Journal Of Turgut Ozal Medical Center
www.jtomc.org

Comparison of Agreement Statistics in Case of Multiple-Raters and Diagnostic Test Being Categorical: A Simulation Study

E. Arzu Kanık[1], Gülhan Örekici Temel[1], Semra Erdoğan[1], İrem Ersöz Kaya[2]

[1]Mersin University Medical School, Department of Biostatistics and Medical Informatics, Mersin, Turkey
[2]Mersin University Technical Education Faculty, Department of Electronic and Computer Education, Mersin, Turkey

Abstract

Aim: When the number of raters and the number of categories of diagnostic tests are two or more, put forward agreement statistics' conditions of being affected by the sample size, the number of raters and the number of categories of scale used.

Material and Methods: AC1 statistic, Fleiss Kappa and Krippendorff's Alpha values belonging to state where there was no agreement between raters and states where agreement was 0.90 for those combinations were recorded for 1000 simulation study.

Results: The expected agreement between raters is 0.90, AC1 statistic and Fleiss Kappa coefficient offer similar results and take equivalent values, to the expected value of agreement in all combinations. When Krippendorff's Alpha coefficient examined, it is not affected by sample size but affected by the number of raters and the number of categories pertaining to diagnostic test.

Conclusion: If prevalence value is known and a bear significant for study, use of AC1 statistic is recommended among agreement statistics, if the existence of missing data is the case in study, Krippendorff's Alpha coefficient is the most appropriate agreement statistic, except these cases mentioned, use of Fleiss Kappa coefficient is recommended.

Key Words: Fleiss Kappa; Gwet's AC1 Statistics; Krippendorff Alpha; Agreement between Raters.

Çoklu Değerlendirici ve Tanı Testinin Kategorik Olması Durumunda Uyum İstatistiklerinin Karşılaştırılması: Bir Simülasyon Çalışması

Özet

Amaç: Değerlendirici sayısının ve tanı testine ait kategori sayısının iki ve daha fazla olduğu durumda, uyum istatistiklerinin, örneklem büyüklüğünden, değerlendirici sayısından ve kullanılan ölçeğin kategori sayısından etkilenme durumlarını ortaya koymaktır.

Gereç ve Yöntem: Değerlendiriciler arasında hiç uyumun olmadığı durum ile uyumun 0.90 olduğu durumlara ait AC1 istatistiği, Fleiss Kappa ve Krippendorff Alpha değerleri 1000 simülasyon denemesi için kaydedilmiştir.

Bulgular: Değerlendiriciler arasındaki beklenen uyumun 0.90 olduğu durumda; AC1 istatistiği ve Fleiss kappa katsayısı, örneklem büyüklüğü, değerlendirici sayısı ve tanı testine ait kategori sayısı ne olursa olsun tüm kombinasyonlarda benzer sonuçlar vermekte ve beklenen uyum değerine eşit değerler almaktadır. Krippendorff Alpha katsayısı incelendiğinde, örneklem büyüklüğünden etkilenmediği ancak değerlendirici sayısından ve tanı testine ait kategori sayısından etkilenmektedir.

Sonuç: Prevelans değeri biliniyor ve çalışma için önem taşıyorsa, Gwet'in AC1 istatistiğinin, eğer çalışmada eksik verilerin varlığı söz konusu ise Krippendorff Alpha katsayısının, bu sözü edilen durumlar dışında Fleiss kappa katsayısının kullanılması önerilmektedir.

Anahtar Kelimeler: Fleiss Kappa; Gwet'in AC1 Istatistiği; Krippendorff Alpha; Değerlendiriciler Arası Uyum.

