BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Homogeneity score test of AC_1 statistics and estimation of common AC_1 in multiple or stratified inter-rater agreement studies

Chikara Honda[1,2*] and Tetsuji Ohyama[3]

Abstract

Background: Cohen's κ coefficient is often used as an index to measure the agreement of inter-rater determinations. However, κ varies greatly depending on the marginal distribution of the target population and overestimates the probability of agreement occurring by chance. To overcome these limitations, an alternative and more stable agreement coefficient was proposed, referred to as Gwet's AC_1. When it is desired to combine results from multiple agreement studies, such as in a meta-analysis, or to perform stratified analysis with subject covariates that affect agreement, it is of interest to compare several agreement coefficients and present a common agreement index. A homogeneity test of κ was developed; however, there are no reports on homogeneity tests for AC_1 or on an estimator of common AC_1. In this article, a homogeneity score test for AC_1 is therefore derived, in the case of two raters with binary outcomes from K independent strata and its performance is investigated. An estimation of the common AC_1 between strata and its confidence intervals is also discussed.

Methods: Two homogeneity tests are provided: a score test and a goodness-of-fit test. In this study, the confidence intervals are derived by asymptotic, Fisher's Z transformation and profile variance methods. Monte Carlo simulation studies were conducted to examine the validity of the proposed methods. An example using clinical data is also provided.

Results: Type I error rates of the proposed score test were close to the nominal level when conducting simulations with small and moderate sample sizes. The confidence intervals based on Fisher's Z transformation and the profile variance method provided coverage levels close to nominal over a wide range of parameter combination.

Conclusions: The method proposed in this study is considered to be useful for summarizing evaluations of consistency performed in multiple or stratified inter-rater agreement studies, for meta-analysis of reports from multiple groups and for stratified analysis.

Keywords: Common AC_1, Consistency evaluation, Gwet's AC_1, Homogeneity test, Inter-rater agreement, Stratified study

Background

To evaluate the reliability when two raters classify objects as either positive (+) or negative (−), Cohen's κ [1] and the intra-class version of κ, which is identical to Scott's π [2], have often been used. Let p_a be the agreement probability, and p_1 and p_2 the probabilities classified as (+) by rater 1 and 2 respectively. Then Cohen's κ (κ_{Cohen}) and Scott's π (κ_{Scott}) are defined as follows:

$$\kappa_{Cohen} = \frac{p_a - p_{e(c)}}{1 - p_{e(c)}}, \quad \kappa_{Scott} = \frac{p_a - p_{e(s)}}{1 - p_{e(s)}},$$

where $p_{e(c)} = p_1 p_2 + (1-p_1)(1-p_2), p_{e(s)} = p_+^2 + (1-p_+)^2$ and $p_+ = (p_1 + p_2)/2$. The $p_{e(c)}$ and $p_{e(s)}$ are the probabilities of agreement expected by chance for Cohen's κ and Scott's π respectively. The $p_{e(c)}$ assumes that the probabilities of positive classification differ between two raters, while the $p_{e(s)}$ assumes that these two

* Correspondence: c.honda@ono.co.jp
[1]Oncology Research Center, Ono Pharmaceutical Co., Ltd., 3-1-1 Sakurai, Mishima-gun, Osaka 618-8585, Japan
[2]Graduate School of Medicine, Kurume University, 67 Asahi-machi, Kurume, Fukuoka 830-0011, Japan
Full list of author information is available at the end of the article

probabilities are the same. Landis and Koch provided benchmarks of the strength of consistency as follows: values ≤0 as poor, 0.00 to 0.20 as slight, 0.21 to 0.40 as fair, 0.41 to 0.60 as moderate, 0.61 to 0.80 as substantial and 0.81 to 1.00 as almost perfect agreement [3]. Although the authors acknowledge the arbitrary nature of their benchmarks, they recommended their benchmark scale as a useful guideline for practitioners.

Many extensions have been made to Cohen's κ including those for agreement in the cases of ordinal data [4], multiple raters [5–9], comparisons of correlated κ 's [10–13] and stratified data [14, 15]. However, as Feinstein and Cicchetti showed, Cohen's κ depends strongly on the marginal distributions and therefore behaves paradoxically [16]. This behavior can be explained by the bias effect and the prevalence effect, on which various discussions have been undertaken [16–18]. A number of alternative measures of agreements have also been proposed, such as Holley and Guilford's G [19], Aickin's α [20], Andres and Marzo's delta [21], Marasini's s^* [22, 23] and Gwet's AC_1 [24] and AC_2 [25].

Gwet showed that AC_1 has better statistical properties (bias and variance) than Cohen's κ, Scott's π and G-index under a limited set of simulations for two raters with binary outcomes [24]. Shanker and Bangdiwala compared Cohen's κ, Scott's π, Prevalence Adjusted Bias Adjusted Kappa (PABAK) [26], AC_1 and B-statistic [27], which is not a kappa-type chance-corrected measure, in the case of two raters and binary outcomes and showed that AC_1 has better properties than other kappa-type measures [28]. In addition, AC_1 has been utilized in the field of medical research over the past decade [29–35]. Therefore, in this study we have limited our discussion to AC_1 in the case of two raters with binary outcomes.

First, a brief review of the concept of Gwet's AC_1 is provided. Consider the situation in which two raters independently classify randomly extracted subject as positive (+) or negative (−). Gwet defined two events: G = {the two raters agree} and R = {at least one rater performs random rating}. The probability of agreement expected by chance is then $p_e = P(G \cap R) = P(G|\,R)P(R)$. A random rating would lead to the classification of an individual into each category with the same probability $\frac{1}{2}$ and it follows that $P(G|R) = 2 \times (\frac{1}{2}) \times (\frac{1}{2}) = \frac{1}{2}$. As for the estimation of $P(R)$, this probability cannot be obtained from data. Therefore, Gwet proposed approximating it with a normalized measure of randomness Ψ, defined as follows:

$$P(R) \simeq \Psi = \frac{\pi_+(1-\pi_+)}{\frac{1}{2}(1-\frac{1}{2})} = 4\pi_+(1-\pi_+), \tag{1}$$

where π_+ is the probability that a randomly chosen rater classifies a randomly chosen subject into the + category. Thus, the approximated probability of chance agreement is represented by

$$p_e^* = P(G|R)\Psi = 2\pi_+(1-\pi_+). \tag{2}$$

AC_1 is thus defined as follows:

$$\gamma = \frac{p_a - p_e^*}{1-p_e^*}, \tag{3}$$

where p_a is the probability of agreement. Although p_e is approximated to p_e^*, Gwet showed that the bias of γ, the difference between γ and the true inter-rater reliability, is equal to or less than Cohen's κ, Scott's π and G-index under some assumption in the case of two raters with binary outcomes. Gwet also provided an estimator $\hat{\gamma}^*$ of γ and its variance for multiple raters and multiple categories based on the randomization approach, which requires the selection of subjects to be random in such a way that all possible subject samples have the exact same chance of being selected. However, it is advantageous to employ a model-based approach when, for example, the evaluation of the effect of subject covariates on agreement is of interest. Therefore, in the case of two raters with binary outcomes, Ohyama [36] assumed the underlying probability that a subject is rated as (+) and its marginal homogeneity of the two raters, and then constructed the likelihood. The maximum likelihood estimator of γ, which is shown to be identical to the estimator given by Gwet, was derived. The likelihood-based confidence intervals for AC_1, inclusion of subject covariates, hypothesis testing and sample size determination were also discussed [36].

In this article, we discuss stratification analyses as another approach to adjust the effect of subject covariates on agreement. For example, a clinical assessment whether a patient has a particular disease symptom may be influenced by overall severity of the disease. In such a case, we consider stratification based on the severity of the disease. Another example is a multicenter inter-rater agreement study, in which the classifications for subjects are conducted independently in each center. These situations require several independent agreement statistics. Then the main purpose of the analyses would be testing whether the degree of inter-rater agreement can be regarded as homogeneous across strata, such as centers and severities of the disease.

