Cumhuriyet Science Journal

e-ISSN: 2587-246X
ISSN: 2587-2680

Cumhuriyet Sci. J., 42(3) (2021) 743-750
http://dx.doi.org/10.17776/csj.898192

How to use adjusted degree of distinguishability and inter-rater reliability simultaneously?

Ayfer Ezgi YILMAZ [1,*]

[1]*Hacettepe University, Faculty of Science, Department of Statistics, Ankara/TURKEY*

Abstract

When the categories of a square contingency table are ordinal, weighted kappa or Gwet's AC2 coefficients are used to summarize the degree of reliability between two raters. In addition, investigating the reliability among raters, the term category distinguishability should be considered. The study aims to assess the inter-rater reliability and category distinguishability in ordinal rating scales together. The weighted kappa, AC2, and adjusted degree of distinguishability coefficients are applied to pathology data. The results are discussed over the pathologist pairs.

Article info

History:
Received:16.03.2021
Accepted:06.08.2021

Keywords:
Ordinal rating scales,
Inter-rater reliability,
Adjusted degree of distinguishability,
Weighted kappa,
AC2 coefficient.

1. Introduction

Square contingency tables are occurred with the same row and column classification [1] and are frequently used in many fields, such as medicine, sociology, and behavioral sciences [2]. When working on these kinds of tables, the inter-rater reliability of row and column variables is investigated. Inter-rater reliability shows the accuracy of the measurement of the data collected in the study, thus it has great importance [3]. It is expected to have reliable results when the severity of the disease is evaluated by several raters during a clinical trial when the radiographs are evaluated by trauma surgeons and radiologists, when two clinicians classified the patients in three categories according to their syndrome type, when the severity of depression is evaluated by two psychiatrists, or when a sample of interview protocols is examined by three evaluators.

The reliability of the raters is to be determined by measuring inter-rater agreement coefficients. Cohen's weighted kappa coefficient which is the most popular coefficient and AC2 coefficients are used to determine the level of agreement between the ordinal classifications of two raters [4,5].

It is also important to determine the distinguishability of the categories (or the severity of the disease). When the categories are not defined clearly or when the raters are not expert enough in their field, the distinguishability of the categories becomes lower. If the reason is unclearly defined categories, then different raters may understand these categories differently or even the same rater may not distinguish the categories correctly. As a result of this indistinguishability, there occurs a low agreement.

In this study, it is purposed to assess the inter-rater agreement coefficients and category distinguishability in ordinal rating scales together. It is aimed to discuss how category distinguishability affects the level of reliability, and the possible solutions of low distinguishability are. Degree of distinguishability, weighted kappa, and AC2 coefficients are applied to a very well-known carcinoma in situ of the uterine cervix data. The results are discussed over the pairs of pathologists. The coefficients to measure inter-rater reliability and adjusted degree of distinguishability are introduced in Section 2. The pathology data is analyzed in Section 3, followed by the conclusions in Section 4.

2. Materials and Methods

2.1. Inter-rater agreement coefficients

Cohen's weighted kappa coefficient [4] is suggested for the analysis of reliability between the classifications of two raters. Suppose two raters rate n observations into R categories, independently. Let π_{ij} is the probability of cell (i,j) where $\pi_{i.}$ indicates the ith row total

*Corresponding author. *e-mail address: ezgiyilmaz@hacettepe.edu.tr*

probability and $\pi_{.j}$ indicates the jth column total probability. The weighted kappa coefficient (κ_w) is

$$\kappa_w = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{i.}\pi_{.j}}{1 - \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{i.}\pi_{.j}}. \quad (1)$$

Gwet's AC2 coefficient [5] is suggested to overcome the prevalence and marginal probability problem of Cohen's kappa. AC2 coefficient is

$$AC2 = \frac{\sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}\pi_{ij} - \frac{w_T}{R(R-1)}\sum_{j=1}^{R} p_i(1-p_i)}{1 - \frac{w_T}{R(R-1)}\sum_{j=1}^{R} p_i(1-p_i)}, \quad (2)$$

where

$$w_T = \sum_{i=1}^{R}\sum_{j=1}^{R} w_{ij}, \quad (3)$$

and

$$p_i = (\pi_{i.} + \pi_{.j})/2. \quad (4)$$

In Equations 1-3, w_{ij} are the weights range $0 < w_{ij} < 1$. Even there are many suggested weighting schemes, linear and quadratic weights are the well-known ones. For different weighting schemes in the literature, see [2].

