Journal of Data Science 12(2014), 697-716

A-Kappa: A measure of Agreement among Multiple Raters

Shiva Gautam [1]*

[1] *Beth Israel Deaconess Medical Center, Harvard Medical School*

Abstract: Medical data and biomedical studies are often imbalanced with a majority of observations coming from healthy or normal subjects. In the presence of such imbalances, agreement among multiple raters based on Fleiss' Kappa (FK) produces counterintuitive results. Simulations suggest that the degree of FK's misrepresentation of the observed agreement may be directly related to the degree of imbalance in the data. We propose a new method for evaluating agreement among multiple raters that is not affected by imbalances, A-Kappa (AK). Performance of AK and FK is compared by simulating various degrees of imbalance and illustrate the use of the proposed method with real data. The proposed index of agreement may provide some insight by relating its magnitude to a probability scale. Existing indices are interpreted arbitrarily. This new method not only provides a measure of overall agreement but also provides an agreement index on an individual item. Computation of both AK and FK may further shed light into the data and be useful in the interpretation and presenting the results.

Key words: Fleiss' Kappa, A-Kappa

1. Introduction

1.1 The problem

Biomedical, social, behavioral and other studies routinely include statistical evaluations of agreement among multiple raters or conditions (Fienstein et al, 1985) and Fleiss' Kappa (FK) is widely used to evaluate agreements (Fleiss, 1981).

This work stems from a real problem which arose when examining agreement among radiologists who were evaluating mammographic breast images. Table 1 shows the results of an experiment where 10 radiologists at Brigham and Women's Hospital, Boston, Massachusetts, independently reviewed 102 breast MRIs obtained between September 2004 and April 2008. Increased breast density may be associated with increased breast cancer risk, and therefore, classification of mammography reports is important both scientifically and clinically. The BI-RADS algorithm developed by American College of Radiology was used to classify breast

* Corresponding author.

composition into four categories. One of the categories indicates whether or not an image exhibits a ‘Fatty’ (< 25% glandular) pattern. Table 1 collapses the three non-Fatty categories into one and denotes them as a ‘1’: A fatty image is denoted as a ‘0’. Each line in the table displays a sequence of ‘1’s and ‘0’s which shows the scoring pattern for each of the 10 raters.

Table 1: Classification of 102 images by 10 raters into Non-Fatty (= 1) versus Fatty (= 0) breast composition categories

Yes = 1, No = 0	Number of ‘Yes’ (= 1)	Frequency	Percent
1001011001	5	1	0.98
1011011111	8	2	1.96
1101110111	8	1	0.98
1111011001	7	1	0.98
1111011110	8	1	0.98
1111011111	9	4	3.92
1111110111	9	1	0.98
1111111001	8	1	0.98
1111111011	9	4	3.92
1111111101	9	1	0.98
1111111111	10	85	83.33
Total	-	102	100

All 10 raters classified 85 (83.33%) of the 102 images into non-Fatty category indicating complete agreement among the raters on these images. An additional 10 images were classified as non-Fatty by 9 (90%) of the raters so that more than 93% of the images were classified into the non-Fatty category by at least 90% of the raters. Despite such a large degree of consensus seen among the raters, Fleiss Kappa (FK) for these data is only 0.119 (95% CI: 0.090, 0.148) indicating a poor or no agreement. An alternate agreement index developed in this paper called A-Kappa (AK) yields a value of 0.906 (95% CI: 0.889, 0.923) for the same data indicating a high agreement among raters. Thus, the widely used method of evaluating agreement index FK yields a counter intuitive result in this instance.

It will be shown later that a data set consisting of a large number of positive (or negative) events may yield a poor FK despite high observed agreement. In the context of Table 1, all the raters classify 83.33% images into the positive (score 1) category. There is not a single image where all the raters classified an image into the negative (score 0) category. Similarly, there are 5 images with eight 1s, but there is not a single image with eight 0s. This type of data set is referred to as unbalanced or asymmetrical data set in this paper. This issue is further addressed in Section 2.3 and then shown that the proposed measure AK is not influenced by the imbalance in a data set.

Medical data are prone to imbalance due to high or low prevalence of a given characteristics or disease (Li, Liu and Hu, 2010). For example, in a screening setting it is likely to have are more healthy individuals, while in a specialty care center there may be a larger number of subjects with a disease. Hence, for many biomedical data Fleiss Kappa (FK) is likely to misrepresent or completely miss the agreement present in a data set. A-Kappa (AK) developed in this paper could be an alternate tool to evaluate agreement in such data sets as it is not influenced by the asymmetry or imbalance.

In two raters' case the phenomenon of high observed agreement with low Cohen's Kappa often found with asymmetrical or unbalanced data has been studied and some remedies have been suggested (Feinstein et al, 1990; Cicchetti et al, 1990; Lantz et al, 1990). These remedies primarily suggest reporting some alternate indices along with Cohen's Kappa. Lantz and Nebenzahl (1990) maintain that Kappa alone has little interpretive value and recommend reporting alternative indices along with Kappa.

In this article we show that Fleiss Kappa, the most widely used agreement index among multiple raters, shows the high agreement low Kappa behavior similar to that of Cohen's Kappa. A new method for evaluating agreement among multiple raters, A-Kappa (AK), is proposed in this article. This method is not affected by the type of imbalances described above and is able to capture the observed agreement. Furthermore, in the case of balanced data set it reduces to FK. It is proposed that both FK and AK be reported with the results of data analysis.

2. An Alternate Measure of Agreement

2.1 Proposed agreement index: A-Kappa

Suppose that each of the $r \ (>2)$ independent raters classifies the *i*th image $(i = 1,2,...,N)$ into one of two categories by assigning a score of 1 or 0 to indicate presence or absence of a disease, respectively. When all the raters agree on a given image, we will have either all 1s or all 0s. Similarly, when the sequence consists of an equal numbers of 1s and 0s (50% of each) it is considered a situation of complete absence of agreement.

