Comput. Biol. Med. Vol. 20, No. 5, pp. 337–340, 1990
Printed in Great Britain

0010–4825/90 $3.00 + .00
© 1990 Pergamon Press plc

A NOTE ON THE USE OF THE INTRACLASS CORRELATION COEFFICIENT IN THE EVALUATION OF AGREEMENT BETWEEN TWO METHODS OF MEASUREMENT

J. M. BLAND* and D. G. ALTMAN†

* Department of Public Health Sciences, St. George's Hospital Medical School, Cranmer Terrace, London SW17 0RE, U.K; † Medical Statistics Laboratory, Imperial Cancer Research Fund, Lincolns Inn Fields, London WC2A 3PX, U.K.

(*Received* 5 *May* 1989; *in revised form* 6 *April* 1990; *received for publication* 29 *May* 1990)

Abstract—The intraclass correlation coefficient (r_I) has been advocated as a statistic for assessing agreement or consistency between two methods of measurement, in conjunction with a significance test of the difference between means obtained by the two methods. We show that neither technique is appropriate for assessing the interchangeability of measurement methods. We describe an alternative approach based on estimation of the mean and standard deviation of differences between measurements by the two methods.

Method agreement Method comparison Intraclass correlation

INTRODUCTION

The analysis of studies comparing two methods of clinical measurement seems to cause tremendous difficulty. Many approaches have been advocated and used, but few answer the question as posed by Lee *et al.* [1], that is, whether two methods can be used interchangeably. Lee *et al.* [1] criticise several approaches and advocate three criteria for agreement: there should be no marked systematic bias, there should be no statistically significant difference between mean readings obtained by the two methods, and the lower limit of the 95% confidence interval of the intraclass correlation coefficient should be at least 0.75. We agree with Lee *et al.* [1] that a plot of difference against subject mean is particularly informative and that the use of either a crude comparison of means or the product moment correlation coefficient, r, is unsatisfactory [2, 3]. However we cannot agree that the intraclass correlation coefficient, r_I, provides a satisfactory alternative, or with their approach to statistical confidence and significance.

THE INTRACLASS CORRELATION COEFFICIENT

Lee *et al.* [1] do not explain why they think that intraclass correlation is suitable, except to state that it is a measure of agreement, corrected for the agreement expected by chance. The intraclass correlation coefficient was devised to deal with the relationship between variables within classes. For example, suppose we wished to see whether the blood pressures of identical twins were related. To calculate the usual (interclass) product moment correlation coefficient, r, we would have to define variables X and Y such that for each pair of twins one measurement was labelled X and the other Y. The assignment of members of a pair of twins to X and Y would be arbitrary and different choices would produce different values of r. The intraclass correlation coefficient is the average correlation across all possible orderings of pairs into X and Y. The intraclass correlation coefficient can be used, for example, as an index of correlation between repeated measures by the same method, i.e. as an index of repeatability, because in that case there is no ordering of the repeated measures and hence no obvious choice of X or Y

Table 1. Measurements of blood pressure (mm Hg) by two methods for five patients (hypothetical data)

Set (i)				Set (ii)			
	Method				Method		
Subject	A	B	A − B	Subject	A	B	A − B
1	90	95	−5	6	120	125	−5
2	95	90	5	7	95	90	5
3	90	90	0	8	80	80	0
4	95	95	0	9	75	75	0
5	90	90	0	10	70	70	0

variables for the calculation of r. In the case of repeatability studies, r_I provides an index of the information content of the measurement, since is essentially a ratio of the variability between subjects to the total variability. This approach has been extended to the comparison of raters, with the raters regarded as a random sample of all possible raters and hence as part of the measurement error.

However, when dealing with measurements by two different methods, we have a very clear ordering, the two variables being the two methods. If we use the intraclass correlation coefficient and ignore the ordering, we are treating our methods as a random sample from a population of methods. Lee *et al.* [1] appear to do this in their two linear models, stating that the method effect is Normally distributed when in fact it must be fixed. We have two specific methods we want to compare, not two chosen at random from some large population. Their model also assumes that the measurement error of both methods is the same, which seems to us a very strong assumption and quite unjustified. Whether the measurement error for one method is greater than that for another is one of the things which we hope to find out in a method comparison study.

