The Statistician 32 (1983) 307–317

Measurement in Medicine: the Analysis of Method Comparison Studies†

D. G. ALTMAN and J. M. BLAND‡

Division of Computing and Statistics, MRC Clinical Research Centre, Watford Road, Harrow HA1 3UJ; and
‡ Department of Clinical Epidemiology and Social Medicine, St George's Hospital Medical School, Cranmer Terrace, London SW17

Summary: Methods of analysis used in the comparison of two methods of measurement are reviewed. The use of correlation, regression and the difference between means is criticized. A simple parametric approach is proposed based on analysis of variance and simple graphical methods.

1 The problem

In medicine we often want to compare two different methods of measuring some quantity, such as blood pressure, gestational age, or cardiac stroke volume. Sometimes we compare an approximate or simple method with a very precise one. This is a calibration problem, and we shall not discuss it further here. Frequently, however, we cannot regard either method as giving the true value of the quantity being measured. In this case we want to know whether the methods give answers which are, in some sense, comparable. For example, we may wish to see whether a new, cheap and quick method produces answers that agree with those from an established method sufficiently well for clinical purposes. Many such studies, using a variety of statistical techniques, have been reported. Yet few really answer the question "Do the two methods of measurement agree sufficiently closely?"

In this paper we shall describe what is usually done, show why this is inappropriate, suggest a better approach, and ask why such studies are done so badly. We will restrict our consideration to the comparison of two methods of measuring a continuous variable, although similar problems can arise with categorical variables.

2 Incorrect methods of analysis

We shall first describe some examples of method comparison studies, where the statistical methods used were not appropriate to answer the question.

Comparison of means

Cater (1979) compared two methods of estimating the gestational age of human babies.

† Paper presented at the Institute of Statisticians conference, July 1981.

Gestational age was calculated from the last menstrual period (LMP) and also by the total maturity score based on external physical characteristics (TMS). He divided the babies into three groups: normal birthweight babies, low birthweight pre-term (<36 weeks gestation) babies, and low birthweight term babies. For each group he compared the mean by each method (using an unspecified test of significance), finding the mean gestational age to be significantly different for pre-term babies but not for the other groups. It was concluded that "the TMS is a convenient and accurate method of assessing gestational age in term babies".

His criterion of agreement was that the two methods gave the same mean measurement; "the same" appears to stand for "not significantly different". Clearly, this approach tells us very little about the accuracy of the methods. By his criterion, the greater the measurement error, and hence the less chance of a significant difference, the better.

Correlation

The favourite approach is to calculate the product-moment correlation coefficient, r, between the two methods of measurement. Is this a valid measure of agreement? The correlation coefficient in this case depends on both the variation between individuals (i.e. between the true values) and the variation within individuals (measurement error). In some applications the "true value" will be the subject's average value over time, and short-term within-subject variation will be part of the measurement error. In others, where we wish to identify changes within subjects, the true value is not assumed constant.

The correlation coefficient will therefore partly depend on the choice of subjects. For if the variation between individuals is high compared to the measurement error the correlation will be high, whereas if the variation between individuals is low the correlation will be low. This can be seen if we regard each measurement as the sum of the true value of the measured quantity and the error due to measurement. We have:

variance of true values $= \sigma_T^2$
variance of measurement error, method A $= \sigma_A^2$
variance of measurement error, method B $= \sigma_B^2$

In the simplest model errors have expectation zero and are independent of one another and of the true value, so that

variance of method A $= \sigma_A ? + \sigma_T^2$
variance of method B $= \sigma_B^2 + \sigma_T^2$
covariance $= \sigma_T^2$ (see appendix)

Hence the expected value of the sample correlation coefficient r is

$$\rho = \frac{\sigma_T^2}{\sqrt{(\sigma_A^2 + \sigma_T^2)\,(\sigma_B^2 + \sigma_T^2)}}$$

Clearly ρ is less than one, and it depends only on the relative sizes of σ_T^2, σ_A^2 and σ_B^2. If σ_A^2 and σ_B^2 are not small compared to σ_T^2, the correlation will be small no matter how good the agreement between the two methods.