Let a_i denote the number of 1s in the sequence for the *i*th image. In other words, a_i raters out of *r* total raters classify the *i*th image into the disease category and remaining *r* - a_i into the non-disease category. We assume that each image is read by all of the *r* readers.

The proposed measure of agreement A-Kappa (AK) is defined as

$$AK = \sum_{i=1}^{N}[(2a_i - r)^2 - r] \big/ [N(r^2 - r)] \qquad (2.1)$$

The derivational argument behind the definition AK in equation (2.1) is provided in Section 3.1 where more than two categories are addressed. More specifically when $k = 2$, equation (3.5) reduces to equation (2.1).

The following results follow from the definition of AK given by equation (2.1). Proofs are given in the Appendix.

Proposition 1. When there are two raters, *AK* reduces to Maxwell's Random Error (*RE*) coefficient (Maxwell, 1977):

$$AK = 2P_0 - 1 = RE, \tag{2.2}$$

where P_0 is the proportion of images on which the two raters agree. Note that, RE was originally proposed as an alternate measure of agreement between two raters to address the issue of high agreement and low Cohen's kappa.

Proposition 2. If w_i is the proportion of pairs of raters who agree and v_i is the proportion of pairs who disagree on the *i*th image, then

$$AK = \sum_{i=1}^{N} (w_i - v_i) / N \tag{2.3}$$

Proposition 3. AK for multiple raters r is the average of *AK* for all possible pairs of raters. Let AK_{ij} denote AK obtained from the *i*th and *j*th raters. Similarly, let $P_{0,ij}$ denote the proportions of images on which both of these raters agree. Then

$$AK = \sum_{i>j}^{r} 2P_{0,ij} / C(r,2) \text{, where } C(r,2) = [r(r-1)]/2 \tag{2.4}$$

2.2 Relationship between A-Kappa (AK) and Fleiss Kappa (FK)

In the section, relationship between AK is developed and shown that for balanced or symmetrical data these indices are identical. Concept of balanced or symmetrical data was introduced at the beginning of this paper. It will be revisited here to establish equality between AK and FK.

Proposition 4. $AK = 1 - 4\bar{p}\bar{q}(1 - FK)$ where $\bar{p}$ is the proportion of one positive classifications (or proportions of '1's) in the entire data set, and $\bar{q} = \bar{p}$.

Proof: Recall that, N, r and a_i $(i = 1,2,...,N)$ denote the total number of images (or objects to be evaluated), total number of raters and number of raters who classify the *i*th image as positive (or who assign '1'), respectively. Let $p_i = a_i / r$ denote the proportion raters who classify the *i*th image into the positive category. Similarly, let $\bar{p}_i = \sum a_i /(rN) = \sum a_i / N$ denote the overall proportion of positive responses, and $\bar{q}_i = 1 - \bar{p}_i = \sum (r - a_i) / N$. Then the Fleiss Kappa (FK) is defined as (Fleiss, 1981)

$$\text{FK}=1-\left[\sum a_i(r-a_i)\right]/\left[Nr(r-1)\overline{p}\overline{q}\right] \qquad (2.5)$$

Multiplying both sides of it by $4\overline{p}\overline{q}$

$$4\overline{p}\overline{q}FK = 4\overline{p}\overline{q} - 4\left[\sum a_i(r-a_i)\right]/\left[Nr(r-1)\right]$$

$$= 4\overline{p}\overline{q} - 1 + 1 - 4\left[\sum a_i(r-a_i)\right]/\left[Nr(r-1)\right] \qquad (2.6)$$

Upon simplification, $1-4\left[\sum a_i(r-a_i)\right]/\left[Nr(r-1)\right] = \left[\sum((2a_i-r)^2-r)\right]/\left[Nr(r-1)\right]$=AK(from the definition of AK)

Therefore, from (2.6)

$$AK = 1-4\overline{p}\overline{q}(1-FK) \qquad (2.7)$$

2.2.1 Symmetrical or balanced data

In the present context, symmetry may be defined in several ways. At the basic level, if 50 % of observations (e.g. images) in a data set come from one population (say healthy) and remaining 50% come from a second population (e.g. disease) then such a data set could be considered a balanced data set. Even with experienced raters, it is likely that there will be instances of misclassifications, and therefore, it is unlikely that an image will be assigned either 1or 0 by all raters. But with large data sets one can expect that they will be evenly misclassified. In other words, one would expect that for every misclassification into positive category there will be a misclassification into the negative category. Imbalance in data set may be due to design (e.g. more positive images in the collected data) or prevalence (e.g. rare disease). For example, if a data set contains considerably more positive (or negative) images, then the data set will be imbalanced. Hence, a data set in which there is an image with a given number of 1s for each image with the same number of zeros will be considered a symmetrical or balanced data. Lack of this gives rise to an unbalanced data set. According to this definition data set presented in Table 1 is an unbalanced data set.

Proposition 5. AK≥FK. When the data set is balanced then AK= FK for balanced or symmetrical data set.

Proof: In the case of balanced data (see definition above) the entire data set be can be presented in terms of $N/2$ pairs of image such that for a pair consisting of ith and $i'th$ images, we have $p_i = a_i/r$ and $p_{i'} = 1-a_i/r$ so that, $p_i + p_{i'} = 1.$ Therefore, in a balanced data set, $\overline{p}_i = \sum a_i/(rN) = \sum a_i/N = N/(2N) = 1/2$. Hence, $\overline{p} = 1/2 = \overline{q}$. Therefore, from (2.7), in balanced (or symmetrical) data set, AK=FK.

