STATISTICS IN MEDICINE, VOL. 13, 2465–2476 (1994)

A CRITICAL DISCUSSION OF INTRACLASS CORRELATION COEFFICIENTS*

REINHOLD MÜLLER AND PETRA BÜTTNER

Institute of Medical Statistics, Free University of Berlin, Hindenburgdamm 30, 12200 Berlin, Germany

SUMMARY

In general, intraclass correlation coefficients (ICC's) are designed to assess consistency or conformity between two or more quantitative measurements. They are claimed to handle a wide range of problems, including questions of reliability, reproducibility and validity. It is shown that care must be taken in choosing a suitable ICC with respect to the underlying sampling theory. For this purpose a decision tree is developed. It may be used to choose a coefficient which is appropriate for a specific study setting. We demonstrate that different ICC's may result in quite different values for the same data set, even under the same sampling theory. Other general limitations of ICC's are also addressed. Potential alternatives are presented and discussed, and some recommendations are given for the use of an appropriate method.

1. INTRODUCTION

While the problem of assessing agreement between quantitative measures is quite common in clinical practice, adequate statistical analysis is rare. In medical journals, linear regression analysis and Pearson's correlation coefficient (r) are often used. However, Pearson's r only measures the association between two variables, and does not provide information about agreement. For example, we may observe a correlation coefficient of nearly 1 when one measure is approximately twice a second measure. The strong correlation allows nearly perfect prediction of one measure from the other, but the actual agreement is non-existent. Good agreement is only obtained when the pairs of readings closely follow the line of equality. Pearson's r may be quite misleading in judging agreement.

By contrast, intraclass correlation coefficients (ICC's) are claimed to be suitable for this purpose. The first ICC's were developed at the end of the last century to measure concordance of items in genetics. Thereafter ICC's gained entry first into psychology and then into medicine in general.

In medicine, ICC's are used to assess agreement of quantitative measurements in the sense of consistency and conformity. The concept of consistency is defined[1] as the agreement of two quantitative measurements in settings where neither one is assumed 'correct'. Thus consistency handles questions of intra- as well as interobserver reproducibility of measurement scales. The different concept of conformity is defined as the agreement of a first measurement with a reference that is established as the 'standard'. Consistency is sometimes cited as reliability, reproducibility

* Presented at the International Society for Clinical Biostatistics Thirteenth International Meeting, Copenhagen, Denmark, August 1992.

CCC 0277–6715/94/232465–12

or repeatability. Alternative expressions for conformity are correctness, accuracy or validity. It seems preferable to replace the ambiguous yet often cited term of agreement by the more precise concepts of consistency and conformity.[1]

Various parametric as well as non-parametric ICC's have been published; they all aim to assess overall consistency or conformity in one single value. The various methods are distributed widely throughout the literature. First, we present a short overview and list important publications. ICC's originate from various disciplines and have been constructed with quite different goals. Therefore, we examine the comparability of the different ICC's with respect to their underlying sampling theories. We develop simple rules to choose a coefficient appropriate to a specific study setting. We point out the general limitations of this kind of correlation coefficient and work through specific problems of ICC's. We critically discuss possible alternatives and finally give some suggestions about how to analyse consistency or conformity studies in general.

METHODS

The concept of intraclass correlation originated from genetics where it was intended to judge sib-ship correlations. First, a special formulation of Pearson's r was defined for this purpose, which was reached by assuming equality of means and variances. In 1925 Fisher[2] showed that this coefficient is equivalent to a one-way analysis of variance (ANOVA), where the total variance is split into within-subject and between-subject variability. We will refer to this approach as Model A.

Later the basic ideas of intraclass correlation were picked up in psychology to assess the reliability of psychometric tests (partially to control for learning effects). More complicated ANOVA models were developed[3] and two models based on two-way ANOVA gained most popularity; one involves the observers as a random sample (Model B), the other assumes the observers to be fixed (Model C). For both models the within-subject variance component is split into a term which yields variability between the observers, an interaction between observers and subjects, and an error term. Thus both models allow differentiation between bias (systematic error) and random error.

Kramer and Feinstein[4] described a parametric coefficient using results from Fleiss and Cohen.[5] By contrast to all other ANOVA-based coefficients, they used sums of squares instead of mean squares. Their coefficient yields an extreme sensitivity with respect to systematic shifts but it is no longer a ratio of a part of the variance to the total variance. Therefore, this coefficient is by definition not a correlation coefficient, despite the assertion in the original paper. It can be shown that this coefficient directly depends on the sample size,[6] a property which impedes clear interpretation. This approach will be referred to as Kramer.