In the extreme case, when we have several pairs of measurements on the same individual, $\sigma_T^2 = 0$ (assuming that there are no temporal changes), and so $\rho = 0$ no matter how close the agreement is. Keim *et al.* (1976) compared dye-dilution and impedance cardiography by finding the correlation between repeated pairs of measurements by the two methods on each of 20 patients. The 20 correlation coefficients ranged from $-0{\cdot}77$ to $0{\cdot}80$, with one correlation being significant at the 5 per cent level. They concluded that the two methods did not agree because low correlations were found when the range of cardiac output was small, even though other studies covering a wide range of cardiac output had shown high correla-

tions. In fact the result of their analysis may be explained on the statistical grounds discussed above, the expected value of the correlation coefficient being zero. Their conclusion that the methods did not agree was thus wrong – their approach tells us nothing about dye-dilution and impedance cardiography.

As already noted, another implication of the expected value of r is that the observed correlation will increase if the between subject variability increases. A good example of this is given by the measurement of blood pressure. Diastolic blood pressure varies less between individuals than does systolic pressure, so that we would expect to observe a worse correlation for diastolic pressures when methods are compared in this way. In two papers (Laughlin *et al.*, 1980; Hunyor *et al.*, 1978) presenting between them 11 pairs of correlations, this phenomenon was observed every time (Table 1). It is not an indication that the methods agree less well for diastolic than for systolic measurements. This table provides another illustration of the effect on the correlation coefficient of variation between individuals. The sample of patients in the study of Hunyor *et al.* had much greater standard deviations than the sample of Laughlin *et al.* and the correlations were correspondingly greater.

Table 1. Correlation coefficients between methods of measurement of blood pressure for systolic and diastolic pressures

	Systolic pressure			*Diastolic pressure*		
	s_A	s_B	r	S_A	S_B	r
Laughlin *et al.* (1980)						
1	13·4[a]	15·3[a]	0·69	6·1[a]	6·3[a]	0·63
2			0·83			0·55
3			0·68			0·48
4			0·66			0·37
Hunyor *et al.* (1978)						
1	40·0	40·3	0·997	15·9	13·2	0·938
2	41·5	36·7	0·994	15·5	14·0	0·863
3	40·1	41·8	0·970	16·2	17·8	0·927
4	41·6	38·8	0·984	14·7	15·0	0·736
5	40·6	37·9	0·985	15·9	19·0	0·685
6	43·3	37·0	0·987	16·7	15·5	0·789
7	45·5	38·7	0·967	23·9	26·9	0·941

[a]Standard deviations for four sets of data combined.

A further point of interest is that even what appears (visually) to be fairly poor agreement can produce fairly high values of the correlation coefficient. For example, Serfontein and Jaroszewicz (1978) found a correlation of 0·85 when they compared two more methods of assessing gestational age, the Robinson and the Dubowitz. They concluded that because the correlation was high and significantly different from zero, agreement was good. However, from their data a baby with a gestational age of 35 weeks by the Robinson method could have been anything between 34 and 39·5 weeks by the Dubowitz method. For two methods which purport to measure the same thing the agreement between them is not close, because what may be a high correlation in other contexts is not high when comparing things that should be highly related anyway. The test of significance of the null hypothesis $\rho=0$ is beside the point. It is unlikely that we would consider totally unrelated quantities as candidates for a method comparison study.

The correlation coefficient is not a measure of agreement; it is a measure of association.

Thus it is quite wrong, for example, to infer from a high correlation that "the methods . . . may be used interchangeably" (Hallman and Teramo, 1981).

At the extreme, when measurement error is very small and correlations correspondingly high, it becomes difficult to interpret differences. Oldham *et al.* (1979) state that: "Connecting [two types of peak flow meter] in series produces a correlation coefficient of 0·996, which is a material improvement on the figure of 0·992 obtained when they are used separately". It is difficult to imagine another context in which it were thought possible to improve materially on a correlaton of 0·992. As Westgard and Hunt (1973) have said: "The correlation coefficient . . . is of no practical use in the statistical analysis of comparison data".

