Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2017, Vol. 26(2) 615–632

DOI: 10.1177/0962280214552881
journals.sagepub.com/home/smm
SAGE

Exact one-sided confidence limits for Cohen's kappa as a measurement of agreement

Guogen Shan[1] and Weizhen Wang[2,3]

Abstract

Cohen's kappa coefficient, κ, is a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative items. In this paper, we focus on interval estimation of κ in the case of two raters and binary items. So far, only asymptotic and bootstrap intervals are available for κ due to its complexity. However, there is no guarantee that such intervals will capture κ with the desired nominal level $1-\alpha$. In other words, the statistical inferences based on these intervals are not reliable. We apply the Buehler method to obtain exact confidence intervals based on four widely used asymptotic intervals, three Wald-type confidence intervals and one interval constructed from a profile variance. These exact intervals are compared with regard to coverage probability and length for small to medium sample sizes. The exact intervals based on the Garner interval and the Lee and Tu interval are generally recommended for use in practice due to good performance in both coverage probability and length.

Keywords

Buehler method, coverage probability, exact confidence interval, order, expected length

1 Introduction

In scientific studies, it is often the case that multiple raters are available to assess subjects. A typical example is that independent raters assess each subject with binary outcomes (e.g. Yes/No and Response/No Response). For the case with two raters A and B, the data can be organized in a 2×2 contingency table as from a matched pairs experiment, see Table 1. The observation vector, $\underline{X} = (n_{11}, n_{10}, n_{01})$, follows a multinomial distribution with N independent and identical trials and probabilities p_{11}, p_{10} and p_{01}, respectively. For example, n_{10} is the number of subjects on which the

[1]Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, USA
[2]College of Applied Sciences, Beijing University of Technology, Beijing, PR China
[3]Department of Mathematics and Statistics, Wright State University, Dayton, USA

Corresponding author:
Guogen Shan, Epidemiology and Biostatistics Program, Department of Environmental and Occupational Health, School of Community Health Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154, USA.
Email: guogen.shan@unlv.edu

Table 1. Observations n_{ij} and associated probabilities (p_{ij}) for two independent raters assessing N subjects with binary outcomes.

	The rater B		
The rater A	1	0	Total
1	n_{11}	n_{10}	$n_{1*}=n_{11}+n_{10}$
	(p_{11})	(p_{10})	($p_{1*}=p_{11}+p_{10}$)
0	n_{01}	n_{00}	$N-n_{1*}$
	(p_{01})	(p_{00})	($1-p_{1*}$)
Total	$n_{*1}=n_{11}+n_{01}$	$N-n_{*1}$	N
	($p_{*1}=p_{11}+p_{01}$)	($1-p_{*1}$)	

two raters' assessment is $(A,B)=(1,0)$ and $p_{10}=P(A=1,B=0)$. The probability mass function of $\underline{X}$ is

$$p_m(\underline{x};p_{11},p_{10},p_{01})=\frac{N!}{n_{11}!n_{10}!n_{01}!n_{00}!}p_{11}^{n_{11}}p_{10}^{n_{10}}p_{01}^{n_{01}}p_{00}^{n_{00}} \tag{1}$$

on a sample space

$$S=\left\{\underline{x}=(n_{11},n_{10},n_{01}):0\le n_{11}+n_{10}+n_{01}\le N,n_{ij}\text{ is a nonnegative integer}\right\} \tag{2}$$

with a total of

$$m=\frac{(N+1)(N+2)(N+3)}{6} \tag{3}$$

sample points, and the parameter space is given as

$$H=\left\{(p_{11},p_{10},p_{01},p_{00}):p_{ij}\in[0,1],(i=0,1;j=0,1);p_{11}+p_{10}+p_{01}+p_{00}=1\right\} \tag{4}$$

The parameter space actually is a three-dimensional space because of the one constraint.

It is often of interest to perform statistical inference regarding the agreement between the two raters, i.e. whether the rater A and the rater B provide the same assessment, (1,1) and (0,0). The simple percent agreement is defined as

$$p_s=P(A=B)=P(A=1=B)+P(A=0=B)=p_{11}+p_{00}$$

Intuitively, we would expect to see a large value of p_s for a high level of agreement between the two raters. However, p_s does not account for the chance agreement that is measured by

$$p_c=P(A=B|\text{A and B are independent})=p_{1*}p_{*1}+(1-p_{1*})(1-p_{*1})$$

where p_{1*} and p_{*1} are marginal probabilities, see Table 1.

So Cohen[1] proposed the Cohen's kappa coefficient

$$\kappa=\frac{p_s-p_c}{1-p_c} \tag{5}$$

This is a measurement combining both the simple percent agreement and the chance agreement and is the most widely used measurement for agreement. It has the range $[-1,1]$: $\kappa = 1$ if and only if $p_s = 1$, i.e. there is complete agreement between the two raters; $\kappa = -1$ if $p_{10} = p_{01} = 0.5$, one case of complete disagreement; $\kappa > 0$ if p_s exceeds p_c; and $\kappa < 0$ if p_s is less than p_c. This coefficient can be alternatively written as

$$\kappa = \frac{2(p_{11} - p_{1*}p_{*1})}{p_{1*} + p_{*1} - 2p_{1*}p_{*1}} \tag{6}$$

Note that κ is a monotone function of p_{11} given the marginal probabilities, and it is obvious that the range of κ depends on the marginal probabilities p_{1*} and p_{*1}. The lower and upper bounds for κ for given marginal probabilities can be found in Lee and Tu.[2] They discussed the relationship between the range of κ and the marginal probabilities of agreement by providing various graphical interpretations. The goal of this paper is to estimate κ using confidence intervals based on $\underline{X}$. To better understand the problem, consider the following two examples.

Example 1. A clinical study (see Kilpikoski et al.[3]) was conducted to examine the inter-examiner reliability using the McKenzie method for assessing subjects with low back pain. In the study, 39 ($=N$) subjects were assessed by two independent physical therapists (Clinician A and Clinician B), and a binary decision was made from each clinician, either low back pain was present or absent. It is often important to test whether the two clinicians have the same diagnostic conclusion. The observation is $\underline{x} = (28, 3, 6)$ that shows an agreement on 30 out of 39 subjects, see Table 2. We will assess the agreement by using confidence intervals for κ in Section 4.

Example 2. A cancer clinical trial with a subpopulation of females ($N = 30$) was described by Hansen et al.[4] Each patient was measured by two commonly used methods to determine the size change of tumors after the treatment. The outcome was either "did not shrink" or "shrink" by definition of the two methods. The first method is an objective method, the result of which is based on a computed tomography imaging scan. The second method is based on the pain score, which is considered a subjective method. It is of interest to establish consistency between the two methods, as the second is much easier to implement but the first is much more reliable. The observation is given in Table 3. This data set was also discussed by Klar et al.,[5] where they used a bootstrap interval to estimate κ. More details for estimating κ are given in Section 4.