Next,

AK< FK $\Rightarrow 1-4\overline{p}\overline{q}(1-FK) < FK$ from equation (2.7)

$\Rightarrow 1-4\overline{p}\overline{q} < FK(1-4\overline{p}\overline{q}) \Rightarrow 1 < FK$. But in fact FK $\leq$ 1. Therefore, AK $\geq$ FK.

So, Fleiss Kappa (FK) yields a smaller value than AK in an unbalanced or asymmetrical data set.

It is worth noting that if the number of 1s is equal to the number of 0s in data set then also AK= FK whether or not the data set meets the definition of symmetry. If 1’s and 0’s are assigned completely randomly then both AK and FK will be equal to zero indicating lack of agreement.

2.3 Simulations

We conducted several simulations to gain some insight into A-Kappa’s performance and to compare it with Fleiss Kappa, and aid in its interpretation.

We based our simulations on 10 raters, 2 categories and 10,000 images. Let βdenote the proportion of diseased images. Let π denote the probability of correctly classifying an image by a rater and is assumed to be the same for π = each rater. Each rater is assumed to assign a score of 1 to a diseased image and 0 to a healthy image (from normal subjects) according to a binomial probability π = P[X =1|diseased image] = P[X =0|healthy image]. One can visualize the entire data set composed of two subsets one consisting of healthy images only and disease image. Simulated data sets were generated with π = 0.50, 0.60, 0.70, 0.80, 0.90, 0.95 and 0.99, and β = 0.50, 0.70, 0.90 and 0.95 where β = proportion of images from the normal (healthy) subjects. It is not necessarily true tah probability π = P[X =1|diseased image] = P[X =0|healthy image] for each rater, but even this simple assumption can be used to generate imbalanced data. Our goal here is to simply demonstrate the vulnerability of FK and robustness of AK in certain types of data. All simulations and subsequent computations were carried out using SAS 9.2.

Table 2: Comparison of A-Kappa and Fleiss’ Kappa using 10,000 images (samples)

	Percent images		Fleiss Kappa with given β				
π	9+ raters agree	A-Kappa	0%	0.50	0.70	0.90	0.95
0.5	3.93	0.011	0.011	0.011	0.011	0.011	0.011
0.6	5.22	0.046	0.009	0.046	0.041	0.023	0.018
0.7	16.08	0.168	0.011	0.168	0.147	0.075	0.046
0.8	37.32	0.365	0.016	0.365	0.327	0.179	0.108
0.85	54.27	0.495	0.019	0.495	0.453	0.269	0.167
0.90	73.25	0.646	0.027	0.646	0.606	0.402	0.268
0.95	91.00	0.813	0.032	0.813	0.785	0.614	0.458
0.99	99.05	0.960	0.044	0.960	0.952	0.896	0.820

π = Probability of classifying an image correctly

β = Proportion of images from normal subjects

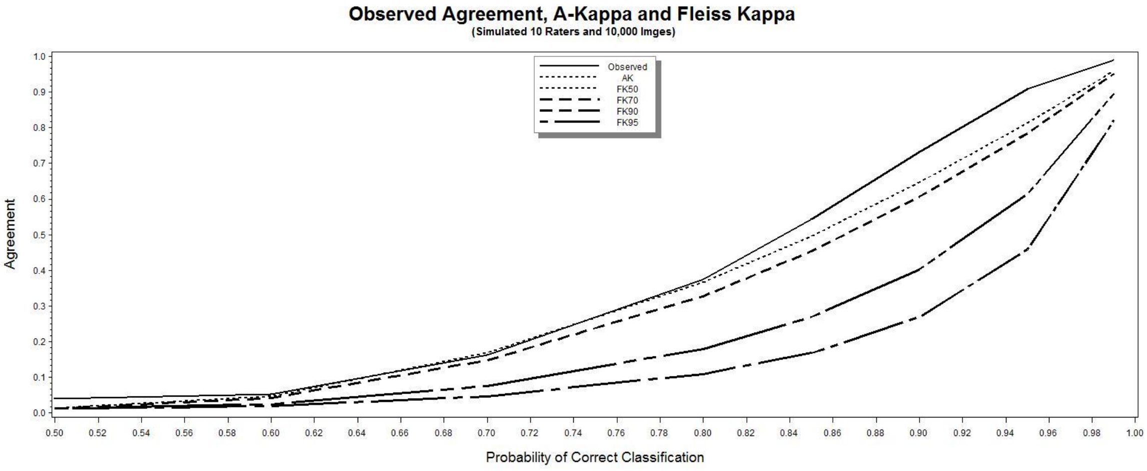
Note that, this is not the unique way to generate columns (or rows) of 0 and 1 in order to show that AK may fail to reflect the observed agreement in ‘unbalanced’ data.

2.4 Performance of A-Kappa vs Fleiss KappaTable 2 presents results of simulations described above.

The first column of Table 2 shows values of π, the second column shows percent to images on which 9 (90%) or more rater agree. This is taken as a measure of crude or observed agreement

A quick glance at Table 2 shows that when $\pi = 0.5$ both AK and FK indicate absence of agreement. This is a situation when raters classify images randomly. However, when π is different from 0.5 then FK varies with the proportion of positive images while AK remains the same for a given π. When the proportion of samples of diseased images is equal to the samples of healthy images then both AK and FK yield the same value for all π. On the other hand, when a majority of images are from diseased patients (or from healthy patients) then FK is further from the observed (or crude) agreement than AK. In such situations AK captures the observed agreement better than FK. For example, when $\pi = 0.90$, then observed agreement = 0.7325 (at least 90% of raters agree on 73.25% images) and A-Kappa = 0.646, but Fleiss Kappa could be as low as 0.03 when almost all images are either positive or negative and could be as high as 0.646 (value of AK) when 50% images are positive.