Orijinal Makale/Original Article

Received: 09.04.2012, **Accepted:** 02.05.2012

Corresponding Author:
Dr. Semra ERDOĞAN, Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, MERSİN/TURKEY
Tel: 0 324 361 00 01/1029
e-mail : semraerdogann@gmail.com

For citing/Atıf için:
Kanık EA, Temel GO, Erdoğan S, Kaya İE. Comparison of agreement statistics in case of multiple-raters and diagnostic test being categorical: A simulation study. J Turgut Ozal Med Cent 2012;19(4):220-7.
DOI:10.7247/jtomc.19.4.4

Introduction

How the data are obtained in scientific researches, which method of measurement is employed, the reliability of method used are the most crucial parts of scientific dimension of a research. In plenty of studies, the data collection instruments such as questionnaires, lab findings or classification systems are used by different people called as raters, observers or deciders. Researchers would like to know whether all raters consistently apply the data collection methods to minimize the effect of rater factor on the data quality (1).

Methods used in measuring inter-rater agreement vary depending on diagnostic test used being constant or categorical and the number of raters. There are many coefficients of agreement employed to assess agreement between two raters in case of diagnostic test in the literature has two categories. First, π-statistic was developed by Scott in 1955 and then, the kappa statistic was developed by Cohen in 1960. Albeit not used quite widespread, G-index agreement coefficient was developed by Holley and Guildford. For cases in which the number of categories pertaining to diagnostic test is two or more, Krippendorff's Alpha coefficient was developed by Klaus Krippendorff in 1970 and generalized kappa coefficient was developed by Fleiss in 1971. Afterwards, AC1 statistic was put forth by Gwet in 2001 as an alternative to these agreement statistics (2-5).

When the number of raters is more than two, comparing raters as binary causes Type I error pertaining to the study to increase. When the number of raters is two and more, AC1 statistic, Fleiss Kappa and Krippendorff's Alpha coefficients of agreement are commonly used in testing inter-rater agreement as well (6).

The goal of this study is to introduce Gwet's AC1 statistic, Fleiss Kappa and Krippendorff's Alpha coefficients of agreement and put forward these agreement statistics' conditions of being affected by the sample size, the number of raters and the number of categories of scale used.

Material and Methods

In reliability studies, both category and the number of raters of measurement instrument can be more than two. Let's acknowledge that R number of raters, N number of patients and K number of categories shall be in the research. In that case, each rater would have N*K number of results and all study would have R*N*K results. In other words, it has a factorial testing order. Each rater in this testing order assesses (rates) more than one existing test results is presented on Table 1 (3,7).

Table 1. The design plan of R rater, N patient and K category

	Diagnostic test											
The number of patient	**1** **Rater**				**2** **Rater**			**....** **...**	**K** **Rater**			
	R_1	R_2	…	R_r	R_1	R_2	R_r	. . . .	R_1	R_2	…	R_r
1	Y_{111}	Y_{121}	…	Y_{1r1}	Y_{112}	Y_{122}	Y_{1r2}	. . . .	Y_{11k}	Y_{12k}	…	Y_{1rk}
2	Y_{211}	Y_{221}	…	Y_{2r1}	Y_{212}	Y_{222}	Y_{2r2}	. . . .	Y_{21k}	Y_{22k}	…	Y_{2rk}
.	.	.	.	.	.	.	. .	. . . .	.	.	…	.
.	.	.	.	.	.	.	. .	. . . .	.	.	…	.
N	Y_{N11}	Y_{N21}		Y_{Nr1}	Y_{N12}	Y_{N22}	Y_{Nr2}	. . . .	Y_{N1k}	Y_{N2k}		Y_{Nrk}

Some agreement coefficients being used in the name of being able to put forward agreement between raters pertaining to this testing order are recommended in the literature.

Fleiss's generalized π-statistic

It is an agreement statistic used for the purpose of measuring agreement of more than two raters in case of diagnostic test being categorical or sequential. It was generalized departing from Scott's π-statistic and propounded by Fleiss in 1971 (8).

Testing order, which will be used for assessing N number of patients and diagnostic test having Q number of categories, is presentedon Table 2 (3).