For κ, Fleiss has been at the forefront of the idea of χ^2 test-based inter-class consistency with large sample variances [37] and further studies by Donner, Eliasziw and Klar [14], Nam [15, 38] and Wilding, Consiglio and Shan [39] have developed the homogeneity test of κ across covariate levels. However, there are no reports on homogeneity tests for AC_1 or on an estimator of common

AC_1. Therefore, in this article, we derive the homogeneity score test for AC_1 from K independent strata and its performance is investigated. An estimation of the common AC_1 between strata and its confidence intervals is also discussed. Finally, an example application of our approach to clinical trial data is provided.

Methods

Homogeneity tests

Score test

Consider K independent strata involving n_k subjects for $k = 1, \ldots, K$. In each stratum, two raters independently classify subjects as either positive (+) or negative (−). Let $X_{kij} = 1$ if subject $i(=1, \ldots, n_k)$ in the k-th stratum is classified as "+" by rater $j(=1, 2)$ and $X_{kij} = 0$ otherwise. Suppose that $P(X_{kij} = 1 | i) = u_{ki}$, $E(u_{ki}) = \pi_k$ and $Var(u_{ki}) = \sigma_k^2$. The γ of the k-th stratum is then expressed as follows [36]:

$$\gamma_k = \frac{1 + 2[\sigma_k{}^2 - 2\pi_k(1-\pi_k)]}{1 - 2\pi_k(1-\pi_k)}. \tag{4}$$

Let the number of observed pairs in the three categories of the k-th stratum be x_{1k}, x_{2k} and x_{3k} and their corresponding probabilities be $P_{1k}(\gamma_k)$, $P_{2k}(\gamma_k)$ and $P_{3k}(\gamma_k)$. The data of the k-th stratum are then given as shown in Table 1.

The log-likelihood function is given by

$$l(\boldsymbol{\gamma}, \boldsymbol{\pi}) = \sum_{k=1}^{K} l_k(\gamma_k, \pi_k), \tag{5}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)'$, $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, $l_k(\gamma_k, \pi_k) = x_{1k} \log P_{1k}(\gamma_k) + x_{2k} \log P_{2k}(\gamma_k) + x_{3k} \log P_{3k}(\gamma_k)$,

$$P_{1k}(\gamma_k) = \pi_k(2-\pi_k) - \frac{1}{2} + \frac{\gamma_k}{2} A_k, \tag{6}$$

$$P_{2k}(\gamma_k) = A_k(1-\gamma_k), \tag{7}$$

$$P_{3k}(\gamma_k) = (1-\pi_k)(1+\pi_k) - \frac{1}{2} + \frac{\gamma_k}{2} A_k, \tag{8}$$

and $A_k = 1 - 2\pi_k(1 - \pi_k)$.

The maximum likelihood estimators of γ_k and π_k are then given by

Table 1 Data layout

Category	Ratings	Frequency	Probability
1	(+, +)	x_{1k}	$P_{1k}(\gamma_k)$
2	(+, −) or (−, +)	x_{2k}	$P_{2k}(\gamma_k)$
3	(−, −)	x_{3k}	$P_{3k}(\gamma_k)$
Total		n_k	1

$$\hat{\gamma}_k = 1 - \frac{2n_k x_{2k}}{n_k^2 + (x_{1k} - x_{3k})^2} \tag{9}$$

and

$$\hat{\pi}_k = \frac{2x_{1k} + x_{2k}}{2n_k}, \tag{10}$$

respectively.

The first and second derivatives of the log-likelihood function and the Fisher information matrix are given in the Appendix. The aim of this study is to test the homogeneity of the agreement coefficients among K strata, and thus the null hypothesis to test is represented by $H_0 : \gamma_k = \gamma_0$ $(k = 1, 2, \ldots, K)$. The score test statistic for the null hypothesis is derived as follows (see Appendix):

$$T(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}}) = \sum_{k=1}^{K} \frac{\tilde{R}_k{}^2 \tilde{D}_k}{n_k(\tilde{B}_k \tilde{D}_k - \tilde{C}_k{}^2)}, \tag{11}$$

where $\tilde{B}_k, \tilde{C}_k, \tilde{D}_k$ and $\tilde{R}_k$ are obtained by substituting the maximum likelihood estimators $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ under the null hypothesis into

$$\begin{aligned}
B_k &= \tfrac{1}{P_{1k}} + \tfrac{4}{P_{2k}} + \tfrac{1}{P_{3k}}, \\
C_k &= \tfrac{1}{P_{1k}} - \tfrac{1}{P_{3k}} + (1-\gamma_k)(1-2\pi_k)B_k, \\
D_k &= \tfrac{1}{P_{1k}} + \tfrac{1}{P_{3k}} + (1-\gamma_k)(1-2\pi_k)(\tfrac{1}{P_{1k}} - \tfrac{1}{P_{3k}} + C_k), \\
R_k &= \tfrac{x_{1k}}{P_{1k}} - \tfrac{2x_{2k}}{P_{2k}} + \tfrac{x_{3k}}{P_{3k}}.
\end{aligned}$$

$T(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}})$ is asymptotically distributed as a χ^2 with $K - 1$ degrees of freedom. The homogeneity hypothesis is rejected at level α when $T(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}}) \geq \chi^2_{(1-\alpha),K-1}$, where $\chi^2_{(1-\alpha),K-1}$ is the $100 \times (1 - \alpha)$ percentile point of the χ^2 distribution with $K - 1$ degrees of freedom.

Note that, since $0 \leq P_{1k}(\gamma_k), P_{2k}(\gamma_k), P_{3k}(\gamma_k) \leq 1$ and $P_{1k}(\gamma_k) + P_{2k}(\gamma_k) + P_{3k}(\gamma_k) = 1$, substituting (6), (7) and (8) into these equations, the admissible range of γ_k with respect to π_k is obtained as follows [36]:

$$\frac{2 - (1 - |1-2\pi_k|)(3 + |1-2\pi_k|)}{2 - (1 - |1-2\pi_k|)(1 + |1-2\pi_k|)} \leq \gamma_k \leq 1. \tag{12}$$

When obtaining the maximum likelihood estimators $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ under the null hypothesis by numerical calculation, initial values need to be set to satisfy this condition.

Goodness-of-fit test

Donner, Eliasziw and Klar proposed a goodness-of-fit approach for testing homogeneity of kappa statistics in the case of two raters with binary outcomes [40]. This procedure can also be applied to AC_1 statistics. Given that the frequencies x_{1k}, x_{2k}, x_{3k}, $k = 1, \ldots, K$ in Table 1 follow a multinomial distribution conditional on n_k, estimated probabilities under H_0 are given by $\hat{P}_{hk}(\tilde{\gamma}_0)$, which is obtained by

replacing π_k by $\hat{\pi}_k$ and γ_k by $\tilde{\gamma}_0$ in $P_{hk}(\gamma_k)$; $h = 1, 2, 3$; $k = 1, \ldots, K$. Then the goodness-of-fit statistic is derived as follows:

$$\chi_G^2 = \sum_{k=1}^{K}\sum_{h=1}^{3}\frac{\left(x_{hk} - n_k\hat{P}_{hk}(\tilde{\gamma}_0)\right)^2}{n_k\hat{P}_{hk}(\tilde{\gamma}_0)}, \tag{13}$$

under H_0, χ_G^2 follows an approximate χ^2 distribution with $K-1$ degrees of freedom. The homogeneity hypothesis is rejected at level α when $\chi_G^2 \geq \chi_{(1-\alpha),K-1}^2$, where $\chi_{(1-\alpha),K-1}^2$ is the $100 \times (1-\alpha)$ percentile point of the χ^2 distribution with $K-1$ degrees of freedom.

