Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2016, Article ID 7141050, 7 pages
http://dx.doi.org/10.1155/2016/7141050

Research Article

Alternative Confidence Interval Methods Used in the Diagnostic Accuracy Studies

Semra Erdoğan and Orekıcı Temel Gülhan

Department of Biostatistics and Bioinformatics, Faculty of Medicine, Mersin University, 33343 Mersin, Turkey

Correspondence should be addressed to Semra Erdoğan; semraerdogann@gmail.com

Received 29 February 2016; Revised 11 May 2016; Accepted 5 June 2016

Academic Editor: Po-Hsiang Tsui

Background/Aim. It is necessary to decide whether the newly improved methods are better than the standard or reference test or not. To decide whether the new diagnostics test is better than the gold standard test/imperfect standard test, the differences of estimated sensitivity/specificity are calculated with the help of information obtained from samples. However, to generalize this value to the population, it should be given with the confidence intervals. The aim of this study is to evaluate the confidence interval methods developed for the differences between the two dependent sensitivity/specificity values on a clinical application. *Materials and Methods*. In this study, confidence interval methods like Asymptotic Intervals, Conditional Intervals, Unconditional Interval, Score Intervals, and Nonparametric Methods Based on Relative Effects Intervals are used. Besides, as clinical application, data used in diagnostics study by Dickel et al. (2010) has been taken as a sample. *Results*. The results belonging to the alternative confidence interval methods for Nickel Sulfate, Potassium Dichromate, and Lanolin Alcohol are given as a table. *Conclusion*. While preferring the confidence interval methods, the researchers have to consider whether the case to be compared is single ratio or dependent binary ratio differences, the correlation coefficient between the rates in two dependent ratios and the sample sizes.

1. Introduction

Today in health field, whether an individual is ill or not, the validation of the existence or nonexistence of any kind of illnesses, information about the prognostics of the diagnosed illness, and in some cases the specification of the response to the treatment may be asked. For this reason, many laboratory methods, clinical observations, or many visualization techniques are needed. Many techniques like biochemical tests and bacteria culture tests are evaluated as diagnostics tests, which are highly important in medical field.

Parallel to the advances in science and technology, in medical services, as an alternative to the existing methods, new diagnostic and treatment methods continue to develop. There is no doubt that every work necessitates different tool, way, and method. If the success needs to be achieved, the one suitable for that specific work has to be chosen [1]. Yet there are still many difficulties in the alternative diagnostics tests and treatment applications in medicine and their comparative evaluation. There may be more than one test used in the diagnostics of an illness; yet the absolute ill or absolute healthy diagnostics for the individuals is found by only one of these tests, which is called gold standard method. However, many cases sometimes occur in the testing of some illnesses. The lack of a reference test, the application of reference test's being difficult and expensive, the physical or emotional negative effects of applied methods on patients, and waiting for reference test results to start treatment are only some of these cases stated above. In such cases, as an alternative to the reference test, imperfect reference tests having a certain rate of mistake are preferred. It is a process of decision making to analyze the questions like whether a suggested new diagnostics test is better than the imperfect diagnostics tests in the recognition of patients, or whether one of the visualization techniques is more practical than the other. In this decision making process, the validity of the diagnosis and also the reliability of it in distinguishing the reality are necessary. Besides, it is also crucially important that the diagnosis power of the diagnostics tests is enhanced and the test is applied at the right time. Only in that way would it be possible not to apply unnecessary treatment to the healthy individuals, to apply the necessary treatment to

the people carrying the disease, and not to take the healthy individual into an unnecessary operation [2, 3].

While evaluating the applicability and accuracy of the diagnostics tests, some statistical criteria about the validity of the enhanced diagnostics are used. In the most basic definition, the term "diagnostic accuracy" is used to define how close the result of the technique used to differentiate ill and healthy individuals and the true situation are to each other. As for the statistical criteria about the "diagnosis accuracy," sensitivity, specificity, positive and negative predictive values, odds rates, likelihood ratios, and receiver operating characteristic curve (ROC) can be given [2, 4]. In this study, as the performance criteria of the diagnosis tests, sensitivity and specificity will be taken into consideration. Sensitivity means true positive rate, which is the ability of the diagnosis test to determine the patients correctly among the true patients. Specificity means true negative rate, which is the ability of the diagnosis test to determine the healthy among the true healthy individuals.