Regression

Linear regression is another misused technique in method comparison studies. Often the slope of the least squares regression line is tested against zero. This is equivalent to testing the correlation coefficient against zero, and the above remarks apply. A more subtle problem is illustrated by the work of Carr *et al.* (1979), who compared two methods of measuring the heart's left ventricular ejection fraction. These authors gave not only correlation coefficients but the regression line of one method, Teichholz, on the other, angiography.

They noted that the slope of the regression line differed significantly from the line of identity. Their implied argument was that if the methods were equivalent the slope of the regression line would be 1. However, this ignores the fact that both dependent and independent variables are measured with error. In our previous notation the expected slope is $\beta = \sigma_T^2/(\sigma_A^2 + \sigma_T^2)$ and is therefore less than 1. How much less than 1 depends on the amount of measurement error of the method chosen as independent. Similarly, the expected value of the intercept will be greater than zero (by an amount that is the product of the mean of the true values and the bias in the slope) so that the conclusion of Floss *et al.* (1982) that "with a slope not differing significantly from unity but a statistically highly significant y-intercept, the presence of a systematic difference . . . is demonstrated" is unjustified.

We do not reject regression totally as a suitable method of analysis, and will discuss it further below.

Asking the right question

None of the previously discussed approaches tells us whether the methods can be considered equivalent. We think that this is because the authors have not thought about what question they are trying to answer. The questions to be asked in method comparison studies fall into two categories:

(a) Properties of each method:
  How repeatable are the measurements?
(b) Comparison of methods:
  Do the methods measure the same thing on average? That is, is there any relative bias?
  What additional variability is there? This may include both errors due to repeatability and errors due to patient/method interactions. We summarize all this as "error".

Under properties of each method we could also include questions about variability between observers, between times, between places, between position of subject, etc. Most studies standardize these, but do not consider their effects, although when they are considered, confusion may result. Altman's (1979) criticism of the design of the study by Serfontein and Jaroszewicz (1978) provoked the response that: "For the actual study it was felt that the fact assessments were made by two different observers (one doing only the Robinson technique and the other only the Dubowitz method) would result in greater objectivity" (Serfontein and Jaroszewicz, 1979). The effects of method and observer are, of course, totally confounded.

We emphasize that this is a question of *estimation*, both of error and bias. What we need is a design and analysis which provide estimates of both error and bias. No single statistic can estimate both.

3 Proposed method of analysis

Just as there are several invalid approaches to this problem, there are also various possible types of analysis which are valid, but none of these is without difficulties. We feel that a relatively simple pragmatic approach is preferable to more complex analyses, especially when the results must be explained to non-statisticians.

It is difficult to produce a method that will be appropriate for all circumstances. What follows is a brief description of the basic strategy that we favour; clearly the various possible complexities which could arise might require a modified approach, involving additional or even alternative analyses.

Properties of each method: repeatability

The assessment of repeatability is an important aspect of studying alternative methods of measurement. Replicated measurements are, of course, essential for an assessment of repeatability, but to judge from the medical literature the collection of replicated data is rare. One possible reason for this will be suggested later.

Repeatability is assessed for each measurement method separately from replicated measurements on a sample of subjects. We obtain a measure of repeatability from the within-subject standard deviation of the replicates. The British Standards Institution (1979) define a coefficient of repeatability as "the value below which the difference between two single test results . . . may be expected to lie with a specified probability; in the absence of other indications, the probability is 95 per cent". Provided that the differences can be assumed to follow a Normal distribution this coefficient is 2·83 σ_r, where σ_r is the within-subject standard deviation. σ_r must be estimated from a suitable experiment. For the purposes of the present analysis the standard deviation alone can be used as the measure of repeatability.

It is important to ensure that the within-subject repeatability is not associated with the size of the measurements, in which case the results of subsequent analyses might be misleading. The best way to look for an association between these two quantities is to plot the standard deviation against the mean. If there are two replicates x_1 and x_2 then this reduces to a plot of $|x_1-x_2|$ against $(x_1+x_2)/2$. From this plot it is easy to see if there is any tendency for the amount of variation to change with the magnitude of the measurements. The correlation coefficient could be tested against the null hypothesis of $r=0$ for a formal test of independence.