Estimation of common AC_1

If the assumption of homogeneity is reasonable, the estimate of γ_0 can be used as an appropriate summary measure of reliability. The maximum likelihood estimator $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ are obtained by maximizing the log-likelihood functions $l_0(\gamma_0, \boldsymbol{\pi}) = \sum_{k=1}^{K} l_k(\gamma_0, \pi_k)$. Since an analytical solution cannot be obtained from this function, numerical iterative calculations are used. The variance $Var(\tilde{\gamma}_0)$ of $\tilde{\gamma}_0$ can be expressed as follows (see Appendix):

$$Var(\tilde{\gamma}_0) = 4\left[\sum_{k=1}^{K} n_k {A_k}^2\left(B_k^{(0)} - \frac{C_k^{(0)2}}{D_k^{(0)}}\right)\right]^{-1} = \left[\sum_{k=1}^{K}\frac{1}{Var_k(\tilde{\gamma}_0)}\right]^{-1}, \tag{14}$$

where $B_k^{(0)}, C_k^{(0)}, D_k^{(0)}$ are values using $\gamma_k = \gamma_0$ in B_k, C_k, D_k respectively, and

$$Var_k(\tilde{\gamma}_0) = \frac{1}{n_k A_k^2}\left[A_k(1-\gamma_0) - \left(A_k^2 - 4A_k + 2\right)(1-\gamma_0)^2 - A_k(2A_k - 1)(1-\gamma_0)^3\right]. \tag{15}$$

A simple $100 \times (1-\alpha)$ % confidence interval using the asymptotic normality of $\tilde{\gamma}_0$ can be expressed as follows:

$$\tilde{\gamma}_0 \pm Z_{\alpha/2}\sqrt{\widehat{Var}(\tilde{\gamma}_0)}, \tag{16}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ upper quantile of the standard normal distribution and $\widehat{Var}(\tilde{\gamma}_0)$ is obtained by substituting $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ into (14). Hereafter, this method is referred to as the simple asymptotic (SA) method. Since Eq. (14) depends on γ_0, SA method may not have the correct coverage rate, and the normality of the sampling distribution of $\tilde{\gamma}_0$ may be improved using Fisher's Z transformation. This method is referred to below as Fisher's Z transformation (FZ) method (see Appendix).

As an alternative method, we employ the profile variance approach, which has been shown to perform well in the case of the intra-class κ for binary outcome data [41–43]. This approach also performs well for AC_1 in the case of two raters with binary outcomes [36]. The confidence interval based on the profile variance can be obtained by solving the following inequality for γ_0:

$$\frac{(\tilde{\gamma} - \gamma_0)^2}{\widetilde{Var}(\tilde{\gamma}_0)} \leq Z_{\alpha/2}^2, \tag{17}$$

where $\widetilde{Var}(\tilde{\gamma}_0)$ is given by substituting $\tilde{\pi}_k$ into π_k in (15). Hereafter, this method is referred to as the profile variance (PV) method (see Appendix).

Numerical evaluations

We conducted Monte Carlo simulations to investigate the performance of the proposed homogeneity tests and to evaluate the estimate of common AC_1 and its confidence intervals under the following conditions: the number of strata in the simulation is $K = 2$ or 3; and random observations are generated from the trinomial distributions according to the probabilities of (6), (7) and (8) by giving the values of γ_k and π_k. The balanced and unbalanced cases were considered for the values of π_k and n_k. The values of γ_k and π_k are set within the theoretical range of Eq. (12) derived in the preceding paragraph. Ten thousand times of iterations were carried out for each parameter combination.

When π_k is close to 0 or 1 and n_k is small, there are cases in which the generated data include zero cells. In such cases, B_k, C_k, D_k and R_k cannot be estimated . Thus, when zero cells were generated, we adopted the approach of adding 0.5 to the frequency of each combination by two raters, (+,+), (+,−), (−,+), (−,−). This simple method was discussed by Agresti [44] and was adopted in a previous study [39].

Results

Empirical type I error rate for the homogeneity test

The type I error rates of the homogeneity tests with a significance level of 0.05 were examined. The sample size was set at $n_k = n$ = 20, 50, 80 for balanced settings and $(n_1, n_2, n_3) = (20, 50, 80)$ for unbalanced settings. The error rate obtained by the score test is expressed as SCORE and the error rate obtained by the goodness-of-fit test is expressed as GOF. Table 2 summarizes the results for $K = 2$.

Overall, the proposed score test did not show any significant type I error rate inflation, but it was very conservative when sample size was small and γ_0 was close to 1.

In the case of n = 20 when γ_0 = 0.1, 0.3 or 0.5, the type I error rates of SCORE were maintained at the nominal level of 0.05 regardless of whether π_k was balanced or unbalanced, but when γ_0 = 0.7 or 0.9, the type I error rates were slightly conservative. Especially when γ_0 = 0.9, the rate was significantly conservative to the extent of being less than 0.01. In the case of n = 50, the type I error rates were maintained at the nominal level of 0.05 except when γ_0 = 0.9. Finally in the case of n = 80, the type I error rates were almost maintained at the nominal level. In

Table 2 Empirical type I error rates of homogeneity tests for $\gamma_1 = \gamma_2 = \gamma_0$ based on 10,000 simulations ($K = 2$ balanced sample size)

Balanced π conditions					Unbalanced π conditions					
$n_1 = n_2$	γ_0	$\pi_1 = \pi_2$	SCORE	GOF	$n_1 = n_2$	γ_0	π_1	π_2	SCORE	GOF
20	0.1	0.5	0.045	0.067	20	0.1	0.5	0.35	0.049	0.097
	0.3		0.046	0.067		0.3			0.049	0.096
	0.5		0.048	0.062		0.5			0.050	0.083
	0.7		0.033	0.041		0.7			0.037	0.049
	0.9		0.002	0.003		0.9			0.003	0.005
	0.1	0.35	0.052	0.121		0.1	0.65	0.35	0.050	0.120
	0.3		0.054	0.126		0.3			0.051	0.120
	0.5		0.052	0.103		0.5			0.051	0.101
	0.7		0.039	0.064		0.7			0.039	0.065
	0.9		0.004	0.006		0.9			0.004	0.007
	0.7	0.2	0.047	0.132		0.7	0.5	0.2	0.038	0.090
	0.9		0.008	0.029		0.9			0.005	0.013
50	0.1	0.5	0.050	0.058	50	0.1	0.5	0.35	0.048	0.117
	0.3		0.047	0.054		0.3			0.051	0.087
	0.5		0.050	0.054		0.5			0.049	0.072
	0.7		0.051	0.053		0.7			0.050	0.060
	0.9		0.026	0.027		0.9			0.024	0.027
	0.1	0.35	0.051	0.172		0.1	0.65	0.35	0.051	0.168
	0.3		0.051	0.126		0.3			0.049	0.117
	0.5		0.052	0.092		0.5			0.051	0.092
	0.7		0.052	0.072		0.7			0.052	0.071
	0.9		0.028	0.033		0.9			0.028	0.033
	0.7	0.2	0.053	0.162		0.7	0.5	0.2	0.051	0.104
	0.9		0.037	0.061		0.9			0.032	0.042
80	0.1	0.5	0.047	0.052	80	0.1	0.5	0.35	0.051	0.120
	0.3		0.047	0.051		0.3			0.053	0.094
	0.5		0.054	0.057		0.5			0.051	0.072
	0.7		0.050	0.052		0.7			0.051	0.061
	0.9		0.037	0.039		0.9			0.047	0.050
	0.1	0.35	0.052	0.173		0.1	0.65	0.35	0.052	0.172
	0.3		0.054	0.123		0.3			0.054	0.124
	0.5		0.053	0.089		0.5			0.055	0.090
	0.7		0.051	0.069		0.7			0.051	0.069
	0.9		0.044	0.051		0.9			0.045	0.052
	0.7	0.2	0.052	0.152		0.7	0.5	0.2	0.050	0.103
	0.9		0.051	0.073		0.9			0.048	0.059

contrast, the type I error rate of GOF tended to be larger than that of SCORE and in many cases it was not maintained at the nominal level.

The results obtained for $K = 3$ are shown Table S1 and Table S2 in Additional file 1.

The Additional file 2 provides the simulation code of empirical type I error rate using R language.

