Assessing agreement with multiple raters on correlated kappa statistics

Hongyuan Cao*,[1], **Pranab K. Sen**[2], **Anne F. Peery**[3], and **Evan S. Dellon**[3]

[1] Department of Statistics, University of Missouri–Columbia, Columbia, MO 65211, USA
[2] Department of Biostatistics, University of North Carolina–Chapel Hill, Chapel Hill, NC 27514, USA
[3] Department of Medicine, University of North Carolina–Chapel Hill, Chapel Hill, NC 27514, USA

Received 3 February 2015; revised 9 October 2015; accepted 13 November 2015

In clinical studies, it is often of interest to see the diagnostic agreement among clinicians on certain symptoms. Previous work has focused on the agreement between two clinicians under two different conditions or the agreement among multiple clinicians under one condition. Few have discussed the agreement study with a design where multiple clinicians examine the same group of patients under two different conditions. In this paper, we use the intraclass kappa statistic for assessing nominal scale agreement with such a design. We derive an explicit variance formula for the difference of correlated kappa statistics and conduct hypothesis testing for the equality of kappa statistics. Simulation studies show that the method performs well with realistic sample sizes and may be superior to a method that did not take into account the measurement dependence structure. The practical utility of the method is illustrated on data from an eosinophilic esophagitis (EoE) study.

Keywords: Contingency table; Dependent kappa statistics; Multinomial distribution.

Additional supporting information including source code to reproduce the results may be found in the online version of this article at the publisher's web-site

1 Introduction

In medical research, analysis of interobserver agreement often provides a useful means of assessing the reliability of a rating system. High measures of agreement would indicate consensus in the diagnosis and reproducibility of the testing measures of interest. The kappa coefficient (κ) is a common index in medical and health research for measuring the agreement of binary (Cohen, 1960) and nominal (Fleiss, 1971) outcomes among raters. Kappa is favored because it corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance.

Many variants and generalizations of kappa have been proposed in the literature, such as stratified kappa (Barlow et al., 1991) and weighted kappa (Cohen, 1968). Kappa can be estimated from multiple (Donner and Klar, 1996), stratified (Graham, 1995), and unbalanced samples (Lipsitz et al., 1994). Kappa may also be modeled with covariate effects (Klar et al., 2000). Davies and Fleiss (1982) developed a large sample theory of the kappa statistic for multiple raters. However, the variance is calculated under the assumption of no rater agreement ($\kappa = 0$), which limits the practical utility of the method. Kraemer et al. (2002) did an extensive overview of the kappa statistic. Other related work can be found in Gwet (2008). Only recently has attention been given to analysis of dependent kappa. Such analysis may arise when the comparison of interest is naturally conducted using the same group

*Corresponding author: e-mail: caohong@missouri.edu, Phone: +1-573-884-8568, Fax: +1-573-884-5524 of subjects. Provided it is feasible to do so, it is clear that using the same sample of subjects rather than two different samples should lead to a more efficient comparison. Most prior work has evaluated the kappa statistic using different group of subjects, in this work, we evaluate the kappa statistic using same group of subjects.

Such an important example, which is in fact the main motivation for the present research, is a reliability study conducted by Peery et al. (2011). This was a prospective study of academic and community gastroenterologists using two self-administered web-based online assessments. Gastroenterologists evaluated endoscopic images twice. First, they evaluated 35 single images obtained with standard white light endoscopy. Next, they examined 35 paired images (from the same patients, but in a random order) of the initial white light image and its narrow band imaging (NBI) counterpart, respectively. The purpose of this study was to determine whether agreement among the gastroenterologists was improved with the addition of NBI. If so, the conclusion would be that this imaging modality would have clinical utility. This comparison suggests a test of equality between two dependent kappa statistics, where each statistic may be regarded as an index of reproducibility.

