A Comparison of the Sensitivity, Specificity and Prevalence Response of Coefficients of Individual Agreement (CIA) with Cohen's Kappa and Gwet's AC1 Statistics

Birey Uyum Katsayısının (CIA) Duyarlılık, Özgüllük ve Prevelanstan Etkilenme Durumunun Cohen Kappa ve Gwet'in AC1 İstatistiği ile Karşılaştırılması

Semra ERDOĞAN,[a]
Gülhan OREKİCİ TEMEL[a]

[a]Department of Biostatistics and Medical Informatics, Mersin University Faculty of Medicine, Mersin

Geliş Tarihi/*Received:* 15.07.2014
Kabul Tarihi/*Accepted:* 05.11.2014

Yazışma Adresi/*Correspondence:* Semra ERDOĞAN
Mersin University Faculty of Medicine, Department of Biostatistics and Medical Informatics, Mersin, TÜRKİYE/TURKEY
semraerdogann@gmail.com

ABSTRACT Objective: In this study, a condition of diagnostic test only two categories as patient/healthy (or positive/negative) is evaluated by two clinicians is considered. Additionally, the aim of this study is exhibiting how the condition of the coefficients of individual agreement (CIA), Cohen's Kappa, AC1 statistics which are some of agreement statistics using for evaluating the agreement between the clinicians is affected by the sensitivity, specificity and prevalence. **Material and Methods:** In this study, two different scenarios are established. In the first scenario, it is aimed to show how they are affected from the prevalence by fixing sensitivity and specificity values. In case of the second scenario, it is considered that the sensitivity and specificity values of X observer are not equal and higher and the sensitivity and specificity values of Y observer are not equal and the above-mentioned 4 combinations. In these two scenarios and eight different combinations, it is aimed to define how the coefficients of individual agreement, Cohen's kappa statistics and AC1 statistics change for different prevalence values and to show how they are affected by the sensitivity, specificity and prevalence values. **Results:** In Scenario 2, in the second and third combinations with, it is an expected situation that in case of a high sensitivity, the agreement coefficients increases when the prevalence value increases and in cases of low sensitivity, they decreases depending on the increase rate of the prevalence value. In these combinations it is observed that only the coefficients of individual agreement give such results while kappa statistics and AC1 statistics have symmetrical results in all combinations. **Conclusion:** At the end of this study, while observing the agreement between the observers in reliability studies including two observers and diagnostic test consists of two categories such as "patients" and "healthy", it is suggested that the researchers should take into account the prevalence and bias concepts and use the coefficients of individual agreement (CIA) since it is not affected by the sensitivity, specificity and prevalence values.

Key Words: Prevalence; agreement; coefficient of individual (CIA); AC1 statistics, Cohen's Kappa statistics

ÖZET Amaç: Bu çalışmada, hasta/sağlam (ya da pozitif/negatif) olmak üzere sadece iki kategorisi olan bir tanı testin iki klinisyen tarafından değerlendirildiği durumlar dikkate alınmıştır. Ayrıca bu çalışmanın amacı, klinisyenler arası uyumu değerlendirebilmek için kullanılan uyum istatistiklerinden birey uyum katsayısı, kappa katsayısı ve AC1 istatistiğinin, duyarlılık, özgüllük ve prevelanstan etkilenme durumları ortaya konulmaya çalışılmıştır. **Gereç ve Yöntemler:** Bu çalışmada iki farklı senaryo tasarlanmıştır. İlk senaryoda, duyarlılık ve özgüllük değerleri sabit tutularak prevelanstan etkilenme durumları ortaya konmaya çalışılmıştır. İkinci senaryoda ise X değerlendiricisinin duyarlılık ve özgüllük değerlerinin yüksek ve eşit olmadığı, Y değerlendiricisinin de duyarlılık ve özgüllük değerlerinin eşit olmadığı ve 4 farklı kombinasyon göz önünde bulundurulmuştur. Bu iki senaryo ve sekiz farklı kombinasyonda, farklı prevelans değerleri için birey uyum katsayıları, Cohen Kappa istatistiği ve AC1 istatistiği hesaplanarak nasıl bir değişim gösterdiği belirlenmeye, duyarlılık, özgüllük ve prevelanstan etkilenme durumları ortaya konulmaya çalışılmıştır. **Bulgular:** Senaryo 2'de ikinci ve üçüncü kombinasyonda, duyarlılığın yüksek olduğu bir durumda prevelans değeri arttıkça uyum katsayılarının artması, duyarlılığın düşük olduğu durumda ise prevelansın artışına bağlı olarak uyum katsayılarının azalması beklenen bir durumdur. Bu kombinasyonlarda sadece birey uyum katsayılarının böyle bir sonuç verdiği gözlenmekte ancak kappa istatistiğinin ve AC1 istatistiğinin tüm kombinasyonlarda simetrik bir durum sergilediği fark edilmektedir. **Sonuç:** Bu çalışma sonunda iki değerlendiricinin bulunduğu ve tanı testinin hasta/sağlam şeklinde iki kategorisinin olduğu güvenilirlik çalışmalarında değerlendiriciler arasındaki uyum incelenirken araştırmacılara, prevelans ve yanlılık kavramlarını dikkate almaları gerektiğini ve duyarlılık, özgüllük ve prevelanstan etkilenmeyen bir yöntem olan birey uyum katsayısını (CIA) kullanmaları önerilmektedir.

Anahtar Kelimeler: Prevalans; uyum; birey uyum katsayısı (CIA); AC1 istatistiği; Cohen Kappa istatistiği

doi: 10.5336/biostatic.2014-41410

Turkiye Klinikleri J Biostat 2015;7(1):25-38

In general, it might be desired to research whether measurements that are taken from the same individuals at different times or measurements that are taken from the same individuals by different observers are compatible with each other in method comparisons and reliability studies. In addition, new methods are proposed, which are day by day considered to be better than the older ones in line with technological advancements. A treatment method, which is considered to be gold standard, or cheaper methods that respond quicker, which are regarded as reliable in case when there is no gold standard method, are used in treatment of any disease. When a newly developed method is found, it might be sought to evaluate whether this method is to what extent compatible with the gold standard method or a reliable method via comparison. If the differentiation between this newly developed method and the standard method is not to an extent that will change clinical interpretation, this newly developed method can be used instead of the older method or both methods can be used interchangeably. Therefore, it is required to establish accuracy and precision of the addressed measurement or the developed method in order to be used as an alternative in method comparisons. Accuracy (systematic bias or bias) represents the closeness of the mean test results to the true value or the accepted reference value while precision (random error) represents the closeness of agreement to the test results. In the method comparison studies, the concept of agreement includes both the terms of precision (random error) and accuracy (systematic bias). Therefore, it is quite important to consider all these concepts.[1-4]

While agreement refers to similarity between measurements obtained through different methods, disagreement refers to how much these obtained measurements are different from each other. When disagreement is the case between methods or observers, it should be investigated whether this disagreement stems from systematic error (bias) or random error. Because, systematic error can be stabilized via calibration in general but on the other hand, taking random error under control is quite difficult since it is a type of error that randomly interferes with measurement results, whose amount and direction are not definitely known and that emerges with luck. If disagreement (differentiation) is due to the random error within a certain method, this method is not recommended to be used in practice. If disagreement is due to the true differences among the methods (systematic bias), this method should be modified. Therefore, assessing agreement generally results with assessment of inter-method agreement and of intra-method agreement. Intra-method agreement measures consistency of measurements obtained by the same method and inter-method agreement measures consistency of true measurements obtained from these methods. Inter-method agreement is defined over true values rather than observed values since it cannot be concealed with random error within the method.[5]