The maximum likelihood estimator for κ is

$$\hat{\kappa} = \frac{\widehat{p_s} - \widehat{p_c}}{1 - \widehat{p_c}} = \frac{(n_{11} + n_{00})/N - [n_{1*}n_{*1} + (N - n_{1*})(N - n_{*1})]/N^2}{1 - [n_{1*}n_{*1} + (N - n_{1*})(N - n_{*1})]/N^2} \tag{7}$$

Table 2. Data from the physical therapy study for low back pain in Kilpikoski et al.[3]

Clinician B			
Clinician A	Present	Absent	Total
Present	$n_{11} = 28$	$n_{10} = 3$	$n_{1*} = 31$
Absent	$n_{01} = 6$	$n_{00} = 2$	$N - n_{1*} = 8$
Total	$n_{*1} = 34$	$N - n_{*1} = 5$	$N = 39$

Table 3. Data from the cancer clinical trial in Hansen et al.[4]

	The subjective method		
The objective method	Did not shrink	Shrink	Total
Did not shrink	$n_{11} = 22$	$n_{10} = 1$	$n_{1*} = 23$
Shrink	$n_{01} = 3$	$n_{00} = 4$	$N - n_{1*} = 7$
Total	$n_{*1} = 25$	$N - n_{*1} = 5$	$N = 30$

i.e. each probability in equation (5) is replaced by the estimator of relative frequency. Following the asymptotic theory, the maximum likelihood estimator $\hat{\kappa}$ approximately follows a normal distribution at each parameter configuration as N becomes large. Asymptotic confidence intervals for κ are then constructed under this fact. Many attempts have been made to improve the estimated variance of $\hat{\kappa}$. Fleiss et al.[6] were among the first to propose an estimator for the variance of $\hat{\kappa}$. Their approach is widely used in commercial statistical software, including SAS. Later, Bloch and Kraemer[7] proposed another estimator of variance based on the first-order Taylor expansion. Garner[8] gave a variance estimator based on a parsimonious log-linear model. However, these improvements were still found to be associated with unsatisfactory performance under certain cases, as pointed out by Jobe and David.[9] Lee and Tu[2] proposed another confidence interval based on a profile variance and conducted an extensive simulation study to show that their interval had better performance than the other three intervals considered in their article. All four asymptotic confidence intervals do not guarantee correct coverage probabilities, especially in small sample settings. This is not a surprise at all. In fact, deriving confidence intervals based on asymptotic normality is fundamentally wrong, even though it has been used in practice for a long time. This is because a confidence interval should capture the parameter of interest on all parameter configurations at a given sample size; however, asymptotic normality only assures reliable capture for a fixed parameter configuration at a large sample size. See Wang and Zhang[10] for more discussion.

There were also efforts to estimate κ using bootstrap confidence intervals.[5,11,12] Since κ is a complicated function of p_{ij}'s, researchers might agree that a bootstrap interval would be an ideal solution for estimating κ. However, Wang[13] recently proved that any bootstrap interval for any function of p_{ij}'s, including κ, has a zero infimum coverage probability (ICP) for any sample size N. So it is highly risky to estimate κ using bootstrap intervals.

In order to have satisfactory coverage probability, deriving exact intervals seems to be the only solution, and with modern computing power, it is feasible. An exact $1-\alpha$ confidence interval $C(\underline{X}) = [L(\underline{X}), U(\underline{X})]$ for κ means that its coverage probability never goes under $1-\alpha$ on the entire parameter space H. i.e.

$$Cover_C(p_{11}, p_{10}, p_{01}) \stackrel{def}{=} \sum_{\underline{x} \in S} I_{C(\underline{x})}(\kappa)\, p_m(\underline{x}; p_{11}, p_{10}, p_{01}) \geq 1 - \alpha, \quad \forall\, (p_{11}, p_{10}, p_{01}) \in H \tag{8}$$

where $I_\Omega(\kappa)$ is the indicator function of set Ω and p_m is the probability mass function as given in equation (1). The ICP over H measures the reliability of using the interval $C(\underline{X})$. For binary data from a match paired experiment, it is difficult to find a pivotal quantity for κ or conduct an exact level $1-\alpha$ test for $H_0 : \kappa = \kappa_0$ for a given value κ_0.[14] Agresti[14] pointed out that it is not possible to calculate exact confidence intervals based on the conditional approach for any measurements which is not a function of the odds ratio, because the nuisance parameters cannot be eliminated by the

conditional approach. So exact confidence intervals are hard to obtain by using pivotal quantity or inversion of tests, the two major confidence interval construction methods, see Casella and Berger.[15] Buehler[16] proposed a direct construction on the smallest exact one-sided confidence intervals provided that an order $\preceq$ on the sample space S is predetermined. In our case, exact $1–\alpha$ one-sided intervals for κ are of the forms: $[L(\underline{X}), 1]$ (lower one-sided) and $[-1, U(\underline{X})]$ (upper one-sided). Then the interval $[L(\underline{X}), U(\underline{X})]$ is of level $1–2\alpha$. One can show that the smallest one-sided intervals are the smallest in terms of set inclusion among all exact $1–\alpha$ confidence intervals whose limits, $L(\underline{X})$ or $U(\underline{X})$, preserve the same order as $\preceq$, see for example, Theorem 4 in Wang.[17] Therefore, the smallest intervals not only have correct coverage probabilities but also are optimal if a good order $\preceq$ is picked. Here, we will use four asymptotic intervals for κ to define eight orders (note there are eight confidence limits and each yields an order) on S and then derive the corresponding exact interval for each order following the Buehler method. This idea has been successfully applied to the risk ratio and the odds ratio in a 2×2 table.[18,19] To the best of our knowledge, no exact confidence intervals are available for κ with binary outcomes. We will use the Buehler method to modify four approximate intervals to four exact intervals and compare their performance on coverage probability and length.

The rest of this article is organized as follows. In Section 2, we briefly review four commonly used asymptotic confidence intervals for κ. We construct exact intervals following the Buehler method using the orders generated from the four asymptotic intervals in Section 3. Examples 1 and 2 are revisited in Section 4 to illustrate the application of exact intervals. In Section 5, we compare the performance of exact and asymptotic intervals by studying their coverage probabilities and lengths of intervals under a wide range of conditions. Section 6 is given to discussion.

2 Four approximate confidence intervals for κ

Traditional confidence intervals are all based on asymptotic normality, and they are widely used in the literature and practice. In particular, the maximum likelihood estimator $\hat{\kappa}$ is nearly unbiased for κ and is approximately normally distributed as N becomes large, and the $1–\alpha$ Wald-type approximate interval has the general form as

$$\hat{\kappa} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}(\hat{\kappa})}$$

where $z_{\alpha/2}$ is the upper $100(\alpha/2)$th percentile of a standard normal distribution and $\widehat{\mathrm{var}}(\hat{\kappa})$ is an estimator of the variance of $\hat{\kappa}$. There were several attempts to estimate $\mathrm{var}(\hat{\kappa})$.

Fleiss et al.[6] used the delta method to estimate the variance of $\hat{\kappa}$ which has the expression as

$$\widehat{\mathrm{var}}_F(\hat{\kappa}) = \frac{U + V - W}{N(1 - \hat{p}_c)^2}$$

where $U = \hat{p}_{00}[1 - (\hat{p}_{*1} + \hat{p}_{1*})(1 - \hat{\kappa})]^2 + \hat{p}_{11}[1 - (2 - \hat{p}_{1*} - \hat{p}_{*1})(1 - \hat{\kappa})]^2$, $V = (1 - \hat{\kappa})^2[\hat{p}_{01}(1 - \hat{p}_{1*} + \hat{p}_{*1}) + \hat{p}_{10}(1 + \hat{p}_{1*} - \hat{p}_{*1})^2]$, $W = [\hat{\kappa} - \hat{p}_c(1 - \hat{\kappa})]^2$, and $\hat{p}_c$ is the estimated chance agreement by plugging in the estimated cell probabilities. We write the interval as

$$C_F(\underline{X}) = [L_F(\underline{X}), U_F(\underline{X})] = \left[\hat{\kappa} - z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}_F(\hat{\kappa})}, \hat{\kappa} + z_{\alpha/2}\sqrt{\widehat{\mathrm{var}}_F(\hat{\kappa})}\right]$$

and name it as the Fleiss interval. This interval is utilized in the PROC FREQ of the statistical software SAS for Cohen's kappa.