- Linear weights [6]:

$$w_{ij} = 1 - \frac{|i-j|}{R-1} \quad (5)$$

- Quadratic weights [7]:

$$w_{ij} = 1 - \frac{(i-j)^2}{(R-1)^2} \quad (6)$$

In the literature, there are several interpretations of the kappa coefficient. The inference shown in Table 1 is the well-known one and can be assigned to the corresponding ranges of kappa [8].

Table 1. Interpretation of the kappa coefficient

Kappa	Strength of Agreement
0.81-1.00	Almost Perfect
0.61-0.80	Substantial
0.41-0.60	Moderate
0.00-0.20	Slight
<0.00	Poor

2.2. Category distinguishability

One of the assumptions of the kappa coefficient is the raters should rate the items independently. Even though the raters rate the items independently, because of the ordinal structure of the table, there occurs a correlation between their decisions. There are two main components of agreement: (1) Marginal homogeneity which corresponds to the differences in the marginal distributions of raters and (2) category distinguishability which is the ability for raters to distinguish the categories [9].

In the agreement studies, it is necessary to determine if the categories of the table are distinguishable from one to another [10]. If the categories are indistinguishable, then there could occur some differences between raters' perceptions. The categories may not be distinguished because of two reasons. The first problem might be due to the definition of the categories. Different raters may understand the categories differently or the same rater may not distinguish the categories correctly. The second problem might be due to the nonexpert raters. The raters may not be experts in their fields and it may be difficult to distinguish the categories. The measure to calculate the distinguishability level of the categories is called the degree of distinguishability.

The degree of distinguishability is suggested to investigate the ability of the raters to distinguish between two categories [9]. The adjusted version of the degree of distinguishability (ADD) is suggested by Yilmaz and Saracbasi [11]. ADD between i and $i+1$ categories is calculated as

$$ADD_{i,i+1} = \begin{cases} 1 - \tau_{i,i+1}^{-1} & \text{if } \tau_{i,i+1} \geq 1, \\ 1 - \tau_{i,i+1} & \text{if } \tau_{i,i+1} < 1, \end{cases} \quad (7)$$

where $0 < ADD_{i,i+1} < 1$, $i = 1,2,\ldots,(R-1)$. The odds ratio for square contingency tables is

$$\tau_{i,i+1} = \frac{\pi_{ii}\pi_{i+1,i+1}}{\pi_{i,i+1}\pi_{i+1,i}}. \quad (8)$$

The interpretation levels of ADD are given in Table 2 [11].

Table 2. Interpretation of ADD

ADD	Strength of Distinguishability
>0.99	Perfect
0.94-0.99	Substantial
0.82-0.93	Moderate
0.57-0.81	Fair
0.00-0.56	Poor

3. The Pathology Data

The pathology data discussed by Holmquist, McMahon, and Williams [12] is used to illustrate the use of the adjusted degree of distinguishability and

inter-rater reliability. To investigate the variability in the classification of carcinoma in situ of the uterine cervix, seven pathologists are classifying 118 biopsy slides into five categories: (1) Negative, (2) Atypical Squamous Hyperplasia, (3) Carcinoma in Situ, (4) Squamous Carcinoma with Early Stromal Invasion, and (5) Invasive Carcinoma.

This data set has also been analyzed in the studies of Landis and Koch [13], Becker and Agresti [14], and Agresti [15]. In their studies, the categories are reclassified into three or four categories as (1), (2), (3)+(4)+(5) or (1), (2), (3), (4)+(5).

It is aimed to investigate carcinoma in situ of uterine cervix data from the point of inter-rater reliability, from the point of category distinguishability, and also from the point of inter-rater reliability and category distinguishability together.