Early work on correlated kappa began with the resampling approach of McKenzie et al. (1996). They proposed a resampling technique for comparing correlated kappa that makes minimum distributional assumptions and does not require large sample approximations. Donner et al. (2000) proposed modeling the joint distribution of the possible outcomes for comparing dependent kappa under two conditions with two raters. Generalized estimating equation (GEE) (Liang and Zeger, 1986; Zeger and Liang, 1986) approaches have been developed for modeling kappa with binary responses (Klar et al., 2000) and categorical responses (Williamson et al., 2000). Barnhart and Williamson (2002) used a least squares approach proposed by Koch et al. (1997) to model dependent kappa under two conditions.

In this paper, we propose a large sample theory based comparison of dependent intraclass kappa statistic with multiple raters and two categories. The method uses the multinomial distribution of the contingency table. By taking into account the correlation structure of dependent kappa statistics, we have improved power at the same level of type I error.

The paper is organized as follows. In Section 2, we set up the model and describe our statistical method. Finite-sample performance of our method is investigated in Section 3. We apply our method to analyze data from the eosinophilic esophagitis (EoE) study in Section 4. Some concluding remarks and a discussion are given in Section 5. Proofs of results from Section 2 are given in the Supporting Information.

2 Statistical method

Suppose that each of N subjects is classified into one of the two categories by each of the same set of n raters under conditions A and B. Let the random vectors $\mathbf{X}^a = (X^a_{ij1}, X^a_{ij2})^T$ and $\mathbf{X}^b = (X^b_{ij1}, X^b_{ij2})^T$ represent the resulting classification of the i-th subject ($i = 1, \ldots, N$) by the j-th rater ($j = 1, \ldots, n$) under conditions A and B. Thus, each X^a_{ijc} and $X^b_{ijc}(c = 1, 2)$ assumes the value 0 or 1, and $\sum_{c=1}^{2} X^a_{ijc} = \sum_{c=1}^{2} X^b_{ijc} = 1$ for all i and j. Let $n_{ica} = \sum_{j=1}^{n} X^a_{ijc}$ denote the number of raters who put the i-th subject into the c-th category under condition A and $n_{icb} = \sum_{j=1}^{n} X^b_{ijc}$ denote the number of raters who put the i-th subject into the c-th category under condition B. If the probability of putting the i-th subject into the c-th category is π_{ica} under condition A and π_{icb} under condition B, then the vectors $\mathbf{n}_{ia} = (n_{i1a}, n_{i2a})$ and $\mathbf{n}_{ib} = (n_{i1b}, n_{i2b})$ have probability density functions

$$f_a(n_{i1a}, n_{i2a}; n, \pi_{i1a}, \pi_{i2a}) = \frac{n!}{n_{i1a}!n_{i2a}!}\pi_{i1a}^{n_{i1a}}\pi_{i2a}^{n_{i2a}} \tag{1}$$

and

$$f_b(n_{i1b}, n_{i2b}; n, \pi_{i1b}, \pi_{i2b}) = \frac{n!}{n_{i1b}!n_{i2b}!}\pi_{i1b}^{n_{i1b}}\pi_{i2b}^{n_{i2b}}, \tag{2}$$

respectively.

Table 1 For *i*-th subject.

	B	1	2	Total
A	1	m_{i11}	m_{i12}	n_{i1a}
	2	m_{i21}	m_{i22}	n_{i2a}
		n_{i1b}	n_{i2b}	n

Intraclass kappa statistic introduced by Fleiss (1971) takes the general form

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3}$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected by chance. For this study design, p_0 and p_e can be obtained as follows.

For each subject, there are a total of $\frac{1}{2}n(n-1)$ pairs of classifications. For the *i*-th subject, the observed number of pairs that are in agreement is $\frac{1}{2}\sum_{c=1}^{2} n_{ica}(n_{ica}-1)$ under condition A and $\frac{1}{2}\sum_{c=1}^{2} n_{icb}(n_{icb}-1)$ under condition B. The observed proportion of agreement is then

$$p_o^a = \frac{1}{Nn(n-1)}\sum_{i=1}^{N}\sum_{c=1}^{2} n_{ica}(n_{ica}-1) = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{c=1}^{2} n_{ica}^2 - Nn\right) \tag{4}$$

under condition A, and a similar expression holds for p_o^b.