Although the intraclass correlation avoids the problem of linear relationship being mistaken for agreement, it does not avoid other problems associated with correlation coefficients in this context. It is dependent on the range of the measurement and it is not related to the actual scale of measurement or to the size of error which might be clinically allowable. For example, we have followed Lee *et al.* [1] in inventing hypothetical data representing five pairs of blood pressures measured by two methods (see Table 1). Data set (i) has a correlation of $r = 0.17$ and an intraclass correlation of $r_I = 0.20$ (obtained using equation 4, [1] after correction of a typographical error in the numerator, which should be $n(\text{ms}S - \text{ms}E)$). This is unacceptable agreement by the criterion of Lee *et al.*, who would want $r_I \geq 0.75$. Yet the methods give identical results for three of the subjects and differ by only 5 mmHg for the other two, better agreement than could be achieved in in practice. The low value of r_I is because the variability between subjects is low, not because the methods do not agree. Consider set (ii), which shows hypothetical data with more variable blood pressures but identical differences between methods. Data set (ii) has $r = 0.99$ and $r_I = 0.99$. The more variable the subjects, the greater will be r_I and so the better the "agreement". Because of this, r is always greater for systolic than for diastolic blood pressures, which vary less [2], and the same will be true for r_I too. This does not imply that diastolic pressure is much harder to measure than systolic. Lee *et al.* [1] accept this dependence on the range of the subjects, and say that "hence it is advantageous to ensure that the study sample is highly heterogeneous" to increase the value of r_I! Surely we want a method of assessing agreement which is not dependent on the particular subjects we choose to investigate. This is why the use of $r_I = 0.75$ as a fixed cutoff point is wrong.

It could be argued that the dependence of r_I on the variation in the sample is not a disadvantage, in that we are concerned with the ability of tests to distinguish between subjects in the population of interest. Indeed, this feature of r_I is quite acceptable in the study of repeatability, where r_I can be regarded as an index of the information content of the measurement. In the context of comparing two methods of measurement this view is less convincing. Perhaps r_I could be thought of as the degree to which the two methods contain the same information. However, if we take this approach, then r_I must be referred to a particular population. The origin of the sample of subjects becomes important and we need a random sample of the population in which the methods are to be used. Ensuring that the study sample is "highly heterogeneous" defeats this object.

We believe that it is impossible to summarize agreement adequately using a single number.

AN ALTERNATIVE APPROACH BASED ON DIFFERENCES

We would not wish to criticise the use of intraclass correlation coefficient without suggesting an alternative which, we think, meets these criticisms. There are several possible approaches, but we will confine our attention to one which we think has the merit of simplicity. This method, which meets the requirement of not depending on the range of the sample, is to consider the differences between the measurements for each subject. The mean difference, $\bar{d}$, is the bias and the standard deviation of the differences, s, enables us to calculate the size of difference likely to arise between the two methods. Approximately 95% of differences will lie between $\bar{d}-2s$ to $\bar{d}+2s$, which we have called the limits of agreement [2, 3]. This approach enables us to separate systematic and random error [4], which r_I combines into a single measure. Standard errors and confidence intervals can be found for these limits in the usual way. If the differences were exactly normally distributed we could use a multiplier of 1.96 rather than 2 to give the best estimate of the interval within which 95% of differences would lie. In practice the more approximate figure of 2 is satisfactory, and does not rely on the assumption of exact normality.

A plot of difference against mean, which Lee *et al.* also recommend [1], enables us to look for a relationship between difference and mean. If the differences diverge as the mean increases, indicating that the measurement error increases with the size of the measurement, a logarithmic transformation can be used leading to limits of agreement in terms of a ratio (e.g. "to within 5%"), a form familiar in biological and medical applications [3].