If the within-subject repeatability is found to be independent of the size of the measurements, then a one-way analysis of variance can be performed. The residual standard deviation is an overall measure of repeatability, pooled across subjects.

If, however, an association is observed, the results of an analysis of variance could be misleading. Several approaches are possible, the most appealing of which is the transformation of the data to remove the relationship. In practice the logarithmic transformation will often be suitable. If the relationship can be removed, a one-way analysis of variance can be carried out. Repeatability can be described by calculating a 95 per cent range for the difference between two replicates. Back-transformation provides a measure of repeatability in the original units. In the case of log transformation the repeatability is a percentage of the magnitude of the measurement rather than an absolute value. It would be preferable to carry out the same transformation for measurement by each method, but this is not essential, and may be totally inappropriate.

If transformation is unsuccessful, then it may be necessary to analyse data from a restricted range of measurements only, or to subdivide the scale into regions to be analysed separately. Neither of these approaches is likely to be particularly satisfactory. Alternatively, the repeatability can be defined as a function of the size of the measurement.

Properties of each method: other considerations

Many factors may affect a measurement, such as observer, time of day, position of subject, particular instrument used, laboratory, etc. The British Standards Institution (1979) distinguish between repeatability, described above, and reproducibility, "the value below which two single test results . . . obtained under different conditions . . . may be expected to lie with a specified probability". There may be difficulties in carrying out studies of reproducibility in many areas of medical interest. For example, the gestational age of a newborn baby could not be determined at different times of year or in different places. However, when it is possible to vary conditions, observers, instruments, etc., the methods described above will be appropriate provided the effects are random. When effects are fixed, for example when comparing an inexperienced observer and an experienced observer, the approach used to compare different methods, described below, should be used.

Comparison of methods

The main emphasis in method comparison studies clearly rests on a direct comparison of the results obtained by the alternative methods. The question to be answered is whether the methods are comparable to the extent that one might replace the other with sufficient accuracy for the intended purpose of measurement.

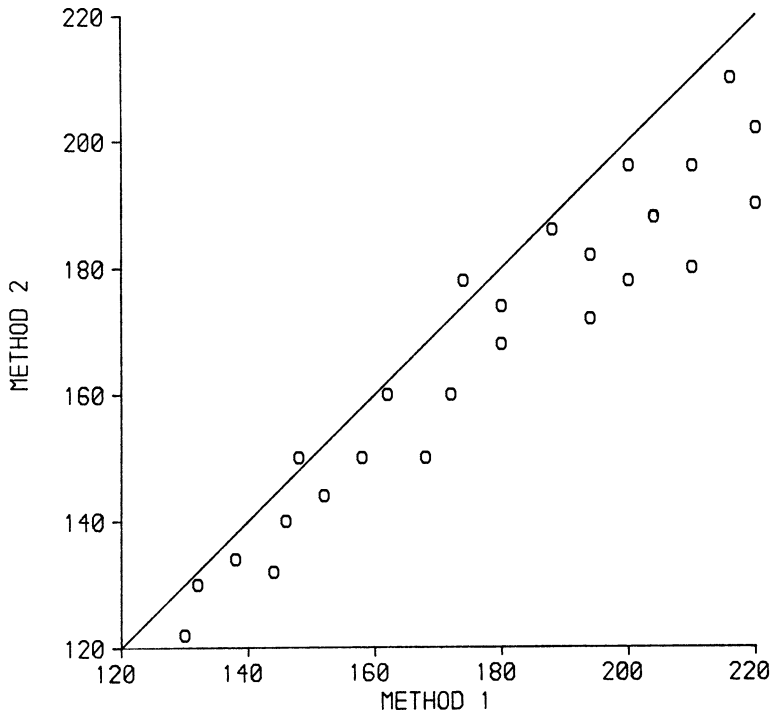

Fig. 1. Comparison of two methods of measuring systolic blood pressure.