3.1. From the point of inter-rater reliability

The estimated values of weighted kappa and AC2 coefficients with linear and quadratic weights, their standard errors are summarized in Figure 1 for each pair of pathologists. The levels of agreement are highlighted by Landis and Koch [8] intervals.

PAIR	Kw_L	Sdt.Error
EF	0.266	0.052
BF	0.320	0.055
AF	0.334	0.052
DE	0.343	0.054
BD	0.406	0.054
FG	0.406	0.055
CF	0.408	0.060
CE	0.429	0.056
AD	0.440	0.052
BC	0.454	0.059
DF	0.462	0.055
CD	0.477	0.058
AC	0.494	0.053
AE	0.509	0.053
DG	0.545	0.052
EG	0.550	0.053
CG	0.557	0.056
AG	0.563	0.050
AB	0.572	0.054
BE	0.586	0.054
BG	0.651	0.055

PAIR	Kw_Q	Sdt.Error
EF	0.365	0.077
BF	0.419	0.076
AF	0.452	0.071
DE	0.471	0.072
CF	0.499	0.087
FG	0.510	0.073
BD	0.523	0.073
CE	0.524	0.082
BC	0.532	0.086
AD	0.549	0.067
CD	0.554	0.080
AC	0.586	0.076
DF	0.597	0.071
CG	0.622	0.079
AE	0.632	0.068
EG	0.643	0.072
DG	0.654	0.071
AB	0.663	0.068
AG	0.667	0.064
BE	0.679	0.073
BG	0.699	0.074

PAIR	AC2_L	Sdt.Error
EF	0.440	0.049
AF	0.463	0.051
BF	0.518	0.045
DE	0.550	0.042
AD	0.579	0.045
BD	0.604	0.040
FG	0.610	0.041
CF	0.623	0.043
AC	0.630	0.044
CE	0.633	0.042
AE	0.655	0.040
BC	0.663	0.041
DF	0.679	0.038
CD	0.684	0.040
AG	0.700	0.039
AB	0.713	0.039
DG	0.713	0.035
CG	0.735	0.037
EG	0.736	0.038
BE	0.764	0.036
BG	0.814	0.034

PAIR	AC2_Q	Sdt.Error
EF	0.628	0.056
AF	0.633	0.058
BF	0.702	0.042
AD	0.732	0.047
DE	0.741	0.042
AC	0.769	0.050
FG	0.774	0.037
CF	0.774	0.045
BD	0.783	0.041
CE	0.788	0.044
BC	0.807	0.037
AE	0.809	0.039
CD	0.818	0.038
DF	0.832	0.034
AG	0.832	0.037
AB	0.837	0.036
CG	0.851	0.035
DG	0.855	0.033
EG	0.859	0.034
BE	0.880	0.032
BG	0.895	0.029

Fair
Moderate
Substantial
Almost perfect

Figure 1. The levels of inter-rater reliability that are highlighted by Landis and Koch's intervals

The results show that the values of quadratically weighted agreement coefficients are higher than the linearly weighted ones. Furthermore, the values of the AC2 coefficient are higher than the weighted kappa. The value of the inter-rater reliability is higher when the quadratically weighted AC2 coefficient is used and is lower when the linearly weighted kappa coefficient is used.

According to the linearly weighted kappa results, there are fair agreements between Pathologists E and F, B and F, A and F, D and E. According to the quadratically weighted kappa results, there is a fair agreement between Pathologists E and F. According to the AC2 coefficient results, there are more than fair agreements between all the pairs of pathologists. In general, Pathologist F has a low agreement with the other pathologists. The highest agreement is observed between Pathologists B and G, B and E.

As the overall agreement coefficient, Light's weighted kappa [16] is used. Linearly weighted Light's kappa is calculated as 0.465 and the quadratically weighted one is calculated as 0.564. There is a moderate agreement between the seven pathologists' decisions.

3.2. From the point of category distinguishability

The levels of ADD that are highlighted by Yilmaz and Saracbasi [11] intervals are given in Figure 2 for the adjacent categories. The results show that six pairs of pathologists cannot classify (1) and (2) well. Three pairs of pathologists cannot classify (2) and (3) well. 14 pairs of pathologists cannot classify (3) and (4) well. 10 pairs of pathologists cannot classify (4) and (5) well. In general, pathologists have difficulties classifying the last three categories.