Tablo 2. The agreement matrix for patient and diagnostic tests

	The number of categories of diagnostic test					
The number of patient	**1**	**…**	**q**	**…..**	**Q**	**Total**
1	r_{11}	…	r_{1q}	…	r_{1Q}	R
2	r_{21}	…	r_{2q}	…	$r_{2\ Q}$	R
.	.	…	.	…	.	.
.	.		.		.	.
n	r_{n1}	…	r_{nq}	…	$r_{n\ Q}$	R
Total	r_{+1}	…	r_{+q}	…	$r_{+\ Q}$	N_r

According to Table 2, n: the total number of patients (i=1, n) Q; the number of categories of diagnostic test (q=1,2, Q) r: the number of raters, r_{iQ}: shows the joint decision of raters

Fleiss's generalized π-statistic is shown as $\hat{\gamma}_{..}$ and calculated as in Equation 1 (1,3).

$$\hat{\gamma}_{..} = \frac{(P_a - P_{e/\pi})}{1 - P_{e/\pi}} \qquad (1)$$

In equation 1, P_a exhibits the overall agreement probability and calculated as in Equation 2.

$$P_a = \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{q=1}^{Q}\frac{r_{iq}(r_{iq}-1)}{r(r-1)}\right\} \quad (2)$$

$P_{e/\pi}$,the change-agreement probability and calculated as in Equation 3.

$$P_{e/\pi} = \sum_{q=1}^{Q}\pi_q \qquad (3)$$

$\hat{\pi}_q$ probability appearing in Equation 3 refers to classification probability of a test subject within category of q by a rater and calculated as in Equation 4.

$$\pi_q = \frac{1}{n}\sum_{i=1}^{n}\frac{r_{iq}}{r} \qquad (4)$$

The variance of Fleiss's generalized π-statistic calculated as in Equation 5 (1,3).

$$V(\gamma_{..}) = \frac{2}{nr(r-1)}\frac{P_{e/\pi} - (2r-3)P_{e/\pi}^2 + 2(r-2)\sum_{q=1}^{Q}\hat{\pi}_q^3}{(1-P_{e/\pi})^2} \quad (5)$$

Gwet's AC1 statistics

It is denominated as Gwet's AC1 statistic and was suggested by Gwet in 2001. It is also called as first order agreement coefficient in the literature and calculated as follows (1,9-10).

$$AC1 = \frac{P_a - P_{e/\gamma}}{1 - P_{e/\gamma}} \qquad (6)$$

The overall agreement probability calculated as in Equation 2, the change-agreement probability follows Equation.7

$$P_{e/\gamma} = \frac{1}{Q-1}\sum_{q=1}^{Q}\pi_q(1-\pi_q) \qquad (7)$$

The value of Equation 7, $\hat{\pi}_{..}$ calculated as in Equation 4. The variance of AC1 statistics given by Equation 8 (1,9).

$$V(\gamma_{..}) = \frac{2}{nr(r-1)}\frac{P_{e/\pi} - (2r-3)P_{e/\pi}^2 + 2(r-2)\sum_{q=1}^{Q}\hat{\pi}_q^3}{(1-P_{e/\pi})^2} \qquad (8)$$

Krippendorff's Alpha coefficient

Krippendorff's Alpha coefficient is an agreement coefficient, which can be used for all scales and calculated as in Equation 9. The most important advantage of this coefficient is that it can present incomplete or missing data (4-5).

$$\alpha = 1 - \frac{D_0}{D_e} \qquad (9)$$

In equation, D_o is the observed disagreement, D_e is the expected disagreement and calculated as in Equation 10 and 11.

$$D_0 = \frac{1}{n} \sum_c \sum_k \; {}_k \, metric\delta \qquad (10)$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k \; n_k \, metric\delta \qquad (11)$$

When $D_o =$, it is inferred that raters have perfect agreement, which in such case reliability coefficient is $\alpha =$. In case of $D_o =$, reliability coefficient will be $\alpha =$ (6).

As the first step in calculating Krippendorff's Alpha coefficient, a data matrix consisting of outcomes of *m* number of raters pertaining to *r* number of cases is generated is presented on Table 3 (4-5).

Table 3. The data matrix of Krippendorff's Alpha coefficient.

