Association or Agreement

K S Chia,**FAMS, MD, MSc (OM)*

Introduction

Association and agreement between two factors are very different concepts, although the methods used to describe them appear similar. For example, the association between levels of aflatoxin serum albumin adducts and dietary aflatoxin intake can be demonstrated in a bivariate scattergram; the agreement between two methods of measuring aflatoxin serum albumin adducts can also be illustrated using a similar diagram (Fig. 1). Similarly, a cross-tabulation is used to assess the association between human papiloma virus (HPV) infection and the occurrence of cervical cancer; it is also used to assess the agreement between two radiologists in classifying mammograms (Table I).

The distinction becomes clear when the underlying research question is clarified. Association deals with the relationship between and exposures and an outcome. Agreement on the other hand assesses the reliability between two methods of assessing exposure (or outcomes). Association answers questions related to aetiology. Agreement answers questions on reliability between assessment methods.

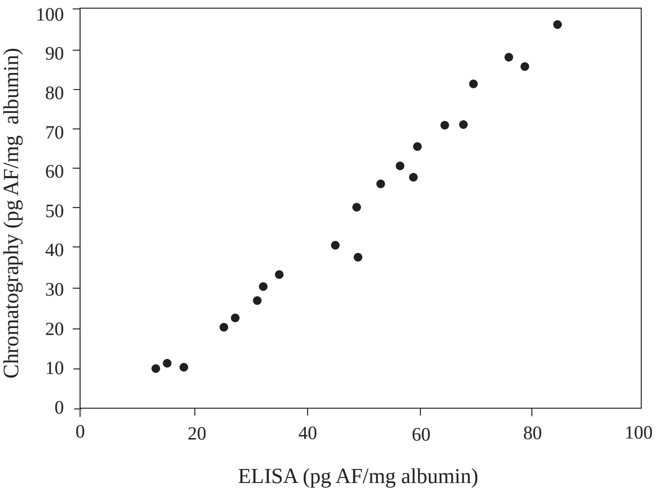


Fig. 1. Scattergram showing agreement between two methods of assaying aflatoxin-albumin adducts.

Although the methods used to describe the association or agreement are similar (scattergrams and cross-tabulations), the summary indices used are different. Relative risks (incidence rate ratios), odds ratios and prevalence rate ratios are used to describe associations in cohort, case-control and cross-sectional studies respectively. Pearson's correlation coefficients (r) with linear regression lines are used to describe associations between variables measured on interval/ratio scales. On the other hand, for agreement, kappa (k), limits of agreement and intraclass correlation coefficient (r_I) are used.

Assessing Agreement from a Cross-tabulation

Two radiologists independently read 102 mammograms and classified them as normal, benign, suspected cancer, definite cancer (Table I). In this case, the c^2 test is not used as the research question is not on association between the radiologists findings. The research question is on how well the radiologists agree in their classification.

The observed agreement (OA) between the two radiologists can be summarised easily as the proportion of mammograms where both radiologists agreed on (i.e. [30+25+20+3]/102 = 0.76). However, some agreement could have occurred by chance. This expected agreement (E_A) can be calculated by deriving the expected numbers in the concordant series by taking into account the corresponding row and column totals. In this case, the

TABLE I: ASSESSING AGREEMENT IN INTERPRETING MAMMOGRAMS BETWEEN TWO RADIOLOGISTS

	Radiologist A				
Radiologist B	Normal	Benign	Suspected	Cancer	Total
Normal	30	6	0	0	36
Benign	4	25	4	0	33
Suspected	0	8	20	2	30
Cancer	0	0	0	3	3
Total	34	39	24	5	102

Percentage agreement = 76.5%
Kappa (k) = 0.66

* *Associate Professor*
Department of Community, Occupational and Family Medicine
National University of Singapore
Address for Reprints: Dr Chia Kee Seng, Department of Community, Occupational and Family Medicine, National University of Singapore, Lower Kent Ridge Road, Singapore 119260.

expected numbers are 12.00, 12.62, 7.06 and 0.15 giving an E_A of 0.31. Intuitively, the agreement beyond chance is 0.76 - 0.31 which is 0.45. However, this index is difficult to interpret because different pairs of O_A and E_A will give the same value (e.g. 0.78, 0.33 and 0.56, 0.11).

Complete agreement occurs when all the observations fall on the concordant series and E_A=1. It can also be seen intuitively that the potential for agreement beyond chance is given by 1-E_A. Kappa then describes the actual agreement beyond chance relative to the potential agreement beyond chance (i.e. $k = [O_A - E_A]/[1-E_A]$).

k varies from -1.0 to +1.0 with the value of 1.0 indicating perfect agreement and 0 as no agreement. As a general rule k>0.8 is considered high agreement, 0.6 to 0.8 as good agreement, 0.4 to 0.6 as fair agreement and below 0.4 is poor agreement.[1] k is only a summary statistic and should always be presented with the cross-tabulation.