According to the results in Figure 2, when Pathologists C and E cannot distinguish (1) and (2) well, Pathologist G substantially distinguishes. When Pathologist F

cannot distinguish (2) and (3) well, Pathologist G substantially distinguishes. When Pathologist F cannot distinguish (3) and (4) well, Pathologist B substantially distinguishes. When Pathologists D, G, and F cannot distinguish (4) and (5) well, Pathologist B substantially distinguishes.

PAIR	ADD12
CE	0.662
BE	0.690
AE	0.720
BC	0.780
AC	0.791
DE	0.811
EF	0.815
DF	0.856
EG	0.872
CF	0.900
BD	0.902
CG	0.905
AB	0.919
BF	0.924
CD	0.938
AF	0.940
AG	0.943
DG	0.944
AD	0.956
FG	0.966
BG	0.977

PAIR	ADD23
BF	0.642
CF	0.687
DF	0.740
DE	0.815
EF	0.824
BC	0.825
AF	0.842
AE	0.858
CD	0.859
AB	0.868
BD	0.869
BE	0.892
FG	0.903
AG	0.933
DG	0.941
AD	0.946
EG	0.960
BG	0.963
AC	0.974
CE	0.979
CG	0.984

PAIR	ADD34
EF	0.061
CG	0.085
CD	0.103
AF	0.174
BF	0.182
AG	0.369
CE	0.533
BD	0.574
DF	0.609
AD	0.609
EG	0.725
AE	0.725
FG	0.769
DE	0.780
BC	0.862
BG	0.881
CF	0.920
AC	0.931
BE	0.948
DG	0.973
AB	0.974

PAIR	ADD45
CD	0.000
AF	0.222
DF	0.400
FG	0.400
BD	0.444
DE	0.667
AG	0.800
BG	0.800
CG	0.800
EG	0.800
AD	0.815
BF	0.857
DG	0.857
EF	0.857
CE	0.880
AE	0.914
CF	0.933
AC	0.945
BE	0.961
BC	0.971
AB	0.990

Poor
Fair
Moderate
Substantial

Figure 2. The levels of ADD that are highlighted by Yilmaz and Saracbasi's intervals

3.3. From the point of reliability and category distinguishability

According to the inter-rater agreement coefficients, it is concluded that there are fair inter-rater reliabilities between Pathologists A and F, B and F, D and E, E and F. One of the reasons for the low agreement is a low ability of distinguishability. The unclearly defined categories or non-expert pathologists may cause a low distinguishability. In this section, the sub-tables that low agreements occur are analyzed in more detail.

Pathologists A and F

The estimated values of linearly weighted inter-agreements and ADD coefficients of Pathologists A and F are summarized in Table 3.

Table 3. The summary of the linearly inter-agreements and ADD coefficients of Pathologists A and F

	ADD				Inter-Rater Agreement	
	12	23	34	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.940	0.842	0.174	0.222	0.334 (0.052)	0.463 (0.051)
Level	Substantial	Moderate	Poor	Poor	Fair	Moderate

The results show that there is poor distinguishability between (3) and (4), and between (4) and (5). A poor distinguishability indicates that pathologists A and F cannot distinguish these categories well. Thus, the categories can be combined as (3+4), (4+5), or (3+4+5). Linearly inter-rater agreement and ADD coefficients for adjacent categories are calculated for the reclassified tables. The results of the three alternatives are:

Alternative 1: 1, 2, (3+4), 5

	ADD			**Inter-Rater Agreement**	
	12	2(3+4)	(3+4)5	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.940	0.908	0.968	0.366 (0.054)	0.479 (0.051)
Level	Substantial	Moderate	Substantial	Fair	Moderate

Alternative 2: 1, 2, 3, (4+5)

	ADD			Inter-Rater Agreement	
	12	23	3(4+5)	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.940	0.842	0.890	0.329 (0.049)	0.331 (0.063)
Level	Substantial	Moderate	Moderate	Fair	Fair

Alternative 3: 1, 2, (3+4+5)

	ADD		Inter-Rater Agreement	
	12	2(3+4+5)	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.940	0.924	0.364 (0.052)	0.272 (0.072)
Level	Substantial	Moderate	Fair	Fair

Alternative 1 is suggested to use because the highest values of inter-rater agreement coefficients are observed. The linearly weighted kappa increases to 0.366 and linearly weighted AC2 is increases to 0.479 after the reclassification 1. For the first alternative, the adjusted degree of distinguishability of (2) and (3+4) increases to moderate, the adjusted degree of distinguishability of (3+4) and (5) increases to a substantial level.