The obvious first step, one which should be mandatory, is to plot the data. We first consider the unreplicated case, comparing methods A and B. Plots of this type are very common and often have a regression line drawn through the data. The appropriateness of regression will be considered in more detail later, but whatever the merits of this approach, the data will always cluster around a regression line by definition, whatever the agreement. For the purposes of *comparing* the methods the line of identity (A = B) is much more informative, and is essential to get a correct visual assessment of the relationship. An example of such a plot is given in Figure 1, where data comparing two methods of measuring systolic blood pressure are shown.

Although this type of plot is very familiar and in frequent use, it is not the best way of looking at this type of data, mainly because much of the plot will often be empty space. Also, the greater the range of measurements the better the agreement will appear to be. It is preferable to plot the difference between the methods (A − B) against (A + B)/2, the average. Figure 2 shows the data from Figure 1 replotted in this way. From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in A − B for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study. It is closely related to the usual plot of residuals after model-fitting, and the patterns observed may be similarly varied. In the example shown (Figure 2) there was a significant relationship between the method difference and the size of measurement ($r = 0{\cdot}45$, $n = 25$, $P = 0{\cdot}02$). This test is equivalent to a test of equality of the total

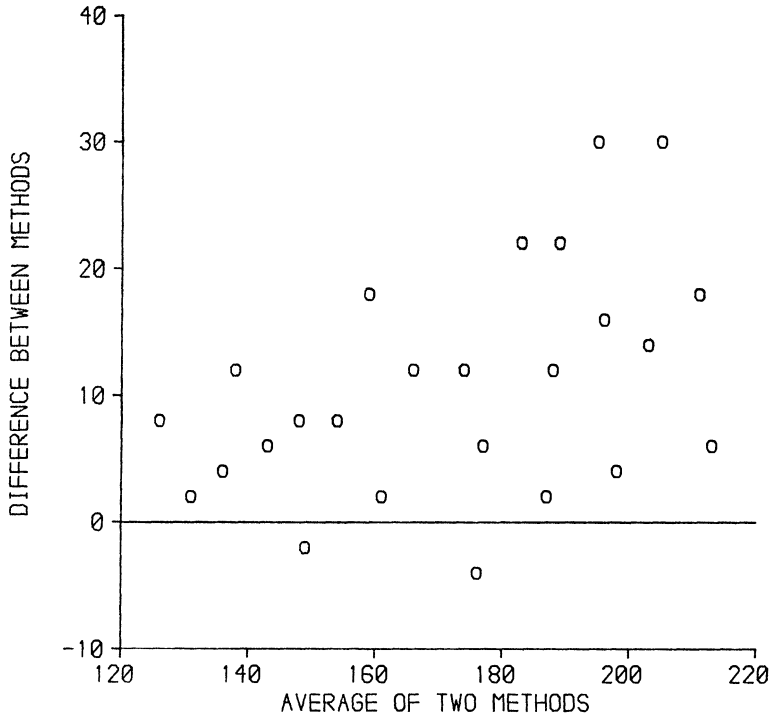

Fig. 2. Data from Figure 1 replotted to show the difference between the two methods against the average measurement.

variances of measurements obtained by the two methods (Pitman, 1939; see Snedecor and Cochran, 1967, pp. 195–7).

As in the investigation of repeatability, we are looking here for the independence of the between-method differences and the size of the measurements. With independence the methods may be compared very simply by analysing the individual A − B differences. The mean of these differences will be the relative *bias*, and their standard deviation is the estimate of *error*. The hypothesis of zero bias can be formally examined by a paired *t*-test.

For the data of Carr *et al.* (1979) already discussed, the correlation of the individual differences with the average value was $-0{\cdot}36$ ($P > 0{\cdot}1$), so that an assumption of independence is not contradicted by the data. Figure 3 shows these data plotted in the suggested manner. Also shown is a histogram of the individual between-method differences, and superimposed on the data are lines showing the mean difference and a 95 per cent range calculated from the standard deviation. A composite plot like this is much more informative than the usual plot (such as Figure 1).

If there is an association between the differences and the size of the measurements, then as before, a transformation (of the raw data) may be successfully employed. In this case the 95 per cent limits will be asymmetric and the bias will not be constant. Additional insight into the appropriateness of a transformation may be gained from a plot of $|A-B|$ against $(A+B)/2$, if the individual differences vary either side of zero. In the absence of a suitable transformation it may be reasonable to describe the differences between the methods by regressing $A-B$ on $(A+B)/2$.