A newly improved test needs to be compared with a gold standard test and in such studies, each test is applied on each individual to reduce variability. This kind of studies is called paired design. To decide whether the new diagnostics test is better than the gold standard test/imperfect standard test, the differences of estimated sensitivity/specificity are calculated with the help of information obtained from samples. However, to generalize this value to the population, it should be given with the confidence intervals. Sensitivity and specificity criteria are a proportion of having binary results. For this reason, confidence intervals methods will be used in this study for paired binomial proportions [3, 5]. Furthermore, the aim in this study is to evaluate the Wald confidence interval methods in literature for paired binomial proportions and other confidence interval methods developed as alternatives to them in a clinical application.

2. Material and Methods

2.1. Statistical Model and General Notations. The methods given for the difference of two dependent sensitivities and the obtained results are similar with the methods given for the two dependent specificities and the obtained results. Thus here, only the difference of two sensitivities and the belonging CI methods will be given. Assume a gold standard test is compared with a newly improved test in the existence of binary results as positive/negative (ill/healthy) in the diagnosis of an illness. When each individual is evaluated with both tests, a two-by-two table representing the results of both tests is below (Table 1). When Table 1 is examined, while cell a gives the number of "positives" in both measurements and cell d gives the number of "negatives" in both measurements, cells b and c show the number of discordances as a result of both measurements. Besides, while the proportions are given in the table as $p_{11} = a/n$, $p_{12} = b/n$, $p_{21} = c/n$, and $p_{22} = d/n$, p_1 and p_2 show the positive proportions and are calculated as $(a + c)/n$ and $(a + b)/n$, respectively [5–7].

$D_i = 1$ and $i = 1, 2, \ldots$ show the true patients for n individuals and $X_{ij} = 1$ $(j = 1, 2)$ show the individuals identified as patient as a result of two diagnostics tests. The

Table 1: A two-by-two table representing concordance/discordance of both diagnostic tests.

	Diagnostic test 1 (gold standard)		Total
	Positive	Negative	
Diagnostic test 2 (new test)			
Positive	a $(p_{11} = a/n)$	b $(p_{12} = b/n)$	$a + b$ $(p_2 = (a + b)/n)$
Negative	c $(p_{21} = c/n)$	d $(p_{22} = d/n)$	$c + d$
Total	$a + c$ $[p_1 = (a + c)/n]$	$b + d$	n

true sensitivity value for the diagnosis test 1 is shown as se_1 and the true sensitivity value for the diagnosis test 2 is shown as se_2. Estimated sensitivity values of the diagnostics tests are calculated as $\widehat{\mathrm{se}}_1 = (a + c)/n$ and $\widehat{\mathrm{se}}_2 = (a + b)/n$. The null hypothesis of equality of two dependent sensitivities and corresponding alternative hypothesis can be written as [5, 6]

$$H_0: \theta = \mathrm{se}_2 - \mathrm{se}_1 = 0,$$
$$H_1: \theta = \mathrm{se}_2 - \mathrm{se}_1 \neq 0. \tag{1}$$

The estimating value of the difference between two sensitivities is calculated as $\widehat{\theta} = \widehat{\mathrm{se}}_2 - \widehat{\mathrm{se}}_1 = (a + b)/n - (a + c)/n = (b - c)/n$ and if the CI includes "0," H_0 hypothesis is accepted; but if the CI does not include "0," H_0 hypothesis is rejected [5].

The probabilities of discordance of both diagnostics tests are denoted as p_{12} and p_{21}, and estimated overall discordance value is calculated as in [5, 6]

$$\widehat{\psi} = p_{12} + p_{21} = \frac{b}{n} + \frac{c}{n} = \frac{b + c}{n}. \tag{2}$$

2.2. Methods. Methods for the CI estimations can be classified as Asymptotic Intervals (Wald Interval, Continuity Corrected Wald Interval, Wald with Agresti-Min pseudo-frequency adjustment, and Wald with Bonett-Price Laplace Adjustment), Conditional Intervals (Exact conditional method, Mid-p Conditional Methods), Unconditional Interval (Unconditional True Profile Likelihood Method), Score Intervals (Wilson Score Interval without Continuity Correction, Wilson Score Interval with Continuity Correction to Score Limits, Wilson Score Interval with Continuity Correction to ψ, and Tango Asymptotic Score), and Nonparametric Methods Based on Relative Effect Intervals (Rank-based CI with Normal Approximation, Rank-based CI with t-Approximation).