Pathologists B and F

The estimated values of linearly weighted inter-agreements and ADD coefficients of Pathologists B and F are summarized in Table 4.

Table 4. The summary of the linearly inter-agreements and ADD coefficients of Pathologists B and F

	ADD				Inter-Rater Agreement	
	12	23	34	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.924	0.642	0.182	0.857	0.320 (0.055)	0.518 (0.045)
Level	Moderate	Fair	Poor	Moderate	Fair	Moderate

The results show that there is poor distinguishability between (3) and (4). A poor distinguishability indicates that pathologists B and F cannot distinguish these categories well. Thus, the categories can be combined as (2+3) or (3+4). Linearly inter-rater agreement and ADD coefficients for adjacent categories are calculated for the reclassified tables. The results of the two alternatives are:

Alternative 1: 1, (2+3), 4, 5

	ADD			Inter-Rater Agreement	
	1(2+3)	(2+3)4	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.988	0.716	0.857	0.421 (0.070)	0.739 (0.037)
Level	Substantial	Fair	Moderate	Moderate	Substantial

Alternative 2: 1, 2, (3+4), 5

	ADD			Inter-Rater Agreement	
	12	2(3+4)	(3+4)5	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.924	0.726	0.990	0.324 (0.053)	0.451 (0.051)
Level	Moderate	Fair	Substantial	Fair	Moderate

Alternative 1 is suggested to use because the highest values of inter-rater agreement coefficients are observed. The linearly weighted kappa increases to 0.421 and linearly weighted AC2 increases to 0.759 after the reclassification 1. For the first alternative, the adjusted degree of distinguishability of (1) and (2+3) is at a substantial level and the adjusted degree of distinguishability of (2+3) and (4) is at a fair level.

Pathologists D and E

The estimated values of linearly weighted inter-agreements and ADD coefficients of Pathologists D and E are summarized in Table 5.

Table 5. The summary of the linearly inter-agreements and ADD coefficients of Pathologists D and E

	ADD				Inter-Rater Agreement	
	12	23	34	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.811	0.815	0.780	0.667	0.343 (0.054)	0.550 (0.042)
Level	Moderate	Moderate	Fair	Fair	Fair	Moderate

The results show that there is fair distinguishability between (3) and (4), and between (4) and (5). A poor distinguishability indicates that Pathologists D and E cannot distinguish these categories well. Thus, the categories can be combined as (3+4), (4+5), or (3+4+5). Linearly inter-rater agreement and ADD coefficients for adjacent categories are calculated for the reclassified tables. The results of the three alternatives are:

Alternative 1: 1, 2, (3+4), 5

	ADD			Inter-Rater Agreement	
	12	2(3+4)	(3+4)5	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.811	0.817	0.971	0.368 (0.054)	0.567 (0.041)
Level	Moderate	Moderate	Substantial	Fair	Moderate

Alternative 2: 1, 2, 3, (4+5)

	ADD			Inter-Rater Agreement	
	12	23	3(4+5)	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.811	0.815	0.900	0.353 (0.053)	0.444 (0.050)
Level	Moderate	Moderate	Moderate	Fair	Moderate

Alternative 3: 1, 2, (3+4+5)

	ADD		Inter-Rater Agreement	
	12	2(3+4+5)	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.811	0.823	0.384 (0.053)	0.389 (0.058)
Level	Moderate	Moderate	Fair	Fair

The highest value of linearly weighted kappa coefficient is observed when Alternative 3 is used and the highest value of linearly weighted AC2 coefficient is observed when Alternative 1 is used. The linearly weighted kappa increases to 0.364 after reclassification 3 and increases to 0.368 after reclassification 1. The linearly weighted AC2 increases to 0.567 after reclassification 1 and decreases to 0.389 after reclassification 3.

