On the Relationship between the Reliability and Accuracy of Bio-Behavioral Diagnoses: Simple Math to the Rescue

Dom Cicchetti*

Department of Biometry, Yale University School of Medicine, New Haven, CT 06520, USA

Abstract: An equivalence between the J statistic (Jack Youden, 1950) and the Kappa statistic (K), Cohen (1960), was discovered by Helena Kraemer (1982). J is defined as: [Sensitivity (Se) + Specificity (Sp)] – 1. The author (2011) added the remaining two validity components to the J Index, namely, Predicted Positive Accuracy (PPA) and Predicted Negative Accuracy (PNA). The resulting D Index or D = [(Se + Sp) + (PPA + PNA) – 1] / 2. The purpose of this research is to compare J and D as estimates of K, using both actual and simulated data sets. The actual data consisted of ratings of clinical depression and self-reports of gonorrhea. The simulated data sets represented binary diagnoses when the percentages of Negative and Positive cases were: (Identical; Slightly varying; Mildly varying; Moderately varying; or Markedly varying diagnostic patterns, For both the diagnosis of clinical depression, and the self-reports of gonorrhea, D produced closer approximations to Kappa. For the simulated data, under both identical and slightly different patterns of assigning Negative and Positive binary diagnoses, K, D and J produced identical results. While J produced acceptably close values to K under the condition of Mild discrepancies in the proportions of Negative and Positive cases, D continued to more closely approximate K. While D more closely estimated K under Markedly varying diagnostic patterns, D produced values under this extreme condition that were closer than would have been predicted. The significance of these findings for future research is discussed.

Keywords: Binary Diagnoses, Diagnostic Reliability, Diagnostic Accuracy.

INTRODUCTION

In 1982, Helena Kraemer [1] applied her keen eye and creative abilities to discover that a statistic published a decade before Cohen's Kappa, bore a mathematical relationship to Kappa. The statistic was named J after the first initial of the first name of the chemist/biostatistician Jack Youden. The statistic was published in a premier medical journal, *Cancer*, and is defined, simply as:

J= [(Sensitivity (Se) + Specificity (Sp))]-1 [2]. Kraemer discovered and reported her finding that when both the judge and the binary criterion are based upon the same or a similar percentage of positive and negative cases, then J=Kappa.

Recently, the author revisited and revised Youden's J statistic, by adding to it the value of both Predicted Positive Accuracy (PPA) and Predicted Negative Accuracy (PNA)(Cicchetti, 2011) [3]. It is called the Dom or D index, defined as [(Se + Sp)-1+ (PPA + PNA)-1] /2.

To review briefly: Se refers to the probability that the test response will be positive whenever the disease is present; conversely, Sp signals the probability that the test result is negative when the disease is absent; PPA signifies the probability that when the test is positive, the disease is present; and, conversely, PPN refers to the probability that when the test is negative, the disease is absent. These four measures of diagnostic accuracy can be derived from the following 2 x 2 contingency table, whereby a rater's positive and negative diagnoses, expressed in proportions, are compared to an experienced clinician's diagnoses, that are being used as a proxy gold standard:

It seems appropriate at this time to briefly review the formulae for K, D, and J, as they will be utilized in this report.

K, J, and D Defined:

Kappa (K), as introduced by Jacob Cohen, in 1960 [10], is defined as (PO-PC)/(1-PC), where:

PO refers to the Proportion of agreement Observed; PC is the Proportion of agreement expected by Chance alone; and (1-PC) refers to the maximum amount that PO can exceed chance expectations. When PO exceeds PC, K is positive; when PO equals PC, K=0; and when PC exceeds PO, K becomes negative. Following the recommendations of Cicchetti & Sparrow (1981), [15], values below 0.40 are typically considered to be of poor quality; those between 0.40 and 0.59 are taken to represent fair agreement; coefficients between 0.60 and 0.74 represent good agreement, and those between 0.75 and 1.00 are viewed as excellent. These are conceptually similar to the earlier guidelines of Landis & Koch (1977) [16], whereby: <0.00= poor; 0.00-0.20=slight; 0.21-0.40=fair; 0.41-0.60=moderate;

*Address correspondence to this author at the Department of Biometry, Yale University School of Medicine, New Haven, CT 06520, USA; Tel: 1 203 488 6563; Fax: 1 203 488 4218; E-mail: dom.cicchetti@yale.edu

E-ISSN: 1929-6029/15

0.61-0.80=substantial; and 0.81-1.00=almost perfect [16].