Empirical power of the homogeneity test

The empirical power of the score test was investigated only for the case of $K = 2$, by setting γ_1 = 0.1, 0.3, 0.5 and $\gamma_2 - \gamma_1$ = 0.3, 0.4. The values of π_k and n_k were set as in the type I error simulation. The results are shown in Table 3. The power tended to be large as the value of γ_1 increased under the fixed values of π and $\gamma_2 - \gamma_1$.

Table 3 Empirical power of homogeneity tests based on 10,000 simulations ($K = 2$ balanced sample size)

Balanced π conditions						Unbalanced π conditions						
$n_1 = n_2$	γ_1	γ_2	$\pi_1 = \pi_2$	SCORE	GOF	$n_1 = n_2$	γ_1	γ_2	π_1	π_2	SCORE	GOF
20	0.1	0.5	0.5	0.243	0.290	20	0.1	0.5	0.5	0.35	0.234	0.304
	0.3	0.6		0.173	0.202		0.3	0.6			0.168	0.224
	0.3	0.7		0.294	0.323		0.3	0.7			0.293	0.351
	0.5	0.8		0.212	0.232		0.5	0.8			0.216	0.258
	0.5	0.9		0.372	0.396		0.5	0.9			0.389	0.430
	0.1	0.5	0.35	0.245	0.357		0.1	0.5	0.65	0.35	0.243	0.355
	0.3	0.6		0.185	0.279		0.3	0.6			0.171	0.266
	0.3	0.7		0.313	0.408		0.3	0.7			0.296	0.398
	0.5	0.8		0.227	0.295		0.5	0.8			0.221	0.291
	0.5	0.9		0.396	0.483		0.5	0.9			0.394	0.475
	0.5	0.8	0.2	0.270	0.411		0.5	0.8	0.5	0.2	0.230	0.278
	0.5	0.9		0.482	0.616		0.5	0.9			0.435	0.473
50	0.1	0.5	0.5	0.538	0.562	50	0.1	0.5	0.5	0.35	0.525	0.595
	0.3	0.6		0.377	0.396		0.3	0.6			0.369	0.428
	0.3	0.7		0.635	0.651		0.3	0.7			0.630	0.673
	0.5	0.8		0.512	0.525		0.5	0.8			0.517	0.548
	0.5	0.9		0.835	0.841		0.5	0.9			0.843	0.855
	0.1	0.5	0.35	0.509	0.652		0.1	0.5	0.65	0.35	0.503	0.648
	0.3	0.6		0.379	0.485		0.3	0.6			0.364	0.470
	0.3	0.7		0.633	0.718		0.3	0.7			0.618	0.711
	0.5	0.8		0.518	0.585		0.5	0.8			0.516	0.581
	0.5	0.9		0.844	0.877		0.5	0.9			0.838	0.874
	0.5	0.8	0.2	0.552	0.711		0.5	0.8	0.5	0.2	0.531	0.598
	0.5	0.9		0.878	0.915		0.5	0.9			0.861	0.863
80	0.1	0.5	0.5	0.757	0.768	80	0.1	0.5	0.5	0.35	0.732	0.786
	0.3	0.6		0.568	0.578		0.3	0.6			0.545	0.596
	0.3	0.7		0.841	0.847		0.3	0.7			0.833	0.858
	0.5	0.8		0.716	0.722		0.5	0.8			0.717	0.741
	0.5	0.9		0.967	0.968		0.5	0.9			0.966	0.970
	0.1	0.5	0.35	0.707	0.819		0.1	0.5	0.65	0.35	0.707	0.816
	0.3	0.6		0.538	0.645		0.3	0.6			0.541	0.644
	0.3	0.7		0.826	0.879		0.3	0.7			0.822	0.884
	0.5	0.8		0.717	0.767		0.5	0.8			0.715	0.764
	0.5	0.9		0.963	0.973		0.5	0.9			0.965	0.974
	0.5	0.8	0.2	0.746	0.872		0.5	0.8	0.5	0.2	0.734	0.787
	0.5	0.9		0.976	0.982		0.5	0.9			0.974	0.970

The empirical power of the GOF test was also examined under the same simulation conditions as the score test. The results are also shown in Table 3. However, the GOF had a large type I error rate inflation (Table 2) and was invalid as a test.

The Additional file 2 provides the simulation code of empirical power using R language.

Bias and mean square error for common AC_1

We evaluated the bias and mean square error (MSE) of the maximum likelihood estimator for the common AC_1, $\tilde{\gamma}_0$. The balanced and unbalanced conditions for π_k and the balanced condition for n_k were considered. The results are shown in Table 4. The bias of $\tilde{\gamma}_0$ tended to be small as γ_0 increased, but $\tilde{\gamma}_0$ was almost unbiased. As expected, the bias and MSE tended to be small as the sample size increased.

The Additional file 3 provides the simulation code of bias and mean square error for common AC_1 using R language.

Confidence intervals for common AC_1

We conducted a simulation study to evaluate the performances of the three confidence intervals presented in the previous section. The coverage rates of the 95% confidence interval were examined. The balanced and unbalanced conditions for π_k and the balanced condition for n_k are considered. The results are shown in Table 5. The coverage rate of the SA method was generally lower than 0.95 under many conditions, with the exception of the value being close to 0.99 in the case of $n_1 = n_2 = 20$ and $\gamma_0 = 0.9$. The FZ method and PV method greatly improved the coverage rates close to the nominal level. However, the coverage rate of the PV method was closer to the nominal level than that of the FZ method in most cases under the conditions examined. The coverage rates of each method were also evaluated in the case of $K = 3$, and the unbalanced n_k conditions and both the FZ method and the PV method achieved coverage rates near 0.95 (results not shown).

Table 4 Bias and mean square error of the maximum likelihood estimator for the common AC_1 based on 10,000 simulations ($K = 2$ balanced sample size)

Balanced π conditions					Unbalanced π conditions					
$n_1 = n_2$	γ_0	$\pi_1 = \pi_2$	Bias	MSE	$n_1 = n_2$	γ_0	π_1	π_2	Bias	MSE
20	0.1	0.5	0.026	0.025	20	0.1	0.5	0.35	0.019	0.027
	0.3		0.023	0.023		0.3			0.018	0.024
	0.5		0.017	0.018		0.5			0.013	0.019
	0.7		0.009	0.012		0.7			0.007	0.012
	0.9		−0.011	0.003		0.9			−0.011	0.003
	0.1	0.35	0.009	0.029		0.1	0.65	0.35	0.010	0.028
	0.3		0.007	0.025		0.3			0.011	0.025
	0.5		0.007	0.019		0.5			0.008	0.019
	0.7		0.004	0.012		0.7			0.004	0.012
	0.9		−0.011	0.003		0.9			−0.011	0.003
	0.7	0.2	−0.007	0.012		0.7	0.5	0.2	0.001	0.012
	0.9		−0.010	0.003		0.9			−0.010	0.003
80	0.1	0.5	0.007	0.006	80	0.1	0.5	0.35	0.006	0.007
	0.3		0.006	0.006		0.3			0.005	0.006
	0.5		0.005	0.005		0.5			0.004	0.005
	0.7		0.003	0.003		0.7			0.003	0.003
	0.9		0.002	0.001		0.9			0.001	0.001
	0.1	0.35	0.004	0.007		0.1	0.65	0.35	0.003	0.008
	0.3		0.002	0.006		0.3			0.001	0.006
	0.5		0.001	0.005		0.5			0.002	0.005
	0.7		0.001	0.003		0.7			0.001	0.003
	0.9		0.000	0.001		0.9			0.000	0.001
	0.7	0.2	−0.001	0.003		0.7	0.5	0.2	0.001	0.003
	0.9		−0.001	0.001		0.9			0.001	0.001

Table 5 Coverage rates of common γ 95% confidence intervals of the three proposed methods based on 10,000 simulations

