Statistical methods for assessing observer variability in clinical measures

Paul Brennan, Alan Silman

The lack of consistency of much medical judgment and decision making has long been appreciated. This is now recognised as an important source of error, and attempts to quantify it have been enhanced by an increasingly sophisticated statistical toolbox at the disposal of the clinical investigator. Variability in measurement and classification may arise from two sources, (*a*) a lack of consistency within an individual observer (or measuring process) when carrying out successive recordings, and (*b*) a lack of consistency between observers. Assessing lack of consistency between observers is important for two reasons. Firstly, in any single study using more than one observer an assessment of the variation between them is essential in interpreting the data derived. Secondly, the observer variation seen in one study may be extrapolated to other studies of the same technique but using different observers—that is, the origin of the variation may be inherent in the method itself. This review attempts to summarise in simple terms the statistical techniques available to quantify the variation within and between observers.

Categorical measurements

Many clinical measures allocate an individual to one of a number of categories either unranked (nominal) or ranked (ordinal). Evaluations of between observer and within observer variation of such measures traditionally relied on the use of the percentage level of agreement.[1] Consider the example in table I, where two rheumatologists have each independently classified the hand *x* ray appearances of 100 patients with rheumatoid arthritis for the presence or absence of erosions. There exists a level of 70% agreement between them (in 50% both scored positive and in 20% negative). An extension of this theme for multiple observers would be the calculation of the mean number of disagreements across all possible observers.[2] However, such measures do not discriminate between actual agreement and agreement which arises due to chance. A measure which attempts to correct for this is the κ statistic.[3,4] This is now the most widely accepted measure of agreement when considering data arising from nominal or ordinal scales.

The mechanics of the κ statistic rely on a comparison between the observed amount of agreement with the expected amount of agreement, the expected amount of agreement representing that due to chance and dependent on the prevalence of the attribute being measured. An illustration of its use from the example described (table I) is as follows. The observed proportion of agreement (p_o) is simply the proportion of *x* ray films agreed on by the two rheumatologists as being positive (50/100) plus the proportion agreed as negative (20/100)—that is, an overall proportion of 0·70. The expected proportions of chance agreement for each of these cells are calculated assuming independence between the observers in an analogous fashion to that employed by the χ^2 analysis of 2×2 tables. Given that both observers scored 65% of the patients as having erosions, then by chance alone the proportion that would be scored positive by both observers is 65/100×65/100=0·42. Similarly, the expected chance proportion for those considered by both to be negative is 35/100×35/100=0·12. This gives the total amount of expected agreement (p_e) as 0·54. The κ statistic is represented by the extra amount of agreement observed after taking into account chance (p_o-p_e) over the maximum amount of such agreement which could theoretically occur ($1-p_e$)—that is, $\kappa=(p_o-p_e)/(1-p_e)$. For the data in table I κ can be calculated to be 0·34.

TABLE I—*Hypothetical agreement between two rheumatologists rating hand radiographs of 100 patients according to presence or absence of evidence of erosions*

	Rheumatologist 2		
Rheumatologist 1	Present	Absent	Total
Present	50	15	65
Absent	15	20	35
Total	65	35	100

TABLE II—*Suggested interpretation of agreement for different values of κ statistic*

κ Statistic	Strength of agreement
<0·20	Poor
0·21-0·40	Fair
0·41-0·60	Moderate
0·61-0·80	Good
0·81-1·00	Very good

INTERPRETATION OF κ STATISTIC

Values for κ will usually lie between zero and 1, zero indicating only chance agreement and 1 indicating perfect agreement. It is actually possible to obtain negative values of κ from situations where there seems to be less than chance agreement. The only meaningful interpretation in this situation is that the level of agreement is what would be expected by chance alone. However, when κ lies between zero and 1 it is not as simple to assign a definite interpretation. The practice of calculating a confidence interval for κ[5] with a view to indicating whether it is significantly greater than zero will not reveal much about the extent of any agreement, only if it may be said to be present or not. Nor is the answer as straightforward as referring the resulting value of κ to levels that have been proposed (table II). Although these have been presented in an arbitrary form,[6] it may be tempting to place more weight on these levels than is intended. Very different 2×2 tables can give rise to the same κ value, and this calls for more emphasis to be placed on the actual raw data.