The above results showing difference between AK and FK are also depicted in Figure 1. When the underlying proportion of positive images is roughly 0.5 ($\beta = 0.5$), the lines for AK and FK are the same for all values of π. That is why lines for AK and FK50 (i.e. FK values when 50% images are positive) are not distinguishable in Fig. 1. Also, AK and FK are the same (and near zero) irrespective the proportion of diseased images when raters assign scores of 0 or 1 to an image randomly (i.e. when $\pi = 0.50$).



Note: FKx= Fliess Kappa from a data set with x% positive images

Figure 1

2.5 An Interpretation of A-Kappa (AK)

Landis and Koch (1977) provided an interpretation of Kappa statistic. However, those interpretations are considered arbitrary. Except for Kappa =1 and 0 implying perfect and chance agreement, respectively, other value fail to convey the degree of agreement in terms of an interpretable scale. The following discussion may help shed some light into interpretation of AK.

Assume that the same image is evaluated by a set of r raters by assigning 0 or 1 to the image for the presence and absence of a disease. Also, assume that some time has elapsed between two evaluations.

Proposition 6. Let a_i and $\hat{p}_i = a_i / r$ denote the number and proportion of raters who assigned 1 during the *i*th evaluation. If it is assumed that the sequence of 0 and 1 are generated by an underlying Bernoulli distribution with a parameter $p = E[\hat{p}_i]$, then

$$E[AK] = (2p - 1)^2 \tag{2.8}$$

(See appendix for a proof of this result).

Equation (2.8) may shed light into the interpretation of an AK value as

$$p \approx (1 + \sqrt{AK})/2 \,. \tag{2.9}$$

Thus for an AK value, we can say one would have obtained the same AK if each rater would classify a diseased image correctly with the probability given by equation (2.9). However, it is not necessarily true that the data in hand was generated with this probability.

Equation (2.8) indicates that when 90% of the raters assign 1 or 90% raters assign 0 (say, to all the images), then AK is about 0.64). In the case of two raters the proportion of images on which two raters agree is expected to be 0.9×0.9+0.1×0.1 = 0.81+0.01 = 0.82 so that AK = 2(0.82) - 1 = 0.64 (see equation 2.2). On the other hand, if data yields AK = 0.64, then one would obtain the same AK from data set where probability of correctly classifying an image by each rater is 0.90. No such interpretation exists for FK where the interpretation is completely arbitrary (Landis et al 1977).

3. Multiple Rates and Multiple Categories

Although the main focus of the paper is evaluation of agreement among multiple of raters (or situations) on two possible classification (or categories), results presented in previous sections are briefly extended to multiple category situation in the following sections. Some additional insights on AK are also presented.

3.1 Derivation of A-Kappa (AK) for multiple categories and multiple raters

Assume that each rater classifies an image into one of the k categories. Let a_{ij} denote the number of raters who classified the ith $(i = 1,2,...,N)$ image into the jth $(j = 1,2,...,k)$ category so that $\sum_{j=1}^{k} a_{ij} = r$ for each image. In case of complete agreement on the ith image, all raters will classify the image into the same category. In the case of complete lack of agreement the ith image will be categorized into each category by an equal number of raters. Therefore in this case, r/k raters are expected to classify such an image into each of the k categories. One can think of agreement as the discrepancy or distance from complete disagreement. This discrepancy may then be expressed as

$$\sum_{j=1}^{k}\left(a_{ij} - r/k\right)^2 = \sum_{j=1}^{k} a_{ij}^{\ 2} - r^2/k \tag{3.1}$$

This quantity can be rescaled by dividing it by its maximum possible value so that the distance between observed data and the state of complete disagreement lies between 0 and 1. The maximum value of the expression given in (3.1) is given by

$$r^2 - r^2/k = r^2(k-1)/k$$

Therefore, the rescaled distance for the ith image is

$$G_i = k\sum_{j=1}^{k}\left(a_{ij} - r/k\right)^2 /[(r^2(k-1)] \tag{3.2}$$

Hence, the mean of G_i across N images can be considered as the 'crude' agreement among raters. This is given by

$$\overline{G} = \sum_{i=1}^{N} G_i / N = k\sum_{i=1}^{N}\sum_{j=1}^{k} a_{ij}^{\ 2} /(Nr^2(k-1)) - 1/(k-1) \tag{3.3}$$

However, some of this agreement may be due to chance. Opinions differ regarding the definition of chance induced agreement. For example, Maxwell uses 0.5 as a chance induced agreement and Cohen uses the marginal probabilities of the 2×2 table under consideration. Another way to quantify the agreement due to chance might be to estimate the agreement expected in a sample that comes from a population lacking agreement among raters. In terms of the notation used above, error due to chance may be given by the expected value of $\overline{G}$ given that there is an absence of agreement in the population.

Proposition 7. The expected value of $\overline{G}$ given that there is an absence of agreement in the population is given by

$$E[\overline{G}] = 1/r \tag{3.4}$$

(See appendix for a proof)

A-Kappa (AK) is the measure $\overline{G}$ adjusted for agreement due to chance and rescaled to yield the maximum possible value of 1 is, given by :

$$AK = (\overline{G} - 1/r)/(1 - 1/r) = (r\overline{G} - 1)/(r - 1) \tag{3.5}$$

The functional forms of AK for two raters and multiple raters seem different, but in fact they are the same as shown below. A-Kappa for two categories was developed first and then it was shown as an extension of Maxwell's Random Error (RE). Since the error due to chance was already imbedded in Maxwell RE, no error adjustment was discussed.

Proposition 8. When k = 2, then

$$(r\overline{G} - 1)/(r - 1) = \sum_{i=1}^{N} [(2a_i - r^2) - r] / N(r^2 - r) \tag{3.6}$$

which is the same as equation 2.1 (A proof is given in the appendix).