The biggest problem while conducting agreement analyses for obtained measurement results is to decide statistical method to be used. It is seen in many agreement studies that classical statistical methods such as Pearson's correlation coefficient, regression analysis and t-tests for dependent groups are used. However, it was observed that the results obtained as a result of these known classical methods are inaccurate and alternative methods have been developed. The statistical method to be used varies depending on with what variable the measurement result is identified or in other words whether outcome variable is constant, discrete, categorical or sequenced, whether the measured property (variable) fulfills the condition for normality, on the number of observers, if used the number of diagnostic tests and the number of categories in diagnostic test.[6]

If the outcome variable is categorized, Scott's p statistics, Cohen's kappa statistics, the G-index, Gwet's AC1 statistics, Fleiss' kappa statistics and Krippendorf's Alpha coefficient are widely used in the literature. If the final variable has a sequential structure, weighed kappa statistics and Kendall's W coefficient are used.[7,8]

In addition, Haber and Barnhart (2007) proposed that disagreement between measurements obtained from the same individuals via different

methods and disagreement between repeated measurements of the same individuals obtained via the same method are similar. In other words, they advocated that when a method replaces another method or methods trade places with each other, this does not increase the amount of disagreement between measurements obtained from the same individuals. Based on this information, they proposed a new approach for estimating and defining agreement coefficient between observers or methods. This approach is the Coefficient of Individual Agreement (CIA) that is described as a special disagreement function, which can be used in cases when repeated measures are also continuous and categorical variables.[9]

In this study, the situations when a diagnostic test with only two categories as of the diseased / not diseased (or positive/negative) are applied on the same individuals by two different clinicians are considered. Additionally, the aim of this study is to show how the coefficients of individual agreement change in order to evaluate the agreement between the clinicians and how these coefficients are affected by Cohen's Kappa, Gwet's AC1 statistics and the prevalence that are available in the literature.

MATERIAL AND METHODS

COEFFICIENT OF INDIVIDUAL AGREEMENT (CIA)

In order to define a coefficient of agreement, we first have to decide how we quantify the agreement between the two methods or clinicians. X and Y show the measurement value of first and second clinician respectively. In cases where there is only two clinicians, measurements of these clinicians are indicated with X and Y. Replicated measurements for the first clinician (X) is indicated with X and X'; disagreement function between two measurements is G(X,X'), two replicated measurements for the second clinician (Y) is indicated with Y and Y' and disagreement function between these two measurements are defined as G(Y,Y'). The quantity of disagreement between measurements obtained from the same individuals is presented with G(X, Y). It is assumed that it is G(X,Y)≥ 0 and G(X,X)=0 for the disagreement functions. Additionally, G(X,Y) disagreement coefficient increases as the disagreement between X and Y increases.[9,10]

N denotes the number of subjects included in the study and stated as i=1,2….N. X_{ik} denotes k.replicated measurement value of X clinician obtained from i. subject (k=1,2,…K_i), Y_{il} denotes l. replicated observations obtained from i. subject (l=1,2……, L_i). For a structure with two results; positive case X and Y values will be equal to 1, negative case X and Y values will be equal to 0. For a positive case of X clinician; $P(X_{ik}=1)=\pi_i$ (k=1,...,K_i), for a positive case of Y clinician $P(Y_{il}=1)=\lambda_i$ (l=1,..,L_i). Disagreement functions specific to individuals are denoted as

$$\begin{aligned} G_i(X,Y) &= P(X_{ik} \neq Y_{il}/i) \\ &= Pr(X_{ik}=1, Y_{il}=0/i) + Pr(X_{ik}=0, Y_{il}=1/i) \\ &= \pi_i(1-\lambda_i) + (1-\pi_i)\lambda_i \\ &= \pi_i + \lambda_i - 2\pi_i\lambda_i \end{aligned} \quad (1)$$

$$\begin{aligned} G_i(X,X') &= P(X_{ik} \neq X_{ik'}/i; k \neq k') \\ &= 2\pi_i(1-\pi_i) \end{aligned} \quad (2)$$

$$\begin{aligned} G_i(Y,Y') &= P(Y_{il} \neq Y_{il'}/i; l \neq l') \\ &= 2\lambda_i(1-\lambda_i) \end{aligned} \quad (3)$$

Total disagreement function, G, mean of disagreement functions for all individuals (G_i) are formulated as below.[10-14]

$$\overline{G} = 1/N \sum_{i=1}^{N} G_i \quad (4)$$

Haber and Barnhart (2007) assessed cases where one of the observers is a reference observer as well as where none is a reference observer while evaluating the agreement between observers.[9] Based on this, if none of the observers is considered as a reference, coefficient of individual agreement (CIA) is formulated as in the Equation 5. This equation's numerator provides the average of disagreements between two repeated measurements taken from the same individuals by the same observer; its denominator provides disagreement between observers X and Y.[9-13]

$$\psi^N = \frac{\left(\overline{G}(X,X') + \overline{G}(Y,Y')\right)}{\overline{G}(X,Y)} = \frac{\sum_i [\pi_i(1-\pi_i) + \lambda_i(1-\lambda_i)]}{\sum_i (\pi_i + \lambda_i - 2\pi_i\lambda_i)} \quad (5)$$

If measurements of an experienced or an reliable observer are to be compared with measurements of a new observer, the observer X should be provided as reference and coefficient of individual agreement (CIA) should be expressed as in the Equation 6 when this new observer is compared with measurements of Y. The equation's numerator provides the disagreement between repeated measurements of the observer that is considered to be reference; its denominator provides disagreement between observers X and Y.[10,13,14]

$$\psi^R = \frac{\overline{G}(X,X')}{\overline{G}(X,Y)} = \frac{2\sum_i \pi_i(1-\pi_i)}{\sum_i(\pi_i + \lambda_i - 2\pi_i\lambda_i)} \quad (6)$$

INTERPRETATION AND PROPERTIES OF THE COEFFICIENTS

Ψ^N coefficient of individual agreement is measured between 0-1 while the agreement coefficient of Ψ^R can exceed 1. Coefficients of individual agreement (CIA) generally take a value between 0 and 1. It is believed that for an acceptable agreement, coefficients of individual agreement should be at least 0.80. A value that is less than 0.80 for Ψ which is regarded as the probability of disagreement between observers, is more or 25% more than probability of disagreement between measurements that were taken twice by the same observer. A very small value of Ψ generally result from almost perfect agreement between repeated measurements obtained from the same observer. If all repeated measurements of the reference observer obtained from the same individual are equal to the same value, the coefficient Ψ^R will be equal to 0. Likewise, when there is no variability within repeated measurements taken from individuals for both observers, Ψ^N will also be equal to 0.[10,14]