Bloch and Kraemer[7] presented another estimator based on the first-order Taylor expansion (referred as the BK interval). The estimator of asymptotic variance is given as

$$\widehat{\text{var}}_{BK}(\hat{\kappa}) = \frac{1-\hat{\kappa}}{N}\left[(1-\hat{\kappa})(1-2\hat{\kappa}) + \frac{\hat{\kappa}(2-\hat{\kappa})}{(\hat{p}_{*1}+\hat{p}_{1*})(1-(\hat{p}_{*1}+\hat{p}_{1*})/2)}\right]$$

They numerically showed that the proposed variance estimate tends to underestimate the true variance and also found that the sample distribution of $\hat{\kappa}$ is not symmetric when κ is near either 0 or 1.

Garner[8] proposed an estimator of variance for $\hat{\kappa}$ for a general $r \times r$ case based on a parsimonious log-linear model and derived an explicit formula for a special case with $r = 2$

$$\widehat{\text{var}}_{G}(\hat{\kappa}) = \frac{4}{(1-\hat{p}_c)^2 N^2\left(\sum_{i=0}^{1}\sum_{j=0}^{1} 1/(n_{ij}+1)\right)} \tag{9}$$

Garner[8] pointed out that the distribution of $\hat{\kappa}$ is not symmetric in small sample settings and a transformation may help to improve the performance of the interval. The relationship between the estimated variance by Garner[8] and the one by Bloch and Kraemer[7] was given in Blackman and Koval.[20]

The last confidence interval is based on a profile variance, and a reparameterization of the κ in Lee and Tu[2] as shown in equation (6). The confidence limits are calculated by solving the following inequality of κ

$$\frac{(\kappa-\hat{\kappa})^2}{\widehat{\text{var}}_{LT}(\kappa)} \leq z^2_{\alpha/2}$$

where $\widehat{\text{var}}_{LT}(\kappa)$ is the estimated variance as in Lee and Tu.[2] They proposed two versions of the profile variance estimator. The one based on profile variance after reparameterization generally has better performance than the other base on the profile variance only. The better version is Method 4 in Lee and Tu[2] and is given below

$$\begin{aligned}\widehat{\text{var}}_{LT}(\kappa) = (\kappa-1)[&-(2\hat{p}_{*1}-1)(2\hat{p}_{1*}-1)(2\hat{p}_{*1}\hat{p}_{1*}-\hat{p}_{*1}-\hat{p}_{1*})\kappa^2 \\ &+2\kappa(6\hat{p}^2_{*1}\hat{p}^2_{1*}-6\hat{p}^2_{*1}\hat{p}_{1*}-6\hat{p}_{*1}\hat{p}^2_{1*}+2\hat{p}^2_{*1}+2\hat{p}^2_{1*}+4\hat{p}_{*1}\hat{p}_{1*}-\hat{p}_{*1}-\hat{p}_{1*}) \\ &-4\hat{p}_{*1}\hat{p}_{1*}(\hat{p}_{*1}\hat{p}_{1*}-\hat{p}_{*1}-\hat{p}_{1*}+1)]/[N(\hat{p}_{*1}+\hat{p}_{1*}-2\hat{p}_{*1}\hat{p}_{1*})^2]\end{aligned}$$

The lower and upper confidence limits would be the two roots of a cubic equation that are closest to $\hat{\kappa}$, as in general, there are three roots for such an equation. Lee and Tu[2] mentioned that the third root is typically outside the range of κ. By conducting extensive simulation studies, they showed that this confidence interval has shorter length and better coverage than others in most settings. Later, Klar et al.[5] compared a bootstrap confidence interval with the one from Method 4 in Lee and Tu[2] and concluded that the bootstrap interval gives slightly better coverage than Lee and Tu's interval, but the length of the bootstrap interval is generally longer.

To draw a coverage probability for an interval $C(\underline{X})$ as a function of κ in a plane, we introduce

$$Cover^*_C(\kappa) = \inf_{(p_{11},p_{10},p_{01})\in D(\kappa)} Cover_C(p_{11},p_{10},p_{01}) \tag{10}$$

where

$$D(\kappa) = \left\{(p_{11}, p_{01}, p_{10}) : (p_{11}, p_{01}, p_{10}, p_{00}) \in H \text{ and } \frac{p_s - p_c}{1 - p_c} = \kappa\right\} \tag{11}$$

is a two-dimensional set for a fixed value of κ by the one constraint (equation 5). Then the new coverage probability function $Cover_C^*$ is reduced to a univariate function from the original coverage probability function $Cover_C$, a trivariate function. The detailed formula for $Cover_C$ is given in equation (8). For a given κ-value, there are only two independent nuisance parameters, say p_{01} and p_{10}, in the function $Cover_C$ due to the constraint (equation 5). Then $Cover_C^*$ is calculated as the minimum of $Cover_C$ over p_{01} and p_{10}. Since these two nuisance parameters are both bounded from 0 to 1, we use the two-stage grid search algorithm to compute the minimum of $Cover_C$ for the given κ. In the first stage, we use a 50 by 50 grid to search the smallest possible value. Once this point of form (p_{01},p_{10}) is identified, then in the second stage an even finer grid around this point is used to find the minimum of $Cover_C$, which is the coverage probability $Cover_C^*(\kappa)$. In this process, the minimum is found by exact probability computation, in particular, by equation (8), and no statistical simulation is involved.

Figures 1 and 2 show the coverage probabilities $Cover_C^*(\kappa)$ for the four asymptotic lower and upper one-sided intervals for sample size $N = 30$ at a 95% confidence level. The κ-values used to evaluate the coverage are not equally spaced in $[-1, 1]$. For Figure 1, the coverage at the lower limits $L(\underline{x})$'s and some points very close to them is computed and plotted against these κ-values; while for Figure 2, the coverage around the upper limits $U(\underline{x})$'s is computed. The plots are typical for other sample sizes as well. If an interval is truly of level 95%, then the coverage probability curve should be always at or above the line of 0.95 for any $\kappa \in [-1, 1]$. However, it can be seen from the plots that none of them are acceptable with regard to the coverage requirement. As can be seen in Figure 1 for the lower asymptotic intervals, all of them do not guarantee the coverage. The Garner interval is generally better than others regarding the coverage. For upper intervals in Figure 2, although the Garner one-sided upper interval has good coverage for positive κ-values, it substantially violates the coverage requirement when κ is negative. The other three upper intervals generally have coverage much lower than $1–\alpha$. In particular, when κ is close to zero, both the Fleiss interval and the Lee–Tu interval have a very low coverage.

The coverage requirement of these asymptotic intervals is not well satisfied even though some previous studies claimed their good performance. This may be due to the fact that the simulation study can not be complete especially when, as in our case, there are too many parameter configurations. The coverage probability for any bootstrap interval for κ would touch the line 0 for any N and any level $1–\alpha$, as being proved by Wang.[13]

3 Exact confidence intervals for κ

We first derive exact $1–\alpha$ lower and upper one-sided intervals, then $1–2\alpha$ two-sided intervals are obtained from them. Exact one-sided intervals due to Buehler[16] attain the nominal level at all observed values. The conservatism of exact one-sided intervals is minimized. We construct the two-sided intervals from exact one-sided intervals. Both one-sided intervals are exact, but the two-sided interval would be conservative due to the discrete observations.