For replicated data, we can carry out these procedures using the means of the replicates. The estimate of bias will be unaffected, but the error will be reduced. We can estimate the standard deviation of the difference between individual measurements from the standard deviation of the difference between means by

$$\text{var}(A-B) = n \text{ var}(\bar{A}-\bar{B})$$

where n is the number of replicates.

Within replicated data it may be felt desirable to carry out a two-way analysis of variance, with main effects of individuals and methods, in order to get better estimates. Such an analysis would need to be supported by the analysis of repeatability, and in the event of the two methods not being equally repeatable the analysis would have to be weighted appropriately. The simpler analysis of method differences (Figure 2) will also need to be carried out to ascertain that the differences are independent of the size of the measurements, as otherwise the answers might be misleading.

Alternative analyses

One alternative approach is least squares regression. We can use regression to predict the measurement obtained by one method from the measurement obtained by the other, and calculate a standard error for this prediction. This is, in effect, a calibration approach and does not directly answer the question of comparability. There are several problems that can arise, some of which have already been referred to. Regression does not yield a single value for relative precision (error), as this depends upon the distance from the mean. If we do try to use regression methods to assess comparability difficulties arise because there no obvious estimate of bias, and the parameters are difficult to interpret. Unlike the analysis of variance model, the parameters are affected by the range of the observations and for the results to apply generally the methods ought to have been compared on a random sample of subjects – a condition that will very often not be met. The problem of the underestimation (attenuation) of the slope of the regression line has been considered by Yates (Healy, 1958), but the other problems remain.

