Current Research in Biostatistics

Original Research Paper

Statistical Tool for Testing Agreement Level on Continuous Datasets

[1]Basavarajaiah Mariyappa Doddagangavadi, [2]B. Narasimha Murthy and [1]Netra Rajpurohit

[1]Karnataka Veterinary Animal and Fisheries Sciences University (B), India
[2]National Health Mission, Government of India, India

Article history
Received: 04-09-2020
Revised: 17-12-2020
Accepted: 28-12-2020

Corresponding Author:
Basavarajaiah Mariyappa Doddagangavadi
Karnataka Veterinary Animal and Fisheries Sciences University (B), India
Email: sayadri@gmail.com

Abstract: Various analytical studies explored the new innovation for testing agreement level in medical and life sciences which can be simulated by Cohen 'κ' based on the practical applications. From the past medical literature, many authors suggested that, there is some disproportion research gaps that exists in the statistical methods for measuring the agreements between two or more observers from 'Cohen 'κ', these methods had some salient properties and analytical characteristics on qualitative data for testing the research hypothetical statements. The '*k*' can be simulated based on few parameters, which can be estimated from the observed data sets at one point of time (*t*). This intervention will be restricted for the experimenter on measuring and comparing the extent of various agreements at varied time intervals *t'*. In this drawback, the present research article attempts to focus on the testing agreement level based on real values by using various mathematical iterations like bootstrap and Thompson (the measurement made on central tendency method). Since above cited methods extrapolate prediction values and Standard Errors (SE) on various agreements with continuous data scale. As per the model results, our formulated model will be able to measure and compare various parameters of our interest, we can also estimate various parameters from agreements between two or more observers by using ranking scale (converted in to random scale of measurement) at the same population in varied time interval '*t'*. The research findings clearly depend heavily on the exact distribution of Binomial and Poisson distribution with same dicatamous classification of the disease conditions. Results of bootstrap technique are more epoch rather than '*k*' and it will provide a very good consistent prediction of different observer agreement level without any biased scale. This model also demonstrated how to examine various kinds of distributions at population level. Importantly, the driven model will explore fiducial limits of the parameters on the basis of agreement drawn with different time intervals t_1, t_2.. t_n (by using real-life datasets from anti-leprosy vaccine trial conducted in south India). We found that, the model results would be coaxial changes between various factors viz (*i*) distribution of the sample estimators is non-Gaussianity, (ii) variance is underestimated and the confidence limits are asymmetric normally distributed data sets.

Keywords: Bootstrap, Kappa, Agreements, Estimates, Parameters

Introduction

In practical sight we determine the prevalence of diseases such as leprosy, tuberculosis and disease investigations that often depend on the results from population screening or new diagnostic tests (Ayoub and Elgammal, 2018; Bakeman and Quera, 2011). Although, none of these methods or procedures can be considered as perfect (Byrt *et al*., 1993), the classification of such results associated from quantum of response variables, classified into two categories namely positive or negative, normal or abnormal etc. (Brennan and Prediger, 1981; Carpentier *et al*.,

2017; Cohen, 1960). Medical research conducted at commonplace at varied time intervals (Cohen, 1968; Feinstein and Cicchetti, 1990). Since, the accuracy of result is undoubtedly important for testing resulted findings and also improving accuracy of the results and standardization of the methodology between any pair of observers which is being usually adopted (Field and Welsh, 2007). This can be evaluated by computing the usual measure of agreement namely concordance or crude agreement (p_o) (Fleiss and Cohen, 1973; Feinstein and Cicchetti, 1990; Hoehler, 2000). However, the well-known measure of agreement Kappa (κ) that would take care of chance of agreements When the rating scale will be multi-categorized (as opposed to the binary positive' or negative'), the weighted 'κ' simulated accurate results by assigning some Weightage to particular parameters (each cell in the $n \times n$ Table 1) (Bakeman and Quera, 2011; Byrt *et al*., 1993). Discussed the different measures of inter observer agreements and their desirable properties of '*k*' at defined time interval '*t*'. While, appraising the use of 'κ' a long list of literature is available relevant to the observation of paradoxes' for its long ranging interpretation on the basis of a few real practical illustrations and useful recommendations to overcome them appeared elsewhere (Carpentier *et al*., 2017; Cohen, 1960; 1968; Feinstein and Cicchetti, 1990; Field and Welsh, 2007; Fleiss and Cohen, 1973). However, discussion on the estimation of 'κ' and its interpretation through a more generalized approach is still to be attempted. Besides with this entire research gap, the present study examines the changes between various levels on continuous data sets (random process) and its interpretation of 'κ' at varied time intervals 't_i'. For example, if the same or different observers will attempt to assess the disease conditions on continuous scale with random process at population level (leprosy screening in selected sites at varied time '*t*' with various level (Rural, Urban and Peri urban districts), the rates ascribe the rates by using ranking scale appended with different geographical location considering with varied time intervals besides with extrapolation of assigned ranking (rates converted in to random scale) example different attributed scale converted in to numerical forms by using time interval (here time is random variable). The extent of improvements of agreement between any pair of observers would be considered as random variables and all the observations are randomly distributed with $x_{ij} \sim N(\mu, \sigma^2)$, it will be ensured that checking of randomness of the raters'. The agreements will takes place between two or more observers which solely depend on the distribution system and classification of attributes at population level (Sim and Wright, 2005; Kraemer and Bloch, 1988; Kang *et al*., 2013; McHugh, 2012). It can provide fixed agreements between the observers. Since, the real tool practiced in various set up, this measure will be usually estimated by newer techniques of bootstrap modified kappa '*k*' approach to continuous data sets in which the same measure of agreement for classifications of un observed groups and extrapolation of predicted outcomes of raters subjects considering with slopes and Root Mean Squared error values (RMS). Assumed the Gaussianity for classifications of parameters of Leprosy population screening, derived an expression of variance (σ^2) and related hypothesis testing (Landis and Koch, 1977). This uncertainty concerning will be extrapolated with the exact distribution of the classifications. As such being the case, the hypothesis testing and confidence intervals can lead to incorrect results. Example, the confidence intervals derived will be based on the methods of derived agreement distribution, the distribution is symmetrical and classifications of observers agreements (population screening) is headed by the traditional Kappa '*k*', there is no adaptation of Gaussianity substantially differing from the raters' scale assumptions, as often to be the expression by traditional Kappa '*k*' is underestimated likelihood of parameters and also the variance (σ^2) is propagated through incorrect Confidence limits (CI) for testing the portion of agreements made by the different raters'. In this paradigm the present study explores and formulates new innovative stochastic model of modified Kappa '*k*' for the estimation of agreement levels of continuous series data sets.