	The number of case										
The number of raters	**1**	**2**	**.**	**.**	**.**	**u**	**.**	**.**	**.**	**.**	**r**
1	C_{11}	C_{12}				C_{1u}					C_{1r}
i	C_{i1}	C_{i2}				C_{iu}					C_{ir}
j	C_{j1}	C_{j2}				C_{ju}					C_{jr}
.											
m	C_{m1}	C_{m2}				C_{mu}					C_{mr}
Total	m_1	m_2				m_u					m_r

According to Table 3, r: the total number of cases, m: the total number of raters, Ciu: the evaluation result of i rater for case u m_u sum of rating values of each raters in unit u

In case of there is no missing data, m_u value shall be equal to the number of raters. As the first step in calculating Krippendorff's Alpha coefficient, the data matrix generated in the first step is converted into an agreement matrix containing frequencies of all assessment pairs matched is presented on Table 4 (4-5).

Table 4. The agreement matrix of Krippendorff Alpha coefficient

Results	**1**	**.**	**k**	**.**	**.**	
1	O_{11}		O_{1k}			n_1
.	.	.	.			.
.	.	.	.			.
c	O_{c1}		O_{ck}	.	.	$n_c = \sum_k O_{ck}$
Total	n_1		n_k			$n \sum_c \sum_k O_{ck}$

Frequencies of assessment pairs matched are displayed by O_{ck} in Table 4 and calculated as in Equation 12. This value indicates the observation frequency of assessment pair c-k at u case.

$$O_{ck} = \sum_u \frac{Number\, of\; c-k\; pair\; in\; unit\; u}{m_u - 1} \qquad 12)$$

Departing from there, Krippendorff's Alpha coefficient is re-formulated as in Equation 13 (4-5).

$$\text{Nominal}\,\alpha = 1 - \frac{D_0}{D_e} = \frac{(n-1)\sum_c O_{cc} - \sum_c n_c(n_c-1)}{n(n-1) - \sum_c n_c(n_c-1)} \quad (13)$$

Simulation study

In this study, a Monte Carlo simulation study was conducted with the aim of examining how AC1 statistic, Fleiss Kappa and Krippendorff's Alpha coefficients are influenced by sample size, the number of raters and outcomes of diagnostic test for two different states in which there is no agreement and agreement is 0.90 among raters. Simulation study, data production and calculation of coefficients were carried out in Matlab 7.0 software package. Data of diagnostic test results for each rater was obtained from integer distribution separately.

A total of 36 different combinations were used, including 3 different situations where the number of raters was 2, 5 and 7, 4 different situations where diagnostic test had 2, 5, 7 and 10 categories and 3 different situations where the sample size was 30, 100 and 1000, in simulation study. AC1 statistic, Fleiss Kappa and Krippendorff's Alpha values belonging to state where there was no agreement between raters and states where agreement was 0.90 for those combinations were recorded for 1000 simulation study. The average and standard deviation values pertaining to 1000 simulations were calculated for each combination. Averages calculated for all coefficients were regarded as population value due to that the number of repetitions was 1000 in simulation tests and comparison was not conducted via hypothesis testing.

Results

Gwet's AC1 statistic, descriptive statistics pertaining to Fleiss Kappa and Krippendorff's Alpha for three different number of raters, 4 different diagnostic tests and two different agreement values for sample sizes 30,100 and 1000 are given on Tables 5, 6 and 7, respectively.

Table 5. Descriptive statistics for agreement coefficients for N= 30.