Balanced π conditions						Unbalanced π conditions						
$n_1 = n_2$	γ_0	$\pi_1 = \pi_2$	SA	FZ	PV	$n_1 = n_2$	γ_0	π_1	π_2	SA	FZ	PV
20	0.1	0.5	0.939	0.959	0.958	20	0.1	0.5	0.35	0.939	0.958	0.958
	0.3		0.936	0.958	0.950		0.3			0.933	0.958	0.955
	0.5		0.931	0.962	0.962		0.5			0.927	0.959	0.960
	0.7		0.924	0.976	0.963		0.7			0.917	0.971	0.961
	0.9		0.998	0.963	0.955		0.9			0.997	0.966	0.955
	0.1	0.35	0.935	0.953	0.956		0.1	0.65	0.35	0.938	0.955	0.957
	0.3		0.934	0.955	0.953		0.3			0.934	0.955	0.956
	0.5		0.926	0.956	0.955		0.5			0.927	0.959	0.955
	0.7		0.918	0.965	0.958		0.7			0.917	0.967	0.960
	0.9		0.996	0.967	0.947		0.9			0.996	0.969	0.950
	0.7	0.2	0.929	0.959	0.950		0.7	0.5	0.2	0.931	0.965	0.956
	0.9		0.970	0.966	0.911		0.9			0.993	0.970	0.943
50	0.1	0.5	0.949	0.953	0.952	50	0.1	0.5	0.35	0.946	0.952	0.953
	0.3		0.945	0.955	0.954		0.3			0.943	0.952	0.951
	0.5		0.945	0.954	0.953		0.5			0.941	0.952	0.952
	0.7		0.936	0.961	0.953		0.7			0.936	0.956	0.953
	0.9		0.920	0.971	0.971		0.9			0.923	0.968	0.965
	0.1	0.35	0.942	0.948	0.952		0.1	0.65	0.35	0.946	0.954	0.954
	0.3		0.942	0.950	0.950		0.3			0.943	0.953	0.952
	0.5		0.940	0.951	0.951		0.5			0.941	0.954	0.953
	0.7		0.937	0.954	0.952		0.7			0.936	0.956	0.953
	0.9		0.926	0.966	0.960		0.9			0.925	0.968	0.960
	0.7	0.2	0.938	0.949	0.948		0.7	0.5	0.2	0.937	0.954	0.952
	0.9		0.927	0.965	0.954		0.9			0.928	0.969	0.960
80	0.1	0.5	0.945	0.952	0.952	80	0.1	0.5	0.35	0.946	0.951	0.951
	0.3		0.947	0.952	0.952		0.3			0.946	0.952	0.952
	0.5		0.943	0.953	0.952		0.5			0.942	0.950	0.950
	0.7		0.944	0.950	0.949		0.7			0.943	0.955	0.953
	0.9		0.901	0.962	0.956		0.9			0.922	0.963	0.957
	0.1	0.35	0.946	0.951	0.951		0.1	0.65	0.35	0.943	0.947	0.946
	0.3		0.944	0.949	0.949		0.3			0.945	0.950	0.950
	0.5		0.944	0.950	0.950		0.5			0.944	0.953	0.952
	0.7		0.941	0.950	0.949		0.7			0.939	0.953	0.952
	0.9		0.931	0.960	0.953		0.9			0.927	0.963	0.954
	0.7	0.2	0.944	0.950	0.949		0.7	0.5	0.2	0.942	0.955	0.954
	0.9		0.929	0.956	0.951		0.9			0.927	0.961	0.955

SA, FZ, and PV refer to 95% confidence intervals for common AC_1 using the simple asymptotic, Fisher's Z transformation, and profile variance methods, respectively

The Additional file 4 provides the simulation code of confidence intervals for common AC_1 using R language.

An example

As an example, we used data from a randomized clinical trial called the Silicon Study, which was conducted to investigate the effectiveness of silicone fluids versus gases in the management of proliferative vitreoretinopathy (PVR) by vitrectomy [45]. The PVR classification, determined at the baseline visit, defines the severity of the disease as a continuum of increasing pathology graded as C3, D1, D2 or D3. The presence or absence of retinal injury in the superior nasal cavity was evaluated clinically by the operating ophthalmic surgeon and photographically by an independent fundus photograph reading center [46].

The data and results are summarized in Table 6. For reference, the results of the homogeneity score test proposed by Nam for the intra-class κ are also provided [15]. The probabilities of agreement in each stratum were from 0.800 to 0.880 and not so different. However, the values of κ in each stratum were from 0.117 to 0.520 and were greatly different. This might be due to the prevalence effect caused by the small values of π. In contrast, the values of γ were 0.723 to 0.861 and did not differ greatly among strata.

The proposed homogeneity score statistic $T(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}})$ was 2.060 (p-value = 0.560) and the homogeneity hypothesis was not rejected. The estimate of common AC_1 was 0.808 and its 95% confidence intervals were 0.743–0.873 (SA method), 0.732–0.864 (FZ method) and 0.730–0.862 (PV method). Also, the score statistic for testing the homogeneity of κ's [15] was 2.700 (p-value = 0.440) and the common κ was 0.352.

The Additional file 5 provides the code for clinical data examples using R language.

To investigate the sensitivity of the indicators to π_k, we hypothetically considered more balanced and less balanced π_k under fixed p_a and n_k in each stratum. The generated data set and analysis results are summarized as Table S3 in the Additional file 1. κ was more sensitive to changes in the value of π, but AC_1 was less sensitive to changes in the value of π than κ. The common AC_1 was not affected as much as the common κ even if the π balance was lost.

Table 6 Agreement between ophthalmologist and reading center classifying superior nasal retinal breaks stratified by PVR grade

	PVR grade			
	C3	D1	D2	D3
Both (x_1)	1	6	5	3
One (x_2)	9	8	11	9
Neither (x_3)	65	46	54	33
Total (n)	75	60	70	45
π	0.073	0.167	0.150	0.167
p_a	0.880	0.867	0.843	0.800
κ (MLE)	0.117	0.520	0.384	0.280
AC_1 (MLE)	0.861	0.815	0.789	0.723

Discussion

It is well known that Cohen's κ depends strongly on the marginal distributions, and Gwet proposed alternative and more stable measures of agreement, AC_1 for nominal data and its extended agreement AC_2 for ordinal data [24, 25]. A number of alternative measures have also been proposed, as in Holley and Guilford's G [19], Aickin's α [20], Andres and Marzo's delta [21] and Marasini's s^* [22, 23]. Gwet [24] and Shankar and Bangdiwala [28] compared some measures and showed that AC_1 has better properties than other kappa-type measures. In addition, AC_1 has been utilized in the field of medical research over the past decade [29–35]. However statistical inference procedures of AC_1 have not been discussed sufficiently. Therefore, Ohyama expressed AC_1 using population parameters to develop a likelihood-based inference procedure and constructed confidence intervals of AC_1 based on profile variances and likelihood ratios. Inclusion of subjects' covariates, hypothesis testing and sample size estimation were also presented [36]. In the present study, the case of stratified data was discussed as one development of Ohyama [36] for two raters with binary outcomes. Furthermore, tests were derived for the homogeneity of AC_1 between K independent strata and the inference of common AC_1 was discussed.

In the numerical evaluation of type I error, both tests were conservative when the sample size was small and γ_0 was 0.9, but the conservativeness was relaxed when the sample size was as large as 80. In other settings of simulation, the score test performed well while GOF sometimes could not achieve the nominal level. Therefore, we recommend using the score test for testing the homogeneity of AC_1 among K strata. Note that, when zero cells are observed, the homogeneity score test statistic cannot be calculated. In such cases in our simulation study, we simply added 0.5 to the data set, which had no serious effect on the performance of the proposed score test in our simulation settings.

If the homogeneity assumption is reasonable, it may be desired to provide an estimate of the common AC_1 as a summary measure of reliability. In the present study, we proposed an estimator of common AC_1 and constructed its confidence intervals based on the SA, FZ, and PV methods. We also evaluated the performance of each numerically. The bias and MSE tended to be small as the sample size increased, and the results were nearly

0 when n = 80. The PV method provides coverage levels close to nominal in most situations, while the SA method tends to provide a shortage of coverage and the FZ method tends to provide excess coverage in some situations. Therefore, we recommend the PV method for calculating confidence intervals.