2.2.1. Asymptotic Wald Interval (Wald). This is a CI method proposed by D. G. Altman in 1991 and is calculated as (3). This

method is also based on Central Limit Theorem and normal distribution approach [5, 6]:

$$\mathrm{CI}_{\alpha/2}\left(\widehat{\theta}\right) = \left[\widehat{\theta} \pm Z_{\alpha/2}\frac{1}{n}\sqrt{b + c - \frac{(b-c)^2}{n}}\right]. \quad (3)$$

2.2.2. Continuity Corrected Wald Interval (Wald.cc). With $1/n$ continuity corrected, Asymptotic Wald CI is calculated as in [5]

$$\mathrm{CI}_{\alpha/2}\left(\widehat{\theta}\right) = \left[\widehat{\theta} \pm \left(Z_{\alpha/2}\frac{1}{n}\sqrt{b + c - \frac{(b-c)^2}{n}} + \frac{1}{n}\right)\right]. \quad (4)$$

2.2.3. Wald with Agresti-Min Pseudo-Frequency Adjustment (Wald + N, Agresti). Before using the Wald interval method, Agresti and Min enhanced the efficiency of this method by adding pseudo numbers ($n = 1, 2, 3, 4$) to the cells observed in Table 1. With this aim, the best performance is observed Wald + 2 according to the results of the simulation applied. According to this, 2 is added overall in every sample and this means adding every cell $1/2$ $(n/4 = 2/4)$. Thus, $\widehat{\theta} = (b-c)/n = ((b + 1/2) - (c + 1/2))/(n + 2) = (b - c)/(n + 2)$ and new CI method is obtained as in [5, 6, 8]

$$\mathrm{CI}_{\alpha/2}\left(\frac{b-c}{n+2}\right) = \left[\left(\frac{b-c}{n+2}\right) \pm \left(Z_{\alpha/2}\frac{1}{n+2}\sqrt{b + c + 1 - \frac{(b-c)^2}{n+2}}\right)\right]. \quad (5)$$

2.2.4. Wald with Bonett-Price Laplace Adjustment (Bonett-Price). This is a method developed by Bonett and Price in 2004. According to this method, Laplace estimations are calculated first and then CI is calculated. Laplace estimators are calculated as $\widehat{p}_{ij} = (f_{ij} + 1)/(n + 2)$. In such a case, $\widehat{p}_{12} = (b+1)/(n+2)$ and $\widehat{p}_{21} = (c+1)/(n+2)$, and Bonett-Price corrected Wald method is calculated as in [8, 9]

$$\mathrm{CI}_{\alpha/2}\left(\frac{b-c}{n+2}\right) = \left[\left(\frac{b-c}{n+2}\right) \pm \left(Z_{\alpha/2}\frac{1}{(n+2)}\sqrt{b + c + 2 - \frac{(b-c)^2}{(n+2)}}\right)\right]. \quad (6)$$

2.2.5. Exact Conditional Intervals (Exact.cond). CI for exact conditional method is defined as in (7). Here, $[\mathrm{exact}_l(\mu), \mathrm{exact}_u(\mu)]$ is Clopper-Pearson Exact CI for $\mu = \pi_2/(\pi_2 + \pi_3)$ [5, 10–12]:

$$\mathrm{CI}_{1-\alpha/2}\left(\widehat{\theta}\right) = \left[\left(2 \cdot \mathrm{exact}_l\left(\mu\right) - 1\right)\widehat{\psi}; \left(2 \cdot \mathrm{exact}_u\left(\mu\right) - 1\right)\widehat{\psi}\right]. \quad (7)$$

*2.2.6. Mid-p Conditional Intervals (Exact.*mid*p).* This approach is similar to exact conditional confidence interval in terms of calculation steps but here, $[\mathrm{mid}p_l(\mu), \mathrm{mid}p_u(\mu)]$ has to be used for μ and thus CI is formulated as in [5, 12, 13]

$$\mathrm{CI}_{1-\alpha/2}\left(\widehat{\theta}\right) = \left[\left(2 \cdot \mathrm{mid}p_l\left(\mu\right) - 1\right)\widehat{\psi}; \left(2 \cdot \mathrm{mid}p_u\left(\mu\right) - 1\right)\widehat{\psi}\right]. \quad (8)$$

