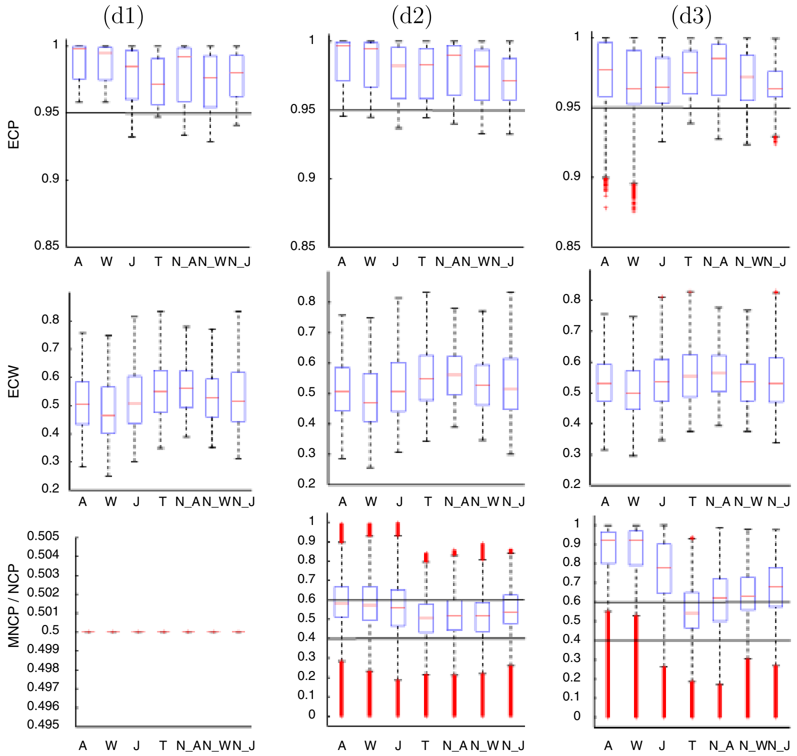

Figure 2. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of Δ for small sample designs (i.e. $10\leqslant n\leqslant 20$): (d1) $\Delta=0$; (d2) $0.01\leqslant\Delta\leqslant 0.05$; and (d3) $0.1\leqslant\Delta\leqslant 0.2$.