As in the PVR example, AC_1 in each stratum is less affected by the prevalence or marginal probability than by the κ. It is suggested that the proposed homogeneity test and the general framework of common AC_1 estimation are also essentially more stable than those of the κ.

There were some limitations in this study. First, as described above, the performance of the proposed score test was very conservative when γ_0 = 0.9 and sample size was small. An exact approach might be an alternative method in such cases.

Next, in this study, the cases were limited to two raters with binary outcomes in each stratum. However, in the evaluation of medical data, it is often the case that multiple raters classify subjects into nominal or ordered categories. Our proposed method may be extended to the case of multiple raters with binary outcomes using the likelihood function for multiple raters. In the cases of two raters with nominal outcomes, Agresti [47] proposed a quasi-symmetry model with kappa as a parameter, and this technique may be extended to AC_1 in the case of stratified data.

Finally, continuous covariates need to be categorized adequately to apply the proposed approach. The regression model proposed by Ohyama [36] can be used to assess the effect of continuous covariates on AC_1, but it is limited to the case of two raters with binary data. Nelson and Edwards [48] and Nelson, Mitani and Edwards [49] proposed a method for constructing a measure of agreement using generalized linear mixed-effect models by introducing continuous latent variables representing the subject's true disease status and for flexibly incorporating rater and subject covariates. These approaches might be applicable to AC_1 and AC_2.

Conclusion

The method proposed in this study is considered to be useful for summarizing evaluations of consistency performed in multiple or stratified inter-rater agreement studies. In addition, the proposed method can be applied not only to medical or epidemiological research but also to assessment of the degree of consistency of characteristics, such as biometrics, psychological measurements, and data in the behavioral sciences.

Appendix

First and second derivatives of the log-likelihood function

The first and second derivatives of $l(\boldsymbol{\gamma}, \boldsymbol{\pi})$, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)'$, $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)'$ with respect to r_k and π_k are obtained as follows:

$$
\begin{aligned}
&\frac{\partial l}{\partial \gamma_k} = \frac{1}{2} A_k R_k, \\
&\frac{\partial l}{\partial \pi_k} = \frac{x_{1k}}{P_{1k}} - \frac{x_{3k}}{P_{3k}} + (1-\gamma_k)(1-2\pi_k) R_k, \\
&\frac{\partial^2 l}{\partial \gamma_k^2} = -\frac{1}{4} A_k^2 S_k, \\
&\frac{\partial^2 l}{\partial \gamma_k \partial \pi_k} = -\frac{A_k}{2}\left[\frac{x_{1k}}{P_{1k}^2} - \frac{x_{3k}}{P_{3k}^2} + (1-\gamma_k)(1-2\pi_k) S_k\right] - (1-2\pi_k) R_k, \\
&\frac{\partial^2 l}{\partial \pi_k^2} = -\frac{x_{1k}}{P_{1k}^2} - \frac{x_{3k}}{P_{3k}^2} - 2\left(\frac{x_{1k}}{P_{1k}^2} - \frac{x_{3k}}{P_{3k}^2}\right)(1-\gamma_k)(1-2\pi_k) - 2(1-\gamma_k) R_k - (1-\gamma_k)^2 (1-2\pi_k)^2 S_k, \\
&\frac{\partial^2 l}{\partial \gamma_k \partial \gamma_{k'}} = \frac{\partial^2 l}{\partial \pi_k \partial \pi_{k'}} = \frac{\partial^2 l}{\partial \gamma_k \partial \pi_{k'}} = 0 \quad \left(k \neq k'\right),
\end{aligned}
$$

where

$$
\begin{aligned}
A_k &= 1 - 2\pi_k(1-\pi_k), \\
R_k &= \frac{x_{1k}}{P_{1k}} - \frac{2x_{2k}}{P_{2k}} + \frac{x_{3k}}{P_{3k}}, \\
S_k &= \frac{x_{1k}}{P_{1k}^2} + \frac{4x_{2k}}{P_{2k}^2} + \frac{x_{3k}}{P_{3k}^2}.
\end{aligned}
$$

Since $E(x_{hk}) = n_k P_{hk}(\gamma_k)$ $(h = 1, 2, 3)$,

$$
\begin{aligned}
E\left(-\frac{\partial^2 l}{\partial \gamma_k^2}\right) &= \frac{n_k}{4} A_k^2 B_k, \\
E\left(-\frac{\partial^2 l}{\partial \gamma_k \partial \pi_k}\right) &= \frac{n_k}{2} A_k C_k, \\
E\left(-\frac{\partial^2 l}{\partial \pi_k^2}\right) &= n_k D_k,
\end{aligned}
$$

where

$$
\begin{aligned}
B_k &= \frac{1}{P_{1k}} + \frac{4}{P_{2k}} + \frac{1}{P_{3k}}, \\
C_k &= \frac{1}{P_{1k}} - \frac{1}{P_{3k}} + (1-\gamma_k)(1-2\pi_k) B_k, \\
D_k &= \frac{1}{P_{1k}} + \frac{1}{P_{3k}} + (1-\gamma_k)(1-2\pi_k)\left(\frac{1}{P_{1k}} - \frac{1}{P_{3k}} + C_k\right).
\end{aligned}
$$

Thus, the Fisher information matrix is given as follows:

$$
I(\boldsymbol{\gamma}, \boldsymbol{\pi}) = \frac{1}{4}\begin{bmatrix} \operatorname{diag}\left(n_k A_k^2 B_k\right) & \operatorname{diag}(2 n_k A_k C_k) \\ \operatorname{diag}(2 n_k A_k C_k) & \operatorname{diag}(4 n_k D_k) \end{bmatrix}.
$$

Score test statistic for the null hypothesis

Define the score function U as

$$
U(\boldsymbol{\gamma}, \boldsymbol{\pi}) = \left(\partial l/\partial \gamma_1, \cdots, \partial l/\partial \gamma_K, \partial l/\partial \pi_1, \cdots, \partial l/\partial \pi_K\right)'
$$

The score statistic for testing the null hypothesis H_0: $\gamma_1 = \cdots = \gamma_K = \gamma_0$ is asymptotically distributed as a χ^2 with K-1 degrees of freedom, and then expressed as

$$
T(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}}) = U(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}})' I(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}})^{-1} U(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}}),
$$

where $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ are the maximum likelihood estimators under H_0, and then score function vector is expressed as $U(\tilde{\gamma}_0, \tilde{\boldsymbol{\pi}}) = \frac{1}{2}(\tilde{A}_1 \tilde{R}_1, \ldots, \tilde{A}_K \tilde{R}_K, 0, \ldots, 0)'$.

The upper left K × K matrix of $I(\tilde{\gamma},\tilde{\boldsymbol{\pi}})^{-1}$ is expressed as follows:

$$I^{\gamma\gamma}(\tilde{\gamma}_0,\tilde{\boldsymbol{\pi}}) = 4(\mathrm{diag}(n_k\tilde{A}_k^2\tilde{B}_k) - \mathrm{diag}(2n_k\tilde{A}_k\tilde{C}_k)\,\mathrm{diag}^{-1}(4n_k\tilde{D}_k)\mathrm{diag}(2n_k\tilde{A}_k\tilde{C}_k))^{-1} = 4\mathrm{diag}\Big[\frac{\tilde{D}_k}{n_k\tilde{A}_k^2(\tilde{B}_k\tilde{D}_k-\tilde{C}_k^2)}\Big],$$

so that, the score test statistic Eq. (11) is derived.