"Gold Standard" Diagnosis

Rater A:	(+)	(-)	Totals:
(+)	A	B	A + B
(-)	C	D	C + D
Totals:	A + C	B + D	1.00

OA= A + D PPA= A/ (A + B)

Se= A/ (A + C) PNA- D/ (C + D)

Sp= D/ (B+D)

Kappa (K) is defined, in proportion, as: K= (PO-PC)/ (1-PC), where:

PO= A + D

PC= [(A + C) X (A + B) + (B + D) X (C + D)]

Youden's J is defined, simply, as: J= [(Se + Sp) -1] and

Dom's D is defined as: D= [(Se + Sp) -1 + (PPA + PNA) -1]/2

In the broader context of establishing the accuracy of a given diagnostic decision, several possible strategies can and will be utilized, depending upon the specific bio-behavioral specialty area. In the ideal case, such as the measurement of the viscosity of fluids, like water or blood, gold standard viscometers are readily available. This enables the assessments of both reliability and validity, or accuracy of measurement (e.g., Schimmel, Kaplan, & Soll [4]). As Feinstein noted in 1987, p. 192 [5]: "…when a clinician uses diagnostic criteria to make a clinical diagnosis of coronary artery disease, we can check the accuracy of the diagnosis against the more definitive findings noted at coronary angiography, surgery, or necropsy." Feinstein (ibid) goes on to state that for other medical diseases, exemplified by rheumatoid arthritis, systemic lupus erythematosus, or rheumatic fever, the criteria themselves act as the definitive standard, simply because no external standard is available to either confirm or refute the accuracy of the diagnostic decision. And finally, in other areas of medical and behavioral science, no standards for measuring accuracy are available, and one is left solely with the task of establishing the reliability of the diagnostic decision as in Fleiss, Levin, & Cho Paik [6]. In situations of this ilk, one is forced to rely upon the reliability of experienced diagnosticians.

OBJECTIVE

The objective of this research is to describe and contrast the two methods (J and D) of comparing the reliability and accuracy/validity of binary diagnoses, such as the presence or absence of arrhythmia in a patient suffering from cardiomyopathy. As correctly noted, more than a decade ago, in the final analysis, a given clinical diagnosis will be expressed as a binary decision. In one's best clinical judgment, does the patient have the disease-yes or no? (Kraemer, Kazdin, Offord, Kessler, Jensen, & Kupfer (1997) [7]. Both actual and simulated data sets will be utilized to accomplish the stated objective.

The simulated data sets that will test the relationship between the reliability and validity/accuracy of binary diagnoses, under controlled conditions, will vary as to the percentages of simulated negative and positive cases, as follows: the percentage of positive and negative cases are the same or similar for each of the two binary conditions; the percentages differ mildly; they are moderately different; or the percentages of positive and negative cases differ markedly for each of the two simulated clinical examiners. Each data set begins with the maximum Proportion of Observed agreement (PO) that is possible within the constraints of the two rater marginals. Each successive data set will decrease incrementally, reaching its "lowest" point when PO and the Proportion of Chance agreement (PC) are at or as close to equal as the rater marginal will permit. This design allows the researcher to examine a full range of patterns of rater marginals and their respective effects upon the values of K, D, and J under controlled conditions, that are not, of course, possible when studying isolated examples of binary diagnostic reliability and accuracy deriving from clinical data. Since PC must, per force, also vary as a function of the pattern of the rater marginals, a wide range of expected levels of agreement will also manifest itself.