Model Formulation

Model considers real-life data sets of anti-leprosy vaccine trial to demonstrate the modified kappa '*k*' by newer bootstrap techniques at greatest epoch with different iterations. Further, we also escalate the bootstrap robustness by using Thompson iteration method to examine the distribution of sampled estimators and likelihoods. The following illustrations were used for the formulation of this model, present study mainly concentrates on two or more observers with varying classifications, a most general situation for more than two observers, we persuaded to convert raters' ranks to continuous scale (data transformation). Suppose, the two doctors diagnose leprosy presenting attributes is (positive) or absent (negative) and the results are presented as follows:

	Observer II		
Observer I	Positive	Negative	Total
Positive	(a)	(b)	a + b
Negative	(c)	(d)	c + d
Total	a + c	b + d	N = (a + b + c + d)

Table 1: Kappa agreements overview

k	Decision
< 0	Poor agreement
0.01-0.2	Slight agreement
0.21-0.40	Fair agreement
0.41-0.60	Moderate agreement
0.61-0.80	Substantial agreement
0.81-1.00	Almost perfect agreement

As per the analytical forms, the crude agreements or concordance:

$$\left(p_o\right)=\frac{a+d}{a+b+c+d} \tag{1.1}$$

The proportion of subjects classified as positive by observer 1 is $\frac{a+b}{a+b+c+d} = p_1$ and by observer 2 is $\frac{a+c}{a+b+c+d} = p_2$. If $b = c$, $p_1 = p_2$ and vice versa.

The expression for the chance agreement between two observers is given by:

$$P_e = p_1p_2 + q_1q_2 \text{where} q_1 = 1-P_1 \text{and} q_2 = 1-P_2 \tag{1.2}$$

$$(\kappa) = \left(\frac{\text{Crude agreement} - \text{Chance agreement}}{1-\text{Chance agreemenet}}\right)$$

$$k=\frac{p_o-p_e}{1-p_e} \tag{1.3}$$

Closeness of rating patterns by different raters', the following matrices was used to test the agreements.

No of raters'	Quantitative data sets	Qualitative data sets
Two raters'	Bland Altman Plot	Cohen's Kappa
Weighted Cohen's Kappa >Two raters'	Krippendorffs alpha	Fleiss Kappa
Two groups of raters'	Based on spearman rank correlations	

Cube Root of Product Measures (CRPm)

The CRPm is used to identify the average agreement between the three values, the Fleiss Kappa and Krippendorffs alpha agreement is specifically used to escalate the agreement at greatest accuracy demonstrated presented in (Table 2):

$$CRPm = \sqrt[3]{Agr\left(A\right) * Agr\left(B\right) * Agr\left(A \cap B\right)}$$
$$Lies between -1 > CRPm < 1 \tag{1.4}$$

The disagreement was measured by the following mathematical Equation:

$$X_{jk} - \left[A_{1k} - B_{jk}, A_{2k} - B_{jk} ... A_{m1} - B_{jk}\right] \tag{1.5}$$

where, $i = 1,2...m_1$; $j = 1,2...m_2$, $k = 1,2...n$.

For subject 1, ratter B_1 and 3 group A raters', the Eq. (1.5) becomes:

$$X_{11} = \left[A_{11} - B_{11}, A_{21} - B_{11}, A_{31} - B_{11}\right]$$
$$\#Vectors = \#ratters' in\ group\ B * \#Subjects = m_2 * n$$

Extent of agreement was determined by the $Q_{jk} = X_{jk}' * S^{-1} * X_{jk}$.

Finally, the property of quadratic form is in the following form of Equation:

$$Agreement\ in\ QF = \frac{X_{jk}' * S^{-1} * X_{jk}}{X' * X} < \lambda$$
$$Dm = \frac{\frac{\sum_{k=1}^{n}\sum_{j=1}^{m_2} X_{jk}' * S^{-1} * X_{jk}}{X' * X} / m_2 * n}{\lambda_1}; 0 < Dm < 1 \tag{1.6}$$

Agreement measures = (Dm-1) applicable for both qualitative and quantitative data sets. Practically we compare and measure hypothetical statements of qualitative data of epilepsy considering group I and group II epilepsy subjects. However, the group I consider scholars who are trained by the physicians to diagnose the epilepsy based on sign and symptoms of 25 subjects in association with (group 2) mentor who is considered as group II. Five diagnoses were done on subjective approach, the implication is subject to test, if any one of the possible five diagnoses to each patients or subjects, on each stages of epilepsy signifies the various attributes. The continuous scale (transformed data sets) was analysed using R-statistical software, the following resulted findings were generated for testing agreement values.

$P\left(x=k\right) = \sum_{k=0}^{n}\binom{n}{k}x^k a^{n-k}$ of subjects classified as positive attributes cited for the whole population from the sample, it will be the mean of p_1 and p_2 i.e., $\bar{P}$. The proportions of subjects were classified as negative by observers 1 and 2 are (1-p_1) and (1-p_2) respectively. Similarly, the best estimates of the proportions of subjects are classified as negative for the population $\left(1-\bar{p}\right)$ drawn from the sample.

Table 2: Kappa agreements of various measures

Methods	Agreement values	Jack-knife Statistics (SE)
Agreement measure Am (1-*dm*)	0.89	0.955±0.00
Proportion agreement measure	0.83	0.722±0.012
Vanbelle's generalized measures	0.86	0.844±0.023
Consensus (median measures)	0.76	0.913±0.01
Pooled agreement measures	0.72	0.711±0.08
Pair wise agreement measures	0.74	0.739±0.006
Cube root of product measures	0.80	0.807±0.003
Consensus (Mode measures)	0.89	0.921±0.02

Minimum and Maximum Values of p_o

The minimax value of p_o is used to derive the minimum value of crude agreement:

$$P_{o\min} = \left|p_1 + p_2 - 1\right| \tag{1.7}$$

$$\textit{Maximum value of crude agreement}, p_{o\max} = 1 - \left|p_1 - p_2\right| \tag{1.8}$$

Prevalence (PI) and Bias Indices (BI) in terms of $p_{o\min}$ and $p_{o\max}$.