In 1989, I-Kuei Lin[7] proposed a concordance correlation coefficient for the case of two observers. The ideas of this approach are quite similar to results already published by Krippendorff[8] in 1970. Assuming independent samples from a bivariate Normal population Lin defines a parametric coefficient which is composed of a product of Pearson's r and a bias correcting factor, which measures how far the readings deviate from the line of equality. Thus this concordance coefficient allows the explicit computation of a location bias term relative to the scale. The concordance coefficient does not fulfil the definition of an ICC. However, simulation studies as well as theoretical considerations show that this approach behaves quite similarly to an ICC,[6] and it is therefore justifiable to treat it like one. We refer to this coefficient as Lin.

The first non-parametric ICC was published in 1949 by Whitfield,[9] who defined a coefficient for two observers using overall ranks. The construction is based on the ideas of Kendall's tau. We refer to this approach as Whitfield.

Table I. Cardiac output measured by two observers using Doppler echocardiography (litres/minute)

Patient	Observer		Patient	Observer	
	A	B		A	B
1	4·8	5·8	13	7·7	8·5
2	5·6	5·1	14	7·7	9·5
3	6·0	7·7	15	8·2	9·1
4	6·4	7·8	16	8·2	10·0
5	6·5	7·6	17	8·3	9·1
6	6·6	8·1	18	8·5	10·8
7	6·8	8·0	19	9·3	11·5
8	7·0	8·1	20	10·2	11·5
9	7·0	6·6	21	10·4	11·2
10	7·2	8·1	22	10·6	11·5
11	7·4	9·5	23	11·4	12·0
12	7·6	9·6			

Another non-parametric ICC which is also applicable to more than two observers was defined in 1979 by Rothery.[10] After overall ranking, his suggestion is to examine every pair of readings from the same subject. He then counts the occurrences where a reading from another subject cannot be found to lie between the original pair. The actual measure is defined as the ratio of the counted triplets without inversions and all possible triplets. This coefficient will be referred to as Rothery. If the readings are drawn from a multivariate Normal distribution it can be shown that this measure of intraclass correlation is a monotone function of the parameter obtained from one-way ANOVA.[11]

A third non-parametric approach is derived from results of Krippendorff[8] and Fleiss and Cohen,[5] who showed independently the approximate equivalence of the squared weighted kappa and the ICC given by Model B. Thus, squared weighted kappa provides for classified data another non-parametric ICC which will be referred to as Kappa.

Finally it is possible to construct a non-parametric ICC of the Spearman type.[6] First the readings are given overall ranks. Then the expected mean ranks can be calculated and put into the classical definition of the Spearman rank correlation coefficient. This approach will be referred to as CR.

CONCURRENCE OF DIFFERENT APPROACHES

Since the above coefficients were developed separately within different disciplines, it is reasonable to ask about their comparability. We now look at an example from cardiology to examine the extent to which the different approaches agree. The aim of this study was to assess the interobserver reproducibility of cardiac output, measured non-invasively by Doppler echocardiography. From the four chamber view of the heart the readings were made by positioning the Doppler sample volume at the mitral anulus plane; the data for 23 ventilated patients are given in Table I and are plotted in Figure 1. From a crude visual examination, we realize that observer B obtained higher values than observer A and would assume the existence of some bias. We would not judge the consistency to be very good.

We now investigate how far different approaches concur in their judgement of this consistency. Table II shows a quite amazing range of values, from Kramer's coefficient of 0·15 to 0·93 in

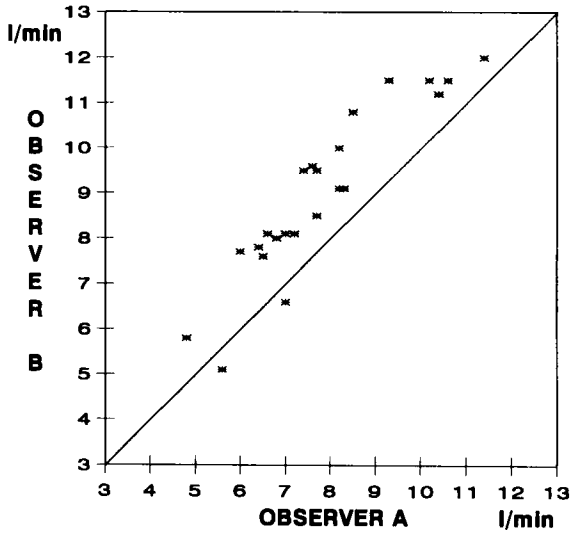

Figure 1. Interobserver reproducibility of non-invasively assessed cardiac output in 23 ventilated patients