RESEARCH DESIGN

The first component of the research design consists of two binary diagnostic areas investigated by the author and clinical research colleagues, namely: (1) Clinical depression (Nelson & Cicchetti, 1991) [8]; and (2) Teen-aged women's self-reports of the presence or absence of sexually transmitted diseases (Niccolai, Kershaw, Lewis, Cicchetti, Ethier, & Ickovics, 2005) [9].

The second leg of the design consists of simulated binary data sets designed to test the effects of varying percentages of positive and negative cases, for each pair of simulated raters, namely, when the percentages of negative and positive cases is IDENTICAL Here each simulated rater diagnoses 80% of the cases as

negative and 20% as positive; SLIGHTLY DIFFERENT. Here the first rater diagnoses 46% of the simulated cases as negative and the remaining 54% as positive, while the corresponding rater diagnoses her cases as 49% negative and 51% positive (a difference of 3%); MILDLY DIFFERENT: Here the first rater's diagnostic pattern is 85% negative cases and 15% positive cases; while the second rater's corresponding diagnostic figures are 90% negative cases and 10% positive cases (a difference of 5%); MODERATELY DIFFERENT: Rater 1's diagnostic distribution is 80% negative and 20% positive, and the corresponding rater 2's pattern is 90% negative and 10% positive (a difference of 10%);and finally diagnostic patterns that are MARKEDLY DIFFERENT, or 65% negative and 35% positive, for the first simulated rater, and 45% negative and 55% positive for the second one (a difference of 20%).

The statistics that will be applied are the Kappa (K) statistic (Cohen 1960 [10]; Fleiss, Cohen, & Everrit, 1969 [11]; Cicchetti & Fleiss, 1977 [12]; Cicchetti, 1981) [13], to test for diagnostic reliability; and the standard model for establishing diagnostic validity: Combined over negative and positive cases (Overall Diagnostic Accuracy); Sensitivity, Specificity, Predicted Positive Accuracy and Predicted Negative Accuracy.

Each simulated 2 X 2 contingency table data set begins with the maximum Proportion of Observed agreement (PO) that is possible within the constraints of the two rater marginals. Each successive data set decreases incrementally, reaching its "lowest" point when PO and the Proportion of Chance agreement (PC) are at or as close to equal as the rater marginals will permit. This design allows the researcher to examine a full range of patterns of rater marginals and their respective effects upon the values of K, D, and J under controlled conditions that are not, of course, possible when studying isolated clinical examples of binary diagnostic reliability and accuracy Since PC must, per force, also vary as a function of the pattern of the rater marginal, a wide range of expected levels of agreement will also manifest itself.

Two examples are given below to illustrate how this would play out for both the highest and that level of rater agreement when PO is equal or nearly equal to PC. Here the most challenging set of rater marginals (most dissimilar) were, as previously stated 65% negative and 35% positive for the first rater and 45% negative and 55% positive for the second rater.

Rater 2:	(-)	(+)	**Totals:**	Rater 2:	(-)	(+)	**Totals:**
(-)	45	0	**45**	(-)	29	16	**45**
(+)	20	35	**55**	(+)	36	19	**55**
Totals:	**65**	**35**	**100**	**Totals:**	**65**	**35**	**100**
PO=0.80				PO=0.48			

PC= .485 or [(.65 x .45) + (.35 x .55)]; and PC= .485 or [(.65 x .45) + (.35 x .55)]; and K= (PO-PC)/(1-PC)= 0.61; K= (PO-PC)/(1-PC)= -0.01.