2.2.7. Unconditional True Profile Likelihood Method (Uncond). In this approach, the maximum likelihood estimator of ψ and inverted likelihood ratio test is given as θ and is shown as ψ_θ. The likelihood function for $\Theta = 0$ is $f_\psi = (1-\psi)^{a+d}(\psi/2)^{b+c}$ and ψ_θ has the biggest value of ψ_θ. The likelihood function for $\Theta \neq 0$ is $f_\psi = (1-\psi)^{a+d}((\psi+\theta)/2)^b((\psi-\theta)/2)^c$ and $\widehat{\psi}$ has the biggest value of $\psi_\theta = B + \sqrt{B^2 - C}$. In this equation, the calculations are as $B = 1/2\widehat{\psi} + 1/2\theta\widehat{\theta}$ and $C = \theta\widehat{\theta} - (a+d)/n\theta^2$. CI is calculated as in [5]

$$\mathrm{CI}_{1-\alpha/2}\left(\widehat{\theta}\right) = \left[\min_{\theta\in[-1;1]}\left(\frac{f\left(\theta\right)}{f\left(\theta\right)} \geq \frac{-z^2_{1-\alpha/2}}{2}\right); \max_{\theta\in[-1;1]}\left(\frac{f\left(\theta\right)}{f\left(\theta\right)} \geq \frac{-z^2_{1-\alpha/2}}{2}\right)\right]. \quad (9)$$

The calculation in this equation is defined as $f(\theta) = (a + d)(\ln(1-\psi_\theta) - \ln(1-\widehat{\psi})) + b(\ln(\psi_\theta+\theta) - \ln(\widehat{\psi}+\widehat{\theta})) + c(\ln(\psi_\theta - \theta) - \ln(\widehat{\psi} - \widehat{\theta}))$ [5].

2.2.8. Wilson Score Interval without Continuity Correction (Wilson). This method is a CI method developed by Newcombe in 1998 and which inverts the score test. The lower limit value of the confidence intervals is given as in (10), and the upper limit value is given as in (11) [5, 8, 12]:

$$L = \widehat{\theta} - \sqrt{\left(\frac{a+b}{n} - l_1\right)^2 + \left(u_2 - \frac{a+c}{n}\right)^2 - 2\psi\left(\frac{a+b}{n} - l_1\right)\left(u_2 - \frac{a+c}{n}\right)}, \quad (10)$$

$$U = \widehat{\theta} + \sqrt{\left(\frac{a+c}{n} - l_2\right)^2 + \left(u_1 - \frac{a+b}{n}\right)^2 - 2\psi\left(\frac{a+c}{n} - l_2\right)\left(u_1 - \frac{a+b}{n}\right)}. \quad (11)$$

Here, ψ is calculated as in (12). Besides, l_1, l_2, u_1, and u_2 are obtained from the quadratic roots of the equations in (13) and (14) [5, 8, 12]:

$$\psi = \begin{Bmatrix} 0 & (a+b, a+c, b+d, c+d = 0) \\ \dfrac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} & (a+b, a+c, b+d, c+d) \neq 0 \end{Bmatrix}, \tag{12}$$

$$(l_1, u_1) = \frac{2(a+b) + Z^2_{\alpha/2} \pm Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4(a+b) + (1-(a+b)/n)}}{2\left(n + Z^2_{\alpha/2}\right)}, \tag{13}$$

$$(l_2, u_2) = \frac{2(a+c) + Z^2_{\alpha/2} \pm Z_{\alpha/2}\sqrt{Z^2_{\alpha/2} + 4(a+c) + (1-(a+c)/n)}}{2\left(n + Z^2_{\alpha/2}\right)}. \tag{14}$$

2.2.9. Wilson Score Interval without Continuity Correction to Score Limits (Wilson.cc). This method is a confidence interval similar to "Wilson Score Interval without Correction" but the quadratic roots of both equations are calculated differently. The correction terms in (13) and (14) are obtained by extracting $1/2n$. Especially in the existence of small samplings, correction term is applied to Wilson score interval method [5].