A LATENT CLASS MODEL FOR DIAGNOSTIC AGREEMENT

The Latent class models are firstly used by Dawid and Skene (1979) for the agreements of diagnostic tests.[13] According to this model, each individual participated to the study is defined either as "diseased" (the patient) or "not diseased" (the healthy). D represents illness status of the individuals as the true patient or the healthy. X and Y represent the individuals engaged in the study. The patients are represented as having positive or 1 value while the healthy is represented as having negative or 0 value. This model includes the following five parameters. ω show the prevalence of illness, η_1 and θ_1 show the sensitivity value of X and Y while $(1-\theta_0)$ and $(1-\eta_0)$ show the specificity value of X and Y. Parameters are indicated in Equation 7-11.[13]

$$\omega = P(D=1) \quad (7)$$

$$\eta_1 = P(X=1/D=1) \quad (8)$$

$$\theta_1 = P(Y=1/D=1) \quad (9)$$

$$1-\eta_0 = P(X=0/D=0) \Rightarrow \eta_0 = P(X=1/D=0) \quad (10)$$

$$1-\theta_0 = P(Y=0/D=0) \Rightarrow \theta_0 = P(Y=1/D=0) \quad (11)$$

Disagreement functions of X and Y observers are indicated in Equation 12-14.[10,12,13]

$$G(X,Y) = \omega(\eta_1 + \theta_1 - 2\eta_1\theta_1) + (1-\omega)(\eta_0 + \theta_0 - 2\eta_0\theta_0) \quad (12)$$

$$G(X,X') = 2\omega\eta_1(1-\eta_1) + 2(1-\omega)\eta_0(1-\eta_0) \quad (13)$$

$$G(Y,Y') = 2\omega\theta_1(1-\theta_1) + 2(1-\omega)\theta_0(1-\theta_0) \quad (14)$$

If X observer is accepted as a reference, then the sensitivity and specificity values should be high. Accordingly, it is assumed that the sensitivity of X observer is higher than 0.50 (η_1>0.50) and the specificity of X observer is higher than 0.50 (η_0<0.50) so that Ψ^R is an increasing function for the sensitivity and a decreasing function for the specificity of Y observer. In other words, it is desired to have high sensitivity and high specificity values without considering how close the values of Y observer to the sensitivity and specificity values of X observer. Therefore, if Ψ^R value reaches a higher value when a new observer is compared with a good reference method, it is expected that sensitivity and specificity values are significantly sensible.[10]

COHEN'S KAPPA STATISTIC

Cohen (1960) suggested Kappa statistic for evaluating of agreement between two raters and formulated as below.[15-18]

$$\kappa = \frac{P_a - P_{e/k}}{1 - P_{e/k}} \quad (15)$$

The overall agreement probability is expressed as P_a. The change agreement probability is ex-

pressed as $P_{e/\kappa}$. Equation 16 show the change agreement probability, Equation 17 show the overall agreement probability.[16]

$$P_{e/\kappa} = 1-(P_{A+}+P_{B+})+2P_{A+}P_{B+} \quad (16)$$

$$P_a = (1-\alpha_A)(1-\alpha_B)+\alpha_A\alpha_B \quad (17)$$

Let P_{A+} and P_{B+} denote respectively the probabilities that raters A and B to classify a participant into the positive category and formulated as Equation 17 and 18 respectively.[16-18]

$$P_{A+} = P_r\alpha_A+(1-P_r)(1-\beta_A) \quad (17)$$

$$P_{B+} = P_r\alpha_B+(1-P_r)(1-\beta_B) \quad (18)$$

Where Pr represents the population trait prevalence. α_A and α_B denote, respectively, raters A and B sensitivity values. Similarity β_A and β_B denote, respectively, raters A and B specificity values. The general equation of Cohen's kappa statistics is given by:[16]

$$\hat{\gamma}_\kappa = \frac{(2\alpha_A-1)(2\alpha_B-1)P_r(1-P_r)}{(2\alpha_A-1)(2\alpha_B-1)P_r(1-P_r)+\frac{(1-P_a)}{2}} \quad (19)$$

GWET'S AC1 STATISTICS

AC1 statistic was proposed in 1991 by Gwet as an alternative to Kappa statistic. Gwet's AC1 statistics is called the first order agreement coefficient or AC1 statistics and formulated as below.[16-18]

$$AC1 = \gamma = \frac{P_a-P_e(\gamma)}{1-P_e(\gamma)} \quad (20)$$

The overall agreement probability for AC1 statistics is defined as Equation 17. The change agreement probability is calculated as $P_e(\gamma) = 2\pi_+(1-\pi_+)$. Where $\pi_+ = \frac{(P_{A+}+P_{B+})}{2} = \lambda P_r+(1-\lambda)(1-P_r)$. and $\lambda = \frac{(\alpha_A+\alpha_B)}{2}$.[16]

SIMULATIONS

In this study two different scenarios are established. In the first scenario, it is aimed to show how they are affected from the prevalence by fixing sensitivity and specificity values. In such a situation, considering three different situations in which the sensitivity and specificity values of X observer are equal and high (η_1=0.90; $(1-\eta_0)$=0.90), medium (η_1=0.50; $(1-\eta_0)$=0.50) and low (η_1=0.30; $(1-\eta_0)$=0.30), 5 different combinations in which the sensitivity and specificity values of Y observer are equal and 0.90; 0.80; 0.50; 0.40 and 0.30.

In case of the second scenario, it is considered 2 different situations that the sensitivity and specificity values of X observer are not equal and high (η_1=0.90; $(1-\eta_0)$=0.80) and low (η_1=0.40; $(1-\eta_0)$=0.30), it is considering 5 different situations in which the sensitivity and specificity values of Y observer are not equal and high (θ_1=0.80; $(1-\theta_0)$=0.70), the sensitivity is high, while the specificity is low (θ_1=0.80; $(1-\theta_0)$=0.40), the sensitivity is low, while the specificity is high (θ_1=0.40; $(1-\theta_0)$=0.80), the sensitivity and specificity values are medium (θ_1=0.50; $(1-\theta_0)$=0.40) and the sensitivity and specificity values are low (θ_1=0.30; $(1-\theta_0)$=0.20). In these two scenarios and 25 (15 for first scenarios, 10 for two scenarios) different combinations, it is aimed to define how the coefficients of individual agreement change for different prevalence values and to show how they are affected by the sensitivity, specificity and prevalence values.

Additionally, it is determined by certain researchers that Cohen's Kappa coefficient is widely used for agreement statistics, however Kappa statistic does not give correct results for agreement analysis since this coefficient is very much affected by the prevalence and bias index. It is determined that when the prevalence indices are high, kappa statistic value decreases; when the prevalence indices decrease, kappa statistic value increases; and also when kappa statistic value is low, bias effect is even higher. It is also suggested that Kappa statistic doesn't properly reflect the agreement between the observers and therefore, AC1 statistic was proposed in 1991 by Gwet as an alternative to Kappa statistic by arguing that AC1 statistic is not affected by sensitivity, specificity and prevalence values and show better performance in comparison with Kappa statistic.[7,15-17,19] Consequently, not only the coefficients of individual agreement but also Kappa statistic and AC1 statistics are calculated in this study in order to show how they are affected by

TABLE 1: Results of Scenario 1 (η_1=0.90; $(1-\eta_0)$=0.90).