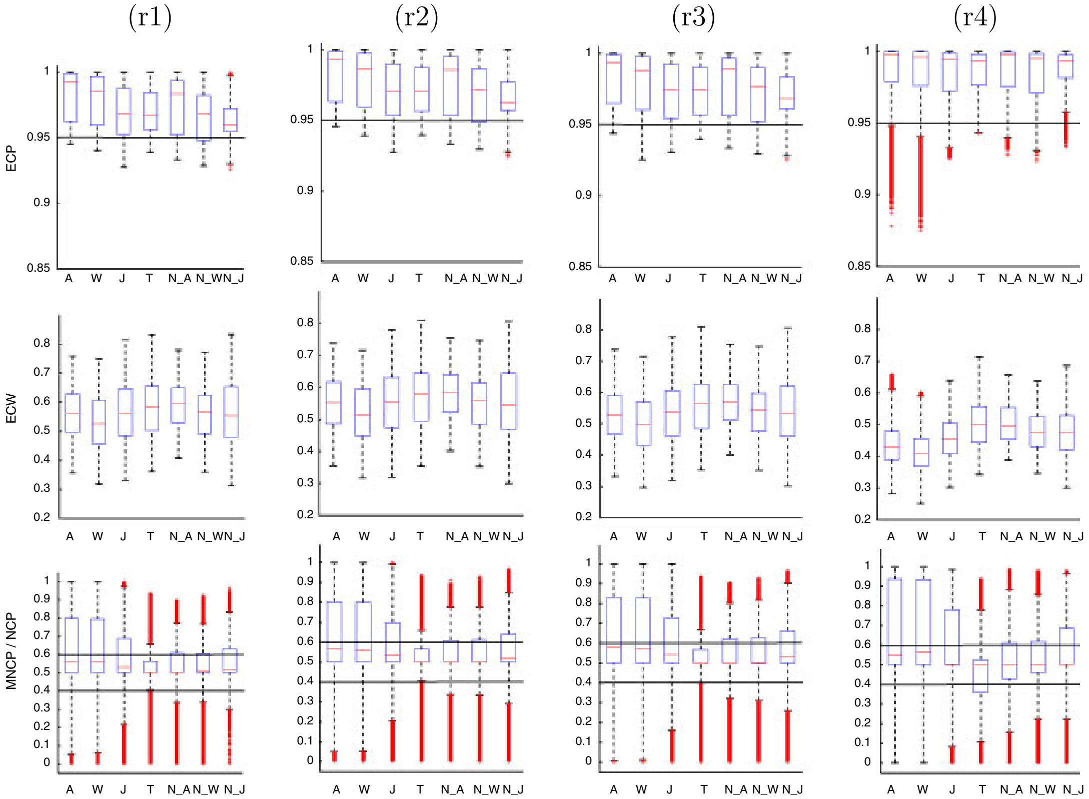

Figure 3. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of ρ for small sample designs (i.e. $10\leqslant n\leqslant 20$): (r1) $-0.1\leqslant\rho\leqslant 0$; (r2) $\rho=0$; (r3) $0\leqslant\rho\leqslant 0.2$; and (r4) $0.4\leqslant\rho\leqslant 0.6$.

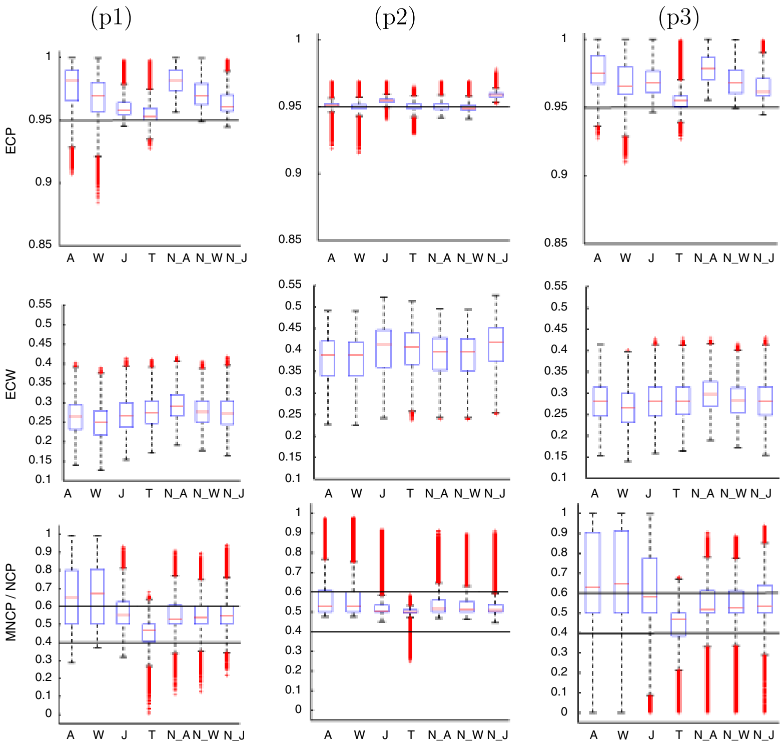

Figure 4. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of π_{+1} for moderate sample designs (i.e. $30\leqslant n\leqslant 50$): (p1) $0.05\leqslant\pi_{+1}\leqslant 0.1$; (p2) $0.4\leqslant\pi_{+1}\leqslant 0.6$; and (p3) $0.8\leqslant\pi_{+1}\leqslant 0.95$.