For the first alternative, the poor distinguishability increase to a substantial level after the reclassification as (3+4). For the third alternative, the adjusted degree of distinguishability of (2) and (3+4+5) increases to a moderate level.

Even though the value of linearly weighted kappa in Alternative 1 is less than the value of kappa obtained from Alternative 3, the values of linearly weighted AC2 and ADD coefficients are higher. Thus, Alternative 1 is suggested to use the interpretation of Pathologists D and E's results.

Pathologists E and F

The estimated values of linearly weighted inter-agreements and ADD coefficients of Pathologists E and F are summarized in Table 6.

Table 6. The summary of the linearly inter-agreements and ADD coefficients of Pathologists E and F

	ADD				Inter-Rater Agreement	
	12	23	34	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.815	0.824	0.061	0.857	0.266 (0.052)	0.440 (0.049)
Level	Moderate	Moderate	Poor	Moderate	Fair	Moderate

The results show that there is poor distinguishability between (3) and (4). A poor distinguishability indicates that Pathologists E and F cannot distinguish these categories well. Thus, the categories can be combined as (2+3) or (3+4). Linearly inter-rater agreement and ADD coefficients for adjacent

categories are calculated for the reclassified tables. The results of the two alternatives are:

Alternative 1: 1, (2+3), 4, 5

	ADD			Inter-Rater Agreement	
	1(2+3)	(2+3)4	45	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.972	0.414	0.857	0.272 (0.066)	0.638 (0.045)
Level	Substantial	Poor	Moderate	Fair	Substantial

Alternative 2: 1, 2, (3+4), 5

	ADD			Inter-Rater Agreement	
	12	2(3+4)	(3+4)5	κ_w (Std. Error)	AC2 (Std. Error)
Estimate	0.815	0.858	0.990	0.281 (0.050)	0.406 (0.049)
Level	Moderate	Moderate	Substantial	Fair	Moderate

The highest value of linearly weighted kappa coefficient is observed when Alternative 2 is used and the highest value of linearly weighted AC2 coefficient is observed when Alternative 1 is used. The linearly weighted kappa increases to 0.272 after reclassification 1 and increased to 0.281 after reclassification 2. The linearly weighted AC2 increases to 0.638 after reclassification 1 and decreases to 0.406 after reclassification 2.

For the first alternative, even though the adjusted degree of distinguishability of (1) and (2+3) increases to a substantial level, the adjusted degree of distinguishability of (2+3) and (4) is still at a poor level. For the second alternative, even though the adjusted degree of distinguishability of (1) and (2) is still at a moderate level, the adjusted degree of distinguishability of (2) and (3+4) increases to moderate and the adjusted degree of distinguishability of (3+4) and (5) increases to a substantial level.

Even though the value of AC2 decreases, because the values of linearly weighted kappa and ADD coefficients increase, Alternative 2 is suggested to use the interpretation of Pathologists E and F's results.

4. Conclusions

In recent studies, inter-rater reliability and category distinguishability have grown impartances. It has been proposing to use agreement coefficients and degree of distinguishability simultaneously [11]. This study is aimed to illustrate how to use inter-rater reliability and degree of distinguishability, together. For this purpose, the carcinoma in situ of uterine cervix data is used. Seven pathologists classify 118 slides into five ordinal categories to investigate the variability in the classification of carcinoma in situ of the uterine cervix. Landis and Koch [13], Becker and Agresti [14], and Agresti [15] reclassify the data into three or four categories, however, the reclassification procedures are made by considering the zero cells or the researcher's personal experience.

Adjusted degree of distinguishability, weighted kappa, and AC2 coefficients are applied to data for 21 pairs of the seven pathologists. The results are discussed together in terms of inter-rater reliability, category distinguishability, and inter-rater reliability and category distinguishability together.