We get $p_e^a = \sum_{c=1}^{2}(\frac{1}{nN}\sum_{i=1}^{N} n_{ica})^2$, where $\frac{1}{nN}\sum_{i=1}^{N} n_{ica}$ is the proportion of all assignments that go to the *c*-th category under condition A, and a similar expression holds for p_e^b.

Therefore, we can calculate the kappa statistics under conditions A and B with the available data:

$$\kappa_a = 1 - \frac{1-p_o^a}{1-p_e^a} \quad \text{and} \quad \kappa_b = 1 - \frac{1-p_o^b}{1-p_e^b}. \tag{5}$$

Our interest is to test whether agreement improves under condition B compared to condition A using kappa statistics. We use the difference of the kappa statistics as our test statistic, calculate its variance, and conduct a hypothesis testing for this purpose. Since the subjects are the same under both conditions, there is strong correlation between κ_a and κ_b. This is exemplified through the relationship between the vectors $\mathbf{n}_{ia}$ and $\mathbf{n}_{ib}$, as shown in the contingency table (Table 1).

Entry $m_{ic_1c_2}$ represents the number of raters who put the *i*-th subject into c_1-th category under condition A and c_2-th category under condition B. The joint probability density function for m_{ic_1,c_2}, $i = 1, \ldots, N$ is

$$g(m_{ic_1,c_2}; n, \theta_{ic_1c_2}) = \frac{n!}{\prod_{c_1=1}^{2}\prod_{c_2=1}^{2} m_{ic_1c_2}!}\prod_{c_1=1}^{2}\prod_{c_2=1}^{2}\theta_{ic_1c_2}^{m_{ic_1c_2}}, \tag{6}$$

where $\theta_{ic_1c_2}$ is the cell probability of the *i*-th subject in the c_1-th category under condition A and the c_2-th category under condition B, which can be estimated through the proportion $\hat{\theta}_{ic_1c_2} = m_{ic_1c_2}/n$. Define the marginal probabilities $\pi_{ica} = \theta_{ic1} + \theta_{ic2}$ and $\pi_{icb} = \theta_{i1c} + \theta_{i2c}$ for $c = 1, 2$. Denote $C_a = 1/N\sum_{i=1}^{N}\sum_{c=1}^{2}\pi_{ica}(1-\pi_{ica})$, $C_b = 1/N\sum_{i=1}^{N}\sum_{c=1}^{2}\pi_{icb}(1-\pi_{icb})$, $C_a^* = \sum_{c=1}^{2}\bar{\pi}_{ca}(1-\bar{\pi}_{ca})$, and $C_b^* = \sum_{c=1}^{2}\bar{\pi}_{cb}(1-\bar{\pi}_{cb})$, where $\bar{\pi}_{ca} = 1/N\sum_{i=1}^{N}\pi_{ica}$ and $\bar{\pi}_{cb} = 1/N\sum_{i=1}^{N}\pi_{icb}$ for $c = 1, 2$.

Next, we present our main result.

Theorem 2.1. *Assuming the raters are i.i.d. and the subjects are independent, there exists a positive definite* Σ *such that,*

$$\sqrt{nN}\left\{(\kappa_a - \kappa_b) - \left(\frac{C_a}{C_a^*} - \frac{C_b}{C_b^*}\right)\right\} \rightarrow N(0, \Sigma), \quad as \quad n \rightarrow \infty, N \rightarrow \infty. \tag{7}$$