Prevalence Index (PI) is given by the difference between the estimates of proportions of subjects classified as positive and negative for the whole population, i.e., $PI = \bar{P} - \left(1 - \bar{P}\right) = 2\bar{P} - 1 = p_1 + p_2 - 1$ (1.4) which takes the values from -1 (when $p_1+p_2 = 0$) to +1. This (+1) happened only when all the screened observations or individuals for both the observers cited positive i.e., $p_1 = 1$ and $p_2 = 1$; and equals 0 when $p_1+p_2 = 1$. It can be seen that $|PI| = p_{o\min}$ discrepancy between the observers, if any, while assessing the frequency of occurrence of a given condition in a study group is denoted as bias. Bias Index (BI) = $|p_1-p_2|$, its minimum value is `0' when $p_1 = p_2$ and maximum value is 1 when $p_1 = 1$ and $p_2 = 0$ or $p_1 = 0$ and $p_2 = 1$. BI can be expressed as = 1- $p_{o\max}$.

Minimum and Maximum Values of κ in Terms of p_1 and p_2

$$\kappa_{\min} = \begin{cases} \dfrac{-2q_1q_2}{p_1q_2 + p_2q_1} \textit{if } p_1 + p_2 \geq 1 \\ \dfrac{-2p_1q_2}{p_1q_2 + p_2q_1} \textit{if } p_1 + p_2 < 1 \end{cases} \tag{1.8}$$

$$\kappa_{\max} = \begin{cases} \dfrac{2p_2q_1}{p_1q_2 + p_2q_1} \textit{if } p_1 \geq p_2 \\ \dfrac{2p_1q_2}{p_1q_2 + p_2q_1} \textit{if } p_1 < p_2 \end{cases} \tag{1.9}$$

$\kappa_{\max}$ in terms of p_e, PI and BI:

$$\kappa_{\max} = \frac{p^2{}_{0\max} - p^2{}_{0\min}}{1 + \left(1 - p_{0\max}\right)^2 - p^2{}_{0\min}} \forall\ p_1 \textit{and } p_2 \tag{1.10}$$

Which is nothing but Eq. (1.9) except that this can assume all values of p_1 and p_2 Eq. (1.10) Standard Error? $\left(S_e\left(\hat{k}\right) of\ \hat{k}\right)$. The naïve estimator of variance of $\hat{k}$ according to is given by $\overset{\sigma^2}{k} = \frac{1}{N\left(1 - \hat{p}_e^2\right)}\left\{\hat{p}_e + \hat{p}_e^{*2} - 2\,p_1p_2\left(p_1 + p_2\right) - 2q_1q_2\left(q_1 + q_2\right)\right\}$

estimated standard error of $\hat{k}$ is given by $S_e = \hat{\sigma}_K$.

Bootstrapping Technique

The nonparametric bootstrap technique is applied as statistical methods or tools for estimating distribution of attributed data sets, the method will help us to draw effective inference in both sample and population level. The technique used for the estimation of agreement between two or more observers is yet to be proposed. The methodological insight for nonparametric bootstrap re-sampling of the estimation of '*k*' and other parameters of the data sets at varying degree of measurements or agreements between the observer. Let us assume, there will be an observed sample of n_1 pairs of classification of the disease *viz* a_1 of '+ +', b_1 of '+ -' c_1 of '- +' and d_1 of '- -' drawn by the two observers, while screening the population for leprosy cases in medical research:

(i) Draw a random sample of 'n_1' pairs $\left(a_1^*, b_1^*, c_1^*, d_1^*\right)$ from the observed sample with replacement

(ii) Derived the related empirical parameters of agreements of our interest:

$$\hat{p}_0^* = \frac{a_1^* + d_1^*}{n_1},\ p_1^* = \frac{a_1^* + b_1^*}{n_1}$$

$$p_2^* = \frac{a_1^* + c_1^*}{n_1},\ p_e^* = p_1^*\,q_1^* + p_2^*\,q_2^*$$

$$p_{0\min}^* = \left|p_1^* + p_2^* - 1\right|, \hat{p}_{0\max}^* = \left\{1 - \left|p_1^* + p_2^*\right|\right\} \tag{1.11}$$

$$\hat{k}^* = \frac{p_0^* - p_e^*}{1 - p_e^*}\hat{k}_{\max}^* = \frac{p_{0k\max}^{*2} - p_{0\min}^{*2}}{1 + \left(1 - p_{0\max}^*\right)^2 - p_{0\ \min}^{*2}} - p_1^* and\ p_2^*$$

$$S_e\left(\hat{k}*\right) = \sqrt{\frac{1}{n_1\left(1 - \hat{p}e^{*2}\right)}\left\{\hat{p}_e^* + \hat{p}_e^{*2} - 2\hat{p}_1^*\hat{p}_2^*\left(\hat{p}_1^* + \hat{p}_2^*\right) - 2q_1^*q_2^*\left(q_1^* + q_2^*\right)\right\}}$$

Where, $q_1^* = 1 - p_1^*$, $q_2^* = 1 - p_2^*$.

(iii) Repeat (i) and (ii) steps we seen 5000 times epoch series to obtain the set of 5000 replications of $\hat{p}_0^*, \hat{p}_e^*, \hat{p}_e^* \min, \hat{p}_e^* \max, \hat{k}, \hat{k}^* \max, and\ S_e\left(\hat{k}^*\right)$

(iv) The estimated mean μ and variance σ^2 on each parameter are simply the average and variance σ^2 of its corresponding sets of 5000 bootstrap sample estimators

(v) The distribution of each sample estimator is studied through the histogram of 5000 bootstrap replications of the Parameters