RESULTS AND DISCUSSION

The relationship between the reliability and accuracy of the diagnosis of clinical depression was investigated in an earlier study by Nelson & Cicchetti (1991). More specifically, the authors tested the accuracy of the venerable Minnesota Multiphasic Psychological Inventory (MMPI) to diagnose the presence or absence of clinical depression. As given in Table **1**: The Overall diagnostic accuracy was 77%. By the suggested criteria of Cicchetti, Volkmar, Klin, & Showalter [14], this represents a Fair or Average level of agreement. The Sensitivity and Specificity of the diagnosis were at similar levels of consensus at respective values of 78% and 75%. Predicted Positive Accuracy was Excellent, at a value of 93%. However, Predicted Negative Accuracy was very poor at a low of only 43%. For predicting levels of the reliability between the MMPI and the clinical diagnosis of depression, Kappa (K) produced a value of 0.41; D was similar at 0.44; but J was appreciably higher at 0.53. These findings are reproduced in Table **1**. Referring to an earlier part of this report, this perhaps somewhat unexpected finding underscores the necessity of obtaining PNA and PPA, in addition to the more widely utilized validity indices of Se and Sp.

The second clinical application concerns the accuracy of teen-aged females' reports of the diagnosis of the sexually transmitted disease (STD), gonorrhea, as reported by Niccolai, Kershaw, Lewis, Cicchetti, Ethier, & Ickovics (2005). Overall Accuracy was Excellent, at 90%; as were Specificity, at 97%; Predicted Positive Accuracy at 91%, Predicted Negative Accuracy at 90%. In distinct contrast, Sensitivity was Poor, at only 69%. In terms of the reliability of diagnosis, K was Good at 0.72, followed very closely by D at 0.73. Once again, J was farther apart than K, at a value of 0.66. Each of these reliability estimates is considered Good, by the criteria of Cicchetti & Sparrow [15]; and Substantial, by the earlier criteria of Landis & Koch [16].

As shown in Tables **3-7**, when the diagnostic distributions of positive and negative cases were the

Table 1: Reliability and Accuracy of Diagnosing Clinical Depression

	CLINICAL SIGNIFICANCE:
RELIABILITY INDEX:	Cicchetti & Sparrow (1981)
Kappa (K) (Cohen, 1960) = 0.41	Fair/Average
Dom Index (D) (Cicchetti, 2011) = .43	Fair/Average
J Index (J) (Youden, 1950) = 0.53	Fair/Average
	CLINICAL SIGNIFICANCE:
DIAGNOSTIC INDEX:	Cicchetti, Volkmar, Klin, & Showalter (1995)
Overall Accuracy = 77%	Fair/Average
Sensitivity = 78%	Fair/Average
Specificity = 75%	Fair/Average
Predicted Positive Accuracy = 93%	Excellent/Superior
Predicted Negative Accuracy = 43%	Poor/Below Average

Depression/MMPI	*Beck Depression Inventory/Clinician Diagnosis*		
	Yes	No	TOTALS
Yes	55	4	59
No	16	12	28
TOTALS:	71	16	87

Table 2: Reliability and Accuracy of Female Teens' Self Reports of the Diagnosis of Gonorrhea

	CLINICAL SIGNIFICANCE:
RELIABILITY INDEX:	Cicchetti & Sparrow (1981)
Kappa (K) (Cohen, 1960) = 0.72	Good/Above Average
Dom Index (D) (Cicchetti, 2011) = .73	Good/Above Average
J Index (J) (Youden, 1950) = 0.66	Good/Above Average
	CLINICAL SIGNIFICANCE:
DIAGNOSTIC INDEX:	Cicchetti, Volkmar, Klin, & Showalter (1995)
Overall Accuracy = 90%	Excellent/Superior
Sensitivity = 69%	Poor/Below Average
Specificity = 97%	Excellent/Superior
Predicted Positive Accuracy = 91%	Excellent/Superior
Predicted Negative Accuracy = 90 %	Excellent/Superior

Self Report	Composite Reference Standard		
	Yes	No	TOTALS
Yes	67	7	74
No	30	257	287
TOTALS:	97	264	361

Table 3: Reliability[1] and Accuracy of Simulated Binary Diagnoses when Both Raters Diagnose 80% of the Cases as Negative and 20% as Positive: Identical Marginals