Corollary 2.2. *The variance in* (7) *can be written as* $\Sigma = \lim_{n\rightarrow\infty, N\rightarrow\infty} V$, *where*

$$\begin{aligned}
V = \frac{1}{C_a^{*4}} & \left\{4C_a^{*2}\frac{1}{N}\sum_{i=1}^{N}\pi_{i1a}(1-\pi_{i1a})(1-2\pi_{i1a})^2 + 4C_a^2(1-2\bar{\pi}_{1a})^2\frac{1}{N}\sum_{i=1}^{N}\pi_{i1a}(1-\pi_{i1a}) + \right. \\
& \left. -8C_a^*C_a(1-2\bar{\pi}_{1a})\frac{1}{N}\sum_{i=1}^{N}\pi_{i1a}(1-\pi_{i1a})(1-2\pi_{i1a})\right\} + \\
& + \frac{1}{C_b^{*4}}\left\{4C_b^{*2}\frac{1}{N}\sum_{i=1}^{N}\pi_{i1b}(1-\pi_{i1b})(1-2\pi_{i1b})^2 + 4C_b^2(1-2\bar{\pi}_{1b})^2\frac{1}{N}\sum_{i=1}^{N}\pi_{i1b}(1-\pi_{i1b}) + \right. \\
& \left. -8C_b^*C_b(1-2\bar{\pi}_{1b})\frac{1}{N}\sum_{i=1}^{N}\pi_{i1b}(1-\pi_{i1b})(1-2\pi_{i1b})\right\} + \\
& - \frac{1}{C_a^{*2}C_b^{*2}}\left\{8C_a^*C_b^*\frac{1}{N}\sum_{i=1}^{N}(1-2\pi_{i1a})(1-2\pi_{i1b})[\theta_{i11}-\pi_{i1a}\pi_{i1b}] + \right. \\
& -8C_a^*C_b(1-2\bar{\pi}_{1b})\frac{1}{N}\sum_{i=1}^{N}(1-2\pi_{i1a})[\theta_{i11}-\pi_{i1a}\pi_{i1b}] + \\
& -8C_aC_b^*(1-2\bar{\pi}_{1a})\frac{1}{N}\sum_{i=1}^{N}(1-2\pi_{i1b})[\theta_{i11}-\pi_{i1a}\pi_{i1b}] + \\
& \left. +8C_aC_b\frac{1}{N}\sum_{i=1}^{N}(1-2\bar{\pi}_{1a})(1-2\bar{\pi}_{1b})[\theta_{i11}-\pi_{i1a}\pi_{i1b}]\right\} + O\left(\frac{1}{n}\right).
\end{aligned}$$

Remark 2.3. Under the stronger assumption that the marginal probabilities $\pi_{ica} = \pi_{icb}$ for $i = 1, \ldots, N$ and $c = 1, 2$, $C_a = C_b$ and $C_a^* = C_b^*$. Theorem 2.1 can be used for testing the equality of kappa statistics.

Remark 2.4. Theorem 2.1 provides a benchmark for the large sample behavior of the difference of kappa statistics. A consistent estimate of Σ can be used for carrying out statistical inference by plugging in observed proportions for corresponding probabilities. Matlab code for such analysis is available in the Supporting Information.

Based on this result, we can construct a confidence interval for the difference of the population kappas. Both the score and Wald statistics can be obtained to test the hypothesis of equality of the population kappa after plugging in consistent estimates of the cell probabilities.

3 Numerical studies

In this section, we investigate finite-sample properties of the estimator, type I error control, and power proposed in Section 2 through Monte Carlo simulation.

3.1 Type I error control

We first study the type I error control and evaluate the accuracy of the proposed variance formula. We take expected value of the kappa statistic and define the population kappa as

$$k_A = 1 - \lim_{N\to\infty} \frac{\sum_{c=1}^{2} \frac{1}{N} \sum_{i=1}^{N} \pi_{ica}(1-\pi_{ica})}{\sum_{c=1}^{2} \bar{\pi}_{ca}(1-\bar{\pi}_{ca})}, \tag{8}$$