Table II. Intraclass correlation coefficients (and 95 per cent confidence intervals) for the echocardiography data

Method	ICC	[95% CI]
Model A	0·73	[0·52, 0·86]
Model B	0·75	[0·02, 0·92]
Model C	0·92	[0·84, 0·96]
Kramer	0·15	[0·00, 0·43]
Lin	0·75	[0·59, 0·86]
Whitfield	0·55	[0·27, 0·83]
Rothery	0·65	[0·29, 0·90]
Kappa	0·73	[0·63, 0·83]
CR	0·67	[0·36, 0·85]

Model C. Thus in using different approaches, estimates from very poor to nearly perfect consistency may be found for the same data set, a result which impedes clear interpretation. Table II also shows approximate 95 per cent confidence limits for the different ICC's, some of which are very conservative.

Bearing in mind the properties of Kramer's coefficient, we exclude this method which gives the lowest value since it is not comparable to the other correlation coefficients.

One explanation for the remaining dissimilarities is that these methods are derived from different sampling theories. For example, Model A treats the observers as random while Model C assumes the observers to be fixed. It can be shown[6] that due to this difference, Model A always yields coefficients lower than Model C when a systematic data shift is present.

Thus, first of all, it is crucial for a valid interpretation of an ICC estimator to choose a coefficient which is appropriate for the underlying sampling theory. In the following we develop a 'decision tree', which leads to a coefficient which is suitable for a specific study setting.

To classify the different study types we abstract from all theoretically possible settings. In general, we face two sets: the set of observers and the set of subjects. Since the observers measure the subjects, we have a one-sided relation between these two.

We first try to classify the statistical properties of the two sets; either may be drawn at random from a larger population or may present the only elements of interest (that is, it may form a fixed set). For statistical inference, however, it is necessary that the subjects are a random sample. Thus, we only have to distinguish between studies where the set of observers is drawn at random and those where the observers are fixed.

Secondly, we classify the possible relations between the two sets. For the moment we consider the 'ideal' case where the maximum information is available; that is, we assume that each observer judges each subject. We then distinguish the case where it is possible to differentiate between the observers (as it is in the cardiac output example discussed above) from the case where it is impossible (or makes no sense) to distinguish between them (as, for example, in a genetic study of twins, see below). This differentiation is also necessary in the setting where maximum information is not available; that is, at least one subject is not judged by each observer.

The considerations above ignore studies where different subjects are measured at different times by single observers. This setting reflects a mixture of two consistency studies (for example, intra- and interobserver reproducibility) or of a consistency and a conformity study performed simultaneously. The analysis can be done separately for each of the two underlying studies. Thus, this mixed study type is reduced to settings already mentioned.

The discussion above suggests there are three major decisions in the differentiation of study types and these enable any consistency or conformity study to be classified into one of six types. In the following we will work through these decisions and clarify them by means of examples. The order in which they are taken is arbitrary; rearrangement leads to the same groups of coefficients.

The first decision is whether the observers are fixed or random and two examples may clarify the difference:

1. Two equally experienced cardiologists use a new echocardiographic unit to measure filling pressure in the left ventricle. Here, the observers are a random sample of physicians who will use the unit in the future and to which the assessed interobserver consistency is intended to relate. Thus using a random approach would yield a suitable coefficient.
2. A new mini peakflow meter and a standard peakflow meter are used to assess peak expiratory flow rate. The resulting intermeasure conformity concerns only these two peakflow meters and thus a fixed approach would be correct.

The second decision is whether or not each observer judges each subject. This is important in the case of random observers. If we want to investigate, for example, the concordance of a quantitative genetic variable in siblings, the families correspond to the 'subjects' and the single sibling to the 'observer'. It is obvious that the number of siblings per family will often vary and this variation has to be taken into account. With fixed observers it is intended to judge the effect of each single observer; when all of the subjects are not assessed by each observer we have a problem of missing values.