	Y observer				Y observer				Y observer				Y observer			
	$(\theta_1=0.90;\ (1-\theta_0)=0.90)$				$(\theta_1=0.80;\ (1-\theta_0)=0.80)$				$(\theta_1=0.50;\ (1-\theta_0)=0.50)$				$(\theta_1=0.40;\ (1-\theta_0)=0.40)$				$(\theta_1=0.30;\ (1-\theta_0)=0.30)$			
Prevalence (ω)	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1
0.00	1	1	0	0.78	0.96	0.69	0	0.76	0.68	0.36	0	0.69	0.57	0.31	0	0.67	0.45	0.27	0	0.64
0.10	1	1	0.39	0.75	0.96	0.69	0.32	0.73	0.68	0.36	0	0.67	0.57	0.31	-0.19	0.66	0.45	0.27	-0.47	0.64
0.20	1	1	0.53	0.71	0.96	0.69	0.46	0.69	0.68	0.36	0	0.66	0.57	0.31	-0.40	0.65	0.45	0.27	-1.32	0.64
0.30	1	1	0.60	0.67	0.96	0.69	0.53	0.67	0.68	0.36	0	0.65	0.57	0.31	-0.60	0.65	0.45	0.27	-2.95	0.64
0.40	1	1	0.63	0.65	0.96	0.69	0.56	0.65	0.68	0.36	0	0.64	0.57	0.31	-0.74	0.64	0.45	0.27	-5.82	0.64
0.50	1	1	0.64	0.64	0.96	0.69	0.57	0.64	0.68	0.36	0	0.64	0.57	0.31	-0.80	0.64	0.45	0.27	-8.0	0.64
0.60	1	1	0.63	0.65	0.96	0.69	0.56	0.65	0.68	0.36	0	0.64	0.57	0.31	-0.74	0.64	0.45	0.27	-5.82	0.64
0.70	1	1	0.60	0.67	0.96	0.69	0.53	0.67	0.68	0.36	0	0.65	0.57	0.31	-0.60	0.65	0.45	0.27	-2.95	0.64
0.80	1	1	0.53	0.71	0.96	0.69	0.46	0.69	0.68	0.36	0	0.66	0.57	0.31	-0.40	0.65	0.45	0.27	-1.32	0.64
0.90	1	1	0.39	0.75	0.96	0.69	0.32	0.73	0.68	0.36	0	0.67	0.57	0.31	-0.19	0.66	0.45	0.27	-0.47	0.64
1.00	1	1	0	0.78	0.96	0.69	0	0.76	0.68	0.36	0	0.69	0.57	0.31	0	0.67	0.45	0.27	0	0.64

ψ^N and ψ^R : Coefficient of individual agreements; κ: Cohen's kappa statistics; AC1: Gwet's AC1 statistics; θ_1: The sensitivity values of Y observer; $(1-\theta_0)$: The specificity values of Y observer.

the sensitivity, specificity and prevalence.

RESULTS

According to Scenario 1, the sensitivity and specificity values of X observer are equal and high (η_1=0.90; $(1-\eta_0)$=0.90) while the sensitivity and specificity values of Y observer are equal and 0.90; 0.80; 0.50; 0.40 and 0.30. In this 5 different combinations, coefficients of individual agreement, Kappa statistic and AC1 statistic values are calculated for different prevalence values (0; 0.10; 0.20; 0.30; 0.40; 0.50; 0.60; 0.70; 0.80; 0.90 and 1), and the results are represented in Table 1.

In 5 different combinations in which the sensitivity and specificity values of X observer are equal and medium (η_1=0.50; $(1-\eta_0)$= 0.50) and the sensitivity and specificity values of Y observer are equal and 0.90; 0.80; 0.50; 0.40 and 0.30, coefficients of individual agreement, kappa statistic and AC1 statistic for different prevalence values (0; 0.10; 0.20; 0.30; 0.40; 0.50; 0.60; 0.70; 0.80; 0.90 ve 1) were calculated and given in Table 2. In 5 different combinations in which the sensitivity and specificity values of X observer are equal and low and the sensitivity and specificity values of Y observer are equal and 0.90; 0.80; 0.50; 0.40 and 0.30, coefficients of individual agreement, kappa statistics and AC1 statistic for different prevalence values (0; 0.10; 0.20; 0.30; 0.40; 0.50; 0.60; 0.70; 0.80; 0.90 ve 1) were calculated and given in Table 3.

TABLE 2: Results of Scenario 1 (η_1=0.50; $(1-\eta_0)$=0.50).

	Y observer $(\theta_1=0.90;\ (1-\theta_0)=0.90)$				Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.80)$				Y observer $(\theta_1=0.50;\ (1-\theta_0)=0.50)$				Y observer $(\theta_1=0.40;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.30;\ (1-\theta_0)=0.30)$			
Prevalence (ω)	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1
0.00	0.68	1	0	0.14	0.82	1	0	0.08	1	1	0	0	0.98	1	0	0	0.92	1	0	0.04
0.10	0.68	1	0	0.09	0.82	1	0	0.05	1	1	0	0	0.98	1	0	0	0.92	1	0	0.02
0.20	0.68	1	0	0.05	0.82	1	0	0.03	1	1	0	0	0.98	1	0	0	0.92	1	0	0.01
0.30	0.68	1	0	0.02	0.82	1	0	0.014	1	1	0	0	0.98	1	0	0	0.92	1	0	0.001
0.40	0.68	1	0	0.006	0.82	1	0	0.004	1	1	0	0	0.98	1	0	0	0.92	1	0	0.01
0.50	0.68	1	0	0	0.82	1	0	0	1	1	0	0	0.98	1	0	0	0.92	1	0	0
0.60	0.68	1	0	0.006	0.82	1	0	0.004	1	1	0	0	0.98	1	0	0	0.92	1	0	0.01
0.70	0.68	1	0	0.02	0.82	1	0	0.014	1	1	0	0	0.98	1	0	0	0.92	1	0	0.001
0.80	0.68	1	0	0.05	0.82	1	0	0.03	1	1	0	0	0.98	1	0	0	0.92	1	0	0.01
0.90	0.68	1	0	0.09	0.82	1	0	0.05	1	1	0	0	0.98	1	0	0	0.92	1	0	0.02
1.00	0.68	1	0	0.14	0.82	1	0	0.08	1	1	0	0	0.98	1	0	0	0.92	1	0	0.04

ψ^N and ψ^R : Coefficient of individual agreements; κ: Cohen's kappa statistics; AC1: Gwet's AC1 statistics; θ_1: The sensitivity values of Y observer; $(1-\theta_0)$: The specificity values of Y observer.

The examination of Table 1 shows that in all combinations where the sensitivity and specificity values of Y observer are equal to each other (0.90; 0.80; 0.70; 0.60; 0.50; 0.40; 0.30; 0.20 ve 0.10), the coefficients of individual agreement are not affected by the prevalence values. However, the agreement between the observers are decreased if the sensitivity and specificity values are decreased and it is always $\Psi^N > \Psi^R$. It is also observed that only if sensitivity and specificity value of Y observer is 0.90, the coefficient of individual agreement is equal to 1, which shows that there is a perfect agreement between these two observers. It is expected that agreement coefficient is equal or almost equal to 1 if both observers precisely distinguish the diseased and the non-diseased groups. However, maximum values for Kappa statistic and AC1 statistic are 0.64 and 0.78 respectively while there is such an agreement between the observers, which is an unexpected situation (Table 1, Figure 1).

Regardless of the prevalence value, when the sensitivity and specificity values are decreased, Kappa statistic value is also decreased dramatically. Especially when the sensitivity and specificity values reduced to 0.50, a case increasing the prevalence value at the range of 0-0.50 and symmetrically decreasing at the range of 0.50-1 takes negative values when the sensitivity and specificity values are 0.40 and lower and prevalence value indicates a case decreasing at the range of 0-0.50, symmetrically increasing at the

TABLE 3: Results of Scenario 1 (η_1=0.30; $(1-\eta_0)$=0.30).