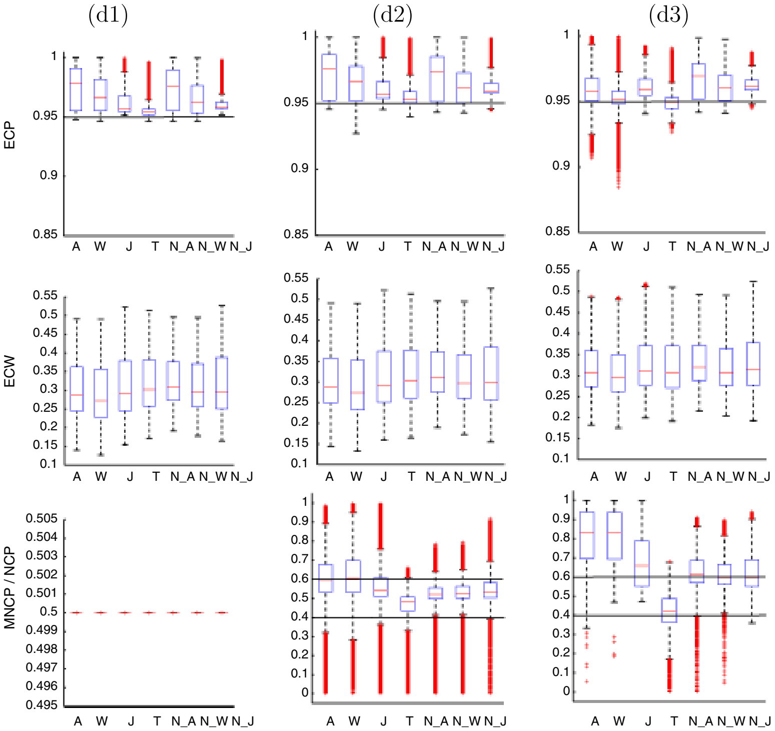

Figure 5. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of Δ for moderate sample designs (i.e. $30\leqslant n\leqslant 50$): (d1) $\Delta=0$; (d2) $0.01\leqslant\Delta\leqslant 0.05$; and (d3) $0.1\leqslant\Delta\leqslant 0.2$.

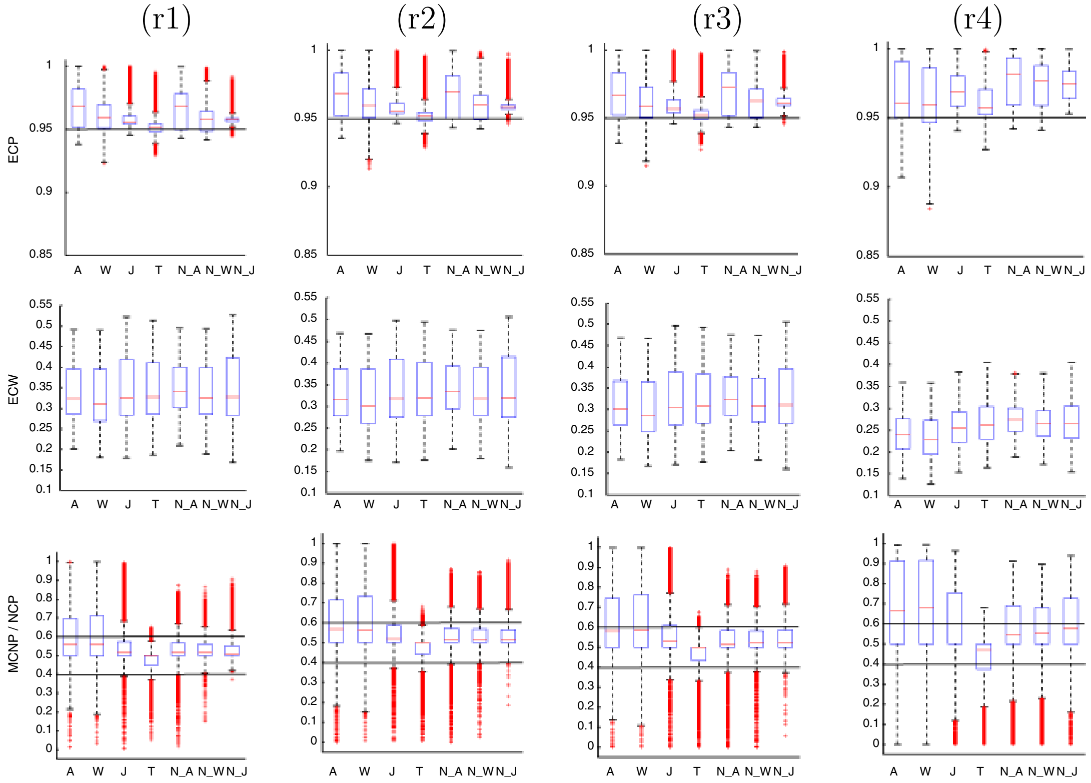

Figure 6. Boxplots of ECP, ECW, and MNCP/NCP for different CIs classified by different zones of ρ for moderate sample designs (i.e. $30\leqslant n\leqslant 50$): (r1) $-0.1\leqslant\rho\leqslant 0$; (r2) $\rho=0$; (r3) $0\leqslant\rho\leqslant 0.2$; and (r4) $0.4\leqslant\rho\leqslant 0.6$.