Confidence interval for γ_0 based on Fisher's Z transformation

Fisher's Z transformation of $\tilde{\gamma}_0$ is defined by

$$\tilde{z} = \frac{1}{2}\log\frac{1+\tilde{\gamma}_0}{1-\tilde{\gamma}_0}.$$

Using the delta method, the asymptotic variance of $\tilde{z}$ is represented by

$$Var(\tilde{z}) = \left(\frac{1}{1-\gamma_0^2}\right)^2 Var(\tilde{\gamma}_0),$$

where $Var(\tilde{\gamma})$ is given by Eq. (14). Then a confidence interval for $z = 0.5\log[(1+\gamma_0)/(1-\gamma_0)]$ is obtained by

$$C_\pm = \tilde{z} \pm Z_{\alpha/2}\sqrt{\widehat{Var}(\tilde{z})},$$

where $\widehat{Var}(\tilde{z})$ is defined by substituting $\tilde{\gamma}_0$ and $\tilde{\pi}_k$ into $Var(\tilde{z})$. Thus the confidence interval for γ_0 based on FZ method is obtained as follows:

$$\left(\frac{\exp(2C_-)-1}{\exp(2C_-)+1}, \frac{\exp(2C_+)-1}{\exp(2C_+)+1}\right).$$

Derivation of Eq. (14)

By the second-order partial derivatives of the log-likelihood function $l_0(\gamma_0,\boldsymbol{\pi})$ and taking expectations, we obtain

$$I_{\gamma\gamma} = E\left(-\frac{\partial^2 l_0}{\partial\gamma_0^2}\right) = \frac{1}{4}\sum_{k=1}^K n_k A_k^2 B_k^{(0)},$$
$$I_{kk} = E\left(-\frac{\partial^2 l_0}{\partial\pi_k^2}\right) = n_k D_k^{(0)},$$
$$I_{\gamma k} = E\left(-\frac{\partial^2 l_0}{\partial\gamma_0\partial\pi_k}\right) = \frac{1}{2}n_k A_k C_k^{(0)},$$

where $B_k^{(0)}, C_k^{(0)}$ and $D_k^{(0)}$ are values using $\gamma_k = \gamma_0$ in B_k, C_k and D_k respectively. Let

$$I = \begin{pmatrix} I_{\gamma\gamma} & I_{\gamma\pi} \\ I_{\gamma\pi}' & I_{\pi\pi} \end{pmatrix},$$

where $I_{\gamma\pi} = (I_{\gamma 1},\ldots,I_{\gamma K})$ and $I_{\pi\pi} = \mathrm{diag}(I_{kk})$. When $P_{1k} \neq 0$, $P_{2k} \neq 0$ and $P_{3k} \neq 0$ for all k, $I_{\pi\pi}$ is non-singular matrix and then the element corresponding to $I_{\gamma\gamma}$ of the inverse matrix of I, which is the variance of $\tilde{\gamma}_0$, is given by

$$I^{\gamma\gamma} = \left(I_{\gamma\gamma} - I_{\gamma\pi}I_{\pi\pi}^{-1}I_{\gamma\pi}'\right)^{-1} = \left(I_{\gamma\gamma} - \sum_{k=1}^K I_{\gamma k}^2 I_{kk}^{-1}\right)^{-1} = 4\left[\sum_{k=1}^K n_k {A_k}^2\left(B_k^{(0)} - \frac{{C_k^{(0)}}^2}{D_k^{(0)}}\right)\right]^{-1}.$$

Since

$$\frac{4D_k^{(0)}}{B_k^{(0)}D_k^{(0)} - C_k^{(0)2}} = \left(1-p_{a,k}\right)p_{a,k} + \left(1-\gamma_0\right)\left(1-2p_{e,k}^*\right) \times \left(1+p_{a,k}\right),$$

where $p_{a,k} = \gamma_0(1-p_{e,k}^*) + p_{e,k}^*$ is the probability of agreement in the k-th stratum and $p_{e,k}^* = 2\pi_k(1-\pi_k)$ is the probability of chance agreement in the k-th stratum, and using the variance formula for the single stratum case given by Ohyama [36], $I^{\gamma\gamma}$ can be reduced to the rightmost expression in Eq. (14).

Profile variance approach

The profile variance of a statistic is defined as the variance similar to the estimated variance but without substituting the estimate for the parameter corresponding to the statistic [41].

In this study, $\tilde{\pi}_k$ is substituted for π_k in Eq. (15), and then we obtain the profile variance $\widetilde{Var}(\tilde{\gamma}_0)$ from (14). Since $\tilde{\gamma}_0$ is distributed asymptotically as the normal distribution with mean γ_0 and variance (14), we have

$$\frac{(\tilde{\gamma}_0-\gamma_0)^2}{\tilde{Var}(\tilde{\gamma}_0)} \to \chi_1^2.$$

Thus we can obtain the confidence limits as the two admissible roots of Eq. (17).

Since Eq. (17) is a cubic equation for γ_0 and is complicated to solve, thus we calculated the confidence limit by numerical calculation. The program is given in additional file. Other examples of the profile variance approach for obtaining confidence intervals can be found in many literatures. For example, Bickel and Doksum [50] reported a confidence interval based on the profile variance for the one-sample binomial proportion. Rothman [51] and Afifi, Elashoff and Lee JJ [52] described profile variance type of confidence intervals for survival probability.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12874-019-0887-5.

Additional file 1. Supplementary tables.

Additional file 2. R code for type I errors and power.

Additional file 3. R code for Bias and MSE.

Additional file 4. R code for coverage rates.

Additional file 5. R code for clinical data examples.

Abbreviations

FZ: Fisher's Z transformation; GOF: goodness-of-fit; MLE: maximum likelihood estimator; MSE: mean squared error; PV: profile variance; PVR: proliferative vitreoretinopathy; SA: simple asymptotic

Acknowledgements

We thank Professor T. Kakuma and Dr. T. Yanagawa for providing useful advice on the model construction and evaluation method. We also thank reviewers and editors for constructive and useful advice for improving this article.

Authors' contributions

CH and TO designed the concept of this research. CH conducted the simulation, analyzed a clinical example and drafted the manuscript. TO supervised this study and critically reviewed the manuscript. Both authors have read and approved the manuscript.

Funding

No grant support or other funding was received.

Availability of data and materials

The program codes are shared in additional files. Clinical data referred to are from Barlow, et al. ref. [45].

Ethics approval and consent to participate

Not applicable because of the study involved the development of statistical methods. Clinical example retrospective data were originally published in Barlow, et al. ref. [45].

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Oncology Research Center, Ono Pharmaceutical Co., Ltd., 3-1-1 Sakurai, Mishima-gun, Osaka 618-8585, Japan. [2]Graduate School of Medicine, Kurume University, 67 Asahi-machi, Kurume, Fukuoka 830-0011, Japan. [3]Biostatistics Center, Kurume University, 67 Asahi-machi, Kurume, Fukuoka 830-0011, Japan.

Received: 6 May 2019 Accepted: 13 December 2019
Published online: 05 February 2020