2.2.10. Wilson Score Interval with Continuity Correction to ψ (Wilson.phi). Especially in the existence of small samplings, a correction term is applied not to the score limits but to ψ. This method is a confidence interval similar to "Wilson Score Interval without Correction", but after ψ is calculated not as in (12) but in (15), the lower and upper limit values of CI are calculated by putting in (10) and (11) [5, 12]:

$$\psi = \begin{Bmatrix} 0 & ad \leq bc \\ \max\left(\dfrac{ad-bc-n}{2}, 0\right) & ad > bc \end{Bmatrix}. \tag{15}$$

2.2.11. Tango Asymptotic Score (Tango). This approach was developed by Tango in 1998 and used as a score test. It has first been used as a dependent samples transformed score test. Later a hypothesis test has been formulized for differences of two rates (as the equality of two rates or the superiority of one of the rates). The confidence interval of the difference between two sensitivity values is given as in [5, 8, 14, 15]

$$T\left(\frac{n}{\theta}\right) = \frac{b - c - n\theta}{\sqrt{n\left[2\widehat{p}_{21} + \theta(1-\theta)\right]}}. \tag{16}$$

Here $\widehat{p}_{21}$ is the maximum likelihood estimator of p_{21} and is formulized as $\widehat{p}_{21} = (\sqrt{B^2 - 4AC} - B)/2A$. It is calculated as $A = 2n$, $B = -b - c + (2n - b + c)\theta$, and $C = -c\theta(1-\theta)$. Lower limit and upper limit values of Tango asymptotic score confidence interval for Θ are calculated as $T(n/L) = Z_{\alpha/2}$ and $T(n/U) = -Z_{\alpha/2}$ [8, 14, 15].

2.2.12. Rank-Based CI with Normal Approximation (np.nv). Nonparametric confidence intervals are based on ranked data and here asymptotic normal approximation is used. The detailed explanations of the method used have been put forward by Lange and Brunner in 2012 and calculated as in [5, 16]

$$\mathrm{CI}_{1-\alpha/2}\left(\widehat{\theta}\right) = \left[\widehat{\theta} \pm Z_{1-\alpha/2}\frac{\widehat{\sigma}_\theta}{\sqrt{N}}\right]. \tag{17}$$

2.2.13. Rank-Based CI with t Approximation (np.t). It has been put forward by Brunner and Munzel in 2000 for small sample sizes and is calculated as in [5, 17]

$$\mathrm{CI}_{1-\alpha/2}\left(\widehat{\theta}\right) = \left[\widehat{\theta} \pm t_{\widehat{f},1-\alpha/2}\frac{\widehat{\sigma}_\theta}{\sqrt{N}}\right]. \tag{18}$$

The degrees of freedom for the quartile of t-distribution are calculated as $\widehat{f} = (\widehat{\sigma}_1^2 + \widehat{\sigma}_2^2)^2/((\widehat{\sigma}_1^2 + \widehat{\sigma}_2)/(N-1))$. $\widehat{\sigma}_1^2$ and $\widehat{\sigma}_2^2$ here show the diagonal elements of covariance matrix [5, 17].

3. Results

3.1. Clinical Application. In this study, the data used by Dickel et al. in diagnostics study has been taken as samples. According to this study, a total of 790 patients have applied the strip patch test (SPT) as an alternative to patch test (PT) accepted as gold standard test in allergies diagnostics and they have been recorded as being nonallergic. Three different main components as Nickel Sulfate (Ni), Potassium Dichromate (Cr), and Lanolin Alcohol (La) have been used to examine the accuracy of the tests and confidence intervals for differences of the two dependent sensitivities have been calculated for these tests [18]. 122 patients have been observed as being sensitive to Ni, 28 patients to Cr, and 8 patients to La and cross tables of situations concordance and discordance according to both diagnostics tests for each allergen are given in Table 2. It has been calculated that the difference of sensitivity of 122 patients sensitive to Ni is 0.164 and 95% CI is 0.087–0.241, the difference of sensitivity of 28 patients sensitive to Cr is 0.250

Table 2: A table representing concordance/discordance of both diagnostic tests of SPT and PT only for patients sensitive to Ni, Cr, and La.