	Y observer $(\theta_1=0.90;\ (1-\theta_0)=0.90)$				Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.80)$				Y observer $(\theta_1=0.50\ (1-\theta_0)=0.50)$				Y observer $(\theta_1=0.40;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.30\ (1-\theta_0)=0.30)$			
Prevalence (ω)	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1
0.00	0.45	0.64	0	0.19	0.60	0.68	0	0.17	0.92	0.84	0	0.19	0.98	0.91	0	0.23	1	1	0	0.28
0.10	0.45	0.64	-0.16	0.18	0.60	0.68	-0.11	0.17	0.92	0.84	0	0.18	0.98	0.91	0.03	0.21	1	1	0.06	0.24
0.20	0.45	0.64	-0.32	0.17	0.60	0.68	-0.22	0.16	0.92	0.84	0	0.17	0.98	0.91	0.06	0.19	1	1	0.11	0.21
0.30	0.45	0.64	-0.47	0.16	0.60	0.68	-0.32	0.16	0.92	0.84	0	0.17	0.98	0.91	0.07	0.17	1	1	0.14	0.18
0.40	0.45	0.64	-0.58	0.16	0.60	0.68	-0.38	0.16	0.92	0.84	0	0.16	0.98	0.91	0.08	0.16	1	1	0.15	0.17
0.50	0.45	0.64	-0.62	0	0.60	0.68	-0.40	0.16	0.92	0.84	0	0.16	0.98	0.91	0.09	0.16	1	1	0.16	0.16
0.60	0.45	0.64	-0.58	0.16	0.60	0.68	-0.38	0.16	0.92	0.84	0	0.16	0.98	0.91	0.08	0.16	1	1	0.15	0.17
0.70	0.45	0.64	-0.47	0.16	0.60	0.68	-0.32	0.16	0.92	0.84	0	0.17	0.98	0.91	0.07	0.17	1	1	0.14	0.18
0.80	0.45	0.64	-0.32	0.17	0.60	0.68	-0.22	0.16	0.92	0.84	0	0.17	0.98	0.91	0.06	0.19	1	1	0.11	0.21
0.90	0.45	0.64	-0.16	0.18	0.60	0.68	-0.11	0.17	0.92	0.84	0	0.18	0.98	0.91	0.03	0.21	1	1	0.06	0.24
1.00	0.45	0.64	0	0.19	0.60	0.68	0	0.17	0.92	0.84	0	0.19	0.98	0.91	0	0.23	1	1	0	0.28

ψ^N and ψ^R : Coefficient of individual agreements; κ: Cohen's kappa statistics; AC1: Gwet's AC1 statistics; θ_i: The sensitivity values of Y observer; $(1-\theta_0)$: The specificity values of Y observer.

range of 0.50-1. Besides, in case that either sensitivity or specificity value of Y observer is 0.50, Kappa statistic has "0" value for all prevalence values (Table 1, Figure 1).

In this scenario, a quite high agreement between the observers are expected in comparison with previous two combinations. In these combinations, Ψ^N coefficient of individual agreement correctly reflects the expected high agreement between the observers by having a higher value than 0.80 that is an acceptable value. Additionally, there is no acceptable value in literature for AC1 statistic and it is expected to be higher than 0.70. It is observed in these combinations that AC1 statistic is not affected by the sensitivity, specificity and prevalence values and it doesn't represent the true values although it is expected to have a higher value than 0.70.

In all these combinations, AC1 statistic interval gets narrow as the sensitivity and specificity values are decreased. It is observed that regardless of the sensitivity, specificity and prevalence values of Y observer, it is minimum 0.64 and maximum 0.78, even in case that sensitivity and specificity values of Y observer is 0.30 and lower, it is 0.64 for all prevalence values. It is observed that AC1 statistic is not affected by the sensitivity, specificity and prevalence values, however it gets a lower value than the true value when it is expected to have a value close to 1 due to the perfect agreement between the observers (Table 1, Figure 1). Considering all these situations, it is concluded that Kappa statistic and AC1 statistics does not provide reliable results but misleading results to the researchers for the agreement studies.

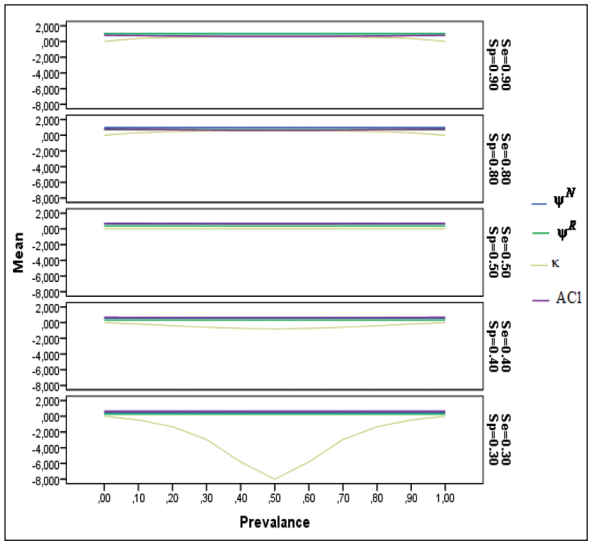

FIGURE 1: Results of Scenario 1 (η_1=0.90; (1−η_0)=0.90).

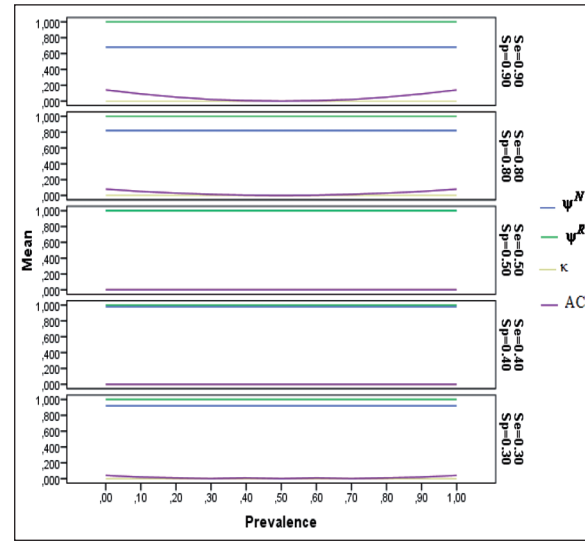

FIGURE 2: Results of Scenario 1 (η_1=0.50; (1−η_0)=0.50).

When Table 2 is examined, that the sensitivity and specificity values of Y observer are equal and are not affected by prevalence values of coefficient of individual agreement in all combinations (0.90; 0.80; 0.70; 0.60; 0.50; 0.40; 0.30; 0.20 and 0.10) and that there is $\Psi^N < \Psi^R$ in all combinations on the contrary to the results in Table 1 is observed. Ψ^R coefficient is coefficient of individual agreement in which X observer is taken as a reference. Therefore, in the case that the sensitivity and specificity values of X observer depends on chance to 0.50, no matter what the sensitivity and specificity values of Y observer are, Ψ^R coefficient was calculated as 1. Ψ^N coefficient indicates a case in which the sensitivity and specificity values of Y observer are increasing at the range from 0.90 to 0.50, and it shows excellent agreement at 0.50 and decreasing symmetrically at the range from 0.50 to 0.10.

When Kappa statistics is examined, since the sensitivity value of X observer is 0.50, it takes the value of "0", no matter what the sensitivity and specificity value of Y observer is. Even in the case in which the sensitivity and specificity values of both the observers are equal, an excellent agreement is expected and it takes the value of "0" for all prevalence values. When AC1 statistics are examined, the sensitivity and specificity of Y observer is at the range from 0.90 to 0.50, it decreasing from 0.14 to 0 is showing, it symmetrically increasing from 0 to 0.14 if it is between 0.40 and 0.10. Besides whichever combination it is, in the case having the prevalence value at 0.50, it is observed to take the value of "0". Besides this, as in kappa statistics in the combination where sensitivity and specificity values of both the observers is 0.50 and 0.40, AC1 statistics value is being calculated as "0" while there is an excellent agreement between the observers (Table 2, Figure 2).