This problem of interpretation is due in part to the dependence of the κ statistic on the prevalence of the attribute being measured.[7] This is because a high underlying prevalence (the vast majority of subjects being of the same state) results in a high level of expected agreement. If the situation in table III is considered, where again two rheumatologists have classified the *x* ray appearances of 100 patients, the percentage of agreement is the same as in table I—that is, 70%. However, the prevalence of a positive

Arthritis and Rheumatism Council Epidemiology Research Unit, Stopford Building, University of Manchester, Manchester M13 9PT
Paul Brennan, *medical statistician*
Alan Silman, *ARC professor of rheumatic disease epidemiology*

Correspondence to:
Professor Silman.

BMJ 1992;304:1491-4

response by both rheumatologists, as shown by the marginal totals of 80%, is much higher than previously. This results in a greater level of expected agreement between the rheumatologists, the data in table III resulting in a κ value of only 0·06. Therefore, in interpreting agreement a more intuitive approach is required with this relation between a high prevalence and a high level of expected agreement being borne in mind. This relation should not be seen to detract from the usefulness of the κ statistic. In effect it penalises against populations in which there is little discrimination and which should therefore be easier to agree on. Also as κ is dependent on the prevalence of the outcome in the population under study values generated from different populations are not easily comparable.

TABLE III—*Hypothetical agreement between two rheumatologists rating hand radiographs of 100 patients for presence of erosions with higher prevalence of positive result than in table I*

	Rheumatologist 2		
Rheumatologist 1	Present	Absent	Total
Present	65	15	80
Absent	15	5	20
Total	80	20	100

The clinical interpretation of κ often relies on whether the source of variation is the result of within or between observer disagreement. As within observer disagreement may explain between observer disagreement (though not necessarily vice versa) finding disagreement between observers should be followed by an investigation within observers. Disagreements from either source may be reduced by standardisation of methods. The nature of any further action depends on how a measuring scale will be used. If a clinical investigation is to be undertaken by a single observer then it is any high within observer disagreement rather than between observer disagreement that needs to be remedied.

BIAS

In studies concerning within observer variation bias would not be expected to be a problem. However, when different observers are being considered there is a possibility that there will exist a clinically important level of systematic difference (bias) in the way they use a measuring scale. Agreement is only one aspect of the variation between observers. Though the κ statistic provides an overall measure of agreement dependent on the prevalence, it does not consider the impact any possible bias may have on the variation. As an illustration, if the situation in table IV is considered, where the *x* ray pictures of 100 patients are classified by two rheumatologists, examination of the two agreement cells on the diagonal seems to indicate a similar level of agreement (70%) as exists in table I. There is, however, a systematic bias, rheumatologist 1 being more likely to score *x* ray appearances as positive (75%) than rheumatologist 2 (55%). However, the value of κ for this situation is 0·37, indicating a level of agreement actually slightly higher than for table I. The reason for this is that the κ statistic recognises that agreement is harder to achieve in the presence of bias, and the slightly higher κ value from table IV reflects this.[8] Bias itself is a form of disagreement with important practical implications and not separately identified by κ. Therefore, the analysis of observer variation must be a dual consideration of both agreement and bias.

TABLE IV—*Hypothetical agreement between two rheumatologists rating hand radiographs of 100 patients for presence of erosions with bias in their evaluation*

	Rheumatologist 2		
Rheumatologist 1	Present	Absent	Total
Present	50	25	75
Absent	5	20	25
Total	55	45	100

TABLE V—*Hypothetical agreement between two rheumatologists rating hand radiographs of 100 patients across three grades of severity of erosions*

	Rheumatologist 2			
Rheumatologist 1	Absent	Minor	Major	Total
Absent	35	12*	5	52
Minor	8*	10	5*	23
Major	5	9*	11	25
Total	48	31	21	100

*Cells with partial agreement.

A procedure for assessing bias is to look for symmetry between the two "off diagonal" or discordant cells. In table IV these cells do not seem to be symmetrical with 25 in the top right hand corner and 5 in the bottom left hand corner. If symmetry were present we would expect 15 in each. A significance test relating to the null hypothesis of no bias is McNemar's test. This computes the z statistic relating to the hypothesis of no bias. z represents the standard normal deviate from which the p value of the null hypothesis of no bias may be derived from standard statistical tables. The calculation of z is simple by using the notation "q" to represent the top right hand corner and "r" the bottom left of the 2×2 table: $z=(q-r-1)/(\sqrt{q+r})$. Applying this formula to the data in table IV gives a value for z of 3·47. This corresponds to a p value of 0·0001.