		Expected agreement =0.90			**Expected agreement =0.00**		
The number of raters	**The number of categories**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**
2	2	0.906±0.055	0.900±0.058	0.906±0.055	0.028±0.185	-0.005±0.186	0.012±0.183
	5	0.903±0.030	0.900±0.031	0.903±0.030	0.004±0.089	0.016±0.089	0.001±0.088
	7	0.901±0.025	0.898±0.026	0.901±0.025	0.002±0.078	-0.018±0.078	-0.001±0.077
	10	0.902±0.021	0.900±0.022	0.902±0.021	0.003±0.062	-0.015±0.062	0.011±0.065
5	2	0.900±0.017	0.895±0.018	0.900±0.017	0.005±0.061	-0.008±0.058	-0.002±0.058
	5	0.900±0.009	0.896±0.010	0.900±0.009	0.004±0.031	-0.005±0.030	0.002±0.030
	7	0.900±0.007	0.897±0.008	0.900±0.007	0.002±0.024	-0.005±0.024	0.002±0.024
	10	0.900±0.006	0.897± 0.006	0.900±0.006	0.002±0.020	-0.006±0.020	0.013±0.026
7	2	0.902±0.013	0.896±0.016	0.902±0.013	0.007±0.042	-0.003±0.041	0.002±0.041
	5	0.901±0.006	0.897±0.007	0.901±0.006	0.003±0.020	-0.003±0.020	0.002±0.020
	7	0.900±0.005	0.897±0.005	0.900±0.005	0.002±0.017	-0.004±0.017	0.001±0.017
	10	0.900±0.004	0.897±0.005	0.900±0.004	0.001±0.014	-0.004±0.014	0.015±0.022

Table 6. Descriptive statistics for agreement coefficients for N=100.

		Expected agreement =0.90			Expected agreement =0.00		
The number of raters	**The number of categories**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**
2	2	0.903±0.031	0.902±0.031	0.902±0.031	0.007±0.096	-0.003±0.096	0.002±0.095
	5	0.902±0.016	0.901±0.016	0.901±0.016	0.002±0.051	-0.005±0.051	0.000±0.051
	7	0.901±0.012	0.90±0.013	0.900±0.012	-0.001±0.041	-0.007±0.041	-0.002±0.041
	10	0.900±0.011	0.899±0.011	0.905±0.014	0.001±0.033	-0.005±0.034	0.009±0.036
5	2	0.901±0.010	0.899±0.011	0.899±0.011	0.002±0.030	-0.003±0.030	-0.001±0.030
	5	0.901±0.005	0.900±0.005	0.900±0.005	0.001±0.017	-0.001±0.017	0.001± 0.017
	7	0.900±0.004	0.899±0.004	0.900±0.004	0.001±0.013	-0.001±0.013	0.001±0.013
	10	0.901±0.004	0.899±0.004	0.934±0.016	0.001±0.011	-0.001±0.011	0.012±0.015
7	2	0.901±0.007	0.899±0.007	0.899±0.007	0.000±0.020	-0.003±0.020	-0.001±0.020
	5	0.900±0.003	0.899±0.004	0.899±0.004	0.001±0.011	-0.001±0.011	0.000±0.011
	7	0.900±0.003	0.899±0.003	0.899 ± 0.003	0.001 ± 0.009	-0.001±0.009	0.001±0.009
	10	0.900±0.002	0.899±0.002	0.946±0.016	0.001±0.007	-0.001±0.007	0.014±0.011

Table 7. Descriptive statistics for agreement coefficients for N= 1000

		Expected agreement =0.90			Expected agreement =0.00		
The number of raters	**The number of categories**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**	**Gwet's AC1 statistics**	**Fleiss Kappa**	**Krippendorff Alpha**
2	2	0.900±0.010	0.900±0.010	0.900±0.010	-0.001±0.031	-0.002±0.031	-0.002±0.031
	5	0.900±0.005	0.900±0.005	0.900±0.005	-0.001±0.015	-0.001±0.015	-0.001±0.015
	7	0.900±0.004	0.900±0.004	0.900±0.004	0.000±0.003	0.000±0.003	-0.000±0.003
	10	0.900±0.003	0.900±0.003	0.905±0.004	-0.001±0.010	-0.001±0.010	0.008±0.011
5	2	0.900±0.003	0.900±0.003	0.900±0.003	0.000±0.010	0.000±0.010	0.000±0.010
	5	0.900±0.002	0.900±0.002	0.900±0.002	0.000±0.005	0.000±0.005	0.000±0.005
	7	0.900±0.001	0.900±0.001	0.900±0.001	0.000±0.004	0.000±0.004	0.000±0.004
	10	0.900±0.001	0.900±0.001	0.933±0.005	0.000±0.003	0.000±0.003	0.012±0.005
7	2	0.900±0.002	0.900±0.002	0.900±0.002	0.000±0.007	0.000±0.007	0.000±0.007
	5	0.900±0.001	0.900±0.001	0.900±0.001	0.000±0.003	0.000±0.003	0.000±0.003
	7	0.900±0.001	0.900±0.001	0.900±0.001	0.000±0.003	0.000±0.003	0.000±0.003
	10	0.900±0.001	0.900±0.001	0.946±0.005	0.000±0.002	0.000±0.002	0.013±0.004