3.1.1 Agreement on individual item (image)

Note that, A-Kappa proposed in this article is the average of $(rG_i - 1)/(r - 1)$ across the observations (images), where G_i is defined by equation (3.2). Therefore, this quantity could be considered as a measure of agreement among raters on the *i*th observation (image).

$$\text{Let } AK_i = (rG_i - 1)/(r - 1), \text{ then } AK = \sum_{i=1}^{N} AK_i / N \tag{3.7}$$

This characteristic of AK is similar to the agreement index proposed by O'Connell and Dobson (1984), but AK is much simpler to compute and uses a different strategy to estimate chance induced agreement. One advantage of obtaining agreement on an individual image (observation) is that the investigators could identify and investigate images with high disagreement. This could especially be useful when training novice raters. Equation (3.6) also points that AK from two or more data sets could be easily combined to yield the overall AK from the combined data set as shown below. Let a data set of sample size $(N_1 + N_2)$ be partitioned into data sets having sample sizes N_1 and N_2. If AK_1, AK_2 and AK_c are AK from the first set, second set and combined set of data, respectively, then $AK_c = (N_1 AK_1)/(N_1 + N_2) + (N_2 AK_2)/(N_1 + N_2)$. Thus, AK from a combined data set is the weighted average of AKs from the component data sets. This is not necessarily true for FK.

3.2 Asymptotic distribution of A-Kappa (AK)

Following the notations of the previous sections, suppose *r* raters classify each of the *N* images into one of the *k* categories. Suppose that the number of ratings $a_{i1}, a_{i2}, \ldots, a_{ik}$

corresponding to the *i*th image (observation) have a multinomial distribution with probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, ..., \pi_{ik})'$. Let $r = a_{i1} + a_{i2} + ... + a_{ik}$, and let $\mathbf{p}_i = (p_{i1}, p_{i2}, ..., p_{ik})'$ denote the sample proportion (proportions of raters), where $p_{ij} = a_{ij} / r$.

Proposition 9. The Asymptotic variance of AK is given by

$$V(AK) = 4rk^2 \sum_{i=1}^{N} \left[\sum_{j=1}^{k} p_{ij}^3 - (\sum_{j=1}^{k} p_{ij}^2)^2 \right] / [N^2 (r-1)^2 (k-1)^2] .$$

(A proof is given in the appendix).

3.3 Simulations for Multiple Categories

Simulations results from earlier section have shown that in the case of two categories, FK and AK are equivalent when there is absence of agreement or when the data are symmetrical. Otherwise FK may fail to reflect the high degree of observed agreement.

Here, we present a few simulations using multiple raters and multiple categories. These simulations show that as with two categories, FK may fail to reflect a high observed agreement in case of multiple categories. We generated 10,000 items (images) and assumed that each of the 10 raters classified each image into one of five categories. Let p_i denote the probability with which a rater assigns an image into the *i*th $(i = 1,2,3,4,5)$ category. When each image is randomly assigned into one of these categories, i.e., when p_i= 0.2 for all *i*, then AK= FK = 0.0004. In this case, both indices truly reflect the absence of agreement among the raters. Next, suppose that 10000 images of category 4 are evaluated by 10 raters, and each rater can correctly classify the images with a probability 0.9. For a simple example, let $p_1 = p_2 = p_3 = p_5 =$ 0.025 and $p_4 = 0.90$, then FK = 0.013 and AK = 0.770. Under this simulated scenario, at least 8 raters (80% or more) are found to classify 9,433 (94.33%) images into the 4th category. Therefore, FK fails to reflect high degree of agreement among the raters.

Next, we simulated that where with 50% of the images 4 and 50% were of category were of category 2, and the raters can correctly classify the image with probability of 0.9 into these two categories (with remaining probability equally distributed over remaining categories). In this case, the simulated data showed that at least 8 raters (80% or more raters) classified 4,486 images in category 2 and 4,699 images in category 4 so that at least 80% of the raters agreed on 9,184 (91.84%) images. FK for this data turned out to be 0.674 while AK = 0.770. Thus this balancing in the data brought FK value almost to the level of AK, while AK remained the same.

Hence, in the case of more than two categories and multiple raters FK may fail to reflect the high degree of observed agreement in asymmetric data, while AK may not be influenced by such asymmetry

3.4 Real Example (multiple categories)

Consider the study described in the introduction section. Ten raters were asked to classify the breast composition of 102 images into the four categories: the breast is almost entirely fat (< 25% glandular), SFD: scattered fibroglandular densities (approximately 25-50% glandular), HD: the breast tissue is heterogeneously dense (approximately 51% – 75% glandular), ED: the breast tissue is extremely dense (> 76% glandular). Table 3 shows AK and FK among multiple raters and multiple categories.

Note that, FK indicates that the raters have poor agreement on whether the images are fatty or not. AK on the other hand shows there is an excellent agreement among the raters. Most of the raters classify images into non-Fatty categories. In Table 3, both AK and FK are 0.403 when classifying an image into heterogeneously dense (HD).

Table 3: Agreement among raters on breast composition classifications

Classification	Agreement Index	
	A-Kappa	Fleiss Kappa
Fatty	0.906	0.119
SFD	0.583	0.466
HD	0.403	0.403
ED	0.715	0.570
Overall	0.534	0.454

This indicates that if the data were re-arranged into HD versus non-HD category, then it indicates perhaps the numbers HD and non-HD images are similar. In all other situations we have AK > FK indicating lack of such symmetry. However, the asymmetry is not substantial except for Category 1 (Fatty vs non-Fatty). In conclusion, if only FK was used we might have been misinformed about the agreement among the raters

4. Discussion

Cohen’s Kappa (CK) is used routinely to evaluate agreement between two raters or two conditions, but has been criticized for being simply a function of prevalence, and counter-intuitive by several investigators. Using simulations, it is shown in this article that Fleiss Kappa (FK) a measure of agreement among multiple raters inherits some of these shortcomings of CK.