The third decision is whether or not the observers are differentiable, in other words whether different measurements taken from the same subject are interchangeable or not. This decision concerns only the case of random observers, for if the observers are fixed, each single measurement can obviously be attributed to the observer it came from. For example, if we compare readings of different instruments each single measurement is clearly traceable to just one, and the assessments obtained from one subject are not interchangeable. This changes if we again want to determine the concordance of a quantitative genetic variable in siblings. We would expect an

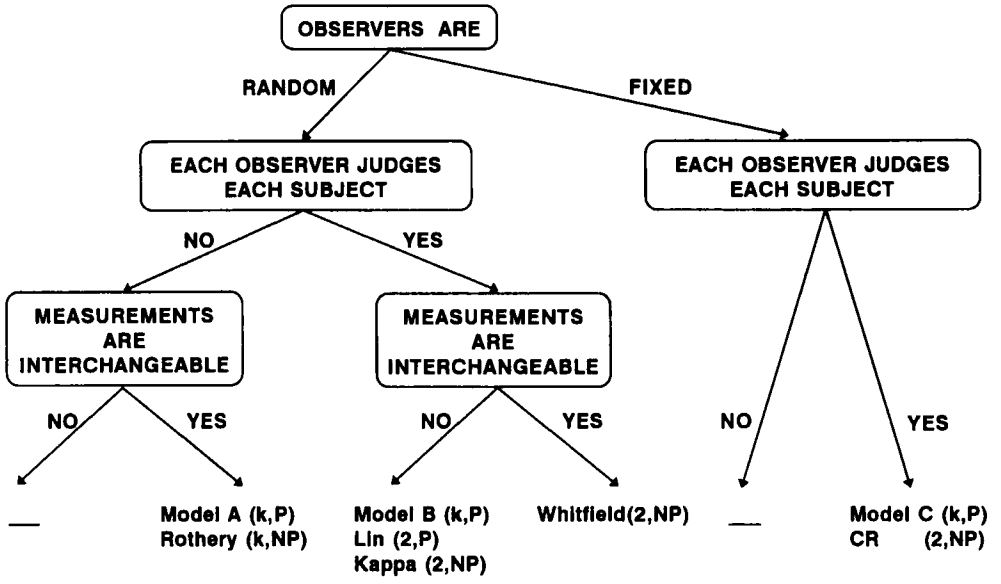

Figure 2. 'Decision tree' to select a suitable approach with respect to the underlying sampling theory; P = parametric coefficient; NP = non-parametric coefficient; k = suitable for $k \geqslant 2$ observers; 2 = only suitable for 2 observers; – = no coefficient available

appropriate ICC to yield the same result if the quite arbitrary ordering of the siblings is changed. In this example the different measurements taken from the same family are interchangeable and a suitable ICC should take this into account.

The three decisions result in the tree displayed in Figure 2, which divides studies into six types according to the underlying sampling theory. In principle, approaches which permit interchangeability of readings could also be used in the corresponding case where the numbering of readings is unambiguous. However, simulation studies show that this practice coincides with loss of information.[6] It is unavoidable for the first group of studies since no special coefficient is available. A similar problem exists with fixed observers when each subject is not judged by every observer; no coefficient is available.

We use the decision tree to choose an appropriate coefficient for our example in Figure 1: the two physicians represent a random sample of all possible users of the echocardiographic unit; each patient was assessed by each observer and a single measurement is clearly traceable to the observer it came from. Thus Model B and Lin provide suitable approaches and in this particular example result in nearly identical values (Model B = 0·753; Lin = 0·751). We also observed some location bias, and both models allow testing of this bias against zero revealing significant results ($p < 0{\cdot}001$).

We can also see that some of the differences between the ICC's in Table II are attributable to different sampling theories. However, some dissimilarities remain; in particular, assuming fixed observers, the Spearman type approach CR and Model C are markedly different even though derived from the same sampling theory. Some theoretical considerations as well as simulation studies indicate that possible remaining differences are ascribable to special data shifts.[6]

For example, a location bias will in general produce a difference between the CR coefficient and Model C. A rotation shift, however, would result in dissimilarities between Model A and Rothery's approach, and so on.

These observations direct our attention to the limitations of those methods. In the following we list some important restrictions of ICC methods which may severely restrict their interpretation.

LIMITATIONS OF ICC APPROACHES

The general problems above illustrate that with certain biases in the underlying data, ICC approaches, even when based on the same sampling theory, may yield conflicting results. These may hamper the interpretation of an ICC estimate.

A second general limitation applies to significance testing for with all the models discussed (except Model A) only tests against zero are as yet well established. Such tests may be useful in genetics, where usually, only very weak concordance is expected. In a clinical setting, however, significance tests against zero are of little interest for such tests may show only that two methods agree more than by chance. It would be quite surprising if measurements obtained from two instruments designed to measure the same thing were not related.