	PT		Total
	+	−	
Nickel Sulfate (Ni)			
SPT			
+	59	23	82
−	3	37	40
Total	62	60	122
Potassium Dichromate (Cr)			
SPT			
+	10	7	17
−	0	11	11
Total	10	18	28
Lanolin Alcohol (La)			
SPT			
+	0	1	1
−	0	7	7
Total	0	8	8

and 95% CI is 0.090–0.410, and the difference of sensitivity of 8 patients sensitive to La is 0.125 and 95% CI is −0.104–0.354.

The result of all CI methods is given in Table 3. When the table is analyzed, for 3 different parameters, the same results have been taken in point estimations of differences with all methods except for Agresti's intervals.

For all three parameters, the methods having the narrowest CI have been observed in conditional confidence interval methods and exact.mid$p$ method. Conditional confidence interval methods are not affected by the sample size like other methods. When examined in terms of Ni, while CI methods supply similar results, Wald.cc is seen to have a relatively wider CI compared to the other methods. When examined from the point of Cr, there are differences among the confidence intervals of the methods and the methods having the wider confidence interval are Wald.cc and Np.t confidence interval methods. However, when examined from the point of La, as the CI includes "0" in all methods, H_0 hypothesis is accepted and the differences between the two sensitivities are not meaningful. No matter which CI method is used, it can be said that SPT is a very good test for Ni and Cr but that it is not such a distinguishing test for testing La.

4. Discussion

It has been noticed by some researchers that Wald method, traditional asymptotic confidence interval for the differences between the values of two sensitivities/specificities, is not enough in terms of some evaluation criteria and many CI methods have been developed as an alternative to this method [5, 6, 8, 9, 12, 14, 19, 20].

Newcombe has designed a simulation study about the asymptotic, conditional, unconditional, and score confidence intervals and the performances of these confidence interval methods have been evaluated. According to this, Newcombe advises the researchers to use conditional confidence interval methods (exact cond. and exact mid$p$.) in the cases of one rate and unconditional confidence interval (uncond) and Wilson score confidence interval methods in the cases of dependent rate differences [12].

Wenzel and Zapf have compared different CI methods for the differences between two dependent sensitivities/specificities. When simulation results are examined according to sample size, for all the cases, it has been observed that Tango and nonparametric confidence interval methods (np.nv, np.t) give the best results. Besides, the performances of confidence interval methods dependent on the correlation coefficient (between 0.20 and 0.80) between the two sensitivity values have been studied in these simulations. According to this, it has been observed that the performances of Wald cc. exact cond. and Wilson.cc methods are conservative in all scenarios and the methods go worse as the correlation value increases and that exact mid$p$, Agresti, and Wilson.phi methods are conservative independent of the correlation. While uncond method is slightly anticonservative in all scenarios, it goes worse as the correlation increases; it has been stated that Wilson, Tango, and np.t methods start as slightly anticonservative but they show a slightly conservative case as the correlation value increases. Wenzel and Zapf, based on all of the possible scenarios they were planning, recommend the researchers to use Wald, Tango, and np.t confidence interval methods among confidence interval methods of difference between two dependent sensitivity studies [5].

Agresti and Min have developed Wald confidence interval method with difference modifications and called that Wald + N method. In their study, they have tried Wald, Wald + 2, and score confidence intervals for different combinations and have expressed that Wald + 2 methods have a narrower confidence interval method compared to score confidence interval for small samples [6].

Fagerland et al. have made a simulation study on the performances of asymptotic, unconditional, and score confidence interval methods. They have showed that when asymptotic confidence intervals are examined, the performances of Wald and Wald.cc methods are very weak in small samples ($n = 15$), and as sample number increases ($n = 40$), though not very successful, they have a slightly better performance. Besides, while Agresti CI performance value falls below the nominal level 95% in small samples, performance value in Bonett-Price CI falls below the nominal level as the sample number increases. They have also emphasized that Wilson and Tango score interval methods are quite conservative in small samples, that Wilson confidence interval has a narrower confidence interval compared to Tango confidence interval, that Bonett-Price confidence interval has a narrower confidence interval compared to Wilson and Tango score confidence interval, and that unconditional confidence interval method has a wider confidence interval compared to the other methods [8].