Table III. Table for different percentage of ECPs with different CIs based on 360 000 runs

		Methods						
Sample size	Percentage of ECP	T (per cent)	A (per cent)	W (per cent)	J (per cent)	N_A (per cent)	N_W (per cent)	N_J (per cent)
Small	>96	69.1	82.5	76.0	69.1	74.5	70.5	72.0
	94–96	30.8	16.1	21.5	27.3	24.2	25.6	27.7
	<94	0.05	1.36	2.48	3.58	1.33	3.84	0.27
Moderate	>96	15.7	60.9	47.9	39.9	67.0	55.0	46.6
	94–96	81.5	37.6	48.7	60.1	33.0	45.0	53.4
	<94	2.85	1.45	3.43	0.00	0.00	0.00	0.00

(0.4, 0.6). It is worth noting that the boxes associated with the Tango score CI are more often lying entirely inside (0.4, 0.6). It is interesting to note that MOVER CIs based on the sample Φ coefficient perform better than those based on the sample correlation coefficient in the sense that their ECPs are more close to the pre-assigned coverage level and their ECWs are shorter.

(2) *Effect of* π_{+1}: According to Figures 1 and 4, it is interesting to see that the median ECPs of all CIs are more close to the pre-asigned confidence level for moderate response group (i.e. $0.4\leqslant\pi_{+1}\leqslant 0.6$). For moderate marginal probabilities, the corresponding ECWs are generally wider, while the symmetry between MNCP and DNCP is more obvious. For small and large marginal probabilities, the ECPs of all CIs tend to be more conservative (i.e. much greater than 0.95) and the ranges for the ratios of MNCPs and NCPs become wider.

(3) *Effect of* Δ: According to Figures 2 and 5, it is interesting to see that all CIs possess symmetric MNCP and DNCP when $\Delta=0$. In this case, CIs on (J) and (N_W) yield the shortest median ECWs. When Δ increases (e.g. $0.1\leqslant\Delta\leqslant 0.2$), it seems that only Tango score CI produces MNCP/NCP ratios close to 0.5. In this case, all CIs except Tango score CI tend to be distally located.

(4) *Effect of* ρ: According to Figures 3 and 6, we observe a significant drop in the ECWs for large values of ρ (e.g. $0.4\leqslant\rho\leqslant 0.6$).

5. Real examples

5.1. The disinfection systems for soft contact lenses study

We first re-visit the cross-over clinical trial of 44 subjects reported in Tango [6]. This study was designed to compare a chemical disinfection system with a thermal disinfection system for soft contact lenses. The outcome of the study was $(n, x_{11}, x_{10}, x_{01}) = (44, 43, 0, 1)$, with an observed difference of -0.023. We reanalyzed this data set using the 95 per cent CI estimates for the difference based on the seven methods and the results are summarized in Table VI. If we compare the 95 per cent CIs with the equivalence limits with $\Delta=0.1$, the Tango score CI, MOVER-Jeffreys CI based on the sample correlation coefficient, and all MOVER CIs based on the sample Φ coefficient are not entirely lying inside the interval $(-\Delta, \Delta)$, in other words, the two disinfection systems are not demonstrated to be equivalent.

5.2. The pain management study

The second example is a pre-/post-test comparison study of 14 women with gynecologic cancers in clinical trials described in Ward *et al.* [25]. In that study, the women were randomized to receive either the 'usual care' or intervention that consisted of individually tailored information about concerns (barriers) and side effect management. Moreover, each subject was assessed by the congruence between severity of pain and medication used at baseline and 1-month post-test. Table IV reports the number of acceptable and unacceptable management between baseline and 1-month post-test in the 14 women who received the intervention. It is of interest to see if there is a difference in the proportion of acceptable management between two different times. The proportion of subjects reporting acceptable pain management was 0.64 at baseline and 0.79 at 1-month post-test. The 95 per cent CI estimates for the difference between the 1-month post-test and baseline, based on various methods, are summarized in Table VI. In this study, we observe that the CI produced from MOVER with any one of six binomial CIs (except the MOVER-Jeffreys CI based on the sample Φ coefficient) is shorter than the Tango score interval.