(vi) 95% confidence limits for each estimator are well defined and found to be the 2.5 and 97.5th percentiles of the corresponding distribution. The bias were corrected and accelerated with confidence limits after adjusting the percentile interval are also replicated, though more complex and better bootstrap confidence intervals are available from the above Eq. (1.10), we used BC_a intervals in this article. Similar procedures were listed in Eqs. (i) to (vi) which are presumably repeated for the data on various classification of diseases by the same two observers in a sample of n_2 subjects vice versa, we test the Null Hypothesis $H_0 : \hat{k}_1 < \hat{k}_2$ against alternative hypothesis; $H_1 : \hat{k}_1 < \hat{k}_2$ by the following algorithms. First, combined samples of $n_1 + n_2$ pairs of classifications has been made by rearranging the n_2 pairs of classifications of the II resurvey to the side of n_1 pairs of classifications of the I resurvey. (i) Draw a sample of $n_1 + n_2$ pairs with replacement from the combined. (ii) Evaluate for each sample $t\left(k^*\right) = k_2^* - k_1^*$

(vii) Repeat (i) and (ii) 5000 times bootstrap values (iv) Examine the $ASL_{Boot} = \#\left\{t\left(k*\right) \geq \left(k_2 - k_1\right)\right\} \Big| 5000$

If ASL_{Boot}<0.05 we conclude that, the agreement between the two observers at II Resurvey is significantly associated with (I-resurvey). We consider only three methods, viz naïve with variance unadjusted, percentile intervals and the bootstrap BC_a confidence intervals.

Results

Evaluation of Agreement through Simulations

We evaluated the statistical accuracy of the unobserved sample estimators from continuous data series of different raters. The results of the bootstrap procedure are presented. From results from first resurvey it is evident that the distribution sample estimators for agreement between two observers are non-normal. Similar findings may be noticed for the data at second resurvey as well as the combined sample (Table 3) presents the results of adopting naïve, percentile interval and BC_a methods to obtain the point and interval estimators of agreement. We can see from this analysis that all the three methods provide the same average estimates. The naïve method for estimating 95% confidence interval always yield symmetrical results unlike the bootstrap method which provides asymmetrical results for all sample estimators of the agreement. The length of the interval is the lowest in the naïve method. However, the lower and upper confidence limits are lower than the corresponding limits in the bootstrap percentile and BC_a methods. We evaluate the performance of the bootstrap for estimating the parameters of agreement; we undertook the similar exercise on the data from II resurvey. The results of the analysis are presented in (Table 3). Here again, we find that all the three methods provide same point estimates for each parameter of agreement. The confidence intervals through naïve method are symmetric vis-à-vis bootstrap percentile and BC_a methods. The findings are otherwise similar to what we observe for first resurvey. The sample estimators of agreement except for k_{max} at II Resurvey are consistently higher. However, the variance of each estimator is consistently lower at second resurvey. The improvement in $\hat{k}$ suggests that the improvement can be as high as 10 percentage points. The sampling distribution of $\hat{k}_{max}$ relative to $\hat{k}$ suggest that the former is more stable than the latter though the former is equally sensitive. Further, the diagnosis that will remain '+ +' or '- -' at both the occasions also confirm the consistency in ratings by two observers.

In the 25 subjects with full set of 5000 replications, the model has yielded good predicted average agreements (0.91±0.06) with positive slope movement (slopes = 2.24 times of average value with ignorable Root Mean Square Errors (RMSE 0.03). The model prediction is perfect level and most robust in nature (Co-efficient of variation (R^2 = 98.0%; SE 0.06) when compared to Cohen's '*k*'. The maximum agreement in case of predicted values eagerly falls on the subject of raters 22 (agreement level was 0.99; RMSE 0.032). For each subjects, the accuracy was maintained (0.80,0.88,0.90,0.98; n = 25) there are 25 simulations falling on each data points. Since, higher the observer accuracy abridged with better agreements levels. The ratio of agreement level in each subjects at different time period 't_i' is more consistent and also maintains greatest accuracy and more precision $\left(\frac{n}{\sigma^2}\right)$ where n = number of subjects σ^2 is the variance of agreements. The newer simulation agreement was determined based on the Bootstrap-Thompson iteration method; the simulated figures were extracted on the basis of central tendency values with average observed rate of 25 subjects. The resulted findings were found to be more epoch (the modal value attained >9 rater scale) rather than traditional '*k*' measurement (Fig. 1). While all predictive subjects achieve a kappa agreements and above (Table 3). The

predicted accuracy influences maximum '*k*' values, as we have shown in the simulation results starting from 1 to 25 subjects, the agreement values on continuous series reached to the normal approximation, model results shown in the form of Bootstrap density plot (Fig. 2). An increasing number of subjects and raters', the simulated agreements will be moved to the positive direction with stronger values of '*k*', simultaneously, the predicted agreement is more epoch and it was found to be statistically significant and different between the observer and different agreement levels with varied time intervals '*t*' (iteration replication 10000 times; $R^2 = 0.99\%$).

Table 3: Agreement values predicted by bootstrap techniques (Qualitative data)

Subjects Raters	Observed	Predicted	Slope	RMSE
1	0.96	0.98	0.220	0.0001
2	0.72	0.94	6.060	0.0003
3	0.81	0.99	4.110	0.0073
4	0.89	0.97	3.740	0.0034
5	0.70	0.74	3.600	0.0060
6	0.96	0.92	1.830	0.0070
7	0.88	0.98	1.600	0.0246
8	0.12	0.87	5.060	0.0813
9	0.74	0.88	3.780	0.0011
10	0.74	0.86	3.210	0.0060
11	0.63	0.81	2.470	0.0085
12	0.65	0.88	1.970	0.0507
13	0.74	0.83	1.510	0.0472
14	0.55	0.86	1.390	0.0134
15	0.63	0.88	1.047	0.0382
16	0.81	0.84	0.250	0.0824
17	0.85	0.92	1.350	0.0176
18	0.84	0.96	2.100	0.0812
19	0.86	0.94	2.050	0.0551
20	0.70	0.92	1.750	0.0312
21	0.77	0.93	1.640	0.0112
22	0.85	0.99	1.340	0.0324
23	0.83	0.91	1.730	0.0423
24	0.84	0.88	1.760	0.0324
25	0.74	0.87	0.590	0.1266
Mean epoch	0.75±0.166	0.91±0.06	2.240	0.0320
Model R^2 (%)	0.77	0.98		