Bonett and Price have studied the performances of asymptotic and score confidence intervals as an alternative in small samples for dependent proportion differences. In

Table 3: The results of all confidence interval methods for different sample sizes.

	Nickel Sulfate (Ni) (n = 122)		Potassium Dichromate (Cr) (n = 28)		Lanolin Alcohol (La) (n = 8)	
	se_1 = 0.508	se_2 = 0.672	se_1 = 0.357	se_2 = 0.607	se_1 = 0.000	se_2 = 0.125
	θ	CI	θ	CI	θ	CI
Asymptotic intervals						
Wald	0.164	(0.087, 0.241)	0.25	(0.090, 0.410)	0.125	(−0.104, 0.354)
Wald.cc	0.164	(0.079, 0.249)	0.25	(0.054, 0.446)	0.125	(−0.229, 0.479)
Agresti	**0.161**	(0.084, 0.238)	**0.23**	(0.068, 0.398)	**0.100**	(−0.170, 0.370)
Bonett-Price	0.164	(0.083, 0.240)	0.25	(0.056, 0.411)	0.125	(−0.234, 0.434)
Conditional intervals						
Exact.cond	0.164	**(0.085, 0.203)**	0.25	**(0.045, 0.250)**	0.125	**(−0.119, 0.125)**
Exact.mid$p$	0.164	**(0.093, 0.200)**	0.25	**(0.076, 0.250)**	0.125	**(−0.113, 0.125)**
Unconditional interval						
Uncond	0.164	(0.090, 0.245)	0.25	(0.111, 0.428)	0.125	(−0.142, 0.445)
Score intervals						
Wilson	0.164	(0.086, 0.238)	0.25	(0.086, 0.388)	0.125	(−0.215, 0.471)
Wilson.cc	0.164	(0.083, 0.241)	0.25	(0.070, 0.400)	0.125	(−0.294, 0.533)
Wilson.phi	0.164	(0.085, 0.239)	0.25	(0.071, 0.400)	0.125	(−0.215, 0.471)
Tango	0.164	(0.090, 0.247)	0.25	(0.099, 0.434)	0.125	(−0.240, 0.471)
Nonparametric methods based on relative effects						
np.nv	0.164	(0.087, 0.241)	0.25	(0.087, 0.413)	0.125	(−0.120, 0.370)
np.t	0.164	(0.087, 0.241)	0.25	(0.083, 0.47)	0.125	(−0.171, 0.421)

their studies, they express that classical Wald method show a weak performance, that Tango method is better than Wald.cc method, and that Agresti method has a better performance than Tango and Wald.cc methods [9].

Tango has compared conditional (uncond), unconditional (exact cond, exact mid$p$), and Tango methods and has put forward that Tango method has a more conservative confidence interval for small samples and that in other conditions it shows a better performance. He has also put forward that, besides Tango method, uncond method has shown a better performance in large samples, too [14].

Tang et al. has compared conditional and score confidence interval methods for single rate and small samples. In their studies, they have put forward the fact that conditional confidence interval methods have shown better performance than score confidence interval methods. Another research of Tang et al. (2010) has found that the performances of Tango and Wilson score methods are better in small sample sizes with the comparison of confidence intervals for the differences between two dependent proportions, advising the researchers to use these methods. Tang et al. have found that the performances of Tango and Wilson score methods are better in small sample sizes and have advised the researchers to use these methods [19, 20].

When studied in terms of point estimations, our clinical data show similar results to all other confidence interval methods except for Agresti method and it has been observed that the only method that is not affected by the sample size is conditional confidence interval methods.

It is true that, from the CI methods, Wald methods are preferred by the researchers because of having easier calculation steps compared to the other methods. While preferring the confidence interval methods, the researchers have to consider whether the case to be compared is single ratio or dependent binary ratio differences, the correlation coefficient between the rates in two dependent ratios and the sample sizes.

According to the simulation results, in the literature, when a single ratio comparison has been made, no matter what the sample size is, in the comparisons belonging to the dependent ratio comparison of the conditional confidence interval methods (exact cond., exact mid$p$.), unconditional confidence interval method and Wilson and Wilson phi confidence interval method of the score confidence interval methods should be preferred [5, 6, 8, 9, 12, 14, 19, 20]. Besides, in the cases where dependent ratio differences exist, without looking at the sample sizes, Tango score confidence interval and nonparametric confidence interval methods (np.nv, np.t) are seen to give the best results. The use of Wald and np.nv methods depending on the correlation coefficient between the two sensitivity values is advised [5].