References

1. Cohen J. Coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37–40. https://doi.org/10.1177/001316446002000104.
2. Scott WA. Reliability of content analysis; the case of nominal scale coding. Public Opin Q. 1955;19:321–5. https://doi.org/10.1086/266577.
3. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. https://doi.org/10.2307/2529310.
4. Cohen J. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213–20. https://doi.org/10.1037/h0026256.
5. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. https://doi.org/10.1037/h0031619.
6. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics. 1977;33(2):363–74. https://doi.org/10.2307/2529786.
7. Kraemer HC. Extension of the kappa coefficient. Biometrics. 1980;36(2):207–16. https://doi.org/10.2307/2529972.
8. Davies M, Fleiss JL. Measuring agreement for multinomial data. Biometrics. 1982;38(4):1047–51. https://doi.org/10.2307/2529886.
9. Berry KJ, Mielke PW. A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. Educ Psychol Meas. 1988;48(4):921–33. https://doi.org/10.1177/0013164488484007.
10. Oden NL. Estimating kappa from binocular data. Stat Med. 1991;10(8):1303–11. https://doi.org/10.1002/sim.4780100813.
11. McKenzie DP, Mackinnon AJ, Péladeau N, Onghena P, Bruce PC, Clarke DM, et al. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? J Psychiatr Res. 1996;30(6):483–92. https://doi.org/10.1016/S0022-3956(96)00033-7.
12. Barnhart HX, Williamson JM. Weighted least-squares approach for comparing correlated kappa. Biometrics. 2002;58(4):1012–9. https://doi.org/10.1111/j.0006-341X.2002.01012.x.
13. Gwet KL. Testing the difference of correlated agreement coefficients for statistical significance. Educ Psychol Meas. 2016 Aug;76(4):609–37. https://doi.org/10.1177/0013164415596420.
14. Donner A, Eliasziw M, Klar N. Testing the homogeneity of kappa statistics. Biometrics. 1996;52(1):176–83. https://doi.org/10.2307/2533154.
15. Nam JM. Homogeneity score test for the intraclass version of the kappa statistics and sample-size determination in multiple or stratified studies. Biometrics. 2003;59(4):1027–35. https://doi.org/10.1111/j.0006-341X.2003.00118.x.
16. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol. 1990;43(6):543–9. https://doi.org/10.1016/0895-4356(90)90158-L.
17. Thompson WD, Walter SD. A reappraisal of the kappa coefficient. J Clin Epidemiol. 1988;41(10):949–58. https://doi.org/10.1016/0895-4356(88)90031-5.
18. Vach W. The dependence of Cohen's kappa on the prevalence does not matter. J Clin Epidemiol. 2005 Jul;58(7):655–61. https://doi.org/10.1016/j.jclinepi.2004.02.021.
19. Holley JW, Guilford JP. A note on the G index of agreement. Educ Psychol Meas. 1964;24(4):749–53. https://doi.org/10.1177/001316446402400402.
20. Aickin M. Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. Biometrics. 1990;46(2):293–302. https://doi.org/10.2307/2531434.
21. Andrés AM, Marzo PF. Delta: a new measure of agreement between two raters. Br J Math Stat Psychol. 2004;57(Pt 1):1–19. https://doi.org/10.1348/000711004849268.
22. Marasini D, Quatto P, Ripamonti E. The ordinal inter-rater agreement for the evaluation of University courses. Stat Appl. 2014;XII(1):5–16. https://doi.org/10.1400/229464.
23. Marasini D, Quatto P. Ripamonti. Assessing the inter-rater agreement through weighted indexes. Stat Methods Med Res. 2016;25(6):2611–33. https://doi.org/10.1177/0962280214529560.
24. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008;61(Pt 1):29–48. https://doi.org/10.1348/000711006X126600.
25. Gwet KL. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among raters. 4th ed. Gaithersburg: Advanced Analytics, LLC; 2014.
26. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. J Clin Epidemiol 1993;46(5):423–429. https://doi.org/10.1016/0895-4356(93)90018-V.
27. Bangdiwala SI. A graphical test for observer agreement, vol. 1. Amsterdam: Proc 45th Int Stats Institute Meeting; 1985. p. 307–8.
28. Shankar V, Bangdiwala SI. Observer agreement paradoxes in 2x2 tables : comparison of agreement measures. BMC Med Res Methodol. 2014;14:100. https://doi.org/10.1186/1471-2288-14-100.
29. Alencar LM, Zangwill LM, Weinreb RN, Bowd C, Vizzeri G, Sample PA, et al. Agreement for detecting glaucoma progression with the GDx guided progression analysis, automated perimetry, and optic disc photography. Ophthalmology. 2010;117(3):462–70. https://doi.org/10.1016/j.ophtha.2009.08.012.
30. Marks D, Comans T, Thomas M, Ng SK, O'Leary S, Conaghan PG, et al. Agreement between a physiotherapist and an orthopaedic surgeon regarding management and prescription of corticosteroid injection for

patients with shoulder pain. Man Ther. 2016;26:216–22. https://doi.org/10.1016/j.math.2016.10.001.
31. Pollock M, Fernandes RM, Hartling L. Evaluation of AMSTAR to assess the methodological quality of systematic reviews in overviews of reviews of healthcare interventions. BMC Med Res Methodol. 2017;17(1):48. https://doi.org/10.1186/s12874-017-0325-5.
32. Veldhoen S, Weng AM, Knapp J, Kunz AS, Stäb D, Wirth C, et al. Self-gated non-contrast-enhanced functional lung MR imaging for quantitative ventilation assessment in patients with cystic fibrosis. Radiology. 2017;283(1):242–51. https://doi.org/10.1148/radiol.2016160355.
33. Zee J, Hodgin JB, Mariani LH, Gaut JP, Palmer MB, Bagnasco SM, et al. Reproducibility and feasibility of strategies for morphologic assessment of renal biopsies using the Nephrotic syndrome study network digital pathology scoring system. Arch Pathol Lab Med. 2018;142(5):613–25. https://doi.org/10.5858/arpa.2017-0181-OA.
34. Hansen D, Hansen E, Retegan C, Morphet J, Beiles CB. Validation of data submitted by the treating surgeon in the Victorian audit of surgical mortality. ANZ J Surg. 2019;89(1–2):16–9. https://doi.org/10.1111/ans.14910.
35. Wennberg S, Karlsen LA, Stalfors J, Bratt M, Bugten V. Providing quality data in health care - almost perfect inter-rater agreement in the Norwegian tonsil surgery register. BMC Med Res Methodol. 2019;19(1):6. https://doi.org/10.1186/s12874-018-0651-2.
36. Ohyama T. Statistical inference of agreement coefficient between two raters with binary outcomes. Commun Stat Theory Methods. 2019. https://doi.org/10.1080/03610926.2019.1576894.
37. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. Hoboken: Wiley; 1981.
38. Nam JM. Testing the intraclass version of kappa coefficient of agreement with binary scale and sample size determination. Biom J. 2002;44:558–70. https://doi.org/10.1002/1521-4036(200207)44:5<558::AID-BIMJ558>3.0.CO;2-5.
39. Wilding GE, Consiglio JD, Shan G. Exact approaches for testing hypotheses based on the intra-class kappa coefficient. Stat Med. 2014;33(17):2998–3012. https://doi.org/10.1002/sim.6135.
40. Donner A, Eliasziw M. A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. Stat Med. 1992;11(11):1511–9. https://doi.org/10.1002/sim.4780130809.
41. Lee JJ, Tu ZN. A better confidence interval for kappa on measuring agreement between two raters with binary outcomes. J Comput Graph Stat. 1994;3:301–21. https://doi.org/10.2307/1390914.
42. Donner A, Zou G. Interval estimation for a difference between intraclass kappa statistics. Biometrics. 2002;58(1):209–15. https://doi.org/10.1111/j.0006-341X.2002.00209.x.
43. Zou G, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics. 2004;60(3):807–11. https://doi.org/10.1111/j.0006-341X.2004.00232.x.
44. Agresti A. Categorical data analysis. 2nd ed. Hoboken: Wiley; 2002. https://doi.org/10.1002/0471249688.
45. Barlow W, Lai MY, Azen SP. A comparison of methods for calculating a stratified kappa. Stat Med. 1991;10(9):1465–72. https://doi.org/10.1002/sim.4780100913.
46. Silicone Study Group. Proliferative vitreoretinopathy. The Silicone Study Group. Am J Ophthalmol. 1985;99(5):593–5. https://doi.org/10.1016/S0002-9394(14)77967-X.
47. Agresti A. An agreement model with kappa as parameter. Stat Prob Lett. 1989;7(4):271–3. https://doi.org/10.1016/0167-7152(89)90104-1.
48. Nelson KP, Edwards D. Measures of agreement between many raters for ordinal classifications. Stat Med. 2015;34(23):3116–32. https://doi.org/10.1002/sim.6546.
49. Nelson KP, Mitani AA, Edwards D. Assessing the influence of rater and subject characteristics on measures of agreement for ordinal ratings. Stat Med. 2017;36(20):3181–99. https://doi.org/10.1002/sim.7323.
50. Bickel PJ, Doksum KA. Mathematical statistics : basic ideas and selected topics. 1st ed. San Francisco: Holden-Day; 1977.
51. Rothman KJ. Estimation of confidence limits for the cumulative probability of survival in life table analysis. J Clin Epdemiol. 1978;31(8):557–60. https://doi.org/10.1016/0021-9681(78)90043-7.
52. Afifi AA, Elashoff RM, Lee JJ. Simultaneous non-parametric confidence intervals for survival probabilities from censored data. Stat Med. 1986;5:653–62. https://doi.org/10.1002/sim.4780050612.

Publisher's Note