When Table 3 is examined, Ψ^N coefficient of individual agreement, the sensitivity and specificity values of Y observer increases between 0.10 and 0.30, while it is showing a case decreasing from 0.98 to 0.45 is at between 0.40 and 0.90, if the excellent agreement is at 0.30. Ψ^R coefficient of individual agreement is equal to 1 in the case that the sensitivity and specificity values of Y observer is 0.10; 0.20 and 0.30, it decreases from 0.91 to 0.64 if it is between 0.40 and 0.90. Kappa statistics and AC1 statistics take the values very close to "0" in all combination, besides it may be misleading results in the cases where both observers are agreement.

According to the scenario 2, considering two different situations that the sensitivity and specificity values of X observer are high and not equal (η_1=0.90; (1−η_0)=0.80) and low (η_1=0.40; (1−η_0)=0.30), in 5 different combinations where the

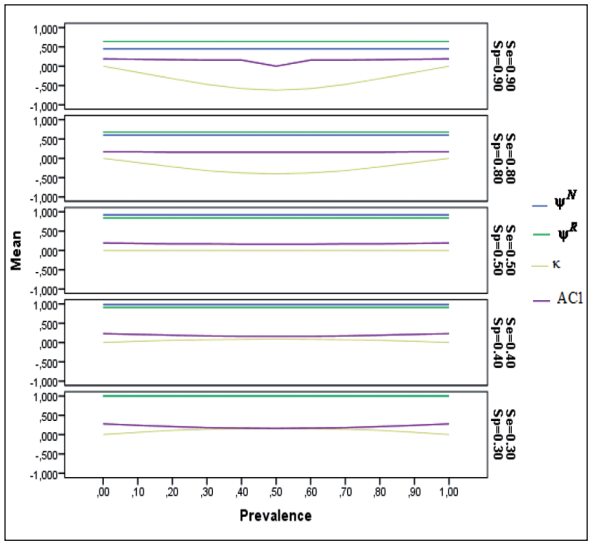

FIGURE 3: Results of Scenario 1 (η_1=0.30; (1–η_0)=0.30).

sensitivity and specificity values of Y observer are high (θ_1=0.80; (1–θ_0)=0.70), the sensitivity value is high and the specificity value is low (θ_1=0.80; (1–θ_0)=0.40), the sensitivity value is low and the specificity value is high (θ_1=0.40; (1–θ_0)=0.80), the sensitivity and specificity values are medium (θ_1=0.50; (1–θ_0)=0.40) and the sensitivity and specificity values is low (θ_1=0.30; (1–θ_0)=0.20), coefficients of individual agreement, Kappa and AC1 statistics results are represented in Table 4 and Table 5 for 11 different prevalence values (0; 0.10; 0.20; 0.30; 0.40; 0.50; 0.60; 0.70; 0.80; 0.90 and 1).

In the first situation where the sensitivity value of X observer is 0.90, specificity value is 0.80, coefficients of individual agreement takes the value of 1 in the case where the sensitivity value of Y observer is 0.90, specificity value is 0.80, while it is taking the value of 0.97 in the case where the sensitivity value of Y observer is 0.80, specificity value is 0.70. In the case where sensitivity value of Y observer is 0.70 and specificity value is 0.60, a decrease from 0.91 to 0.88 can be observed while in the case where sensitivity value of Y observer is 0.60 and specificity value is 0.50, a decrease from 0.82 to 0.79 can be observed. As a conclusion while sensitivity and specificity values of Y observer is getting decreased, a slight decrease in coefficients of individual agreement can be observed (Table 4).

In case that sensitivity is high and specificity is low, prevalence value gets closer to 1 Ψ^N value is even bigger. On the other hand, in case that sensitivity is low and specificity is high, Ψ^N value is decreased as the prevalence value increases (Figure 4). In the case where the sensitivity is 0.80 and the specificity is 0.40, Ψ^N value decreases from 0.71 to 0.96 while prevalence value is getting increased and in the case with the specificity of 0.30, it increases from 0.60 to 0.96, and in the case having specificity of 0.20, it increases from 0.47 to 0.96 and in the case having specificity of 0.10, it increases from 0.34 to 0.96. Despite this, in the value of Ψ^R, an increase from 0.57 to 0.69, from 0.52 to 0.69, from 0.47 to 0.69 and from 0.43 to 0.69 respectively while the value of specificity is getting decreased. It is expected that the acceptable level of agreement between the observers is at least 0.80. However, it is observed that there is no acceptable agreement between the observers for the combination with low value of sensitivity and specificity (Table 4, Figure 4).

The examination of Kappa statistic it is determined that Kappa statistic gets 0 value when prevalence is low (ω = 0) and high (ω = 1) in all combinations and in the first two combinations, it increases when prevalence value is between 0-0.50, it symmetrically decreases when the prevalence is between 0.50-1 and gets the highest value when the prevalence is equal to 0.50. It is also observed that in the third combination with low sensitivity and high specificity, the value of Kappa statistic decreases (having negative value) when the prevalence is between 0-0.50, and it symmetrically increases when the prevalence is between 0.50-1. However, it is on the contrary for the other two combinations. Additionally, it is also observed that regardless of the prevalence value, Kappa statistic gets 0 value in the last combination where the specificity value of one of the observers is 0.50 (Table 4, Figure 4). Besides, in a case where there is an excellent agreement between two observers, Kappa statistics takes a value between 0 and 0.51, AC1 statistics takes a value between 0.49 and 0.68. While the sensitivity and specificity values of Y observer are getting decreased, the results belonging

TABLE 4: Results of Scenario 2 (η_1=0.90; $(1-\eta_0)$=0.80).

	Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.70)$				Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.40;\ (1-\theta_0)=0.80)$				Y observer $(\theta_1=0.50;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.30;\ (1-\theta_0)=0.20)$			
Prevalence (ω)	Ψ^N	Ψ^R	κ	AC1	Ψ^N	Ψ^R	κ	AC1	Ψ^N	Ψ^R	κ	AC1	Ψ^N	Ψ^R	κ	AC1	Ψ^N	Ψ^R	κ	AC1
0.00	0.97	0.84	0	0.65	0.71	0.57	0	0.65	1	1	0	0.52	0.71	0.57	0	0.55	0.47	0.47	0	0.50
0.10	0.97	0.83	0.25	0.60	0.73	0.58	0.25	0.60	0.93	0.88	-0.13	0.51	0.71	0.55	0	0.53	0.47	0.45	-0.28	0.49
0.20	0.97	0.82	0.37	0.56	0.74	0.58	0.37	0.56	0.87	0.79	-0.25	0.50	0.71	0.53	0	0.51	0.47	0.43	-0.65	0.49
0.30	0.97	0.81	0.44	0.52	0.76	0.59	0.44	0.52	0.81	0.70	-0.35	0.49	0.71	0.51	0	0.49	0.47	0.41	-1.07	0.48
0.40	0.97	0.80	0.47	0.49	0.77	0.60	0.47	0.49	0.76	0.62	-0.42	0.48	0.70	0.49	0	0.48	0.46	0.39	-1.44	0.48
0.50	0.97	0.78	0.48	0.48	0.79	0.61	0.48	0.48	0.72	0.56	-0.44	0.48	0.70	0.47	0	0.48	0.46	0.37	-1.6	0.48
0.60	0.97	0.77	0.47	0.49	0.82	0.62	0.47	0.49	0.69	0.50	-0.42	0.48	0.70	0.45	0	0.48	0.46	0.35	-1.44	0.48
0.70	0.97	0.75	0.44	0.52	0.84	0.63	0.44	0.52	0.65	0.44	-0.35	0.49	0.69	0.43	0	0.49	0.46	0.33	-1.07	0.48
0.80	0.97	0.73	0.37	0.56	0.88	0.65	0.37	0.56	0.62	0.39	-0.25	0.50	0.69	0.41	0	0.51	0.46	0.31	-0.65	0.49
0.90	0.96	0.71	0.25	0.60	0.91	0.67	0.25	0.60	0.59	0.35	-0.13	0.51	0.68	0.38	0	0.53	0.46	0.29	-0.28	0.49
1.00	0.96	0.69	0	0.65	0.96	0.69	0	0.65	0.57	0.31	0	0.52	0.68	0.36	0	0.55	0.45	0.27	0	0.50