The inter-rater reliability results showed that the value of the quadratically weighted kappa is higher than the value of the linearly weighted kappa. Besides, the value of the quadratically weighted AC2 is higher than the value of the linearly weighted AC2, as well. Pathologist F has the lowest, Pathologists B and G have the highest agreement with the others.

The adjusted degree of distinguishability results showed that Pathologist F cannot distinguish the categories except categories 1 and 2. The reason is Pathologist F may have less experience than the other pathologists. Pathologists C and E cannot distinguish the categories 1 and 2. In general, because Pathologist F has a lower agreement between the other pathologists, it may be excluded from the study.

The results showed that the pathologists have some problems distinguishing the categories (3) Carcinoma in Situ, (4) Squamous Carcinoma with Early Stromal Invasion, and (5) Invasive Carcinoma, and the incorrect classifications affect the level of the agreement in this respect. It is suggested to recollect the data or to combine the categories as considering the category distinguishability. According to the poor and fair inter-rater reliability between Pathologists A and F, B and F, D and E, E and F, the degrees of distinguishability of these sub-tables are analyzed in more detail. To get more reliable results for Pathologists A and F and Pathologists D and E, it is

suggested to combine (3) Carcinoma in Situ and (4) Squamous Carcinoma with Early Stromal Invasion. Besides, it is suggested to combine (2) Atypical Squamous Hyperplasia and (3) Carcinoma in Situ for Pathologists B and F and Pathologists E and F. As a result of reclassifications, an increase in the level of inter-rater reliability is observed.

Conflicts of interest

The authors state that did not have a conflict of interests.

References

[1] Altun G., Aktaş, S., Karesel Olumsallık Tablolarında Asimetri ve Çarpık Simetri Modelleri, *Turkiye Klinikleri J Biostat*, 8 (2) (2016) 152–161.

[2] Yilmaz, A.E., Saracbasi, T., Assessing Agreement between Raters from the Point of Coefficients and Log-linear Models, *Journal of Data Science*, 14 (1) (2017) 1–24.

[3] McHugh, M.L., Interrater Reliability: the Kappa Statistic, *Biochemia Medica*, 22 (3) (2012) 276–282.

[4] Cohen, J., Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit, *Psychological Bulletin*, 70 (4) (1968) 213–220.

[5] Gwet, K.L. Handbook of Inter-rater Reliability, The Definitive Guide to Measuring the Extent of Agreement among Raters. 3rd ed. Maryland: Advanced Analytics, LLC, (2002).

[6] Cicchetti, D., Allison, T., A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings, *American Journal EEG Technology*, 11 (3) (1971) 101–109.

[7] Fleiss, J.L., Cohen, J., The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measure of Reliability, *Educational and Psychological Measurement*, 33 (3) (1973) 613–619.

[8] Landis, J.R., Koch, G.G., The Measurement of Observer Agreement for Categorical Data, *Biometrics*, 33 (1) (1977a) 159–174.

[9] Darroch, J.N., McCloud, P.I., Category Distinguishability and Observer Agreement, *Australian Journal of Statistics*, 28 (3) (1986) 371–388.

[10] Perkins, S.M., Becker, M.P., Assessing Rater Agreement using Marginal Association Models, *Statistics in Medicine*, 21 (12) (2002) 1743–1760.

[11] Yilmaz, A.E., Saracbasi, T., Agreement and Adjusted Degree of Distinguishability for Square Contingency Tables, *Hacettepe Journal of Mathematics and Statistics*, 48 (2) (2019) 592–604.

[12] Holmquist, N.D., McMahon, C.A., Williams, O.D., Variability in Classification of Carcinoma in Situ of the Uterine Cervix, *Archives of Pathology*, 84 (4) (1967) 334–345.

[13] Landis, J.R., Koch, G.G., An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers, *Biometrics*, 33 (2) (1977b) 363–374.

[14] Becker, M.P., Agresti, A., Log-linear Modelling of Pairwise Interobserver Agreement on a Categorical Scale, *Statistics in Medicine*, 33 (1) (1992) 101–114.

[15] Agresti, A., Categorical Data Analysis. New York: John Wiley and Sons, (2002).

[16] Light, R.J., Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternatives, *Psychological Bulletin*, 76 (5) (1971) 365-377.