Other methods which have been proposed include principal component analysis (or

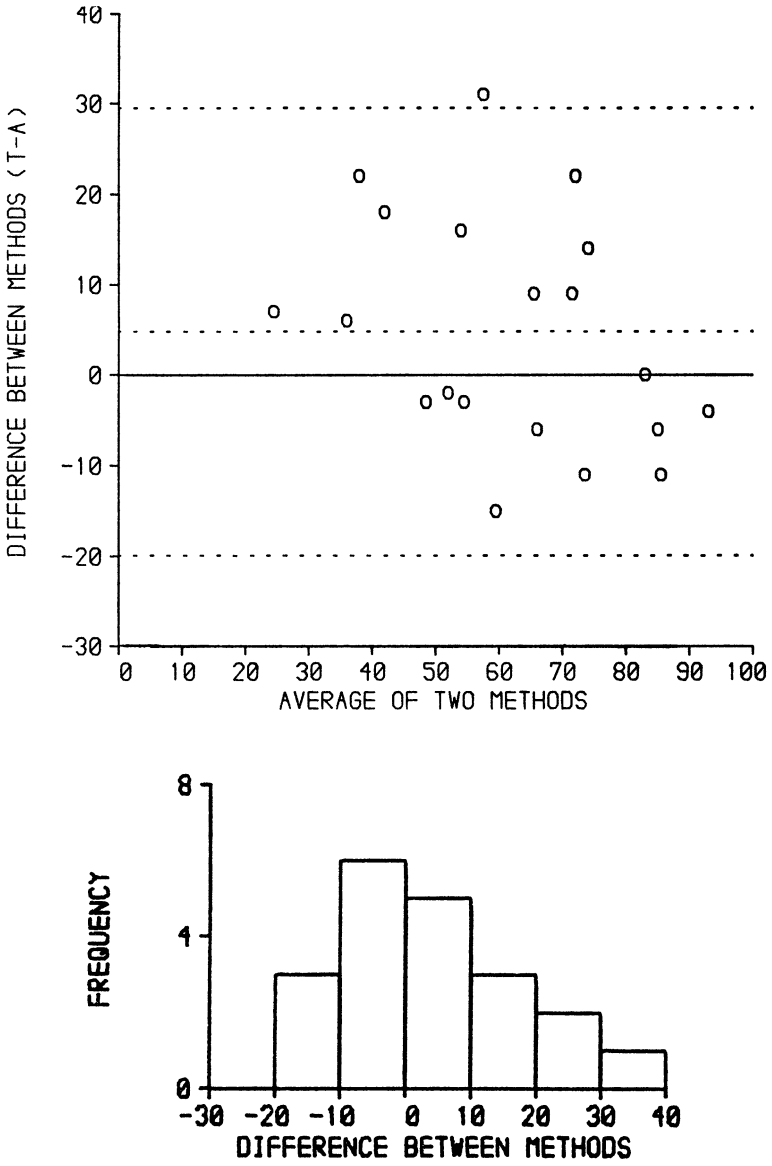

Fig. 3. Comparison of two methods of measuring left ventricular ejection fraction (Carr *et al.*, 1979) replotted to show error and bias.

orthogonal regression) and regression models with errors in both variables (structural relationship models) (see for example Carey *et al.*, 1975; Lawton *et al.*, 1979; Cornbleet and Gochman, 1979; Feldmann *et al.*, 1981). The considerable extra complexity of such analysis will not be justified if a simple comparison is all that is required. This is especially true when the results must be conveyed to and used by non-experts, e.g. clinicians. Such

methods will be necessary, however, if it is required to *predict* one measurement from the other – this is nearer to calibration and is not the problem we have been addressing in this paper.

Why does the comparison of methods cause so much difficulty?

The majority of medical method comparison studies seem to be carried out without the benefit of professional statistical expertise. Because virtually all introductory courses and textbooks in statistics are method-based rather than problem-based, the non-statistician will search in vain for a description of how to proceed with studies of this nature. It may be that, as a consequence, textbooks are scanned for the most similar-looking problem, which is undoubtedly correlation. Correlation is the most commonly used method, which may be one reason for so few studies involving replication, since simple correlation cannot cope with replicated data. A further reason for poor methodology is the tendency for researchers to imitate what they see in other published papers. So many papers are published in which the same incorrect methods are used that researchers can perhaps be forgiven for assuming that they are doing the right thing. It is to be hoped that journals will become enlightened and return papers using inappropriate techniques for reanalysis.

Another factor is that some statisticians are not as aware of this problem as they might be. As an illustration of this, the blood pressure data shown in Figures 1 and 2 were taken from the book *Biostatistics* by Daniel (1978), where they were used as the example of the calculation coefficient. A counter-example is the whole chapter devoted to method comparison (by regression) by Strike (1981). More statisticians should be aware of this problem, and should use their influence to similarly increase the awareness of their non-statistical colleagues of the fallacies behind many common methods.

Conclusions

1. Most common approaches, notably correlation, do not measure agreement.
2. A simple approach to the analysis may be the most revealing way of looking at the data.
3. There needs to be a greater understanding of the nature of this problem, by statisticians, non-statisticians and journal referees.

Acknowledgements

We would like to thank Dr David Robson for helpful discussions during the preparation of this paper, and Professor D. R. Cox, Professor M. J. R. Healy and Mr A. V. Swan for comments on an earlier draft.

Appendix

Covariance of two methods of measurement in the presence of measurement errors

We have two methods A and B of measuring a true quantity T. They are related T by $A = T + \epsilon_A$ and $B = T + \epsilon_B$, where ϵ_A and ϵ_B are experimental errors. We assume that the errors have mean zero and are independent of each other and of T, and define the following variances:

$$\text{var}(T) = \sigma_T^2, \text{ var}(\epsilon_A) = \sigma_A^2, \text{ and var}(\epsilon_B) = \sigma_B^2$$

Now the covariance of A and B is given by

$$E(AB)-E(A)E(B)$$
$$=E\{(T+\epsilon_A)(T+\epsilon_B)\}-E(T+\epsilon_A)E(T+\epsilon_B)$$
$$=E\{(T^2+\epsilon_AT+\epsilon_BT+\epsilon_A\epsilon_B)\}-\{E(T)+E(\epsilon_A)\}\{E(T)+E(\epsilon_B)\}$$

But $E(\epsilon_A)=E(\epsilon_B)=0$, and the errors and T are independent, so

$$E(\epsilon_A)E(T)=E(\epsilon_B)E(T)=0$$

and

$$E(\epsilon_A\epsilon_B)=E(\epsilon_A)E(\epsilon_B)=0$$

Hence $\text{cov}(A,B)=E(T^2)-\{E(T)\}^2=\sigma_T^2$

References

Altman, D. G. (1979). Estimation of gestational age at birth – comparison of two methods. *Archives of Disease in Childhood* **54,** 242–3.