For the data of the Table, either set (i) or set (ii), we have $\bar{d}=0$, $s=3.5355$ mmHg, so the limits are -7.1 to $+7.1$ mmHg. We would expect 95% of pairs of measurements would be within 7 mmHg of one another. For the serum copper data given by Lee *et al.* [1], their Fig. 2, a plot of difference against mean, shows that the difference is unrelated to the mean. We have $\bar{d}=0.121$, $s=0.126\,\mu\text{g ml}^{-1}$ and the limits of agreement (AAS − IC) are -0.131–$0.373\,\mu\text{g ml}^{-1}$. Thus it is unlikely (a probability of less than 0.05) that measurements by the two methods on the same individual would differ by more than $0.37\,\mu\text{g ml}^{-1}$. The two methods could be used interchangeably if differences in measurements of the order of $0.37\,\mu\text{g ml}^{-1}$ did not matter, or in other words did not affect the clinical decision made on the basis of the measurement. The magnitude of the difference which is acceptable is not a statistical decision, but a clinical one. We should ask whether the agreement is good enough for a particular purpose, not whether it conforms to some absolute, arbitrary criterion. Methods which may agree well enough for one purpose may not agree well enough for another. Further details are given elsewhere [2, 3].

CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS

We must also take issue with the approach to confidence intervals and significance testing adopted by Lee *et al.* [1]. Their criterion for agreement is that the lower limit of the 95% confidence interval for r_I is greater than 0.75. They conclude that the agreement is inadequate for the copper data on the grounds that the lower limit of the 95% confidence interval for r_I is 0.42, although the point estimate for r_I is 0.84. The sample is small and a larger sample would give a narrower confidence interval and hence it is possible that these methods could meet the r_I criterion. If the wide confidence interval shows anything it shows that the sample size is too small to draw any meaningful conclusions. Further they reject agreement because there is a mean difference significantly different from zero. It is not statistical significance that matters but magnitude, as we think Lee *et al.* would agree. We have previously given an example showing excellent agreement for clinical purposes where there is significant, but small, bias [3].

CONCLUSION

Correlation coefficients have a rôle to play in validity studies, particularly for questionnaire scales and other subjective assessments. They cannot answer the question as to whether methods of measurement can be used interchangeably.

REFERENCES

1. J. Lee, D. Koh and C. N. Ong, Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput. Biol. Med.* **19**, 61–70 (1989).
2. D. G. Altman and J. M. Bland, Measurement in medicine: the analysis of method comparison studies. *The Statistician* **32**, 307–317 (1983).
3. J. M. Bland and D. G. Altman, Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310 (1986).
4. L. M. Irwig and J. M. Simpson, Assessing agreement. *Med. J. Aust.* **151**, 235–236 (1989).

About the Author—JOHN MARTIN BLAND received the B.Sc. degree in mathematics and the M.Sc. in statistics from Imperial College, London, in 1968 and 1969. He worked for Imperial Chemical Industries for three years before joining St. Thomas's Hospital Medical School, London, in 1972 and St. George's Hospital Medical School, London, in 1976. He was awarded the degree of Ph.D. (London) in 1980. His main medical research has been in the study of respiratory disease in children and of birthweight, with numerous other clinical and epidemiological publications. In the statistical field his main interest is in medical measurement. He is the author of *An Introduction to Medical Statistics* (Oxford University Press, 1987).

About the Author—DOUGLAS G. ALTMAN received the B.Sc. degree in statistics from the University of Bath in 1970, and joined the Department of Community Medicine at St. Thomas's Hospital Medical School, London. In 1976 he moved to the Medical Research Council's Clinical Research Centre, where he stayed for eleven years. In 1988 he became head of the Medical Statistics Laboratory at the Imperial Cancer Research Fund in London. His research interests include statistics in medical journals and survival analysis. He is coauthor of *Statistics in Practice* (British Medical Association, 1982) and *Statistics with Confidence* (British Medical Journal, 1989), and is author of the forthcoming *Practical Statistics for Medical Research* (Chapman and Hall, 1990).