under condition A, where $\bar{\pi}_{ca} = \frac{1}{N}\sum_{i=1}^{n} \pi_{ica}$ and we use the fact that $\sum_{c=1}^{2} \pi_{ica} = 1$. We vary the number of raters from small ($n = 30$), moderate ($n = 70$) to large ($n = 200$), and the number of subjects from small ($N = 40$) to large ($N = 120$). In order to generate data with population kappa equal to 0.49, from Table 1, half of the two-by-two table has cell probabilities $\theta_{i11} = 0.05, \theta_{i12} = \theta_{i21} = 0.1, \theta_{i22} = 0.75$ with $\pi_{i1a} = \pi_{i1b} = 0.15$; and the other half of the two-by-two table has cell probabilities $\theta_{i11} = 0.75, \theta_{i12} = \theta_{i21} = 0.1, \theta_{i22} = 0.05$ with $\pi_{i1a} = \pi_{i1b} = 0.85$. The population kappa is calculated from (8). For i-th subject, data in the cell counts of Table 1 are generated from a multinomial distribution with parameters $(n; \theta_{i11}, \theta_{i12}, \theta_{i21}, \theta_{i22})$. We then calculate $n_{i1a} = m_{i11} + m_{i12}, n_{i1b} = m_{i11} + m_{i21}$, and subsequently, the difference of kappa statistic and its estimated variance $\hat{\Sigma}$ by plugging in consistent estimates of relevant parameters. z-Value is constructed and we compare its absolute value with 97.5% quantile of standard normal distribution, which is 1.96. We replicate this 1000 times and calculate how many times rejection occurs for type I error control. Empirical variance of the kappa difference can be obtained through Monte Carlo to compare with our proposed estimate. Similar generation is used for other kappa values in Table 2.

The results are summarized in Table 2. In the table, σ_t^2 is the variance calculated from the Monte Carlo, σ_e^2 is the estimated variance using the formula we propose, and σ_i^2 is the variance calculated ignoring the dependence. RB_e represents the relative bias of the variance estimate based on the method we propose and RB_i denotes the relative bias of the variance estimate ignoring the dependence. $\mathrm{Rejection}_e$ means empirical rejection based on the method we propose and $\mathrm{Rejection}_i$ means empirical rejection ignoring the dependence. We use a two-sided test at significance level 5%.

As we can observe in Table 2, the method ignoring the dependence is overly conservative; on the other hand, the method we propose controls the type I error at the nominal level when both number of raters and number of subjects are large. Variance estimates based on our approach have smaller bias compared to the approach that ignores the dependence completely. When the number of raters is large enough, the relative bias of the variance estimate is controlled at around 3%. The $\sqrt{nN}$ rate of convergence pattern is very clear and consistent across different scenarios, verifying our theoretical prediction.

3.2 Power comparison

Next, we compare the empirical powers for different number of raters, number of subjects, and population kappas. The population kappa is calculated by (8). Suppose half of the two-by-two table has cell probabilities $\theta_{i11} = 0.05, \theta_{i12} = 0.1, \theta_{i21} = 0.11$, and $\theta_{i22} = 0.74$, then $\pi_{i1a} = 0.15$ and $\pi_{i1b} = 0.16$; the other half of the two-by-two tables has cell probabilities $\theta_{i11} = 0.74, \theta_{i12} = 0.11, \theta_{i21} = 0.1$, and $\theta_{i22} = 0.05$ with $\pi_{i1a} = 0.85$ and $\pi_{i1b} = 0.84$. The population kappas are $k_A = 0.49$ and $k_B = 0.46$, respectively. Samples are drawn from these two-by-two tables. The procedures are exactly the same as in type I error control, except that now the population kappas are different. The results are summarized in Table 3. In Table 3, we can observe that as kappa decreases, the power gets smaller. The power using our proposed method is better than the method that ignores the dependence, especially when sample sizes are small. The sample size for the EoE study is relatively small. For moderate sample sizes

Table 2 Type I errors for testing $H_0 : k_A = k_B = k$ at $\alpha = 0.05$ (two-sided).