Ψ^N and Ψ^R : Coefficient of individual agreements; κ: Cohen's kappa statistics; AC1: Gwet's AC1 statistics; θ_1: The sensitivity values of Y observer; $(1-\theta_0)$: The specificity values of Y observer.

to these statistics, a slight decrease can be noticed.

The examination of AC1 statistic shows that AC1 statistic value is calculated around 0.50 when both sensitivity and specificity values of Y observer is low, in other words, when there is 50% possibility of having correct diagnosis for the diseased and non-diseased groups. In the first two combinations with high sensitivity value, high specificity value and high sensitivity value, low specificity value, AC1 statistic results are similar to Kappa statistics results. In the first combination where there is a high capacity for both observers to correctly distinguish the patients and the healthy groups, AC1 value changes between 0.48-0.65 although it is expected to have a high agreement between the observers. It is also observed that it is not affected by the sensitivity and specificity values and regardless of the sensitivity and specificity values, it even gets the same value when the prevalence is equal to 0.50 and it decreases when the prevalence is between 0-0.50 and it symmetrically increases when the prevalence is between 0.50-1 (Table 4, Figure 4).

In the second case where the sensitivity value of X observer is 0.40 and specificity value is 0.30, an increase in coefficient of individual agreement can be observed while the sensitivity and specificity values are getting decreased. When sensitivity is 0.90, specificity is 0.80, Ψ^N value takes the value of nearly 0.60, Ψ^R value is 0.68-0.83, sensitivity is 0.80, specificity is 0.70, Ψ^N value takes the value of 0.72, Ψ^R value is 0.72-0.86, when the sensitivity is 0.70, specificity is 0.60, Ψ^N value takes the value of 0.83, Ψ^R value takes a value at the range of 0.78-0.89. When the sensitivity is 0.60, specificity is 0.50, value of

TABLE 5: Results of Scenario 2 (η_1=0.40; $(1-\eta_0)$=0.30).

	Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.70)$				Y observer $(\theta_1=0.80;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.40;\ (1-\theta_0)=0.80)$				Y observer $(\theta_1=0.50;\ (1-\theta_0)=0.40)$				Y observer $(\theta_1=0.30;\ (1-\theta_0)=0.20)$			
Prevalence (ω)	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1	ψ^N	ψ^R	κ	AC1
0.00	0.72	0.72	0	0.12	0.98	0.91	0	0.12	0.60	0.68	0	0.12	0.98	0.91	0	0.09	0.97	1	0	0.16
0.10	0.72	0.74	-0.05	0.10	0.95	0.91	-0.05	0.10	0.63	0.70	0.02	0.10	0.98	0.92	0	0.09	0.97	1	0.03	0.13
0.20	0.72	0.75	-0.09	0.09	0.92	0.90	-0.09	0.09	0.66	0.73	0.03	0.09	0.98	0.92	0	0.08	0.97	1	0.05	0.11
0.30	0.72	0.76	-0.12	0.09	0.89	0.89	-0.12	0.09	0.70	0.76	0.04	0.09	0.98	0.93	0	0.08	0.98	1	0.07	0.09
0.40	0.72	0.78	-0.14	0.08	0.86	0.89	-0.14	0.08	0.73	0.79	0.04	0.08	0.98	0.93	0	0.08	0.98	1	0.08	0.08
0.50	0.72	0.79	-0.15	0.08	0.83	0.88	-0.15	0.08	0.77	0.82	0.04	0.08	0.98	0.94	0	0.08	0.98	1	0.08	0.08
0.60	0.72	0.80	-0.14	0.08	0.81	0.88	-0.14	0.08	0.81	0.85	0.04	0.08	0.98	0.94	0	0.08	0.98	1	0.08	0.08
0.70	0.72	0.82	-0.12	0.09	0.78	0.87	-0.12	0.09	0.86	0.89	0.04	0.09	0.98	0.95	0	0.08	0.98	1	0.07	0.09
0.80	0.72	0.83	-0.09	0.09	0.76	0.87	-0.09	0.09	0.90	0.92	0.03	0.09	0.98	0.95	0	0.08	0.98	1	0.05	0.11
0.90	0.72	0.84	-0.05	0.10	0.74	0.86	-0.05	0.10	0.95	0.96	0.02	0.10	0.98	0.96	0	0.09	0.98	1	0.03	0.13
1.00	0.72	0.86	0	0.12	0.71	0.86	0	0.12	1	1	0	0.12	0.98	0.96	0	0.09	0.98	1	0	0.16

ψ^N and ψ^R : Coefficient of individual agreements; κ: Cohen's kappa statistics; AC1: Gwet's AC1 statistics; θ_1: The sensitivity values of Y observer; $(1-\theta_0)$: The specificity values of Y observer.

Ψ^N takes the value of 0.92, the value of Ψ^R takes the value of 0.84-0.92, when the sensitivity is 0.50 and specificity is 0.40, the value of Ψ^N takes the value of 0.98 and the value of Ψ^R takes the value of 0.91-0.95 and when sensitivity is 0.40, and specificity is 0.30, both coefficients of individual agreement take the value of 1. In the case where sensitivity is high but specificity is low, Ψ^N and Ψ^R values decrease while prevalence value is accessing to 1, in the case where sensitivity is low, specificity is high; it increases while prevalence is getting increased on the contrary.

Until the sensitivity and specificity values of Y observer reach to the sensitivity and specificity values of X observer, an increase is observed in coefficient of individual agreement. Besides this, in the combinations having equal the sensitivity and specificity values of both observers, both coefficients take the value of 1. When the sensitivity value of Y observer takes the values below 0.40 and the specificity value is below 0.30, the value of Ψ^N decreases according to 1, the value of Ψ^R keeps the value of 1 (Table 5, Figure 5).

When Kappa statistics was examined, it is observed that the prevalence decreases at the range of 0-0.50 and it increases at the range of 0.50-1 symmetrically and takes negative values according to the first two combinations in which the sensitivity value of Y observer and the specificity value are high and low. In other combinations in which the sensitivity is low and the specificity is high and

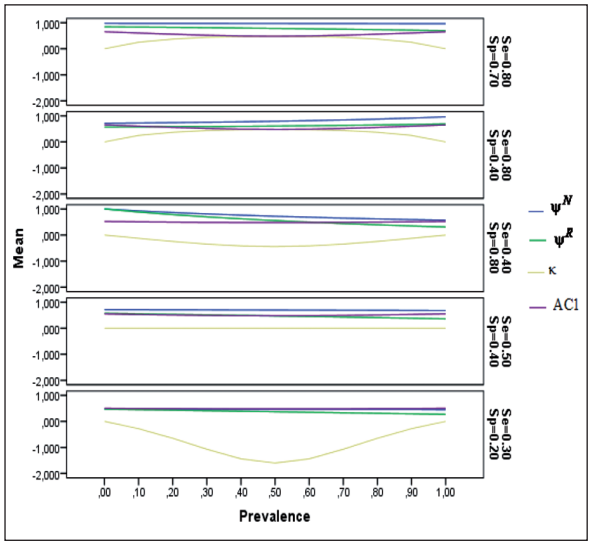

FIGURE 4: Results of Scenario 2 (η_1=0.90; $(1-\eta_0)$=0.80).