The construction of one-sided intervals by the Buehler method requires an order on the entire sample space S. This order on S provides an order by the confidence limit $L(\underline{X})$ (and $U(\underline{X})$). In particular, when the sample space is discrete, the order reveals which sample point(s) yields the

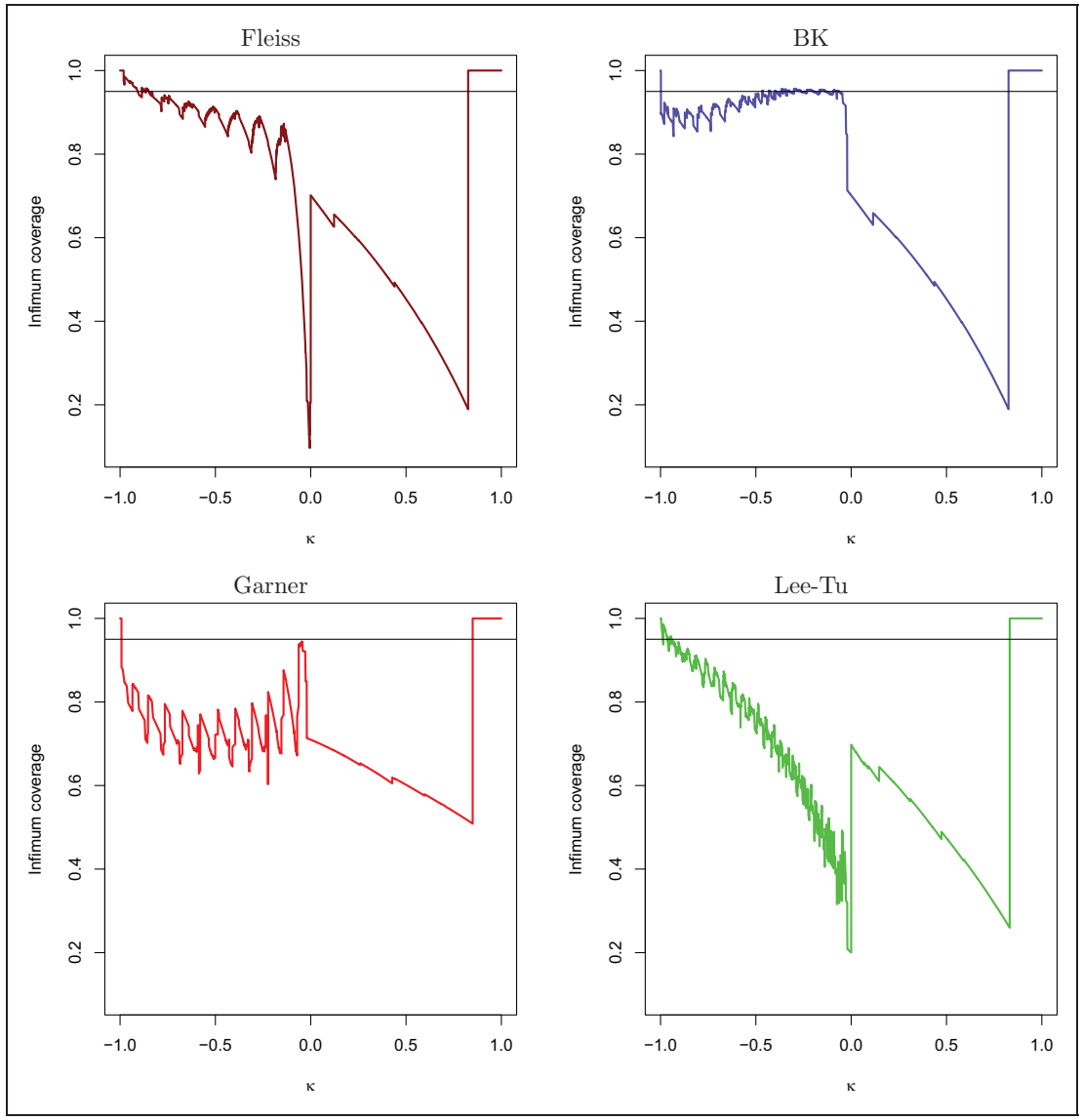

Figure 1. Coverage for the four asymptotic 95% lower one-sided confidence intervals when $N = 30$ ($Cover_C^*$ versus κ).

largest confidence limit, which sample point(s) yields the second largest confidence limit and so on. Intuitively, $\underline{x}_0 = (N, 0, 0)$ that has a complete agreement and would yield the largest L (also the largest U), i.e. $L(\underline{x}_0) = \max_{\underline{x} \in S} L(\underline{x})$. Ties are allowed, for example, $\underline{x} = (N-1, 0, 0)$ and $\underline{x}' = (1, 0, 0)$ should be tied by intuition. Therefore, the confidence limits at these two points should be equal. However, an order with more ties yields a wider interval.

The Buehler method has obvious advantage and disadvantage. For any given order on S, this method automatically produces the smallest exact lower (upper) one-sided interval among all exact

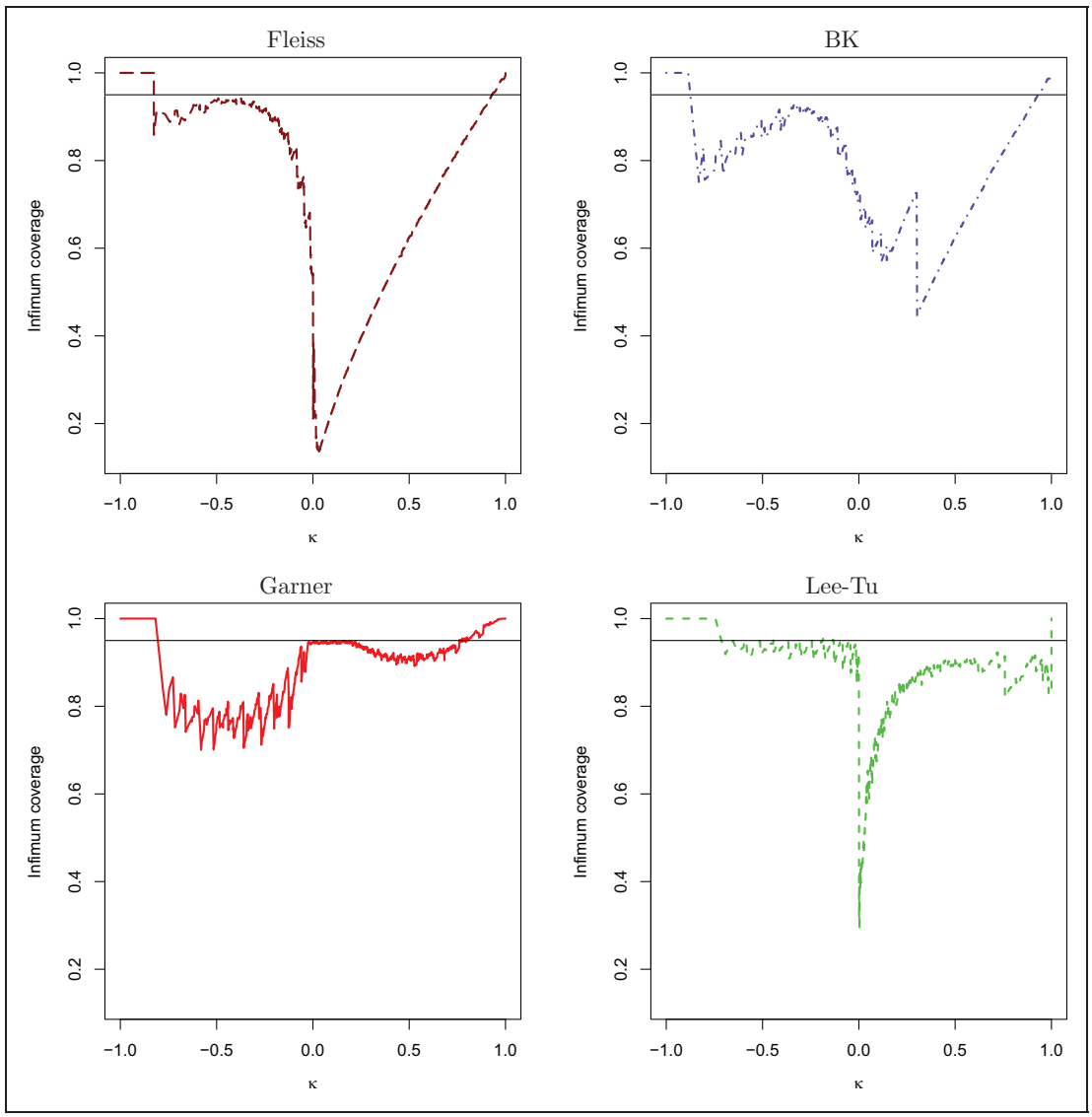

Figure 2. Coverage for the four asymptotic 95% upper one-sided confidence intervals when $N = 30$ ($Cover_C^*$ versus κ).