Disclosure

A part of this study was presented as a poster at the 16th National Congress of Biostatistics Antalya, September 2014.

Competing Interests

The authors declare that they have no competing interests.

References

[1] E. Erdoğan, *Bilim ve Metafizik Üzerine Tarihsel Bir Soruşturma*, Arkeoloji ve Sanat Yayınları, İstanbul, Turkey, 1st edition, 2011.

[2] X. H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine*, Wiley Series in Probability and Statistics, Wiley-Interscience, New York, NY, USA, 1st edition, 2002.

[3] G. Yasemin, "Tani testi çalişmalarinda metodolojik standartlarin kullanilmasi," *Ankara Üniversitesi tip Fakültesi Mecmuasi*, vol. 56, no. 4, pp. 259–264, 2003.

[4] D. M. Hawkins, J. A. Garrett, and B. Stephenson, "Some issues in resolution of diagnostic tests using an imperfect gold standard," *Statistics in Medicine*, vol. 20, no. 13, pp. 1987–2001, 2001.

[5] D. Wenzel and A. Zapf, "Difference of two dependent sensitivities and specificities: comparison of various approaches," *Biometrical Journal*, vol. 55, no. 5, pp. 705–718, 2013.

[6] A. Agresti and Y. Min, "Simple improved confidence intervals for comparing matched proportions," *Statistics in Medicine*, vol. 24, no. 5, pp. 729–740, 2005.

[7] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, New York, NY, USA, 1st edition, 1990.

[8] M. W. Fagerland, S. Lydersen, and P. Laake, "Recommended tests and confidence intervals for paired binomial proportions," *Statistics in Medicine*, vol. 33, no. 16, pp. 2850–2875, 2014.

[9] D. G. Bonett and R. M. Price, "Adjusted wald confidence interval for a difference of binomial proportions based on paired data," *Journal of Educational and Behavioral Statistics*, vol. 37, no. 4, pp. 479–488, 2012.

[10] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934.

[11] A. Agresti and B. A. Coull, "Approximate is better than 'exact' for interval estimation of binomial proportions," *The American Statistician*, vol. 52, no. 2, pp. 119–126, 1998.

[12] R. G. Newcombe, "Improved confidence intervals for the difference between binomial proportions based on paired data," *Statistics in Medicine*, vol. 17, no. 22, pp. 2635–2650, 1998.

[13] S. E. Vollset, "Confidence intervals for a binomial proportion," *Statistics in Medicine*, vol. 12, no. 9, pp. 809–824, 1993.

[14] T. Tango, "Equivalence test and confidence interval for the difference in proportions for the paired-sample design," *Statistics in Medicine*, vol. 17, no. 8, pp. 891–908, 1998.

[15] Z. Yang, X. Sun, and J. W. Hardin, "A non-iterative implementation of Tango's score confidence interval for a paired difference of proportions," *Statistics in Medicine*, vol. 32, no. 8, pp. 1336–1342, 2013.

[16] K. Lange and E. Brunner, "Sensitivity, specificity and ROC-curves in multiple reader diagnostic trials—a unified, nonparametric approach," *Statistical Methodology*, vol. 9, no. 4, pp. 490–500, 2012.

[17] E. Brunner and U. Munzel, "The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation," *Biometrical Journal*, vol. 42, no. 1, pp. 17–25, 2000.

[18] H. Dickel, B. Kreft, O. Kuss et al., "Increased sensitivity of patch testing by standardized tape stripping beforehand: a multicentre diagnostic accuracy study," *Contact Dermatitis*, vol. 62, no. 5, pp. 294–302, 2010.

[19] M.-L. Tang, M.-H. Ling, L. Ling, and G. Tian, "Confidence intervals for a difference between proportions based on paired data," *Statistics in Medicine*, vol. 29, no. 1, pp. 86–96, 2010.

[20] M.-L. Tang, N.-S. Tang, and I. S. F. Chan, "Confidence interval construction for proportion difference in small-sample paired studies," *Statistics in Medicine*, vol. 24, no. 23, pp. 3565–3579, 2005.