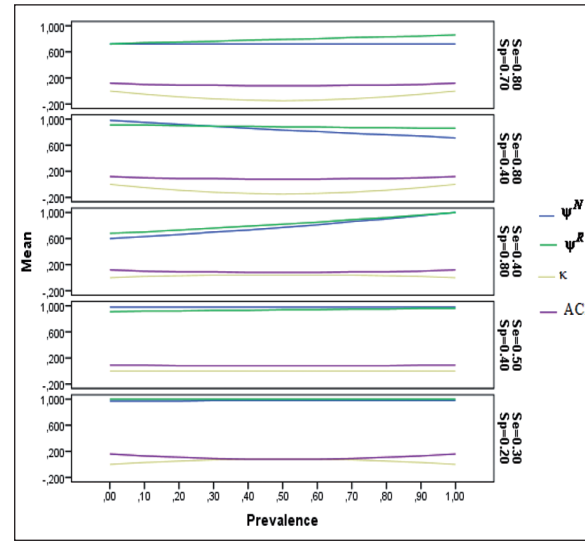

FIGURE 5: Results of Scenario 2 (η_1=0.40; $(1-\eta_0)$=0.30).

sensitivity and specificity are low and sensitivity and specificity are low, it was observed that the prevalence increases at the range of 0-0.50, and decreases at the range of 0.50-1 symmetrically and takes positive values (Table 5, Figure 5). When Table 4 and 5 are examined, it is observed that kappa statistics takes positive values when the sensitivity values of both observers are high/low, and in the case in which one of the observers is high/low and the other one is low/high, the results of kappa statistics were found to be negative.

When the sensitivity and specificity values of X observer are low, the values of AC1 statistics take the value such as at least 0.08 and at the most 0.16 no matter what the sensitivity and specificity values of Y observer. Besides, in the combination having equal sensitivity and specificity values in both observers, AC1 statistics takes at the range of 0.08-0.12, while kappa statistics take the value in the range of 0-0.04. In such a situation, these two agreement statistics researchers give misleading results in the studies examining the agreement between the observers.

DISCUSSION AND CONCLUSION

In Scenario 2, in the second and third combinations with high (low) sensitivity and low (high) specificity values, it is an expected situation that in case of a high sensitivity, the agreement coefficients increases when the prevalence value increases and in cases of low sensitivity, they decreases depending on the increase rate of the prevalence value. In these combinations it is observed that only the coefficients of individual agreement give such results while kappa statistics and AC1 statistics have symmetrical results in all combinations.

It is known by the researchers that Kappa statistic that is commonly used to measure the agreement between the observers is affected by bias and prevalence; it gets 0 value when the prevalence is both high and low; and if there is a high bias situation, it gets higher values than the situations where the bias is low or almost 0.[7,15,17] As a result of our study, if the prevalence is low and high, it takes the value of 0, that the observers are affected by the sensitivity more than specificity values in the case in which the sensitivity value of X observer is high/low, the sensitivity of Y observer is high/low, the prevalence increases at the range of 0-0.50 and decreases of 0.50-1, and in the case in which it is equal to 0.50, it takes the highest value. In conclusion, since kappa statistic is highly affected by these concepts, it is quite difficult to correctly interpret it without considering the prevalence indices and bias and to get a diagnosis by clinicians.

Although AC1 statistic takes its place in the literature since it is not affected by the sensitivity, specificity and prevalence values,[7,19] no matter

what sensitivity, specificity and prevalence value in our study is, it is observed that AC1 statistic value is not yet high while a high agreement rate is expected between the observers in the event that both observers have a high capacity to distinguish the patients and the healthy group. In other words, AC1 statistics do not precisely reflect the true value. Therefore, the results achieved from the agreement studies are possible to be misleading.

At the end of this study, while observing the agreement between the observers in reliability studies including two observers and diagnostic test consists of two categories such as "patients" and "healthy", it is suggested that the researchers should take into account the prevalence and bias concepts and use the coefficients of individual agreement (CIA) since it is not affected by the sensitivity, specificity and prevalence values.

REFERENCES

1. Barnhart HX, Haber MJ, Lin LI. An overview on assessing agreement with continuous measurements. J Biopharm Stat 2007;17(4): 529-69.
2. Barnhart HX, Lokhnygina Y, Kosinski AS, Haber M. Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. J Biopharm Stat 2007;17(4):721-38.
3. Watson PF, Petrie A. Method agreement analysis: a review of correct methodology. Theriogenology 2010;73(9):1167-79.
4. Szmytkowski J, Kapala A, Dabrowiecki S. A comparison of statistical methods for the evaluation of diagnostic tests shown on the example of the methods of blood recirculation measurements in dialysis Access. Polish J Surg 2009;81(4):186-92.
5. Barnhart HX, Song J, Haber MJ. Assessing intra, inter and total agreement with replicated readings. Stat Med 2005;24(9):1371-84.
6. Lin L, Hedayet AS, Wu W. Statistical Tools for Measuring Agreement. 1sted. New York: Springer; 2012. p.1-109.
7. Kanık EA, Erdoğan S, Orekici Temel G. [Agreement statistics impacts of prevalence between the two clinicians in binary diagnostic tests]. İnönü Üniversitesi Tıp Fakültesi Dergisi 2012;19(3):153-8.
8. Kanık EA, Orekici Temel G, Ersöz Kaya I. [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.] Turkiye Klinikleri J Biostat 2010;2(2):74-81.
9. Haber M., Barnhart HX. A general approach to evaluating agreement between two observers or methods of measurement from quantitative data with replicated measurement. Stat Methods Med Res 2008;17(2):151-69.
10. Haber M, Gao J, Barnhart HX. Assessing observer agreement in studies involving replicated binary observations. J Biopharm Stat 2007;17(4):757-66.
11. Pan Y, Gao J, Haber M, Barnhart HX. Estimation of coefficients of individual agreement (CIA's) for quantitative and binary data using SAS and R. Comput Methods Programs Biomed 2010; 98(2): 214-9.
12. Gao J, Pan Y, Haber M. Assessment of observer agreement for matched repeated binary measurements. Computational Statistics and Data Analysis 2012;56(5):1052-60.
13. David AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. Applied Statist 1979;28(1):20-8.
14. Pan Y, Haber M, Barnhart HX. A new permutation-based methods for assessing agreement between two observers making replicated binary readings. Statistics in Medicine 2011;30(8):839-53.
15. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation and sample size requirements. Phys Ther 2005;85(3):257-68.
16. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Mathem Stat Psychol 2008;61(Pt 1):29-48.
17. Gwet K. Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. Series: Statistical Methods for Inter-Rater Reliability Assessment 2002;2:1-9.
18. Haley DT, Thomas P, Petre M, Roeck AD. Using a new inter-rater reliability statistics. Technl Rep 2008;15:14-23.
19. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol 2008;61(1):29-48.