There therefore exists a significant amount of bias between the two rheumatologists in this second example. An interpretation of this result would be that there exists a similar yet only moderate level of reproducibility between the two examples in tables I and IV. However, as the level of variation is dependent on both agreement and bias there is an intrinsically higher level of agreement in table IV, with much of the variation there being explained by bias. Any attempt to improve on this would be most fruitful if concentrated on the reasons for the bias between the two rheumatologists.

EXTENSIONS OF κ STATISTIC

The κ statistic was originally proposed for 2×2 tables—that is, two observers scoring individuals as either positive or negative. It has subsequently been extended for multiple observers and for observations with more than two categories—for example, radiological evidence of erosions being classified as absent, minor, or major. For this example calculation of κ will necessarily result in a lower value than if erosions were classified simply as being absent or present. This is because the opportunities for error and disagreement increase as the numbers of categories increase. To overcome this problem a weighted κ statistic (κ_w) has been proposed to adjust for the seriousness of different levels of disagreement.[9] Its use may be illustrated from the example in table V. As a disagreement between the absent and minor categories or minor and major categories is substantially less serious than one between the absent and major categories the first two disagreements may be considered "partial agreements."

In effect this is done by calculating the agreement after weighting these partial agreement cells with a value between zero and 1, this weight reflecting the seriousness of the disagreement. A weight of zero indicates total disagreement and a weight of 1 indicates

total agreement. When considering all cells, only the weighted value of the observed and expected levels of agreement is used in the calculation of κ_w. For the example in table V the four partial agreement cells were given an arbitrary weight of 0·25 and the two extreme disagreement cells a weight of zero. This resulted in a weighted κ value of 0·33 compared with an unweighted κ value (when considering all disagreements as equal) of 0·30. Note that in this example the categories were ordered from absent to major. However, the use of κ_w is also applicable to situations where no ordering is present, such as different diagnostic groups.[9]

Use of κ_w has been criticised, given that the choice of weights which largely determines the result is subjective. If different weights are used the results from similar studies are rarely comparable,[10] and standard weighting systems have been proposed. However, the subjective approach has the advantage that clinicians can select the weights appropriate for a particular situation. In addition, comparisons of κ values from different study populations are often invalid owing to the variation in prevalence of the attribute studied. One strength of the weighted κ is its ability to determine where the largest source of disagreement is occurring. This is achieved by weighting out various disagreements, effectively considering them as agreements by giving them a weight of 1. The resulting increase from the unweighted κ is a relative measure of the seriousness of that particular disagreement. In table V, if all disagreements between "absent" and "minor" are considered as agreement this results in a κ_w of 0·32, only a slight increase from the unweighted value of 0·30. However, if disagreements between minor and major are considered as agreements this results in a κ_w of 0·40, a larger increase from the null value, indicating that the greatest source of disagreement is between these two categories. Inspection of the raw data in table V suggests that this is the case.

As before, it is of interest to consider not just the internal agreement but also any bias which may be occurring. However, an assessment of bias when more than two categories exist is substantially more difficult than with a 2×2 table and involves a consideration of various forms of symmetry.[11] A suitable analysis requires an iterative approach, which is far beyond the scope of a calculator. A computer program which considers bias, Cohen's κ, and the weighted κ may be obtained by writing to us direct.

Continuous measurements

Methods which have been widely used to assess levels of variation in observations recorded on a continuous scale—for example, blood pressure or grip strength—contain flaws relating to their general lack of interpretation and even the appropriateness of their use. Consider the data in figure 1, which is a plot between two rheumatologists (A and B) of grip strength measured on 20 patients with rheumatoid arthritis. The measurements seem to be closely related with a Pearson correlation coefficient (r) of 0·96. However, to draw conclusions from this regarding the variation between the rheumatologists would be incorrect. This is because r simply assesses the association between the two observers. This association, however, is constant under deviations of scale or bias—for example, r would be the same if the first consistently records twice the value of the second observer. It is therefore not applicable as a measure of between observer variability. This point was made clearly by Bland and Altman in their important paper on the subject.[12] A summary of the more appropriate methodology which they proposed is given below.