British Standards Institution (1979). Precision of test methods, part 1: guide for the determination of repeatability and reproducibility for a standard test method. BS 5497, Part 1. London.

Carey, R. N., Wold, S. and Westgard, J. O. (1975). Principal component analysis: an alternative to "referee" methods in method comparison studies. *Analytical Chemistry* **47,** 1824–9.

Carr, K. W., Engler, R. L., Forsythe, J. R., Johnson, A. D. and Gosink, B. (1979). Measurement of left ventricular ejection fraction by mechanical cross-sectional echocardiography. *Circulation* **59,** 1196–1206.

Cater, J. I. (1979). Confirmation of gestational age by external physical characteristics (total maturity score). *Archives of Disease in Childhood* **54,** 794–5.

Cornbleet, P. J. and Gochman, N. (1979). Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical Chemistry* **25,** 432–8.

Daniel, W. W. (1978). *Biostatistics: a Foundation for Analysis in the Health Sciences*, 2nd edn. Wiley, New York.

Feldmann, U., Schneider, B., Klinkers, H. and Haeckel, R. (1981). A multivariate approach for the biometric comparison of analytical methods in clinical chemistry. *Journal of Clinical Chemistry and Clinical Biochemistry* **19,** 121–37.

Hallman, M. and Teramo, K. (1981). Measurement of the lecithin/sphingomyelin ratio and phosphatidylglycerol in amniotic fluid: an accurate method for the assessment of fetal lung maturity. *British Journal of Obstetrics and Gynaecology* **88,** 806–13.

Healy, M. J. R. (1958). Variations within individuals in human biology. *Human Biology* **30,** 210–8.

Hunyor, S. M., Flynn, J. M. and Cochineas, C. (1978). Comparison of performance of various sphygmomanometers with intra-arterial blood-pressure readings. *British Medical Journal* **2,** 159–62.

Keim, H. J., Wallace, J. M., Thurston, H., Case, D. B., Drayer, J. I. M. and Laragh, J. H. (1976). Impedance cardiography for determination of stroke index. *Journal of Applied Physiology* **41,** 797–9.

Laughlin, K. D., Sherrard, D. J. and Fisher, L. (1980). Comparison of clinic and home blood-pressure levels in essential hypertension and variables associated with clinic-home differences. *Journal of Chronic Diseases* **33,** 197–206.

Lawton, W. H., Sylvestre, E. A. and Young-Ferraro, B. J. (1979). Statistical comparison of multiple analytic procedures: application to clinical chemistry. *Technometrics* **21,** 397–416.

Oldham, H. G., Bevan, M. M. and McDermott, M. (1979). Comparison of the new miniature Wright peak flow meter with the standard Wright peak flow meter. *Thorax* **34,** 807–8.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika* **31,** 9–12.

Ross, H. A., Visser, J. W. E., der Kinderen, P. J., Tertoolen, J. F. W. and Thijssen, J. H. H. (1982). A comparative study of free thyroxine estimations. *Annals of Clinical Biochemistry* **19,** 108–13.

Serfontein, G. L. and Jaroszewicz, A. M. (1978). Estimation of gestational age at birth – comparison of two methods. *Archives of Disease in Childhood* **53,** 509–11.

Serfontein, G. L. and Jaroszewicz, A. M. (1979). Estimation of gestational age at birth – comparison of two methods. *Archives of Disease in Childhood* **54,** 243.

Snedecor, G. W. and Cochran, W. G. (1967). *Statistical Methods*, 6th edn. University Press, Iowa.

Strike, P. W. (1981). *Medical Laboratory Statistics*. P. S. G. Wright, Bristol.

Westgard, J. O. and Hunt, M. R. (1973). Use and interpretation of common statistical tests in method-comparison studies. *Clinical Chemistry* **19,** 49–57.