intervals that have the same order on the confidence limit as the given order. From a mathematical point of view, it specifies a class of exact intervals, shows the existence of the best interval in this class and also successfully identifies the best interval. On the other hand, for a sample space S with m sample points, there are 2^m possible orders on S, and each order yields a best interval under that order. In Example 2 with $N = 30$, then $m = 5456$ by equation (3), there are 2^{5456} possible orders. How to choose an optimal one from the best intervals that are from 2^m different orders is generally unknown, and the Buehler method does not resolve this problem. Wang[17] proposed an inductive method to construct an order on any finite sample space that yields an admissible interval under the

set inclusion criterion. But his interval computation seems heavy, especially when there are multiple nuisance parameters.

In this paper, we apply the Buehler method to some given orders specified by the four asymptotic intervals in Section 2. For a given statistic $O(\underline{X})$ on S in (equation 2), we define an order $\preceq_O$ on S as follows. For any two sample points $\underline{x}$ and $\underline{x}'$ in S, $\underline{x}$ is said to be less than or equal to $\underline{x}'$ if $O(\underline{x})$ is less than or equal to $O(\underline{x})$; and they are equal if $O(\underline{x}) = O(\underline{x}')$. This relationship is denoted by

$$\underline{x} \prec_O (\equiv_O) \underline{x}' \quad \text{if } O(\underline{x}) \leq (=) O(\underline{x}')$$

Therefore, any statistic on S can be used to introduce an order on S. Let $\mathcal{B}_O^L$ be a class of exact 1–α lower one-sided intervals for κ of form $[L(\underline{X}), 1]$ satisfying

(1) $L(\underline{x}) \leq L(\underline{x}')$ if $O(\underline{x}) \leq O(\underline{x}')$
(2) $L(\underline{x}) = L(\underline{x}')$ if $O(\underline{x}) = O(\underline{x}')$

i.e. the lower limit L is a nondecreasing function with respect to the order $\preceq_O$. The best interval $[L_O(\underline{X}), 1]$ in $\mathcal{B}_O^L$ is given in the following lemma.

Lemma 3.1 *Assume* $\alpha \in (0,1)$. *For a given function* $O(\underline{X})$ *on* S *and any* $\underline{x} \in S$, *let*

$$h_{\underline{x}}(\kappa) = \inf_{(p_{11}, p_{10}, p_{01}) \in D(\kappa)} \sum_{\{\underline{x}' \in S: O(\underline{x}') < O(\underline{x})\}} p_m(\underline{x}'; p_{11}, p_{10}, p_{01}) \tag{12}$$

Let

$$H_{\underline{x}} = \left\{\kappa \in [-1, 1] : h_{\underline{x}}(\kappa) = 1 - \alpha\right\} \tag{13}$$

Define

$$L_O(\underline{x}) = \begin{cases} \inf\ H_{\underline{x}}, & \text{if } H_{\underline{x}} \neq \emptyset; \\ -1, & \text{otherwise} \end{cases} \tag{14}$$

Then we have

(i) $[L_O(\underline{X}), 1]$ *belongs to* $\mathcal{B}_O^L$ *(i.e. it is of level 1–*α *and satisfies 1) and 2));*
(ii) *for any interval* $[L(\underline{X}), 1] \in \mathcal{B}_O^L$, $L(\underline{X}) \leq L_O(\underline{X})$

The proof can be found in Lloyd and Kabaila[21] and Wang.[17] The interval $[L_O(\underline{X}), 1]$ is the best in $\mathcal{B}_O^L$ because it has the largest lower limit. Hence it is a subset of any interval in $\mathcal{B}_O^L$, and it is the smallest interval under the set inclusion criterion. On the other hand, if the class $\mathcal{B}_O^L$ is small (i.e. it does not have enough intervals), then the best interval in $\mathcal{B}_O^L$ is not equal to a good interval. An extreme case is that if $O(\underline{X})$ is a constant, then this $\mathcal{B}_O^L$ only contains one interval and $[L_O(\underline{X}), 1]$ that also assumes a constant value is still the best in $\mathcal{B}_O^L$, but is not good at all. Therefore, a much more challenging issue is to identify a large class of intervals by identifying an appropriate function $O(\underline{X})$.

For upper one-sided intervals, let $\mathcal{B}_O^U$ be a class of exact 1–α upper one-sided intervals for κ of form $[-1, U(\underline{X})]$ satisfying

(1) $U(\underline{x}) \leq U(\underline{x}')$ if $O(\underline{x}) \leq O(\underline{x}')$

(2) $U(\underline{x}) = U(\underline{x}')$ if $O(\underline{x}) = O(\underline{x}')$

The best interval $[-1, U_O(\underline{X})]$ in $\mathcal{B}_O^U$ is given as follows.

Lemma 3.2 *Assume* $\alpha \in (0,1)$. *For a given function* $O(\underline{X})$ *on* S *and any* $\underline{x} \in S$, *let*

$$j_{\underline{x}}(\kappa) = \inf_{(p_{11}, p_{10}, p_{01}) \in D(\kappa)} \sum_{\{\underline{x}' \in S: O(\underline{x}') > O(\underline{x})\}} p_m(\underline{x}'; p_{11}, p_{10}, p_{01}) \tag{15}$$

Let

$$J_{\underline{x}} = \left\{\kappa \in [-1, 1] : j_{\underline{x}}(\kappa) = 1 - \alpha\right\} \tag{16}$$

Define

$$U_O(\underline{x}) = \begin{cases} \sup\ J_{\underline{x}}, & \text{if } J_{\underline{x}} \neq \emptyset; \\ B, & \text{otherwise} \end{cases} \tag{17}$$

Then we have

(i) $[-1, U_O(\underline{X})]$ *belongs to* $\mathcal{B}_O^U$ *(i.e. it is of level* $1-\alpha$ *and satisfies 1) and 2));*
(ii) *for any interval* $[-1, U(\underline{X})] \in \mathcal{B}_O^U$, $U_O(\underline{X}) \leq U(\underline{X})$

In this paper, the function $O(\underline{X})$ is equal to each of the confidence limits for the four asymptotic intervals in Section 2. More precisely, consider the 90% Garner interval when $N = 30$

$$C_G(\underline{X}) = [L_G(\underline{X}), U_G(\underline{X})] \stackrel{def}{=} \left[\hat{\kappa} \pm z_{0.05}\sqrt{\widehat{\text{var}}_G(\hat{\kappa})}\right] \tag{18}$$

where $\hat{\kappa}$ and $\widehat{\text{var}}_G(\hat{\kappa})$ are given in equations (7) and (9), respectively. The order by $L_G(\underline{X})$ also depends on the confidence level. We derive $L_{G,O}(\underline{X})$ with $O(\underline{X}) = L_G(\underline{X})$ and $\alpha = 0.05$ following Lemma 3.1 and $U_{G,O}(\underline{X})$ with $O(\underline{X}) = U_G(\underline{X})$ and $\alpha = 0.05$ following Lemma 3.2. Then $[L_{G,O}(\underline{X}), 1]$ and $[-1, U_{G,O}(\underline{X})]$ are the smallest exact 95% intervals under the order by $L_G(\underline{X})$ and $U_G(\underline{X})$, respectively, and $[L_{G,O}(\underline{X}), U_{G,O}(\underline{X})]$ is an exact 90% interval for κ. Figure 3 illustrates a coverage probability comparison between the smallest interval $[L_{G,O}(\underline{X}), 1]$ and the asymptotic interval $[L_G(\underline{X}), 1]$ as well as a comparison between $[-1, U_{G,O}(\underline{X})]$ and $[-1, U_G(\underline{X})]$. It is clear that the exact intervals always have a coverage at least 0.95, while the asymptotic ones have a coverage as low as 0.5. Similar to Figures 1 and 2, we do not evaluate the coverage using equal spacing in κ. Instead, we compute the coverage near the lower (or upper) confidence limits and obtain irregular curves.