A new and simple method for evaluating agreement A-Kappa (AK) among multiple raters is proposed. This method reduces to Maxwell’s Random Error (RE) proposed to address the high agreement low kappa paradox in case of two raters. In this article it is shown, by simulations, that Fleiss’ kappa (FK) may also yield low kappa although there is a high degree of agreement among the raters. This is especially true in the case of imbalanced data where one class of items is relatively less than the other.

AK, proposed in this paper, may be used as an alternate or an additional index in the case of multiple raters. The proposed measure does not have the seemingly paradoxical characteristic of FK. Computing both AK and FK may provide additional insight. The difference between the two values may indicate whether the data are dominated by one kind of classification of image. As indicated by simulations, FK coincides with AK when proportions of positive and negative image are the same. A small FK may not really be an indication of low agreement, while a small AK is indication of low agreement. Also, AKs from two data sets can be easily combined to yield the AK from the combined data set. We recommend calculating both AK and FK.

In the case of two categories, A-Kappa may have a meaningful interpretation in a more familiar scale of probability as discussed in this paper. Existing interpretation of Kappa values are considered somewhat arbitrary. Computation of both AK and FK may further shed light into the data and be useful in the interpretation and presenting the results. This is similar to recommendation by of computing both maximum and minimum Kappa in two raters two categories situation.

Though not the focus of the paper, there exists a body literature with model based approaches to evaluate agreement (Agresti, 1992 and 2002; Tanner et al 1985). Similar to the kappa-like indices, most of the model based methods have also dealt with the situation of two raters, and the number of parameters to be estimated increases exponentially with number of raters creating computational challenges. Estimating equation approaches are also proposed to model agreement in data with multiple raters having binary and multiple categories (Williamson et al, 2000, Klar et al, 2000). AK will also be examined from repeated measures viewpoint in a future study. However, investigators especially in biomedical studies still routinely use CK and FK to evaluate agreement. This paper highlights some situations where these methods may fail to capture the agreement and propose an alternative method which reduces to existing methods proposed in two raters’ situations.

Current limitations of AK include its inability to incorporate raters’ characteristics. However, this is also the case with FK and CK. Evaluation AK with repeated measures (or hierarchical) approach is being explored. This will allow us to adjust for confounding factors. Despite such limitations its simplicity and intuitive nature, it provides some insights into the nature of agreement and the data set itself. It is very simple to calculate.

Appendix A

Proposition 1. In the case of two raters, $AK = 2P_0 - 1 = RE$.

Proof: From equation 2.1, in the case of two categories and r raters,

$$AK = \sum_{i=1}^{N} [(2a_i - r)^2 - r] \big/ [N(r^2 - r)]$$

Noting that when *r* =2 then,

$$[(2a_i - r)^2 - r]/(r^2 - r) = (2a_i - r)^2 \big/ [(r(r-1)] - 1/(r-1) \;= 2(a_i - 1)^2 - 1$$

Also, when there are only two raters, then a_i (number of raters who agree on the *i*th image) takes values 0, 1 or 2. So, the term $(a_i - 1)^2$ = 1 when the two raters agree, and $(a_i - 1)^2$ = 0 when the two raters disagree.

Therefore, when *r* = 2,

$$AK = \sum_{i=1}^{N} [(2a_i - r)^2 - r]/[N(r^2 - r)] = 2P_0 - 1 = AE$$

where P_0 is the proportion of images on which both raters agree.

Proposition 2. If w_i is the proportion of pairs of raters who agree and v_i is the proportion of pairs who disagree on the *i*th image, then $AK = \sum_{i=1}^{N} (w_i - v_i)/N$

Proof: Assume that a_i raters out of *r* classify the *i*th image into disease category (by assigning a score of 1) and remaining *r* - a_i into the non-disease category (assigning a score of 0). Let $C(x,2) = x(x-1)/2$.

Then both members of $C_2^{a_i}$ pairs of raters will assign 1 and both members of $C_2^{r-a_i}$ pairs will assign 0 to the *i*th image. Similarly, one member of $a_i \times (r - a_i)$ pairs will assign 1 while the other member will assign 0 to the *i*th image. Therefore,

$$w_i = [C(a_i,2) + C(r - a_i,2)]/C(r,2) \text{ and } v_i = [a_i(r - a_i)]/C(r,2)$$

Hence difference in proportion of images on which there is pair-wise agreement and disagreement on the *i*th image is given by

$$\begin{aligned} w_i - v_i &= [C(a_i,2) + C(r - a_i,2) - a_i(r - a_i)]/C(r,2) \\ &= [a_i(a_i - 1) + (r - a_i)(r - a_i - 1) - 2a_i(r - a_i)]/[r(r-1)] \\ &= \sum_{i=1}^{N} [(2a_i - r)^2 - r]/[r(r-1)] \end{aligned}$$

Thus,

$$\sum_{i=1}^{N}\left(w_i - v_i\right)/N = \sum_{i=1}^{N}[(2a_i - r)^2 - r]/[N(r^2 - r)] = AK$$

Proposition 3. AK for multiple raters r is the average of *RE* for all possible pairs of raters. Let RE_{ij} denote Maxwell's Random Error from the *i*th and *j*th raters. Similarly, let $P_{0,ij}$ denote the proportions of images on which both raters agree. Then

$$AK = \sum_{i>j}^{r}(2P_{0,ij} - 1)/C(r,2) = \sum_{i>j}^{r} RE_{ij}/C(r,2)\text{, where } C(r,2) = r(r-1)/2\,.$$

Proof: Assume that readings of *N* images by *r* raters are presented in *N* rows and *r* columns. Let the *r* columns be denoted by $X_1, X_2, \ldots X_r$. Also assume X_h (representing the *h*th rater) takes values either 1 or 0. Next consider $r(r-1)/2$ variables $V_{12}, V_{13}, \ldots V_{r,r-1}$ such that

$$V_{st} = X_s X_t + (1 - X_s)(1 - X_t)\text{, } t>s$$

Note that, V_{st} = 1 when X_s and X_t both take 1 or both take 0 otherwise $V_{st} = 0$. The AK coefficient from the *s*th and *t*th rater is then $(2\overline{V}_{st} - 1)$ where $\overline{V}_{st}$ is the average of V_{st} taken across *N* images (or observations), and can be expressed as $\overline{V}_{st} = \sum_{i=1}^{N} V_{sti}/N$.