n	N	k	$\sigma_t^2(10^{-4})$	$\sigma_e^2(10^{-4})$	RB$_e$(%)	Rejection$_e$(%)	$\sigma_i^2(10^{-4})$	RB$_i$(%)	Rejection$_i$(%)
30	40	0.85	4.44	3.84	−13.61	6.70	7.36	65.51	1.16
		0.74	7.87	6.78	−13.87	7.18	10.97	39.37	2.12
		0.64	8.63	7.43	−13.96	7.26	13.23	53.26	1.78
		0.49	13.97	11.53	−17.47	7.52	14.59	4.41	4.54
	120	0.85	1.47	1.28	−12.64	6.64	2.45	66.99	1.22
		0.74	2.54	2.25	−11.38	6.52	3.65	43.53	2.28
		0.64	2.92	2.47	−15.48	7.48	4.40	50.67	1.58
		0.49	4.46	3.84	−13.97	7.26	4.85	8.82	4.04
70	40	0.85	1.89	1.81	−4.12	5.36	3.47	83.80	0.78
		0.74	3.32	3.17	−4.62	5.52	5.15	55.15	1.30
		0.64	3.70	3.44	−6.98	5.66	6.18	66.86	1.06
		0.49	5.84	5.31	−9.14	6.46	6.75	15.54	3.72
	120	0.85	0.63	0.60	−3.91	5.42	1.15	84.15	0.78
		0.74	1.16	1.06	−8.76	5.90	1.72	48.22	1.72
		0.64	1.20	1.15	−4.64	5.26	2.06	71.04	1.08
		0.49	1.98	1.77	−10.78	6.34	2.25	13.40	3.68
200	40	0.85	0.68	0.66	−2.25	4.96	1.27	87.51	0.56
		0.74	1.22	1.16	−5.37	5.26	1.88	53.96	1.38
		0.64	1.31	1.25	−4.46	5.54	2.25	71.81	1.04
		0.49	1.99	1.92	−3.13	5.54	2.45	23.41	2.82
	120	0.85	0.22	0.22	−0.25	5.14	0.42	91.32	0.66
		0.74	0.40	0.39	−3.45	5.32	0.63	57.02	1.58
		0.64	0.43	0.42	−1.64	5.52	0.75	76.76	1.00
		0.49	0.66	0.64	−2.44	5.08	0.82	24.22	2.80

($n = 70, N = 40$), the powers are at least 25% when the difference between two population kappas is only 0.03.

4 Application to EoE data

In this section, we return to the motivating example using our newly proposed methods of variance estimation. In Peery et al. (2011), the variance was calculated using the jackknife method rather than results based on asymptotics. In clinical practice, findings of endoscopic mucosal abnormalities are used for supporting a diagnosis of EoE, direct esophageal biopsies to sample tissue, and to assess a response to treatment. This was a prospective study of 77 gastroenterologists using self-administered web-based online assessments of endoscopic images in patients with suspected EoE. The endoscopic findings of interest included the presence or absence of three key endoscopic features: esophageal rings, linear furrows, and white plaques. Under the missing completely at random assumption (Little and Rubin, 2002), we eliminated five gastroenterologists who did not have complete data. Analysis was based on 72 gastroenterologists' assessment of 35 images under white light endoscopy and then 35 paired images using white light endoscopy enriched with NBI. Our interest was to see whether agreement is improved by adding NBI to the standard white light endoscopy.

Table 3 Empirical power for testing $H_0 : k_A = k_B = k$ at $\alpha = 0.05$ (two-sided).