Overall Accuracy	Sensitivity	Specificity	Predicted Positive Accuracy	Predicted Negative Accuracy	K=D=J
1.00	1.00	1.00	1.00	1.00	1.00
.98	.95	.99	.95	.99	.94
.96	.90	.98	.90	.98	.88
.94	.85	.96	.85	.96	.81
.92	.80	.95	.80	.95	.75
.90	.75	.94	.75	.94	.69
.88	.70	.93	.70	.93	.63
.86	.65	.91	.65	.91	.56
.84	.60	.90	.60	.90	.50
.82	.55	.89	.55	.89	.44
.80	.50	.88	.50	.88	.38
.78	.45	.86	.45	.86	.31
.76	.40	.85	.40	.85	.25
.74	.35	.84	.35	.84	.19
.72	.30	.83	.30	.83	.13
.70	.25	.81	.25	.81	.06
.68	.20	.80	.20	.80	.00

[1]The level of inter-rater agreement expected by chance alone is held constant across each simulated case as 0.68.

Table 4: Reliability[1] and Accuracy of Simulated Binary Diagnoses when Rater 1 Diagnoses 46% of the Cases as Negative and 54% as Positive and Rater 2 Diagnoses the Same Cases as 49% Negative and 51% Positive: Slightly Different Marginals

Overall Accuracy	Sensitivity	Specificity	Predicted Positive Accuracy	Predicted Negative Accuracy	Kappa=D Index=J Index
.97	.94	1.00	1.00	.94	.94
.95	.93	.98	.98	.92	.90
.93	.91	.96	.96	.90	.86
.91	.89	.94	.94	.88	.82
.89	.87	.91	.92	.86	.78
.87	.85	.89	.90	.84	.74
.85	.83	.87	.88	.82	.70
.83	.81	.85	.86	.80	.66
.81	.80	.83	.84	.78	.62
.79	.78	.80	.82	.76	.58
.77	.76	.78	.80	.73	.54
.75	.74	.76	.78	.71	.50
.73	.72	.74	.76	.69	.46
.71	.70	.72	.75	.67	.42
.69	.69	.70	.73	.65	.38
.67	.67	.67	.71	.63	.34
.65	.65	.65	.69	.61	.30
.63	.63	.63	.67	.59	.26
.61	.61	.61	.65	.57	.22
.59	.59	.59	.63	.55	.18
.57	.57	.57	.61	.53	.14
.55	.56	.54	.59	.51	.10
.53	.54	.52	.57	.49	.06
.51	.52	.50	.55	.47	.02
.49	.50	.48	.53	.45	-.02

[1]The level of inter-rater agreement expected by chance alone is held constant across each simulated case as 0.50.

Table 5: Reliability[1] and Accuracy of Simulated Binary Diagnoses when Rater 1 Diagnoses 85% of the Cases as Negative and 15% as Positive and Rater 2 Diagnoses the Same Cases as 90% Negative and 10% Positive: Mildly Different Marginals

Overall Accuracy	Sensitivity	Specificity	Predicted Positive Accuracy	Predicted Negative Accuracy	Kappa	D Index	J Index
.95	.67	1.00	1.00	.94	.77	.81	.67
.93	.60	.99	.90	.93	.68	.71	.59
.91	.53	.98	.80	.92	.59	.62	.51
.89	.47	.96	.70	.91	.50	.52	.43
.87	.40	.95	.60	.90	.41	.42	.35
.85	.33	.94	.50	.89	.32	.33	.27
.83	.27	.93	.40	.88	.23	.24	.20
.81	.20	.92	.30	.87	.14	.14	.12
.79	.13	.91	.20	.86	.05	.05	.04
.77	.07	.89	.10	.84	-.05	-.05	-.04

[1]The level of inter-rater agreement expected by chance alone is held constant across each simulated case as 0.78.