A sound distribution theory exists only for Model A,[12-14] and a significance test against a pre-specified value is available only for this model. However, calculation of at least approximate confidence intervals for all cited ICC's is feasible. Interested readers can find the formulae for the confidence intervals for Model A, Model B and Model C in Reference 3, for Lin, Whitfield and Rothery in the original articles, References 7, 9 and 10, for Kappa in Reference 15 and for CR in Reference 6.

Another problem is that some estimators may be negative, whereas their corresponding parameters are strictly positive. How such negative values should be interpreted is quite unclear, and the suggestion to redefine them as zero does not really solve the problem.

There are other limitations with parametric approaches. In using a parametric ICC, we have to consider possible violations of the underlying assumptions; that is, multivariate Normal distributions and equality of variances. We would expect the problem of different variances to arise more often in conformity than in consistency studies, especially when a new instrument is compared with an accepted standard, or when a less-experienced observer is compared with an experienced one. Consistency studies are concerned with repeatability of the same measurement or reliability of a random sample of observers; in both cases we tend to expect similar variances.

Another severe restriction concerns interpretation of any parametric ICC, for the estimates (as with all correlation coefficients) are dependent on the range of the measuring scale; the wider the range, the better the result. A popular illustration of this limitation is the judgement of interobserver consistency of diastolic and systolic blood pressures. Since systolic blood pressure measurements have a wider range than diastolic measurements, the ICC for the former is higher than that for the latter, apparently implying that diastolic blood pressure is more difficult to assess than systolic. This is an artefact which is not supported by any real evidence.

With more than two observers or methods of measurement ($k > 2$) the parameters of the ICC's are asymmetrically distributed. The range of the values as well as the estimates themselves depend on k. Thus an ICC for three methods is not comparable to one for four methods. The interpretation is correspondingly unclear.

One special limitation applies to studies with fixed observers for which sampling theory supplies Model C as a suitable parametric ICC. This approach originates from psychometric theory and was constructed for test–retest analysis. In this context a systematic data shift would only reflect a learning effect. This bias is estimated explicitly by Model C and the measure of

consistency is adjusted for it. In medicine a systematic data shift may be of special interest especially when assessing consistency between two observers, and consequently Model C may be inappropriate.

The non-parametric alternative with fixed observers, coefficient CR, handles bias like a random error. Thus keeping the decision tree in mind, we realize that with fixed observers there is no coefficient which is sensitive to bias and simultaneously allows explicit estimation of it. Consequently, no convincing coefficient exists for conformity studies where the main interest lies in detailed analysis of possible bias. With random observers, however, both Model B and Lin allow an adequate analysis of bias.

In listing limitations of non-parametric approaches, we first note a severe practical disadvantage: while ANOVA based ICC's may be computed using common statistical software, non-parametric approaches are not implemented in such programs. A second limitation is that such coefficients are insensitive to outliers. By contrast, parametric approaches react with an increase of the error component and, therefore, outliers are reflected in the estimates.

A third problem is the underlying sampling theory; except for the CR method, all non-parametric ICC's are developed under the assumption of interchangeability of readings from the same subject. Such study settings, however, are quite uncommon in practice, and, therefore, the use of non-parametric ICC's remains correspondingly limited.

POTENTIAL ALTERNATIVE APPROACHES

As a result of the criticism of ICC's, we discuss below some potential alternative approaches to judge consistency or conformity.

One straightforward solution appears to be the combination of detailed linear regression analysis and a paired t-test.[16] However, this kind of analysis assumes that only one method, the dependent, is subject to measurement errors; the other, the independent method, has to be exact. Thus linear regression analysis may only be used for conformity studies where the accepted standard is measured without error. Assuming this quite rare study setting we may test Pearson's r against zero to prove that a significant correlation exists. We may test the slope of the regression line against 1 and the intercept against zero to rule out a significant rotation or location shift. An additional paired t-test may also exclude a significant location shift.