Apart from relying on k, it is important to study the distribution of discordant pairs in the cross-tabulation. For example, if radiologist A classified all the 10 discordant mammograms as cancer whereas radiologist B maintained that they were normal, it shows that the agreement is poor even though the k would have changed marginally and still suggests good agreement.

Assessing Agreement from Scattergrams

Many authors would compare the means derived by the two methods and conclude that if the difference between the two means are not statistically significant, then the two methods agree with each other. However, a small difference with a large number of data points can give a statistically significant result even though intuitively, the difference is of no practical significance. Conversely, a large difference may not be statistically different because of small number.

Another common error is to calculate a Pearson's correlation coefficient (r) and conclude that a r value close to +1 suggests that the two methods agree with each other. However, if method A consistently gives a reading which is twice that of method B, the two methods are clearly not in agreement. Therefore, we can only conclude that two methods are agreeable if the data points are very close to the regression line, y=x. The intraclass correlation coefficient (r_I) is a summary index which measures how well the data points fall on the y=x regression line. A r_I of 0.75 and above suggests high agreement.[2]

Just like the k, the r_I is a useful index for agreement but it should not be used alone. Even with a high r_I, there is a need to evaluate if one method consistently gives a higher reading than the other. The best way is to plot a graph with the difference in readings on the Y-axis and the mean of the readings on the X-axis. Ideally, all the points should cluster on either side of the line Y=0. Figure 2 shows that ELISA gives higher readings than chromatography when the readings are low and the reverse when the readings are high. This suggests that even though the r_I is high, the two methods are not in agreement.

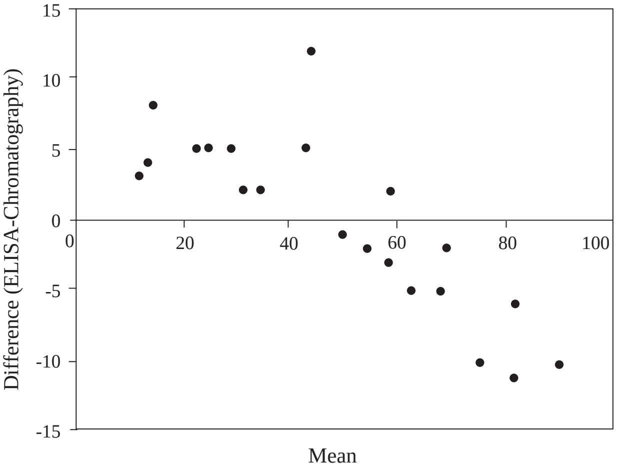


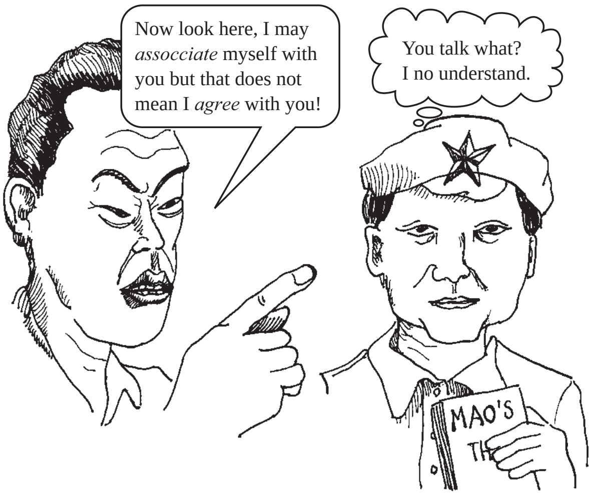
Fig. 2. Relationship between mean and difference in aflatoxin-albumin adducts as measured by ELISA and chromatography.

Conclusion

Examining the scattergram and cross-tabulation intuitively often provides much information about agreement between two methods. The summary indices, k (for agreement between nominal scale variables) and r_I (for agreement between interval/ratio scale variables) provide additional supportive evidence but should not replace careful study of the data.

REFERENCES

1. Landis J R, Koch G G. The measurement of observer agreement for categorical data. Biometrics 1977; 33:159-74.
2. Burdock E L, Fleiss J L, Hardesty A S. A new view of interobserver agreement. Perspect Psychol 1963; 16:373-84.


Now look here, I may *assocciate* myself with you but that does not mean I *agree* with you!
You talk what? I no understand.
MAO'S TH

Agreement and association are very different concepts.