Table 6: Reliability[1] and Accuracy of Simulated Binary Diagnoses when Rater 1 Diagnoses 80% of the Cases as Negative and 20% as Positive and Rater 2 Diagnoses the Same Cases as 90% Negative and 10% Positive: Moderately Different Marginals

Percent Overall Accuracy	Sensitivity	Specificity	Predicted Positive Accuracy	Predicted Negative Accuracy	Kappa	D Index	J Index
.90	.50	1.00	1.00	.89	.62	.69	.50
.88	.45	.99	.90	.88	.54	61	.44
.86	.40	.98	.80	.87	.46	.53	.38
.84	.35	.96	.70	.86	.38	.43	.31
.82	.30	.95	.60	.84	.31	.35	.25
.80	.25	.94	.50	.83	.23	.26	.19
.78	.20	.93	.40	.82	.15	.17	.13
.76	.15	.91	.30	.81	.08	.09	.06
.74	.10	.90	.20	.80	.00	.00	.00

[1]The level of inter-rater agreement expected by chance alone is held constant across each simulated case as 0.74.

same or similar for each dichotomous/binary outcome, both J and D were equal to K, as expected. However, the D statistic was consistently closer to Kappa than was J, under both moderate and markedly different distributions of positive and negative cases. This latter diagnostic pattern, when the rater marginal are markedly different, is the condition that holds the most interest from a clinical research stand point. This is because it would be expected to produce the most stringent test of the relative similarity of D and J to K.

In earlier publications, it was hypothesized that under the same or very similar diagnostic patterns, K, J, and D would produce very similar values (Cicchetti, 2011; Kraemer, 1982). This hypothesis has been confirmed. However, what was not expected and therefore not hypothesized was that under the condition of markedly different diagnostic patterns, the same equivalencies would occur. In fact, the D statistic, rather surprisingly, manifested a remarkably close relationship with Kappa. J, in contrast, was more and consistently different than K when the diagnostic patterns or rater marginals were markedly different.

In conclusion, the following caveat seems apt. Subsequent to a successful assessment of diagnostic

Table 7: Reliability[1] and Accuracy of Simulated Binary Diagnoses when Rater 1 Diagnoses 65% of the Cases as Negative and 35% as Positive and Rater 2 Diagnoses the Same Cases as 45% Negative and 55% Positive: Markedly Different Marginals

Overall Accuracy	Sensitivity	Specificity	Predicted Positive Accuracy	Predicted Negative Accuracy	Kappa	D Index	J Index
.80	1.00	.69	.64	1.00	.61	.66	.69
.78	.97	.68	.62	.98	.57	.63	.65
.76	.94	.66	.60	.96	.53	.58	.60
.74	.91	.65	.58	.93	.50	.54	.56
.72	.89	.63	.56	.91	.46	.50	.52
.70	.86	.62	.55	.89	.42	.45	.47
.68	.83	.60	.53	.87	.38	.41	.43
.66	.80	.58	.51	.84	.34	.37	.39
.64	.77	.57	.49	.82	.30	.33	.34
.62	.74	.55	.47	.80	.26	.29	.30
.60	.71	.54	.45	.78	.22	.24	.25
.58	.69	.52	.44	.76	.18	.20	.21
.56	.66	.51	.42	.73	.15	.16	.16
.54	.63	.49	.40	.71	.11	.12	.12
.52	.60	.48	.38	.69	.07	.07	.08
.50	.57	.46	.36	.67	.03	.03	.03
.48	.54	.45	.35	.64	-.01	-.01	-.01

[1]The level of inter-rater agreement expected by chance alone is held constant across each simulated case as 0.485.

accuracy, the clinical research scientist is advised to select K as the gold standard reliability statistic of choice, while realizing that D will provide an acceptable proxy under most nosologic clinical research conditions. It should finally be noted that whenever Se, Sp PPA, and PNA are given, in the *absence* of an accompanying 2 X 2 contingency table, as is often the case, D will more closely approximate K than will J.