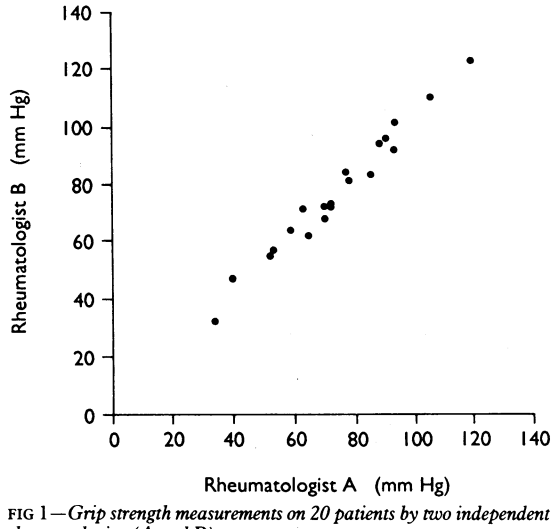

FIG 1—*Grip strength measurements on 20 patients by two independent rheumatologists (A and B)*

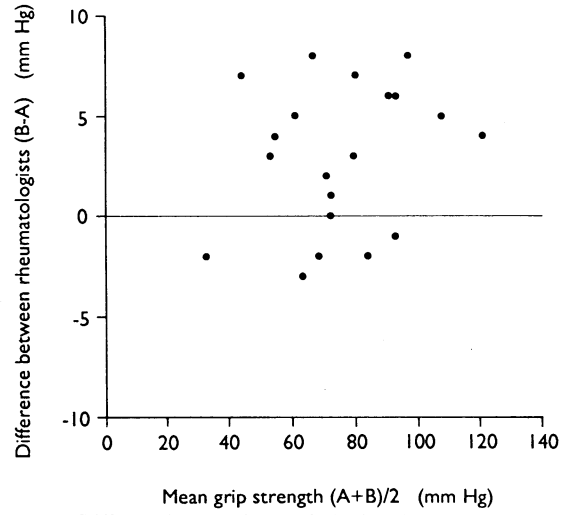

FIG 2—*Difference between rheumatologist A and rheumatologist B's readings plotted against mean measurements for 20 patients*

AGREEMENT AND BIAS

Figure 2 shows the difference between each of the observers' two readings (B−A) plotted against the corresponding mean for each patient ((A+B)/2) and gives a more meaningful representation of the level of variability. We see that the differences between the two observers lie between +10 and −4 mm, with a tendency for observer B to rate higher than observer A. A more accurate assessment of the magnitude of these discrepancies is now desirable.

As shown by figure 2 the level of precision is not related to the patient's mean score—that is, higher mean (average) readings do not result in larger discrepancies. This point is important as the analytical procedures discussed assume a constant level of error. The proposed measure for the level of agreement between the two observers is the calculation of the range within which most of their disagreements occurred. This range is based on the mean difference between the observers ($\bar{d}$) and the standard deviation of these differences (s_{diff}). A range can therefore be defined as $\bar{d} \pm t_{n-1} s_{diff}$, where t_{n-1} is the appropriate probability point of the t distribution on n−1 degrees of freedom. (For large samples (n>50) the 95% range= $\bar{d} \pm 2 s_{diff}$. For smaller sample sizes it is more accurate to use the t distribution with the appropriate number of degrees of freedom.)

In the example shown the mean difference between the observers was +3·0 (s_{diff} 3·6, t_{19}=2·1). This results in a 95% range for agreement between the observers of −4·6 to 10·6. It is also of interest to estimate the "true" value of $\bar{d}$ based on the sample studied. This is a

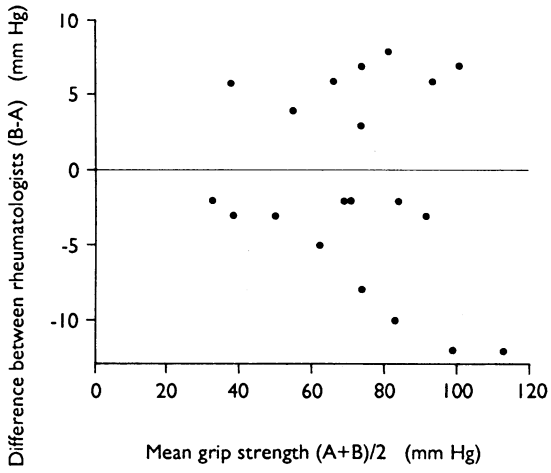

FIG 3—*Difference between two other rheumatologists' readings plotted against mean reading for 20 patients: difference increased with increasing patient mean*