If it is needed to assess all agreement statistics in the state where the expected agreement between raters is 0.90, AC1 statistic and Fleiss Kappa coefficient offer similar results and take equivalent values to the expected value of agreement in all combinations regardless of the sample size, the number of raters and the number of categories pertaining to diagnostic test, Krippendorff's Alpha coefficient gets a value above the expected agreement when the sample size is 100 and 1000, the number of raters is 5 and the number of categories pertaining to diagnostic test is 10 (Table 5-7). In case of the expected agreement is 0 between raters, when all agreement statistics assessed, it was observed that all agreement statistics exhibit similar results and get quite close values to the expected agreement (Table 5-7).

Discussion

While diagnostic studies are carried out in the clinic, if a single rater is referred especially when there is no gold standard, results for the case can include subjectivity. Therefore, reports of more than one rater are considered in the practice. Especially in the areas such as radiology and pathology, cases dependent on the decision of more than one rater are frequently viewed. Moreover, category level of the diagnosis test is as important as the number of rater. Category of the diagnosis test can be at nominal level rather than binary structure (patient/healthy). Increase of category level of diagnosis test complicates making decision and agreement between raters.

It is known that agreement statistics used in the clinic have relationship with many factors such as the number of rater, experience and education of rater, category level of the diagnosis test, number of case that is observed and current status of the case (stage of the disease). Therefore these agreement statistics should be evaluated according to the number of raters, sample size and category level of diagnosis test.

Dorfman et al. (1992) suggested test plan including more than one rater in order to make diagnostic decision. For this aim, they have developed multi reader multi case (MRMC) models in which the impact of decisions of more than one rater exists. In these statistical models, the effect of agreement statistics between raters is considered (11). Then Obuchowski (2000) thought that when more than one rater are considered making the decision of patient- healthy, the agreement among raters has important impact on the sample size and therefore created sample size table. On this table it was calculated what the minimum sample size shall be, taking 80% power and Type I error 5%, in order to make diagnostic studies on conditions where the compliance between raters is low, medium and high; and the number of raters is 4, 6 and 10. As a result of this study, it was stated sample size, number of raters and agreement between them should be balanced in the study to be planned (12). Eye et al. (2006) thought that agreement statistics are influenced from sample size and the level of diagnosis test therefore carried out simulation study. As a result of the study it was stated that as the sample size decrease so does the power of agreement statistics. In the agreement study among raters; Bogartz (2010) has made a simulation study in order to clarify category level of diagnosis tests and optimum number of raters (13-14).

Gwet (2008) has made a simulation study in order to compare the condition of being influenced from of Fleiss Kappa and Gwet's AC1 statistics from sample size and the number of raters. In this study conditions were regarded where the number of raters is 5, 7, 9, 11 and 13; and the sample size is 20, 30, 40 and 50.[1] As a result of simulation, it was stated that both agreement statistics were not influenced from sample size; and as the number of raters increase standard errors of both statistics decrease. When two statistics are compared, it was concluded that Gwet's AC1 statistics works better than Fleiss kappa value.

Conclusion

When diagnostic decisions are made in the clinic, when there is no gold standard accuracy of the evaluation made by raters group is required to be estimated. Repeatability of the evaluations is measured with high agreement between raters. High compliance is the measurement of consistency of repeatability of results at different times and laboratories.

Inconsistency of doctors about diagnosis in practice is a common and serious problem. Results of many statistical analyses conducted are influenced by the sample size taken into research, the state of inter-rater agreement, high or low prevalence of the disease, inter-diagnostic test relationship and the level of category pertaining to diagnostic test. Therefore, what the number of raters, the number of categories pertaining to diagnostic test and sample size shall be is a frequently discussed issue in inter-rater agreement calculations.