n	N	k_A	k_B	$\sigma_t^2(10^{-4})$	$\sigma_e^2(10^{-4})$	RB$_e$(%)	R$_e$(%)	$\sigma_i^2(10^{-4})$	RB$_i$(%)	R$_i$(%)
30	40	0.85	0.81	5.41	4.67	−13.71	39.81	8.04	48.48	19.92
		0.74	0.71	8.69	7.44	−14.40	26.29	11.41	31.37	14.08
		0.64	0.61	9.17	7.96	−13.24	22.58	13.47	46.90	9.14
		0.49	0.46	14.29	11.77	−17.64	15.70	14.58	1.99	11.02
	120	0.85	0.81	1.79	1.56	−13.12	80.78	2.68	49.16	62.67
		0.74	0.71	2.80	2.47	−11.72	57.51	3.80	35.60	40.37
		0.64	0.61	3.11	2.64	−14.99	49.97	4.48	44.04	29.57
		0.49	0.46	4.54	3.92	−13.68	29.37	4.85	6.89	21.76
70	40	0.85	0.81	2.27	2.20	−3.16	67.91	3.79	66.78	44.67
		0.74	0.71	3.65	3.47	−4.87	45.29	5.35	46.76	28.13
		0.64	0.61	3.94	3.69	−6.31	39.41	6.29	59.72	19.68
		0.49	0.46	5.97	5.41	−9.38	23.30	6.73	12.77	16.82
	120	0.85	0.81	0.77	0.73	−5.05	98.74	1.26	63.44	94.72
		0.74	0.71	1.26	1.16	−8.17	87.82	1.78	41.50	75.72
		0.64	0.61	1.31	1.23	−6.38	81.30	2.09	59.58	61.23
		0.49	0.46	2.03	1.80	−11.04	54.47	2.24	10.67	45.49
200	40	0.85	0.81	0.82	0.80	−2.27	98.24	1.39	68.37	93.26
		0.74	0.71	1.35	1.27	−5.74	84.84	1.96	45.48	70.79
		0.64	0.61	1.41	1.34	−4.56	77.92	2.29	63.09	56.51
		0.49	0.46	2.03	1.96	−3.20	50.17	2.44	20.69	41.19
	120	0.85	0.81	0.27	0.27	−1.05	100.00	0.46	70.44	100.00
		0.74	0.71	0.44	0.42	−3.29	99.96	0.65	49.19	99.82
		0.64	0.61	0.45	0.45	−0.87	99.70	0.76	69.30	98.16
		0.49	0.46	0.67	0.65	−2.18	92.92	0.81	21.90	89.00

Table 4 Test results of the agreement on endoscopic mucosal abnormalities.

Symptom	p_o^w	p_e^w	p_o^n	p_e^n	κ_w	κ_n	$\kappa_w - \kappa_n$	z_e	p_e	z_i	p_i
Rings	0.817	0.593	0.812	0.628	0.550	0.493	0.056	3.073	0.002	2.474	0.013
Furrow	0.756	0.533	0.747	0.501	0.478	0.493	−0.015	−0.769	0.442	−0.662	0.508
Plaque	0.724	0.613	0.701	0.607	0.287	0.241	0.046	2.538	0.011	2.486	0.013
None	0.853	0.784	0.909	0.884	0.318	0.217	0.101	3.627	0.000	3.168	0.002

We summarize our results in Table 4. We use p_o^w to denote observed proportion of agreement under white light endoscopy, p_e^w to denote proportion of agreement expected by chance under white light endoscopy, p_o^n to denote observed proportion of agreement with the addition of NBI, p_e^n to denote proportion of agreement expected by chance with the addition of NBI, κ_w to denote the kappa statistic under white light endoscopy, and κ_n to denote the kappa statistic with the addition of NBI. z_e represents the z-value of the difference based on our approach, z_i represents the z-value of the difference obtained ignoring the dependence, p_e represents the p-value of the difference based on our approach, and p_i represents the p-value of the difference obtained ignoring the dependence.

The results show that except for furrows, the agreement is better using white light endoscopy alone. Overall, we conclude that it is better to use white light endoscopic based on the statistical testing results. Whether the difference is meaningful in clinic is subject to clinical experts' opinions, but these results suggest that there is not added utility to performing examination with NBI.

5 Concluding remarks

Kappa statistic is the most commonly reported measure of interobserver agreement in the medical literature. It does not assess agreement with the gold standard as the true values are usually not available. This is very different from the multireader multicase studies considered in Chen et al. (2014), where the probabilities of agreement with the reference standard are compared. Chen et al. (2014) used a measurement of accuracy, while we use kappa statistic as a measure of reliability. The purpose of our study is to provide support to a diagnosis of EoE and as such there is no gold standard. More discussion on accuracy and reliability can be found in Viera et al. (2005).