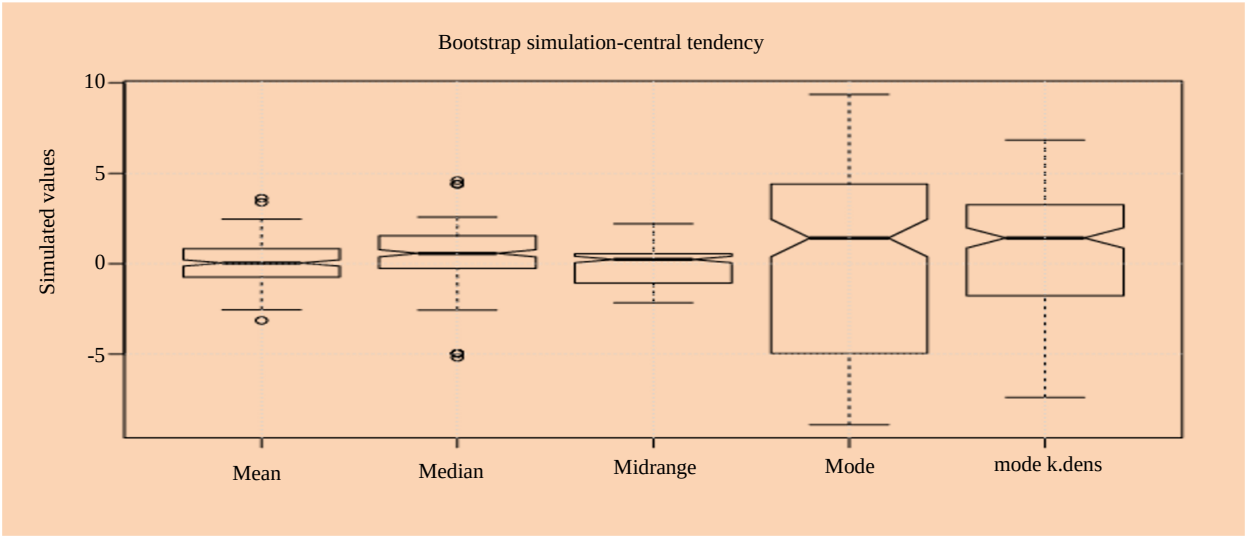

Fig. 1: Bootstrap simulation predicted based on central tendency

From the research findings we notice that, the agreement level shifts lower when data points become lower. At observer accuracy level substantially very low, the subject agreement does not produce accurate results because the iteration points would not be generated with lower accuracy. However, many factors that affect values of Kappa include observer accuracy and number of subjects, as well as when observer distributes the subjects equally. There is no one value of modified kappa that can be (produced Bootstrap techniques) regarded as universally acceptable and propagation of simulation figures is highly associated with level of observer *s*, accuracy, precision and the number of subjects. With a fewer number of subjects (k<5), epically in binary classifications, modified kappa values need to be interpreted with extra caution. Eventually, in case of binary classification, predicted variability has the strongest impact on Kappa '*k*' values and leads to the heterogeneous and also strongest impact on the observer agreements. On the other hand, when there are more observers (>25 subjects), the increment *t* of expected kappa value becomes flat. Hence simply determine the percentage of agreement. Moreover, the increment to f values of the performance matrices apartment form sensitivity also reaches the asymptote from more than 25 subjects (Fig. 3 to 5).