And, finally, it should be stressed that even overall diagnostic accuracies that are excellent, as in the case of the accuracy of self-reports of gonorrhea, specific and important components of diagnostic accuracy, such as Sensitivity, can be quite poor and unacceptable (here a value of only 69%). In fact, to further stress the point being made here, Specificity, Predicted Positive Accuracy and Predicted Negative Accuracy were also at a level of excellent diagnostic accuracy. This phenomenon is discussed in further detail in a diagnostic context in which multiple clinical examiners assessed the reliability of PTSD symptomatology in a Vietnam era veteran, published recently in this Journal [17].

REFERENCES

[1] Kraemer HC. Estimating false alarms and missed events from interobserver agreement: Comment on Kaye. Psychol Bull 1982; 92: 749-754. http://dx.doi.org/10.1037/0033-2909.92.3.749

[2] Youden WJ. J Index for rating diagnostic tests. Cancer 1950; 3: 32-35. http://dx.doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3

[3] Cicchetti DV. On the reliability and accuracy of the Evaluative Method for identifying evidence-based practices in Autism. In: Reichow B, Doehring P, Cicchetti DV, Volkmar F, Eds. Evidence-based practices and treatments for children with Autism. New York, NY: Springer, 2011; pp. 41-51. http://dx.doi.org/10.1007/978-1-4419-6975-0_3

[4] Schimmel MS, Kaplan M, Soll Rf. Blood transfusion in the neonate- Where are we today? In: Peterson BR, Ed. New developments in blood transfusion research. New York, NY: Nova Science 2006; pp. 1-15.

[5] Feinstein AR. Clinimetrics. New Haven CT: Yale University Press, 1987.

[6] Fleiss JL, Levin B, Cho Paik M. Statistical methods for rates and proportions. New York, NY: Wiley, 2003. http://dx.doi.org/10.1002/0471445428

[7] Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Coming to terms with the terms of risk. Arch Gen Psychiat 1982; 54: 337-343. http://dx.doi.org/10.1001/archpsyc.1997.01830160065009

[8] Nelson L, Cicchetti DV. Validity of the MMPI Depression Scale for outpatients. Psychol Assess 1991; 3: 55-59. http://dx.doi.org/10.1037/1040-3590.3.1.55

[9] Niccolai LM, Kershaw TS, Lewis JB, Cicchetti DV, Ethier KA, Ickovics J. Data collection for sexually transmitted disease diagnoses: A comparison of self-reports, medical record reviews, and state health department reports. Annals Epidemiol 2005; 15: 236-242. http://dx.doi.org/10.1016/j.annepidem.2004.07.093

[10] Cohen J. A coefficient of agreementfor nominal scales. Educ Psychol Meas 1960; 23: 37-46. http://dx.doi.org/10.1177/001316446002000104

[11] Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. Psychol Bull 1969; 72: 323-327. http://dx.doi.org/10.1037/h0028106

[12] Cicchetti DV, Fleiss JL. Comparison of the null distributions of kappa and the C ordinal statistic. Applied Psychol Meas 1977; 1: 195-201. http://dx.doi.org/10.1177/014662167700100206

[13] Cicchetti DV. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. Applied Psychol Meas 1981; 5: 101-104. http://dx.doi.org/10.1177/014662168100500114

[14] Cicchetti DV, Volkmar F, Klin A, Showalter D. Diagnosing Autism using ICD-10 criteria: A comparison of neural networks and standard multivariate procedures. Child Neuropsychol 1995; 1: 26-37. http://dx.doi.org/10.1080/09297049508401340

[15] Cicchetti DV, Sparrow SS. Developing criteria for establishing interrater reliability of specific items: Applications to assessments of adaptive behavior. Amer J Mental Deficiency 1981; 86: 127-137.

[16] Landis JR, Koch GG. The measure of observer agreement for categorical data. Biometrics 1977; 33: 159-174. http://dx.doi.org/10.2307/2529310

[17] Cicchetti DV, Fontana A, Showalter D. Establishing reliability when multiple examiners evaluate a single case- Part II: Applications to symptoms of Post-Traumatic Stress Disorder (PTSD). Internat J Stat Med Research 2014; 3.

Received on 16-02-2015 Accepted on 14-04-2015 Published on 21-05-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.02.2