4 Examples

Example 1 (continued). For the data from the physical therapy example in Table 2, the estimated simple percent agreement is $\hat{p}_s = (28 + 2)/39 = 0.7692$ and the estimated chance agreement is $\hat{p}_c = 0.7193$, and these yield an estimated Cohen's kappa coefficient $\hat{\kappa} = 0.1778$, which indicates a slight agreement between Clinician A and Clinician B by the criterion from Landis and Koch.[22] Both asymptotic and exact confidence intervals are presented in Table 4 at $\alpha = 0.05$. It takes less than 1 min to calculate the interval limits for this example by using a personal computer (Intel Core i7 = 2640M CPU@2.80 GHz and 6 GB RAM). The asymptotic intervals are unreliable, as evidenced

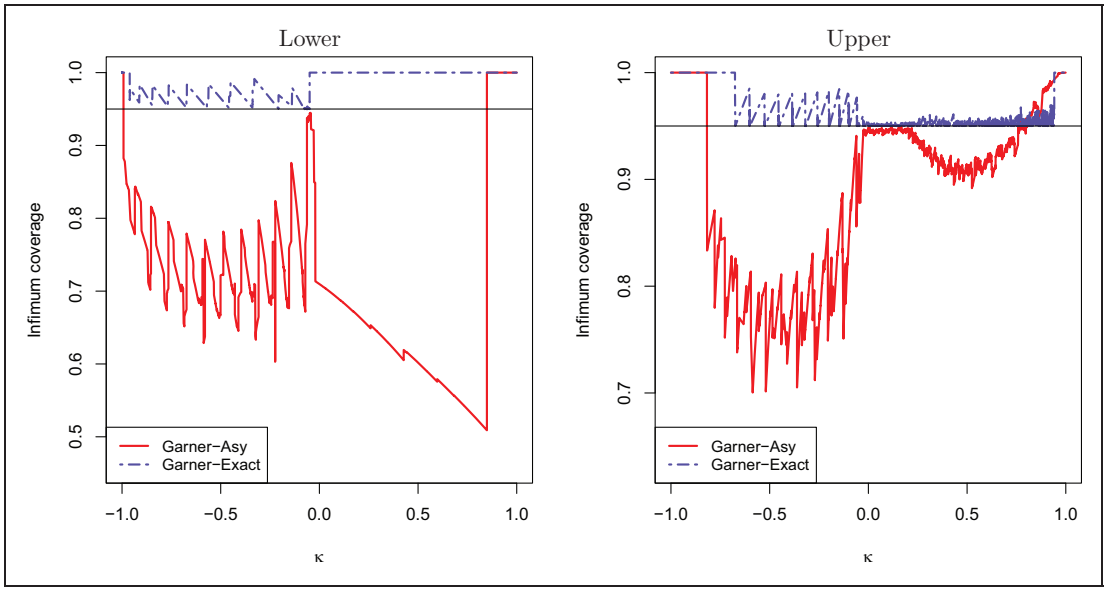

Figure 3. Coverage probabilities for the 95% asymptotic Garner one-sided interval and the corresponding exact 95% interval when $N = 30$.

Table 4. Asymptotic and exact confidence intervals and their lengths at $\alpha = 0.05$ for Example 1.

	Fleiss			BK			Garner			Lee–Tu		
	Lower	Upper	Length	Lower	Upper	Length	Lower	Upper	Length	Lower	Upper	Length
Asy	−0.1237	0.4797	0.6034	−0.1331	0.4891	0.6221	−0.1665	0.5225	0.6890	−0.0505	0.4790	0.5295
Exact	−0.1971	0.9312	1.1283	**−0.1363**	0.9312	1.0675	−0.2578	0.5734	0.8312	−0.1401	**0.5569**	**0.6973**

The best intervals and the shortest length are in bold.

in Figures 1 and 2, and should not be directly compared with the corresponding exact intervals due to the lack of a common ICP. Exact one-sided intervals based on the BK lower limit $L_{BK}(\underline{X})$ and the Lee–Tu upper limit $U_{LT}(\underline{X})$ have the largest (best) lower limit and the smallest (best) upper limit. The two-sided interval based on the Lee–Tu interval is the shortest among the four exact intervals for this data set. These intervals do not support an agreement between the two clinicians, as they all include zero.

To make this example more concrete, we next detail the computation of $L_{G,O}(\underline{x})$ and $U_{G,O}(\underline{x})$ at observation $\underline{x} = (28, 3, 6)$ with $\alpha = 0.05$ using two orders by two functions $L_G(\underline{X})$ and $U_G(\underline{X})$ given in equation (18). First, we identify the set of points $\underline{x}'$ with $L_G(\underline{x}') < L_G(\underline{x})$ by calculating $L_G(\underline{x}')$ on all 11,480 $(= (N+1)(N+2)(N+3)/6$ for $N = 39)$ sample points and name it $GL_{\underline{x}}$. The set $GL_{\underline{x}}$ contains 6263 sample points. Second, we compute

$$h_{\underline{x}}(\kappa) = \inf_{(p_{11}, p_{10}, p_{01}) \in D(\kappa)} \sum_{\underline{x}' \in GL_{\underline{x}}} p_m(\underline{x}'; p_{11}, p_{10}, p_{01}) \tag{19}$$

For each given value of κ, $D(\kappa)$ is a two-dimensional set defined in equation (6), and the infimum is calculated by a two-step grid search. We choose a reasonable, equally spaced, partition for $D(\kappa)$ using p_{*1} and p_{1*} as nuisance parameters, and a size of 50 by 50 would be fine in the first stage. The set that yields the smallest value of $h_{\underline{x}}(\kappa)$ is identified in this stage. At the second stage, an even finer partition in that set is then used to search for the minimum of the function again. Last, solve the equation

$$h_{\underline{x}}(\kappa) = 1 - 0.05$$

and the smallest solution is $L_{G,O}(\underline{x}) = -0.2578$. Similarly, for computing $U_{G,O}(\underline{x})$, we introduce a function

$$j_{\underline{x}}(\kappa) = \inf_{(p_{11}, p_{10}, p_{01}) \in D(\kappa)} \sum_{\underline{x}' \in GU_{\underline{x}}} p_m(\underline{x}'; p_{11}, p_{10}, p_{01})$$

where $GU_{\underline{x}} = \{\underline{x}' \in S : U_G(\underline{x}') > U_G(\underline{x})\}$ and the largest solution of $j_{\underline{x}}(\kappa) = 1 - 0.05$ is $U_{G,O}(\underline{x}) = 0.5734$. Therefore, we obtain two exact 95% one-sided confidence intervals [−0.2578, 1] and [−1, 0.5734] and an exact two-sided 90% interval [−0.2578, 0.5734] for κ based on the asymptotic Garner interval.