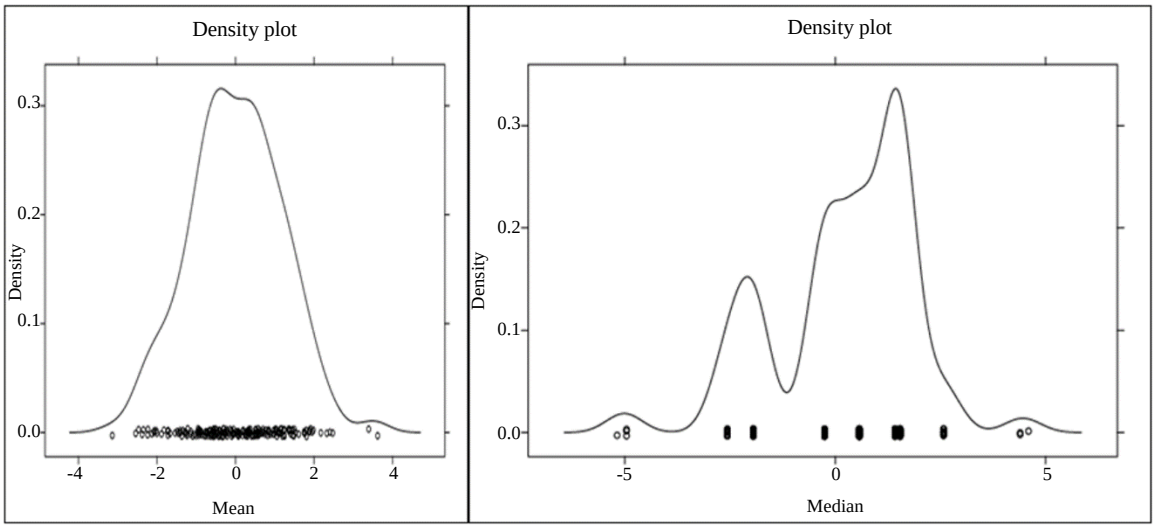

Fig. 2: Density plot mean and median demonstrated by bootstrap simulation

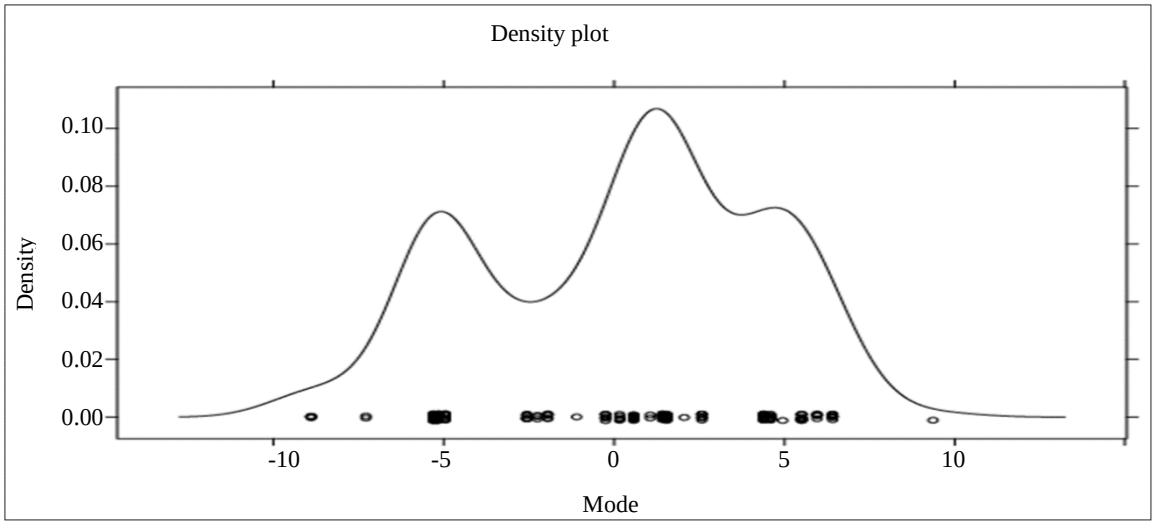

Fig. 3: Density plot of mode demonstrated by bootstrap simulation

The observer's agreement data points were plotted on the Box Cox and QQ plot from modified kappa statistics-bootstrap techniques to know the normality originated from the population with a common distribution. The results showed that, the observer agreement points is normally distributed and to optimize the bootstrap techniques with various methods of integration (Thompson). Optimal lambda value ranged from 1-2, that means each individual agreement can take a weight of raters value of 2.0 in the ordinal scale for estimation of density plot and also, the model has produced the QQ plot for weibull distribution to know the exact confidence limit of kappa '*k* '. Figure 6 showed that the confidence limit of waited modified kappa '*k*' was 6.88 with scale of 0.81 agreement level.

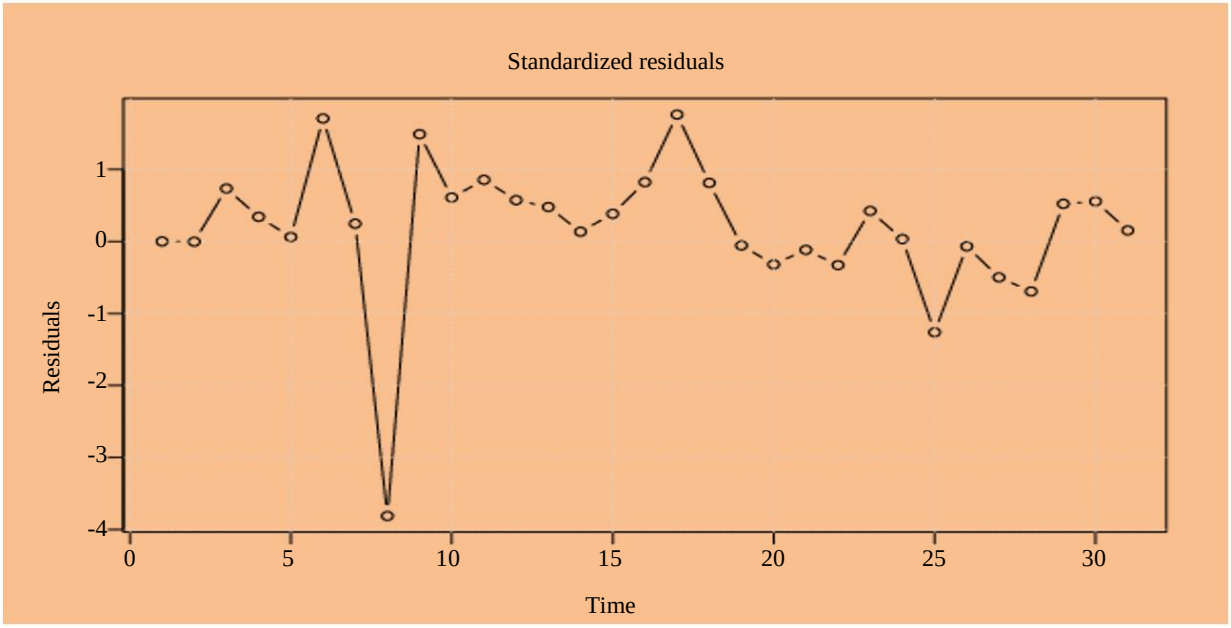

Fig. 4: Standard residual of bootstrap simulation varying values of agreements of '*k*'

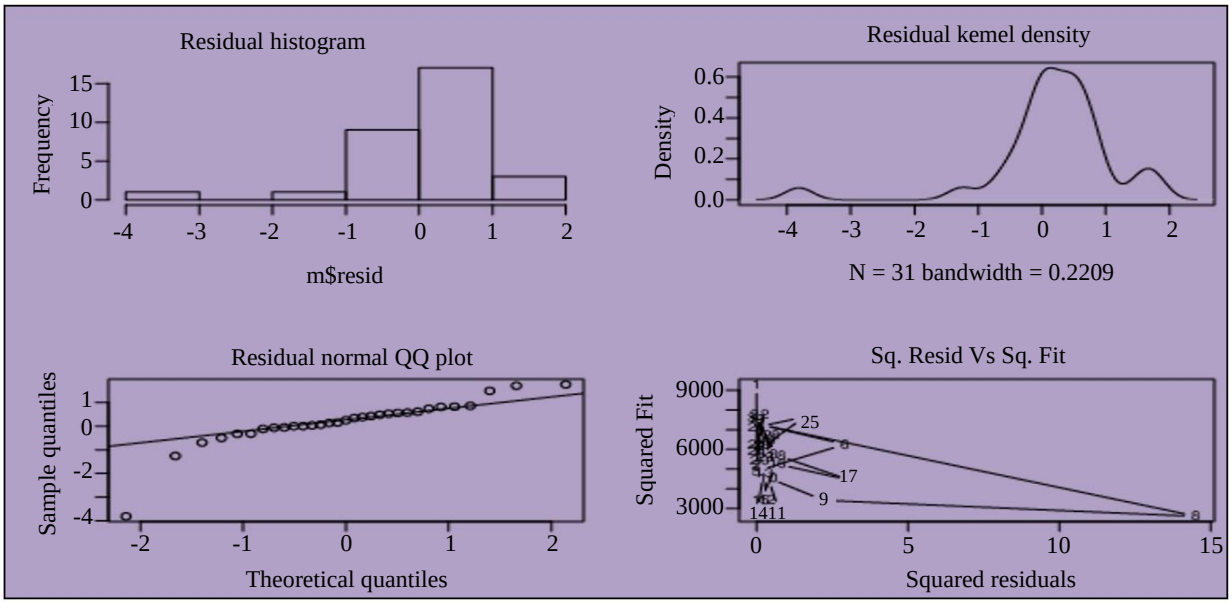

Fig. 5: Residual graphs of '*k*' by bootstrap simulation '*k*'