We need to show that $\sum_{t>s}^{r}(2\overline{V}_{st} - 1)/C(r,2) = AK = \sum[(2a_i - r)^2 - r]/[N(r^2 - r)]$

$$\text{Since} \sum_{t>s}^{r}\overline{V}_{st} = (\sum_{t>s}^{r}\sum_{i=1}^{N} V_{sti})/N = (\sum_{i=1}^{N}\sum_{t>s}^{r} V_{sti})/N = \sum_{i=1}^{N}[C(a_i,2) + C(r - a_i)]/N$$

$$= \sum_{i=1}^{N}\left[a_i(a_i - 1) + (r - a_i)(r - a_i - 1)\right]/2N$$

$$\text{We have, } \sum_{t>s}^{r}(2\overline{V}_{st} - 1)/C(r,2) = [4\sum_{t>s}^{r}\overline{V}_{st}]/[r(r-1)] - 1$$

$$= 4/[Nr(r-1)]\sum_{i-1}^{N}[a_i(a_i - 1) + (r - a_i)(r - a_i - 1)]/2 - 1$$

$$= \sum_{i=1}^{N}[(2a_i - r)^2 + r^2 - 2r]/[Nr(r-1)] - 1 = \sum[\{(2a_i - r)^2 + r^2 - 2r\}/\{N(r(r-1)\} - 1]$$

$$= \sum_{i=1}^{N} [(2a_i - r)^2 - r]/[N(r^2 - r)] = AK$$

Proposition 6. $E[AK] = (2p-1)^2$

Proof: Note that, by definition, $AK = \sum_{1}^{N} [(2a_i - r)^2 - r]/[N(r^2 - r)]$.

Also, $[(2a_i - r)^2 - r]/(r^2 - r) = [r/(r-1)][(2a_i/r-1)^2 - 1/(r-1)$

$= [r/(r-1)](4a_i^2/r - 4a_i + 1) - 1/(r-1)$

Under the assumption that each rater will assign a score of 1 with probability $p, a_i \sim B(r,p)$. Therefore, $E[a_i] = rp$, $Var(a_i) = E[a_i^2] - \{E[a_i]\}^2$ so that $E[a_i^2] = rp(1-p) + (rp)^2$.

Using this information, we have $E[AK] = (2p-1)^2$.

Proposition 7. The expected value of $\overline{G}$ given that there is an absence of agreement in the population is given by $E[\overline{G}] = 1/r$

Proof: From equation (3.2),

$$G_i = [k\sum_{i=1}^{N} a_{ij}{}^2]/[r^2(k-1)] - 1/(k-1) = k\sum_{i=1}^{N} (a_{ij} - r/k)^2 /[r^2(k-1)].$$

So that

$$r(k-1)G_i = \sum_{i=1}^{N} (a_{ij} - r/k)^2 /(r/k)$$

Note that, under the assumption of random classification of images by the raters, *i*th image will be classified into each category by r/k raters. Therefore, $r(k-1)G_i$ is distributed as χ^2 with $(k-1)$ degrees of freedom, and its expected value will be $(k-1)$. Hence, $E[r(k-1)G_i] = k-1$ which implies that $E[G_i] = 1/r$. Therefore,

$$E[\overline{G}] = \sum_{i=1}^{N} E[G_i]/N = 1/r \cdot$$

Proposition 8. When k = 2,

$$(r\overline{G} - 1)/(r-1) = \sum_{i=1}^{N} [(2a_i - r^2) - r]/N(r^2 - r),$$

where $\overline{G} = \sum_{1=1}^{N} G_i/N = k\sum_{i=1}^{N}\sum_{j=1}^{k} a_{ij}{}^2/(Nr^2(k-1)) - 1/(k-1)$.

Proof: Note that, from equation (3.2), $G_i = k\sum_{j=1}^{k}\left(a_{ij} - r/k\right)^2/[(r^2(k-1)]$.

When k = 2, then $G_i = 2(a_{i1}^2 + a_{i2}^2)/r^2 - 1$. Noting that $a_{i1} + a_{i2} = r$, and dropping the second subscript, we have

$$G_i = 2[a_i^2 + (r - a_i)^2]/r^2 - 1 = (2a_i - r)^2/r^2.$$

So that,

$$(rG_i - 1)/(r-1) = [(2a_i - r)^2 - r]/r(r-1).$$

Therefore,

$$(r\overline{G} - 1)/(r-1) = \sum_{i=1}^{N}[(2a_i - r^2) - r]/N(r^2 - r).$$

Proposition 9. Asymptotic variance of AK is given by

$$V(AK) = 4rk^2\sum_{i=1}^{N}[\sum_{j=1}^{k}p_{ij}^3 - (\sum_{j=1}^{k}p_{ij}^2)^2]/[N^2(r-1)^2(k-1)^2]$$