All these results, however, are answers to the wrong questions. The absence of a significant difference does not imply good conformity. Only equivalence tests would be correct in this context. The common principle of inclusion within the confidence interval (see Mau,[17] Westlake[18]) works as follows. To test, for instance, that an intercept lies between $-0{\cdot}05$ and $+0{\cdot}05$, we construct a $(1-2\alpha)\cdot 100$ per cent confidence interval for the computed intercept of the regression line. If we find both limits of this interval to lie inside of $[-0{\cdot}05, +0{\cdot}05]$, we reject the null hypothesis of non-equivalence by holding a significance level of α (not 2α!). The problem of this method as well as of other equivalence approaches is that the studies concerned may lack the necessary sample size.[19]

A second alternative was proposed by Altman and Bland.[20-22] The authors stress the importance of visual investigations. They propose a residual-like plot of the differences of the observed pairs of readings against their mean values. They define 'limits of agreement' by combining the mean d and the standard deviation s of the differences as $d \pm 2s$. An example of such a plot is given for the cardiac output data in Figure 3. This type of plot allows identification of outliers as well as an examination for trend by means of linear regression analysis. One particular advantage of this more visual approach is that it can be easily explained to non-statisticians.

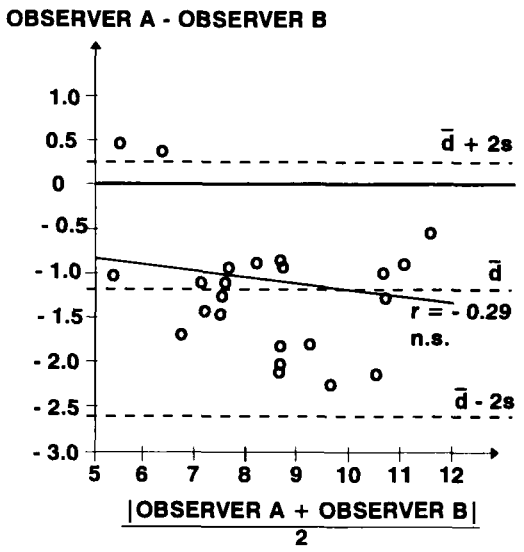

Figure 3. Residual-like plot[20] of the cardiac output data

Some critical aspects should also be noted. The concepts of consistency or conformity require an equivalence test for the mean of the differences as well as for the correlation coefficient against zero ± delta. Altman and Bland, however, test the differences for significance and therefore the estimation of the goodness of consistency or conformity seems rather subjective. This approach also lacks a single measure which would be preferable especially when more than two methods are compared.

A third alternative method frequently used in clinical chemistry is the method of structural regression analysis (STRUCREG).[23,24] Although the concept of STRUCREG is closely related to the well-known model of linear structural relationship,[25,26] STRUCREG offers some advantages. Bivariate structural regression analysis is not restricted to multivariate Normal distributions and without additional constraints the model remains identifiable in the bivariate case.

Figure 4 presents an example of STRUCREG using the cardiac output data. The figure shows the bivariate calibration line, a straight line which allows simultaneously the prediction of Y given an x-value and of X given a y-value. The slope of the bivariate calibration line is given by the ratio of the variances of the readings, and is nothing other than the geometric mean of the slopes of the two ordinary regression lines.

Feldmann and Schneider[23,24] developed maximum likelihood estimators as well as robust estimators for the slope and the intercept. In contrast to the concept of linear structural relationship in the case of maximum likelihood estimation, STRUCREG allows unbiased estimation of standard errors of the slope and the intercept of the bivariate calibration line.

Feldmann and Schneider also give confidence limits for their robust estimators. The calculated confidence intervals for the estimators of the slope and the intercept of the bivariate calibration line directly show whether proportional or additive bias exists. Calculating the 95 per cent confidence intervals (CI) of the slope and the intercept for the cardiac output data yields intercept = 0·23, 95 per cent CI = [− 0·07, 0·53] and slope = 1·12, 95 per cent CI = [0·95, 1·30]; STRUCREG does not detect a significant bias.

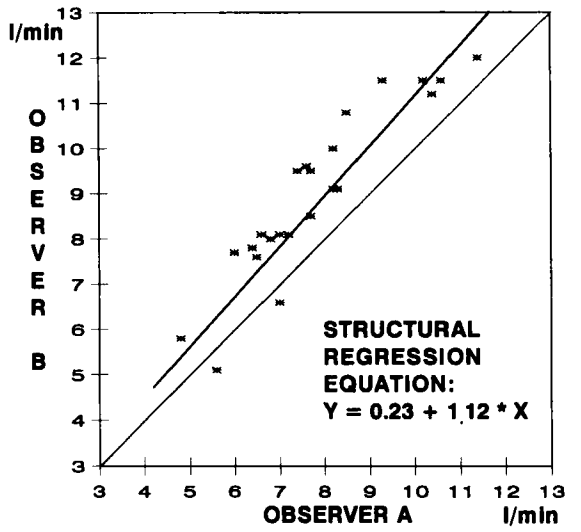

Figure 4. STRUCREG example of the cardiac output data

DISCUSSION

Many of the cited ICC methods were developed in genetics and psychology to examine questions of, for example, concordance of a genetic variable in twins, or to assess the reliability of a psychological questionnaire by means of repeating similar questions about the same subject. In these disciplines, most study settings assume that measurements taken from the same subject are interchangeable.