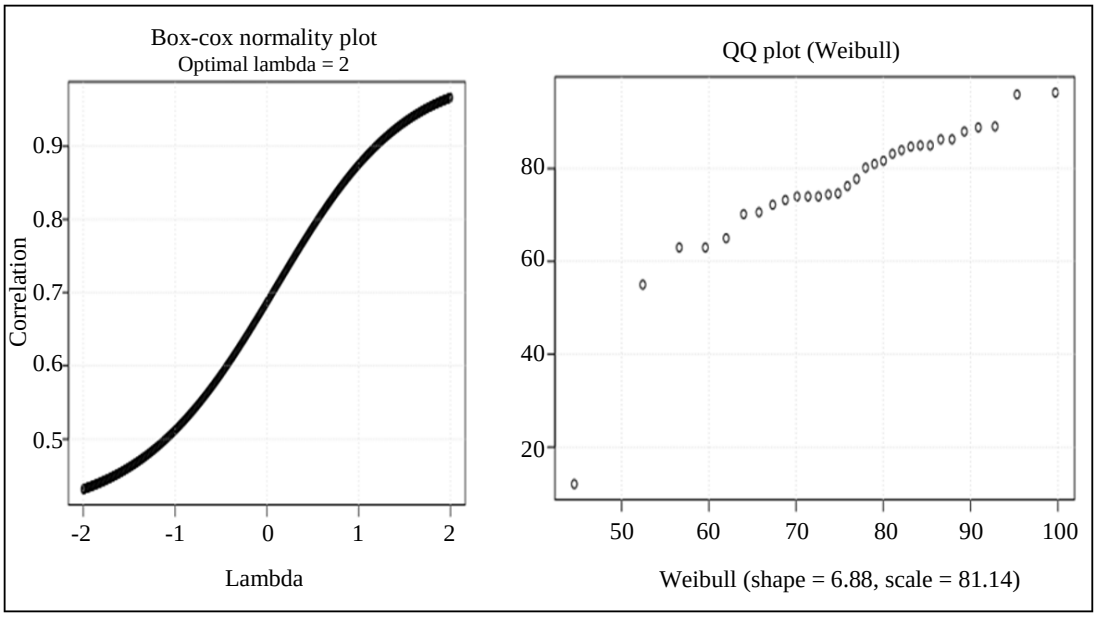

Fig. 6: Box and Cox normality and QQ plot of agreements of '*k*'

Discussion

The first point of interest emerges from the study; the estimates of agreement between two observers follow non-Gaussian distribution. Consequently, the sample estimators though unbiased provide lower variance and narrow confidence limits. Further the 95% confidence limits on each parameter are symmetrical. This emphasizes the need for estimating the agreement measures and its standard errors that allows its appropriate distribution status. Adjustment to these estimates as well as hypothesis testing also perform poorly as they are based on the assumption that the observations from population screening for disease conditions follow asymptotic normal distribution. The second point of interest in this study is that we can demonstrate through bootstrap that we can obtain accurate estimates of $\hat{k}$ and its standard error. Further, this approach is better than currently used methods for non-normal data. This approach also allows for testing the hypothesis whether the $\hat{k}$ values at I and II Resurvey are significantly different. Consequently we observe that the $\hat{k}$ at second Resurvey is significantly more than that at I Resurvey indicating that $\hat{k}$ should not be treated as static over a period of time if the same pair of observers are repeatedly employed for classification of the disease condition while screening the given population. This may ensure that the estimation of near true incidence of the disease. On the other hand if we use $\hat{k}_{max}$ as a single index in terms of chance agreement, prevalence and bias indices, we can be at a safer side because this index is more stable while equally sensitive vis-à-vis $\hat{k}$. We therefore propose $\hat{k}_{max}$ to be employed as a measure of agreement for comparison and interpretation. Studied inter ratter reliability of categorical data sets, the '*k*' will be used to verify the presence of the theme that were presented. The *k* co efficient is a statistical measures the inter ratter reliability of agreement that is used to assess qualitative documents and determine agreement between two raters. An important assumption underlying the use of the kappa co efficient is that the errors associated with clinicians rating are independent (Brennan and Prediger, 1981; Hoehler, 2000; Sim and Wright, 2000; Cohen, 1960; Richards *et al*., 2003). This requires the patients or subjects to be independent and ratings to be independent, so that each observer should generate a rating without knowledge and thus without influence, of the other observer's rating. The fact that the ratings are related in the sense of pertaining to the same intervention, however does not contravene the assumption of independence. Sim and Wright (2005) Reliability of clinicians rating is an important consideration in areas such as diagnosis and the interpretation of examination findings. Often, these ratings lie on a nominal or an ordinal scale. For such data, the kappa co efficient is an appropriate measure of reliability. Important factors that can influence the magnitude of kappa (prevalence, bias and non independent ratings) are key indicators for assessing the kappa statistics. The present research, the co efficient can be used for scaling with more than two or more categories for assessing the agreement levels inclusion qualitative and quantitative traits similar study reported by (Rigby, 2000) he assessed the intra observer and inter observer agreement of radiographic classification of scoliosis in relation to the king classification system. Emphasized the value of

multiple regression method and the importance of power and measuring effects rather than testing significance. For more than two raters, the mathematics is such that the two raters' are not considered unique. For instance, if there are three raters', there is no assumption that the three raters' who rate the first subject are the same as the three raters' who rate the second. Although, we call this more than two raters' case it can be used with two raters' when the raters' identified vary. The '*k*' is the generalization for weights reflecting to the relative seriousness (Cohen, 1960) of each possible disagreement due to attributable factors. The analysis of variance approach for $k = 2$ and $m > = 2$ is due to (Landis and Koch, 1977).