Proof: Recall that *r* raters classify each of the *N* images into one of the *k* categories. Suppose that the number of ratings $a_{i1}, a_{i2}, \ldots, a_{ik}$ corresponding to the *i*th image (observation) have a multinomial distribution with probabilities $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik})'$. Let$r = a_{i1} + a_{i2} + \cdots + a_{ik}$, and let $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{ik})'$ denote the sample proportion (proportions of raters), where $p_{ij} = a_{ij}/r$. Then it follows that

$$\sqrt{r}(\mathbf{p}_i - \boldsymbol{\pi}_i) \xrightarrow{d} N[\mathbf{0}, \mathbf{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}_i'],$$

where $\mathbf{diag}(\boldsymbol{\pi}_i)$ is a diagonal matrix with matrix with elements of $\boldsymbol{\pi}_i$ on the main diagonal. Under the assumption that images are independent, the variance covariance matrix for the entire sample will be a block diagonal matrix with each block being of the form expressed as in (3.6). The following results can be used to estimate asymptotic variance of AK.

From equation (3.2), for the *i*th image,

$$G_i = (k\sum_{i=1}^{N}a_{ij}{}^2)/[r^2(k-1] - 1/(k-1) = (k\sum_{i=1}^{N}p_{ij}{}^2)/(k-1) - 1/(k-1)$$

$$= (k\mathbf{p}_\mathbf{i}'\mathbf{p}_\mathbf{i} - 1)/(k-1) = f(\mathbf{p}_i)$$

$$\left.\frac{\partial f}{\partial \mathbf{p}_i}\right|_{\mathbf{p}_\mathbf{i} = \boldsymbol{\pi}_\mathbf{i}} = (2k\boldsymbol{\pi}_\mathbf{i})/(k-1)$$

Therefore, from the Delta method,

$$\sqrt{r}(G_i - f(\boldsymbol{\pi}_i)) \xrightarrow{d} N[0, \{4k^2/(k-1)^2\}\{\boldsymbol{\pi}'_{\mathbf{i}}(\mathbf{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}'_i)\boldsymbol{\pi}_{\mathbf{i}}\}]$$

Hence the variance of $\sqrt{r}G_i$ denoted as $V(\sqrt{r}G_i) = \{4k^2/(k-1)^2\}\{\boldsymbol{\pi}'_{\mathbf{i}}(\mathbf{diag}(\boldsymbol{\pi}_i) - \boldsymbol{\pi}_i\boldsymbol{\pi}'_i)\boldsymbol{\pi}_{\mathbf{i}}\}$

Or $rV(G_i) = 4k^2/(k-1)^2[\sum_{i=1}^{k} \pi_{ij}^3 - (\sum_{j=1}^{k} \pi_{ij}^2)]$

Therefore, the variance (asymptotic) A-Kappa statistic for the *i*th image is given by

$$V(AK_i) = (rG_i - 1)/(r-1)$$

$$= V(G_i)[r^2/(r-1)^2 = 4rk^2[\sum_{j=1}^{k} \pi_{ij}^3 - (\sum_{j=1}^{k} \pi_{ij}^2)^2]/[(r-1)^2(k-1)^2]$$

Assuming independent images and noting that that the overall A-Kappa for the given data set is the average of AK across the images (observations) we have,

$$V(AK) = V(\sum_{i=1}^{N} AK_i)/N = V(\sum_{i=1}^{N} (AK_i)/N^2$$

$$= 4rk^2 \sum_{i=1}^{N} [\sum_{j=1}^{k} p_{ij}^3 - (\sum_{j=1}^{k} p_{ij}^2)^2]/[N^2(r-1)^2(k-1)^2]$$

Since parameters π_{ij} are generally unknown, they are replaced by their sample estimates $p_{ij} = a_{ij}/r$ in the above equation.

References

[1] Agresti, A. (1992). Modeling patterns of agreement and disagreement. *Statistical Methods in Medical Research* **1**, 201-218.

[2] Agresti, A. (2002). *Category Data Analysis* (Second Edition). Wiley: New York.

[3] Cicchetti, D.V. and Feinstein, A.R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal Clinical Epidemiology* **43**, 551-558.

[4] Feinstein, A.R. (1985). A bibliography of publications on observer variability. *J Chron Dis* **38**, 619-632.

[5] Feinstein, A.R. and Cicchetti, D.V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal Clinical Epidemiology* **43**, 543-549.

[6] Fleiss, J.L. (1981) *Statistical methods for rates and proportions*. Wiley: New York, pp. 38-46.

[7] Klar, N.. Lipsitz, S.R. and Ibrahim, J. (2000). Estimating equation approach for modeling kappa. *Biometrical Journal* **42**, 45-58

[8] Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**,159-74.

[9] Lanz, C.A. and Nebenzahl, E. (1996) Behavior and interpretation of the kappa statistic: resolution of the two paradoxes. *Journal Clinical Epidemiology* **49**, 431-434.

[10] Li, D.C. Liu,C.W. and Hu, S.C. (2010). A learning method for the class imbalance problem with medical data sets. *Comput Biol Med* **40**,509-18.

[11] Maxwell, A.E. (1977). Coefficients of agreement between observers and their interpretation. *The British Journal of Psychiatry* **130**, 79-83.

[12] O'Connell, D. L. and Dobson, A.J. (1984). General Observer-Agreement Measures on Individual Subjects and Groups of Subjects. *Biometrics* **40**, 973-983.

[13] Tanner, M.A. and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American statistical Association* **80**, 175-180.

[14] Williamson, J.M., Manatunga, A.K. and Liptsitz, S.R.(2000). Modeling kappa for measuring dependent categorical agreement data. *Biostatistics* **1**,191-202.

Received October 24, 2013; accepted April 26, 2014.

Shiva Gautam
Beth Israel Deaconess Medical Center
Harvard Medical School
330 Brookline Avenue, Boston, MA 02215
sgautam@bidmc.harvard.edu