For studies of this kind, however, neither detailed linear regression analysis nor the methods proposed by Bland and Altmann, nor structural regression analysis, are applicable since single measurements are not clearly attributable to one method or observer. For those study settings, ICC approaches provide a useful methodology for analysing consistency or conformity. ICC's, moreover, allow tests against zero, and, therefore, these methods are adequate for study settings where the main interest lies in assessing whether the concordance is greater than by chance.

Thus ICC's are quite helpful for certain studies in genetics and psychology. However, care must be taken over the choice of an appropriate coefficient with respect to the underlying sampling theory. The proposed decision tree may help to avoid the misuse of inadequate approaches.

In medicine the main aim is to judge conformity or consistency of different measurement methods or observers. Thus we are confronted with a quite different study setting since the single measurements are clearly traceable to the unit or observer they came from. Moreover, the aim is to judge how far the observations differ from an ideal conformity or consistency, and it is of little interest whether two instruments concur more than by chance.

For study settings with just two observers or methods, all of the alternative approaches discussed here are applicable. They provide more detailed information than a single ICC. Using these alternative methods, we are also able to avoid the problems of interpreting one overall measure. Unfortunately ICC's do not provide tests against unity. Therefore, assessing overall consistency or conformity in medicine by solely computing an ICC without further information must be regarded as insufficient and use of the ICC's discussed here seems limited. Two exceptions can be made. First, if more than two methods are to be compared at once, all the

mentioned alternatives to ICC fail to provide any adequate analysis. Thus, an ICC may be the only possibility for getting at least a relative measure of consistency or conformity. Second, in a study in which the conformities of two or more new measurements are to be compared with a 'gold standard', the ICC approach, resulting in one single measure, is the easiest method to judge which of the new measurements yields the best conformity.

Only studies of this kind profit from the relative nature of an ICC. All other study settings face problems of interpretation; the dependence of the ICC on the variance of the population illustrates that a calculated ICC has no absolute meaning. Consequently there is no reason to judge an absolute ICC greater than 0·75 as indicating good conformity or consistency as proposed by Burdock *et al.*[27] and Lee *et al.*[28]

When single measurements are unambiguously attributable to one method or observer, we recommend first plotting a square scatter diagram including the line of equality. This type of visual investigation is imperative to get an initial impression of the data. An additional residual-like plot of the differences of the observed pairs of readings against their mean values, as proposed by Bland and Altman, reveals further important information; using their 'limits of agreement' we are able to identify outliers. Using linear regression analysis of the differences of the pairs of readings on their mean values, it is possible to rule out a significant bias. However, as already discussed in the context of linear regression analysis, we recommend the use of equivalence tests.

The different approach of structural regression analysis, recently developed by Feldmann,[23] represents a powerful instrument to judge consistency or conformity of two methods. The main advantages of this bivariate calibration technique are that parametric ML estimators as well as robust estimators are available, bivariate residual analysis as well as outlier detection are provided and tests concerning the slope and the intercept of the calibration line are possible. By contrast to Feldmann and Schneider, however, we again recommend equivalence tests.

The STRUCREG method assumes that both measurements are subject to error. Thus structural regression analysis is an adequate model for nearly any kind of bivariate conformity as well as consistency studies. Only one exception has to be made, the case where the gold standard is assumed to be measured without any error. For this and only this special case of conformity studies, detailed linear regression analysis using equivalence tests of the described form is recommended.

All the above methods result in quite specific, interpretable measures, so in most consistency or conformity studies it is not necessary to use an ICC. The calculation of such an additional coefficient only provides a relative measure which, moreover, has considerable problems in interpretation, as already discussed.