Example 2 (continued). Using the data in Table 3, we obtain $\hat{p}_s = (22 + 4)/30 = 0.8667$, $\hat{p}_c = 0.6778$ and $\hat{\kappa} = 0.5862$, which indicates a moderate agreement between the subjective and objective methods by the criterion from Landis and Koch.[22] In this example, the 90% exact interval based on the Garner interval is [−0.0497, 0.9054] and is the shortest among the four exact intervals. All the four intervals include zero and then do not suggest an agreement between the subjective and objective methods.

5 Numerical study

We now compare the performance of four asymptotic intervals, the Fleiss interval (C_F), the BK interval (C_{BK}), the Garner interval (C_G) and the Lee–Tu interval (C_{LT}). We will also examine their corresponding exact intervals, $C_{F,O}$, $C_{BK,O}$, $C_{G,O}$ and $C_{LT,O}$ through the ICP, the average length (AL) of intervals over the entire sample space and the expected length (EL) over the parameter space. All numerical calculation results in the paper are based on exact probability computation, and no simulation is involved. One such example is given in Example 1 (continued) when we compute $h_{\underline{x}}(\kappa)$ in equation (19).

First, the ICP of a confidence interval $C(\underline{X})$ for κ is defined as

$$ICP(C) = \inf_{(p_{11}, p_{10}, p_{01}) \in H} Cover_C(p_{11}, p_{10}, p_{01}) = \inf_{\kappa \in [-1,1]} Cover_C^*(\kappa)$$

where $Cover_C(p_{11}, p_{10}, p_{01})$ and $Cover_C^*(\kappa)$ are given in equations (8) and (10). This measures the reliability of using the interval $C(\underline{X})$, and we wish it to be equal to the given nominal level 1–α. If $ICP(C)$ is less than 1–α, then there is no guarantee that $C(\underline{X})$ captures κ with the desired probability which, from a mathematical point of view, makes applying the confidence interval baseless. Table 5 provides the ICP for the four asymptotic two-sided intervals for N up to 30. It can be seen that the ICP values for asymptotic intervals are very low, less than 30% for all of them and some of them are extremely low, less than 1%. This is not surprising as we see in Figure 3 that the ICP values for upper one-sided intervals are not well satisfied. On the other hand, exact intervals guarantee the

Table 5. Infimum coverage probability ICP of the 90% asymptotic two-sided intervals.

Method	$N = 10$	20	30
Fleiss	<0.01	<0.01	<0.01
BK	<0.01	<0.01	<0.01
Garner	0.0966	0.1838	0.2626
Lee–Tu	<0.01	<0.01	0.0322

Table 6. Average length AL_C of the 95% exact and asymptotic lower one-sided intervals.

	$N = 10$	20	30	40	50
Exact interval					
Fleiss	1.6182	1.4277	1.3507	1.3091	**1.2787**
BK	**1.5363**	**1.3985**	**1.3407**	**1.3073**	1.2805
Garner	1.7126	1.5256	1.4347	1.3814	1.3378
Lee–Tu	1.7968	1.4318	1.3592	1.3188	1.2870
Asy interval					
Fleiss	1.2740	1.2326	1.1971	1.1722	1.1539
BK	1.4093	1.3042	1.2495	1.2152	1.1912
Garner	1.4844	1.3015	1.2258	1.1844	1.1578
Lee–Tu	1.2919	1.2593	1.2090	1.1793	1.1589

The shortest average lengths are in bold.

nominal level by mathematical proofs and the ICP values for all exact intervals are greater than or equal to 90%.

Second, the efficiency of a confidence interval is measured by its length. We compute the AL of $C(\underline{X}) = [L(\underline{X}), U(\underline{X})]$ over the sample space S

$$AL_C = \frac{1}{m}\sum_{\underline{x}\in S}(U(\underline{x}) - L(\underline{x}))$$

where m is the number of all sample points in S as given in equation (3). The smaller the AL, the better the interval. Table 6 provides the ALs for the exact and asymptotic lower one-sided intervals of form $[L(\underline{X}), 1]$ for the four methods, and Table 7 provides a similar comparison for the upper intervals. The exact intervals generally have longer lengths as compared to the corresponding asymptotic intervals; however, it is not a meaningful comparison since the asymptotic intervals have ICPs smaller than the nominal level. For a fixed value of N, the exact interval with the shortest AL in the two tables is in bold. The exact lower one-sided interval based on the BK interval performs better than the other three. The exact upper intervals based on the Lee–Tu interval and the Garner interval are competitive, and they are better than the other two.

We also compare the AL for the exact two-sided confidence intervals $[L(\underline{X}), U(\underline{X})]$ for the four methods. The result is presented in Table 8 and the shortest ones are in bold. However, pointed out by a reviewer, a shorter two-sided interval can be constructed with lower and upper intervals from different methods. For example, if $N = 10$, let $[L_{BK,O},1]$ denote the exact 95% BK interval in Table 6 and let $[-1, U_{G,O}]$ denote the exact 95% Garner interval in Table 7. Then $[L_{BK,O},U_{G,O}]$ is an exact 90% two-sided interval with an $\mathrm{AL} = 1.5363 + 1.4829 - 2 = 1.0192$, which is smaller than the AL ($= 1.1955$) for the Garner interval in Table 8.

Table 7. Average length AL_C of the 95% exact and asymptotic upper one-sided intervals.

	$N = 10$	20	30	40	50
Exact interval					
Fleiss	1.7734	1.7099	1.6789	1.6484	1.6283
BK	1.7098	1.6645	1.6367	1.6154	1.5974
Garner	**1.4829**	**1.3542**	**1.3028**	1.2727	1.2496
Lee–Tu	1.5103	1.3737	1.3102	**1.2714**	**1.2415**
Asy interval					
Fleiss	1.3344	1.2819	1.2447	1.2192	1.2006
BK	1.4634	1.3520	1.2967	1.2620	1.2378
Garner	1.5124	1.3413	1.2696	1.2295	1.2034
Lee–Tu	1.3466	1.2793	1.2405	1.2152	1.1970

The shortest average lengths are in bold.

Table 8. Average length AL_C of the 90% exact two-sided intervals.

	$N = 10$	20	30	40	50
Fleiss	1.3915	1.1376	1.0296	0.9575	0.9070
BK	1.2462	1.0630	0.9775	0.9227	0.8779
Garner	**1.1955**	0.8798	0.7376	0.6542	0.5874
Lee–Tu	1.3071	**0.8055**	**0.6694**	**0.5902**	**0.5286**

The shortest average lengths are in bold.

In addition, we compute the EL of $C(\underline{X})$

$$EL_C(p_{11}, p_{10}, p_{01}) = E(U(\underline{X}) - L(\underline{X})) = \sum_{\underline{x} \in S} (U(\underline{x}) - L(\underline{x}))\, p_m(\underline{x}; p_{11}, p_{10}, p_{01})$$

At different sample sizes, the EL of four 90% exact two-sided intervals at 10,660 parameter configurations is calculated. These 10,660 parameter points are uniformly picked over the three-dimensional parameter space (p_{11}, p_{10} and p_{01}). The proportions of these configurations on which each interval has the shortest EL are reported in Table 9. For example, when $N = 20$, the exact interval $C_{LT,O}(\underline{X})$ based on the Lee–Tu interval $C_{LT}(\underline{X})$ has the shortest EL among the four exact intervals at 71.4% of parameter configurations. It can be seen in Table 9 that $C_{LT,O}(\underline{X})$ has the short EL in general.