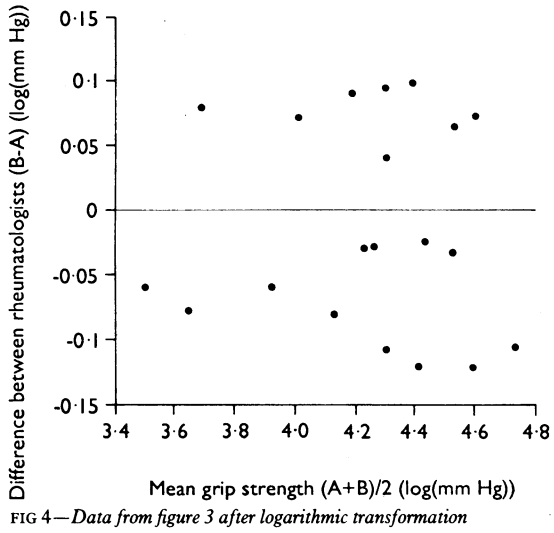

FIG 4—*Data from figure 3 after logarithmic transformation*

measure of the bias between observers, and an indication of the strength of any possible bias is gained by calculating a 95% confidence interval for $\bar{d}$. The appropriate measure of the variation of $\bar{d}$ is the standard error of the mean difference ($SE=S_{diff}/\sqrt{n}$). Thus from the data above, the 95% confidence interval for $\bar{d}$ ($\bar{d}\pm t_{n-1}SE$) is 1·3 to 4·7. As zero lies outside this interval it is concluded that there seems to exist bias between the two observers.

An important cause of the variation between the two observers depicted in figure 2 was due to bias. A different problem is depicted in figure 3, where the difference between the two observers increases with the mean level of grip strength. When such a relation between the mean and the variation occurs the calculation of confidence intervals and ranges of agreement from the data are not appropriate. This is because such procedures assume a constant variance with increasing mean level. An alternative approach is to transform the data so that this relation between the mean and variance no longer exists.

Figure 4 represents the corresponding plot of the above data after they have been logarithmically transformed with the relation between mean and variance now no longer being so apparent. The mean and standard deviation of the logarithmic differences were −0·01 and 0·08 respectively. This results in a range of agreement of −0·18 to 0·16 on the logarithmic scale or transforming back by taking antilogs (0·84 to 1·18). The interpretation of this range is not that the differences between the first and second observer lie between −0·84 and 1·18 (95% of the time) but, instead, that the first observer usually gives a reading between 84% and 118% of the second observer's reading—that is, between 16% below and 18% above. The most likely need for transforming data will be increasing discrepancies with increasing mean. The logarithmic transformation will usually be able to help correct this, although its appropriateness should still be checked.

Extensive variation between observers will usually be partly explained by variation within an observer, as explained above. A more comprehensive design could therefore be incorporated to consider both within observer and between observer variation, this achieved by each observer repeating the readings. A measure of the within observer variation for each may be gained using the methods described above, although bias should not be of concern. Also by considering the mean value for each patient by each observer an assessment of the variability between the observers may be carried out. However, this approach, by not taking into account the fact that repeated measurements were used, will result in too small a standard deviation and hence result in limits of agreement which are too conservative. An approximate correction for the standard deviation is to multiply it by $\sqrt{2}$. No such correction is necessary for the standard error of the mean difference between the two observers, as this is unaffected.[13]

Conclusion

It should be recognised that there exist similarities between studies assessing variation whether data are of a categorical or a continuous form, both entailing a consideration of agreement and bias. They differ in that the question of "How variable is a certain measure?" is more easily answered when considering continuous data through the use of 95% ranges of agreement. With categorical data this is not as easily answered simply by calculating values of the κ statistic. A more pragmatic approach is often necessary, which may involve placing more weight on the raw data than on any summary measure.

1 Cochrane AL, Garland LH. Observer error in the interpretation of chest films. An international investigation. *Lancet* 1952;ii:505-9.
2 Schilling RSF, Hughes JPW, Dingwall-Fordyce I. Disagreement between observers in an epidemiological study of respiratory disease. *BMJ* 1955;i: 65-8.
3 Cohen J. A coeficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960;**20**:37-46.
4 Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly* 1955;**19**:321-5.
5 Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol Bull* 1969;**72**:323-7.
6 Altman DG. *Practical statistics for medical research.* London: Chapman and Hall, 1991.
7 Thompson WG, Walter DW. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988;**41**:949-58.
8 Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problem of two paradoxes. *J Clin Epidemiol* 1990;**43**:543-9.
9 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213-20.
10 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;**126**:161-9.
11 Bishop YMM, Fienberg SE, Holland PW. *Discrete multivariate analysis.* Cambridge: MIT Press, 1978.
12 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983;**32**:307-17.
13 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;i:307-10.

(Accepted 23 February 1992)