REFERENCES

1. Feinstein, A. R. *Clinical Epidemiology*, W.B. Saunders Company, Philadelphia, 1985.
2. Fisher, R. A. *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., Edinburgh, 1925.
3. Shrout, P. E. and Fleiss, J. L. 'Intraclass correlations: Uses in assessing rater reliability', *Psychological Bulletin*, **86**, 2, 420–428 (1979).
4. Kramer, M. S. and Feinstein, A. R. 'Clinical biostatistics LIV. The biostatistics of concordance', *Clinical Pharmacology and Therapeutics*, **29**,(1), 111–123 (1981).
5. Fleiss, J. L. and Cohen, J. 'The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability', *Educational and Psychological Measurement*, **33**, 613–619 (1973).
6. Müller, R. 'Intraklassenkorrelationsanalyse – ein Verfahren zur Beurteilung der Reproduzierbarkeit und Konformität von Meßmethoden', Ph.D. thesis, 1992.
7. I-Kuei Lin, L. 'A concordance correlation coefficient to evaluate reproducibility', *Biometrics*, **45**, 255–268 (1989).

8. Krippendorff, K. 'Bivariate agreement coefficients for reliability of data', E. F. Borgatta (ed.) *Sociological Methodology*, 139–150 Jossey-Bass, San Francisco (1970).
9. Whitfield, J. W. 'Intra-class rank correlation', *Biometrika*, **36**, 463–467 (1949).
10. Rothery, P. 'A nonparametric measure of intraclass correlation', *Biometrika*, **66**, 3, 629–639 (1979).
11. Kendall, M. G. and Stuart A. *The Advanced Theory of Statistics*, 2nd edn., Volume 1, C. Griffin Ltd., London, 1961.
12. Kraemer, H. C. 'Improved approximation to the non-null distribution of the correlation coefficient', *Journal of the American Statistical Association*, **68**, 344, 1004–1008 (1973).
13. Kraemer, H. C. 'The non-null distribution of the spearman rank correlation coefficient', *Journal of the American Statistical Association*, **69**, 345, 114–117 (1974).
14. Kraemer, H. C. 'On estimation and hypothesis testing problems for correlation coefficients', *Psychometrika*, **40**, 4, 473–485 (1975).
15. Fleiss, J. L. *Statistical Methods for Rates and Proportions*, 2nd edn., Wiley, New York, 1981.
16. Westgard, J. O. and Hunt, M. R. 'Use and interpretation of common statistical tests in method-comparison studies', *Clinical Chemistry*, **19**, 1, 49–57 (1973).
17. Mau, J. 'A statistical assessment of clinical equivalence', *Statistics in Medicine*, **7**, 1267–1277 (1988).
18. Westlake, W. J. 'Response to T. B. L. Kirkwood: bioequivalence testing – a need to rethink', *Biometrics*, **37**, 591–593 (1981).
19. Wellek, S. 'Statistische Methoden zum Nachweis von Äquivalenz' Habil. schrift, Johannes-Gutenberg-Universität Mainz, 1991.
20. Altman, D. G. and Bland, J. M. 'Measurement in medicine: the analysis of method comparison studies', *The Statistician*, **32**, 307–317 (1983).
21. Bland, J. M. and Altman, D. G. 'Statistical methods for assessing agreement between two methods of clinical measurements', *The Lancet*, 307–310 (1986).
22. Bland, J. M. and Altman, D. G. 'A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement', *Computers in Biology and Medicine*, **20**, 5, 337–340 (1990).
23. Feldmann, U. and Schneider, B. 'Bivariate structural regression analysis: a tool for the comparison of analytic methods', *Methods of Information in Medicine*, **26**, 205–214 (1987).
24. Feldmann, U. 'Robust bivariate errors-in-variables regression and outlier detection', *European Journal of Clinical Chemistry and Clinical Biochemistry*, **30**, 405–414 (1992).
25. Haeckel, R. and Schneider, B. 'Statistische Modelle und Verfahren beim Vergleich von Analysenmethoden', *GiT Labor-Medizin*, **2**, 97–102 (1980).
26. Feldmann, U., Schneider, B., Klinkers, H. and Haeckel, R. 'A multivariate approach for the biometric comparison of analytical methods in clinical chemistry', *Journal of Clinical Chemistry and Clinical Biochemistry*, **19**, 121–137 (1981).
27. Burdock, E. I., Fleiss, J. L. and Hardesty, A. S. 'A new view of interobserver agreement', *Personnel Psychology*, **16**, 373–384 (1963).
28. Lee, D., Koh, D. and Ong, C. N. 'Statistical evaluation of agreement between two methods for measuring a quantitative variable', *Computers in Biology and Medicine*, **19**, 61–70 (1989).