In practice, negative kappa estimates do not have practical interpretation and small kappa estimates are also not interested to clinicians. According to the standard for strength of agreement for the kappa coefficient by Landis and Koch,[22] the kappa estimate between 0.4 and 0.6 is considered as moderate agreement between two raters and the kappa estimate above 0.6 is considered as substantial and almost perfect agreement. For this reason, we also compare the AL of exact intervals for the sample points with kappa estimate between 0.4 and 0.6 in Table 10 and above 0.6 in Table 11. In both cases, the exact Garner interval is the best for other large sample sizes considered, $N = 30$, 40 and 50. For small sample size, the exact BK interval has the best performance when kappa estimate is moderate and the exact intervals based on the Lee–Tu interval are shorter than those based on other intervals for the case with kappa above 0.6.

Table 9. Proportions of the 10,660 parameter configurations from which each of four exact intervals has the smallest expected length.

	$N = 10$	20	30	40	50
Exact interval					
Fleiss	1.7	0.1	2.3	3.9	5.1
BK	38.4	0.1	0.3	0.4	0.7
Garner	**55.3**	28.4	32.6	32.3	28.5
Lee–Tu	4.6	**71.4**	**64.8**	**63.4**	**65.7**
Total	100	100	100	100	100

The largest proportions are in bold.

Table 10. Average length AL_C of the 90% exact two-sided intervals for the data with $\hat{\kappa}$ between 0.4 and 0.6.

	$N = 10$	20	30	40	50
Fleiss	1.2216	1.0144	0.9828	0.9689	0.9526
BK	**1.1179**	1.0061	0.9815	0.9685	0.9524
Garner	1.1417	**0.8960**	**0.8315**	**0.7903**	**0.7527**
Lee–Tu	1.2930	0.9142	0.8577	0.8115	0.7680

The shortest average lengths are in bold.

Table 11. Average length AL_C of the 90% exact two-sided intervals for the data with $\hat{\kappa} \geq 0.6$.

	$N = 10$	20	30	40	50
Fleiss	1.1437	1.0524	1.0075	0.9862	0.9657
BK	1.1483	1.0523	1.0074	0.9860	0.9658
Garner	1.2111	1.0589	**0.9921**	0.**9580**	**0.9236**
Lee–Tu	**1.1324**	**1.0375**	1.0025	0.9857	0.9619

The shortest average lengths are in bold.

6 Conclusion

Cohen's kappa coefficient, κ, is the most important measurement of agreement between two raters. In this paper, we estimate it using confidence intervals in a matched pair experiment. Our numerical study shows that four widely used asymptotic intervals, C_F, C_{BK}, C_G and C_{LT}, have ICPs much less than the nominal level, which indicates that the statistical inferences based on these intervals are not reliable. This is the motivation for deriving exact confidence intervals for κ.

The two traditional confidence interval construction methods, the pivotal quantity and the inversion of tests, cannot be applied efficiently to κ. So we utilize the Buehler method to construct the smallest exact one-sided intervals for κ and obtain two-sided intervals using the intersection of two one-sided intervals. The Buehler method needs a predetermined order on the sample space S for a valid interval construction. In this paper, we use the lower and upper limits of

the four asymptotic intervals to define orders on S. The exact intervals $C_{G,O}$ and $C_{LT,O}$ which are based on C_G and C_{LT}, respectively, are generally better than others and recommended for practice. These intervals are reliable due to correct coverage probabilities. The program to implement the exact one-sided or two-sided intervals is written in the statistical software R and is available from the first author.

Can the derived intervals be further improved? That is, are there intervals that are of level $1–\alpha$ and are subset of the derived intervals? The quick answer is "yes". This is because the orders of the asymptotic confidence limits generate too many ties. One can obtain a uniform improvement by breaking ties, see Proposition 2 in Wang.[17] However, the computation is tedious.

As mentioned before, there are 2^m possible orders on S, where $m=(N+1)(N+2)(N+3)/6$. Which order yields an optimal interval among these orders? An admissible interval under the set inclusion criterion may be the possible answer, i.e. any subinterval of this interval is of level strictly less than $1–\alpha$. Wang[17] developed such an order on any finite sample space using an inductive construction. Two such intervals were successfully derived for the difference of two proportions, see Wang,[17] Wang[23] and Shan and Wang.[24] From a mathematical point of view, his result can be extended to the case of κ. The implementation of the interval in programming is challenging.

Acknowledgements

The authors would like to thank the Editor and two referees for their valuable comments and suggestions that helped to improve this manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

1. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20**: 37–46.
2. Lee JJ and Tu ZN. A better confidence interval for Kappa on measuring agreement between two raters with binary outcomes. *J Comput Graphical Stat* 1994; **3**: 301–321.
3. Kilpikoski S, Airaksinen O, Kankaanpää M, et al. Interexaminer reliability of low back pain assessment using the McKenzie method. *Spine* 2002; **27**: 207–214.
4. Hansen RM, Ryan L, Anderson T, et al. Phase III study of bolus versus infusion fluorouracil with or without cisplatin in advanced colorectal cancer. *J Natl Cancer Inst* 1996; **88**: 668–674.
5. Klar N, Lipsitz SR, Parzen M, et al. An exact bootstrap confidence interval for in small samples. *J R Stat Soc D* 2002; **51**: 467–478.
6. Fleiss JL, Cohen J and Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969; **72**: 323–327.
7. Bloch DA and Kraemer HC. 2×2 kappa coefficients: measures of agreement or association. *Biometrics* 1989; **45**: 269–287.
8. Garner JB. The standard error of Cohen's Kappa. *Stat Med* 1991; **10**: 767–775.
9. Jobe JM and David HT. Buehler confidence bounds for a reliability-maintainability measure. *Technometrics* 1992; **34**: 214–222.
10. Wang W and Zhang Z. Asymptotic infimum coverage probability for interval estimation of proportions. *Metrika* 2014; **77**: 635–646.
11. Fung KP and Lee J. Bootstrap estimate of the variance and confidence interval of kappa. *Br J Indust Med* 1991; **48**: 503–504.
12. Reichenheim ME. Confidence intervals for the kappa statistic. *Stata J* 2004; **4**: 421–428.
13. Wang W. A note on bootstrap confidence intervals for proportions. *Stat Probab Lett* 2013; **83**: 2699–2702.
14. Agresti A. A survey of exact inference for contingency tables. *Stat Sci* 1992; **7**: 131–153.
15. Casella G and Berger RL. *Statistical inference*, 2nd ed. Belmont, CA: Cengage Learning, 2001.
16. Buehler RJ. Confidence intervals for the product of two binomial parameters. *J Am Stat Assoc* 1957; **52**: 482–493.

17. Wang W. On construction of the smallest one-sided confidence interval for the difference of two proportions. *Ann Stat* 2010; **38**: 1227–1243.
18. Lloyd CJ and Moldovan MV. Exact one-sided confidence bounds for the risk ratio in 2 2 tables with structural zero. *Biom J* 2007; **49**: 952–963.
19. Lloyd CJ and Moldovan MV. Unconditional efficient one-sided confidence limits for the odds ratio based on conditional likelihood. *Stat Med* 2007; **26**: 5136–5146.
20. Blackman NJ and Koval JJ. Interval estimation for Cohen's kappa as a measure of agreement. *Stat Med* 2000; **19**: 723–741.
21. Lloyd CJ and Kabaila P. On the optimality and limitations of Buehler bounds. *Aust N Z J Stat* 2003; **45**: 167–174.
22. Landis JR and Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
23. Wang W. An inductive order construction for the difference of two dependent proportions. *Stat Probab Lett* 2012; **82**: 1623–1628.
24. Shan G and Wang W. ExactCIdiff: an R package for computing exact confidence intervals for the difference of two proportions. *R J* 2013; **5**: 62–71.