According to simulation findings, in case of there was no agreement between raters, it was observed that it was not affected giving similar results with regard to all agreement statistics, the sample size, the number of raters and the number of categories pertaining to diagnostic test and displayed the expected agreement. In case of inter-rater agreement was high, it was observed that Gwet's AC1 statistic and Fleiss Kappa offered similar results, were not affected by the sample size, the number of raters and the number of categories pertaining to diagnostic test. Krippendorff's Alpha coefficient is not influenced by sample size but it is observed that it makes estimations above the expected value of agreement in case of the number of raters is 5 minimally, the number of categories pertaining to diagnostic test is 10 at minimum. Accordingly, in case of using Krippendorff's Alpha coefficient in measuring inter-rater agreement, it can be said that the number of raters and the number of categories of diagnostic test should be taken into consideration. In addition to this disadvantage of that coefficient, it is known that it can also be used in cases where there are lacking data in the literature.

In conclusion, if prevalence value is known in conducted researches and bears significant for study, use of Gwet's AC1 statistic is recommended among agreement statistics.

Besides, it was put forward that Gwet's AC1 statistic is not affected by sensitivity, specify and prevalence values belonging to raters as a result of calculations performed (15). If the existence of lacking data is the case in study, it can be said that in such case, Krippendorff's Alpha coefficient is the most appropriate agreement statistic. Except these cases mentioned, use of Fleiss Kappa coefficient is recommended. Thus, it can be argued that these three agreement statistics have a crucial place in calculation of inter-rater agreement.

References

1. Gwet K. Computing inter-rater reliability and its variance in the presence of high agreement. Brit J Mathematic Stat Psychol 2008;61:29-48.
2. Gwet K. Kappa statistics is not satisfactory for assessing the extent of agreement between raters. Series: Stat Met Inter-Rater Reliab Asses 2002;1:1-5.
3. Gwet K. Handbook of Inter-Rater Reliability; 1st rev ed. USA: STATAXIS Publishing Company; 2001.
4. Krippendorff K. Reliability in content analysis some common misconceptions and recommendations. Hum Commun Res 2004;30:411-33.
5. Hayes AF, Krippendorff K. Answering the call for a standard reliability measure for coding data. Com Method Measur 2007;1:77-89.
6. Kanık EA, Orekici Temel G, Ersöz Kaya İ. Effect of sample size, the number of raters and the category levels of diagnostic test on Krippendorff Alpha and the Fleiss Kappa statistics for calculating inter-rater agreement: a simulation study. Türkiye Klinikleri J Biostat 2010;2:74-81.
7. Zhou X, Obuchowski N, McClish D. Statistical Methods in Diagnostic Medicine, 1st rev ed; New York: Wiley. 2002.
8. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull 1971;76:378-82.
9. Haley DT, Thomas P, Petre M, Roeck AD. Using a new inter-rater reliability statistics. Technical Report 2008; 15.
10. Blood E, Spratt KF. Disagreement on Agreement: Two Alternative Agreement Coefficients. Statistics and Data Analysis. SAS Global Forum 2007.
11. Dorfman D, Berbaum K, Metz C. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with Jackknife method. Invest Radiol 1992;27:723-31.
12. Obuchowski NA. Sample size tables for receiver operating characteristic studies. AJR Am J Roentgenol 2000;175:603-8.
13. Eye VA, Mair P, Schauerhuber M. Significance tests for the measure of raw agreement. 2007. Working Paper.http://epub.wu.ac.at/1336/
14. Bogartz RS. Interrater agreement and combining Ratings. 2010.http://ebookbrowse.com/interrater-agreement-pdf-d15137096
15. Kanık EA, Erdoğan S, Temel Orekici G. İkili değişkenler için iki değerlendirici arasındaki uyum istatistiklerinin prevelanstan etkilenme durumları. XIII. Ulusal Biyoistatistik Kongresi, 12-14 Eylül 2011, Ankara-Kızılcahamam.