The kappa '*k*' was first proposed by (Cohen, 1960). The generalization for weights reflecting to the relative seriousness of each possible disagreement is due to (Cohen, 1968). The analysis-of-variance approach for $k = 2$ and $m \geq 2$ is due to (Landis and Koch, 1977). or (for an introductory treatment and chap. 18) for a more detailed treatment. All formulas below are as presented Let *m* be the number of raters' and let *k* be the number of rating outcomes. Carpentier *et al*. (2017) demonstrated the free response kappa in *a* computed form that the total numbers of discordant (*b* and *c*) and concordant positive (*d*) observations made in all patients, as $2d/(b + c + cd)$. In 84 full body magnetic resonance imaging procedures in children that were evaluated by two independent raters', the free-response kappa statistics was 0.820. Aggregation of results within regions of interest resulted in overestimation agreement beyond chance. The free response kappa provides an estimate of agreement beyond chance in situation where we only have positive findings and are reported by raters'. Kang *et al*. 2013) when the observations are independent, confidence intervals can be computed using several methods, in case of clustered data, a common situation radiology, the present study we opted is a bootstrap based approach for testing various subjects on the basis different attributes. We sampled subjects (with cluster approach) and used all observations from any selected patient. Yang and Zhou (2014) '*k*' is widely used to assess the agreement between two procedures in the independent matched pair data. For matched pair the data is collected in clusters, on the basis of the data method and sampling techniques, that propose a non parametric variance estimator for the kappa statistics without cluster correlation structure or distributional assumptions, further result demonstrated by the extensive Monte Carlo simulation that the proposed '*k*' provides consistent estimation and the proposed variance estimator behaves reasonably well for at least moderately large number of clusters our study compliments this and we have derived this model based on Thompson iteration optimization methods applied various subjects to escalate the agreement levels of both qualitative and quantitative data sets.

Conclusion

The new statistical intervention tool is very useful to know the different association measure for testing the agreement between two or more observers on continuous scale with varied time intervals. The newer tool increases the precision of rater's confidence level in a single real number without any substantial loss of information. It is capable to reproduce the accurate predicted agreement levels and slopes of continuous data series. There is no complexity between inter observer agreement testing considering with the various multiple degrees on different time intervals.

Acknowledgement

Heartfelt thanks to Dr. Ayanendranath Basu, Associate Professor, Indian Statistical Institute, Calcutta for his valuable comments and suggestions. Thanks to Mr. A. Murugarasan for his secretarial assistance.

Funding Information

There is no external funding used for this research, it is fully oriented primary and secondary data.

Author's Contribution

Basavarajaiah Mariyappa Doddagangavadi: Computation, writing editing and model formulation with support of experts.

Narasimha Murthy: Data collection and Problem formulation.

Netra Rajpurohit: Technical Proof reading and assisted write-up work.

Ethics

The present research study conducted on secondary data sets, it is an Observational empirical study, on conceptual point of view. There is no ethical consideration aroused from the study design.

References

Ayoub, A., & Elgammal, A. (2018, October). Utilizing Twitter data for identifying and resolving runtime business process disruptions. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 189-206). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-02610-3_11

Bakeman, R., & Quera, V. (2011). Sequential analysis and observational methods for the behavioral sciences. Cambridge University Press. ISBN-10: 1139504606.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses and alternatives. Educational and Psychological Measurement, 41(3), 687-699. https://journals.sagepub.com/doi/abs/10.1177/001316448104100307

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. Journal of Clinical Epidemiology, 46(5), 423-429. https://www.sciencedirect.com/science/article/abs/pii/089543569390018V

Carpentier, M., Combescure, C., Merlini, L., & Perneger, T. V. (2017). Kappa statistic to measure agreement beyond chance in free-response assessments. BMC Medical Research Methodology, 17(1), 1-8. https://link.springer.com/article/10.1186/s12874-017-0340-6

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37-46. https://journals.sagepub.com/doi/pdf/10.1177/001316446002000104

Cohen, J. (1968). Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70(4), 213. https://psycnet.apa.org/record/1969-00069-001

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology, 43(6), 543-549. https://www.sciencedirect.com/science/article/abs/pii/089543569090158L

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(3), 369-390.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement, 33(3), 613-619. https://journals.sagepub.com/doi/pdf/10.1177/001316447303300309

Hoehler, F. K. (2000). Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. Journal of Clinical Epidemiology, 53(5), 499-503. https://doi.org/10.1016/S0895-4356(99)00174-2

Kang, C., Qaqish, B., Monaco, J., Sheridan, S. L., & Cai, J. (2013). Kappa statistic for clustered dichotomous responses from physicians and patients. Statistics in Medicine, 32(21), 3700-3719.

Kraemer, H. C., & Bloch, D. A. (1988). Kappa coefficients in epidemiology: an appraisal of a reappraisal. Journal of Clinical Epidemiology, 41(10), 959-968. https://www.sciencedirect.com/science/article/abs/pii/0895435688900327

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 159-174. https://www.jstor.org/stable/2529310?seq=1

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. Biochemia Medica, 22(3), 276-282. https://hrcak.srce.hr/89395

Richards, B. S., Sucato, D. J., Konigsberg, D. E., & Ouellet, J. A. (2003). Comparison of reliability between the Lenke and King classification systems for adolescent idiopathic scoliosis using radiographs that were not premeasured. Spine, 28(11), 1148-1156. https://journals.lww.com/spinejournal/fulltext/2003/06010/comparison_of_reliability_between_the_lenke_and.12.aspx

Rigby, A. S. (2000). Statistical methods in epidemiology. v. Towards an understanding of the kappa coefficient. Disability and Rehabilitation, 22(8), 339-344. https://www.tandfonline.com/doi/abs/10.1080/096382800296575

Sim, J., & Wright, C. (2000). Research in health care: Concepts, designs and methods. Nelson Thornes. ISBN-10: 0748737189.

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation and sample size requirements. Physical Therapy, 85(3), 257-268. https://academic.oup.com/ptj/article-abstract/85/3/257/2805022

Yang, Z., & Zhou, M. (2014). Kappa statistic for clustered matched-pair data. Statistics in Medicine, 33(15), 2612-2633. https://doi.org/10.1002/sim.6113