5.3. The epidemiological study of sleeping difficulty

Finally, we consider the data analyzed by Karacan *et al.* [26]. The study was to compare a group of 32 marijuana users with 32 matched controls with respect to their sleeping difficulties (see, Table V). In that study, we have $(n, x_{11}, x_{10}, x_{01}) = (32, 4, 9, 3)$. The difference in proportions experiencing sleeping difficulties between marijuana users and the matched control group is 0.19. The 95 per cent CI estimates for the difference, based on various methods, are also summarized in Table VI. We observe that the Tango score interval is longer than CIs based on MOVER (except the MOVER-Jeffreys CI based on the sample Φ coefficient).

From the above examples, it is believed that the confidence width of Tango score interval is generally longer than CIs based on MOVER in small to moderate sample sizes.

Table IV. Number of women having acceptable versus unacceptable analgesics at baseline and 1-month post-test for the intervention in Example 5.2

	Baseline	
1-month post-test	Acceptable	Unacceptable
Acceptable	8	3
Unacceptable	1	2

Table V. Number of patients with (+) or without (−) sleeping difficulties among marijuana users and matched controls in Example 5.3

	Marijuana group	
Control group	+	−
+	4	9
−	3	16

Table VI. 95 per cent confidence intervals for several data examples.

Data example				Methods			
n	x_{11}	x_{10}	x_{01}	TANGO	A	W	J
44	43	0	1	$(-0.1181, 0.0594)$	$(-0.0997, 0.0617)$	$(-0.0906, 0.0488)$	$(-0.1063, 0.0401)$
14	8	3	1	$(-0.1670, 0.4327)$	$(-0.1379, 0.3621)$	$(-0.1331, 0.3573)$	$(-0.1373, 0.4048)$
32	4	9	3	$(-0.0271, 0.3897)$	$(-0.0274, 0.3622)$	$(-0.0255, 0.3604)$	$(-0.0270, 0.3852)$
Data example				Methods			
n	x_{11}	x_{10}	x_{01}		N_A	N_W	N_J
44	43	0	1		$(-0.1289, 0.0757)$	$(-0.1181, 0.0597)$	$(-0.1063, 0.0401)$
14	8	3	1		$(-0.1639, 0.4178)$	$(-0.1574, 0.4136)$	$(-0.1721, 0.4323)$
32	4	9	3		$(-0.0296, 0.3823)$	$(-0.0273, 0.3807)$	$(-0.0368, 0.3934)$

6. Discussion

In this article, we investigate several CIs based on hybridizing two individual Agresti–Coull, Wilson score, or Jeffreys CIs for single proportion based on the MOVER. The proposed MOVER CIs are simple and non-iterative, which make them more appealing to general applications. In the case of matched-pair designs, the sample correlation and Φ coefficients are used to estimate the correlation between the sample marginal probabilities. In general, the MOVER CIs based on the sample Φ coefficient (i.e. N_W) perform better than those based on the sample correlation coefficient in the sense that their ECPs are more close to the pre-assigned coverage level and their ECWs are shorter.

Our simulation results suggest that the MOVER-Wilson score and Jeffreys CIs are more preferable as they generally possess shorter ECWs while the their ECPs are well controlled around the desired coverage level. When the difference between the two marginal probabilities is very small (e.g. $|\Delta| \leqslant 0.1$), their MNCPs and DNCPs are very close. Therefore, the MOVER-Wilson score and Jeffreys CIs are highly recommended for those applications in which $|\Delta|$ is small (e.g. equivalence trials). On the other hand, the famous Tango score CI is recommended for applications in which the difference between the two marginal probabilities is anticipated to be large (e.g. pre- and post-treatment comparison studies). It is also worth pointing out that Tango's method disregards the distinction between x_{00} and x_{11}, and thus is essentially based on a trinomial distribution. On the other hand, all other CIs considered in this article give different results according to the split of x_{00} and x_{11}. An Excel spreadsheet, which implements the N_W CI, is freely downloadable from http://tinyurl.com/7mr754. Matlab programs that implement the proposed methodologies can be available from the second author upon request.