We propose a large sample based testing procedure using kappa statistics by taking into account the measurement dependence on the subjects. This newly proposed procedure is shown to improve power while controlling type I error in large samples. For small-to-moderate samples, the type I error is slightly inflated. The advantage of this approach is that it can test the equality of kappa statistics taking values different from zero.

An important assumption is that the subjects are independent and the raters are independent and identically distributed so that each rater generates a rating without knowledge, and thus without influence, of the other rater's rating. Equally, ratings on the first occasion may sometimes influence those given on the second occasion, which will threaten the assumption of independence. Thus, apparent agreement may reflect a recollection of the previous decision compared to a genuine judgement. In our study, we sent out the second survey at least 14 days after the first survey with random ordering of the images in order to overcome the bias due to memory.

Acknowledgments The authors thank Rosalie Dominik and Jianwen Cai for helpful discussions. We are indebted to the editor and reviewers for helping in improving this paper. This research was supported in part by the University of Missouri Research Board and National Institute of Health.

Conflict of interest
The authors have declared no conflict of interest.

References

Barlow, W., Lai, M. and Azen, S. (1991). A comparison of methods for calculating a stratified kappa. *Statistics in Medicine* **10**, 1465–1472.

Barnhart, H. and Williamson, J. (2002). Weighted least-squares approach for comparing correlated kappa. *Biometrics* **58**, 1012–1019.

Chen, W., Wunderlich, A., Petrick, N. and Gallas, B. D. (2014). Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing. *Journal of Medical Imaging* **1**, 031011.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**, 37–46.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220.

Davies, M. and Fleiss, J. (1982). Measuring agreement for multinomial data. *Biometrics* **38**, 1047–1051.

Donner, A. and Klar, N. (1996). The statistical analysis of kappa statistics in multiple samples. *Journal of Clinical Epidemiology* **49**, 1053–1058.

Donner, A., Shoukri, M., Klar, N. and Bartfay, E. (2000). Testing the equality of two dependent kappa statistics. *Statistics in Medicine* **19**, 373–387.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378–382.

Graham, P. (1995). Modelling covariate effects in observer agreement studies: the case of nominal scale agreement. *Statistics in Medicine* **14**, 299–310.

Gwet, K. (2008). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika* **73**, 407–430.

Klar, N., Lipsitz, S. R. and Ibrahim, J. G. (2000). An estimating equations approach for modelling kappa. *Biometrical Journal* **42**, 45–58.

Koch, G., Landis, J., Freeman, J., Freeman, D. and Lehnen, R. (1997). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* **33**, 133–158.

Kraemer, H., Periyakoil, V. and Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine* **21**, 2109–2129.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

Lipsitz, S. R., Laird, N. M. and Brennan, T. A. (1994). Simple moment estimates of the κ-coefficient and its variance. *Applied Statistics* **43**, 309–323.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. *Wiley Series in Probability and Statistics*. John Wiley & Sons, New York, NY.

McKenzie, D., MacKinnon, A., Peladeau, N., Onghena, P., Bruce, P., Clarke, D., Harrigan, S. and McGorry, P. (1996). Comparing correlated kappas by re-sampling: is one level of agreement significantly different from another? *Journal of Psychiatric Research* **30**, 483–492.

Peery, A., Cao, H., Dominik, R. C., Shaheen, N. and Dellon, E. (2011). Variable reliability of endoscopic findings with white-light and narrow-band imaging for patients with suspected eosinophilic esophagitis. *Clinical Gastroenterology and Hepatology* **9**, 475–480.

Viera, A. J. and Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. *Family Medicine* **37**, 360–363.

Williamson, J., Manatunga, A. and Lipsitz, S. (2000). Modelling kappa for measuring dependent categorical agreement data. *Biostatistics* **1**, 191–202.

Zeger, S. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–130.