Acknowledgements

The authors thank Prof. R. G. Newcombe, another referee, and the associate editor for their constructive suggestions. M. L. Tang thanks Chow, Daisy Hoi-Sze for her endless encouragement. M. L. Tang's research was fully supported by two grants from the Research Grant Council of the Hong Kong Special Administrative Region (Project Nos. HKBU261007 and HKBU261508).

References

1. Miyanaga Y. Clinical evaluation of the hydrogen peroxide SCL disinfection system (SCL-D). *Japanese Journal of Soft Contact Lenses* 1994; **36**:163–173 (in Japanese).
2. Newcombe RG. Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine* 1998; **17**:873–890.
3. Wilson EB. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association* 1927; **22**:209–212.
4. Miettinen O, Nurminen M. Comparative analysis of two rates. *Statistics in Medicine* 1984; **4**:213–226.
5. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1998; **17**:2635–2650.
6. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine* 1998; **17**:891–908.
7. Newcombe RG. Confidence intervals for the mean of a variable taking the values 0, 1 and 2. *Statistics in Medicine* 2003; **22**:2737–2750.
8. Tango T. Letter to the editor: improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine* 1999; **18**:3511–3513.
9. Donner A, Zou GY. Interval estimation for a difference between intraclass kappa statistics. *Biometrics* 2002; **58**:209–215.
10. Zou GY, Donner A. Construction of confidence limits about effect measures: a general approach. *Statistics in Medicine* 2008; **27**:1693–1702.
11. Zou GY. On the estimation of additive interaction by use of the four-by-two table and beyond. *American Journal of Epidemiology* 2008; **168**:212–224.
12. Zou GY, Taleban J, Huo CY. Confidence interval estimation for lognormal data with application to health economics. *Computational Statistics and Data Analysis* 2009; **53**:3755–3764.
13. Howe WG. Approximate confidence limits on the mean of $X+Y$ where X and Y are two tabled independent random variables. *Journal of the American Statistical Association* 1974; **69**:789–794.
14. Graybill FA, Wang CM. Confidence intervals on nonnegative linear combinations of variances. *Journal of the American Statistical Association* 1980; **75**:869–873.
15. Lee Y, Shao J, Chow SC. Modified large-sample confidence intervals for linear combinations of variance components: extension, theory, and application. *Journal of the American Statistical Association* 2004; **99**:467–478.
16. Zou GY. Toward using confidence intervals to compare correlations. *Psychological Methods* 2007; **12**:399–413.
17. Zou GY, Huo CY, Taleban J. Simple confidence intervals for lognormal means and their differences with environmental applications. *Environmetrics* 2008; **20**:172–180.
18. Ramasundarahettige CF, Donner A, Zou GY. Confidence interval construction for a difference between two dependent intraclass correlation coefficients. *Statistics in Medicine* 2009; **28**:1041–1053.
19. Zou GY, Huo CY, Taleban J. A note on confidence interval estimation for a linear function of binomial proportions. *Computational Statistics and Data Analysis* 2009; **53**:1080–1085.
20. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statistical Science* 2001; **16**:101–133.
21. Agresti A, Coull BA. Approximate is better than 'exact' for interval estimation of binomial proportions. *American Statistician* 1998; **52**:119–126.
22. Piegorsch WW. Sample sizes for improved binomial confidence intervals. *Computational Statistics and Data Analysis* 2004; **46**:309–316.
23. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine* 1998; **17**:857–872.
24. Newcombe RG. Measures of location for confidence intervals for proportions. *Technical Report*, Department of Primary Care and Public Health, Cardiff University, 2009.
25. Ward S, Donovan HS, Owen B, Grosen E, Serlin R. An individualized intervention to overcome patient-related barriers to pain management in women with gynecologic cancers. *Research in Nursing and Health* 2000; **23**:393–405.
26. Karacan I, Fernandez SA, Coggines WS. Sleep electrocephalographic–electrooculographic characteristics of chronic marijuana users: part 1. *New York Academy of Science* 1976; **282**:348–374.