Eye (2006) 20, 749–754
© 2006 Nature Publishing Group All rights reserved 0950-222X/06 $30.00
www.nature.com/eye

Statistical strategies to assess reliability in ophthalmology

N Patton[1], T Aslam[2] and G Murray[3]

Abstract

Reliability of measurements and measurers is important so that we can trust the measurements we record. However, the statistical techniques used to assess reliability of measurements or measurers in the ophthalmic literature are often inappropriate, and not able to evaluate reliability between measurements/measurers. We review the techniques used in reliability studies for both continuous and categorical data, and describe appropriate statistical methods for particular study designs. We also highlight current techniques that are not appropriate in the analysis of reliability, but that are still commonly used in the ophthalmic literature. We hope that by highlighting these, we shall discourage their future use.

Eye (2006) **20,** 749–754. doi:10.1038/sj.eye.6702097; published online 2 December 2005

Keywords: reliability; agreement; reproducibility; test–retest; statistics

Introduction

Instruments to record measurements (eg intraocular pressure, corneal thickness, axial length, etc) should only be used if we know they are reliable. In addition, instruments developed for newly quantifiable measurements (eg posterior capsular opacification,[1–3] ocular blood flow[4–6]) must also be shown to be reliable before they can be applied either in clinical or research settings.[7] Reliability means that the measurements that the instrument records are reproducible at different time intervals (test–retest reliability) and that those observers making the measurements produce repeatable results, both for the same observer over a period of time (intraobserver reliability) and between different observers on the same subject (interobserver reliability).[8–15] In addition, reliability is used in the context of assessing agreement between one method of measurement and another (method comparison or parallel reliability). Thus, reliability (as well as sensitivity and specificity) is a prerequisite to using any instruments of measurement and forms a major component of ophthalmic research.[16]

However, techniques of data analysis employed in studies assessing reliability in the ophthalmic literature vary tremendously[17–22] and studies often use techniques that are inappropriate for the task they are set.[17,23–28] In this paper, we review current statistical techniques employed in reliability/agreement studies and provide a framework to help the ophthalmologist decide on the most appropriate statistical method.

Continuous *vs* categorical data

Statistical techniques for agreement studies depend on whether data are continuous (derived from a possible range of values or an underlying continuum) or categorical.

Analyses of continuous data in agreement studies

Correlation

This is a very commonly performed technique used to assess level of agreement, but is inappropriate as it measures association and not agreement. A highly significant and large value for the correlation coefficient (r) can coexist with gross bias.[17,23,25,26,29–31] For example, when comparing the performance of two observers, observer A may consistently overestimate the result when compared to observer B (Figure 1). A highly significant value for r would be achieved, and this could be misinterpreted as revealing good agreement between both observers. Inspection of Figure 1 reveals a fixed systematic bias between observers A and B, that is, observer A consistently measures a higher

[1]Lions Eye Institute, Nedlands, Western Australia, WA, Australia

[2]Manchester Royal Eye Hospital, Manchester, UK

[3]Medical Statistics, University of Edinburgh, Teviot Place, Edinburgh, UK

Correspondence: N Patton, Lions Eye Institute, 2 Verdun Street, Nedlands, Western Australia, WA 6009, Australia
Tel: +44 61 8 63891216;
Fax: +44 61 8 93463333.
E-mail: niallpatton@hotmail.com

None of the authors have any financial or proprietary interest in any products included in this review

Published online:
2 December 2005

reading by a fixed amount that does not change according to the size of the reading measured.

The intraclass correlation coefficient (ICC) (ratio of between-groups variance to the total variance[32]) is another correlation statistic often used to assess agreement. The ICC varies from +1 (perfect agreement) to 0 (no agreement). The ICC is designed to assess agreement when there is no intrinsic ordering between two variables (ie the measurements are interchangeable, such as test–retest reliability using the same method[33]). However, when dealing with method comparison studies, there is a very clear ordering of the two variables (the two methods under comparison).

While the ICC is better able to avoid the confusion of mistaking linear association for agreement, it suffers from being highly dependent on the range of values measured, that is, the greater the variability between subjects, the greater the value of the ICC. Consider a hypothetical group of five subjects who have IOP recorded by two different techniques (Goldmann tonometry *vs* tonopen) (Table 1). For study 1, the ICC for the two techniques is $r=0.167$ ($P=0.38$). When we repeat the study on a different set of subjects (study 2), the ICC for the two techniques becomes $r=0.95$ ($P=0.002$). However, despite such extreme differences in the value of the ICC, the actual level of agreement in study 1 and 2 look on inspection to be approximately equal (both studies have the same differences recorded). The reason for the disparity in the ICC values is that in study 1, the range of IOPs is much narrower than study 2.

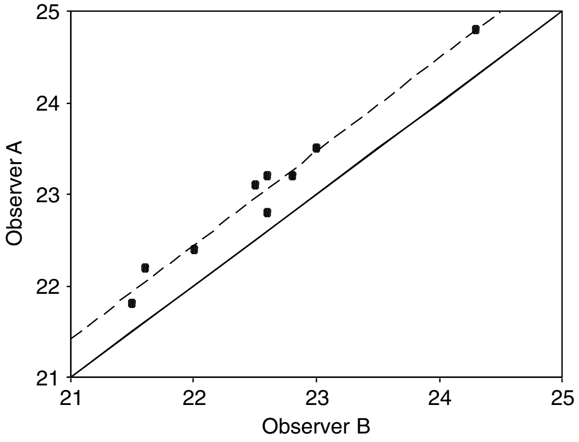

Figure 1 Scatterplot diagram of the results of axial length measurements using B-scan ultrasound from observers A and B. The dotted line represents the ordinary least squares (OLS) regression line ($r=0.98$, $P<0.0001$). The solid line represents the line of equality ($y=x$).

However, the ICC may be used to measure agreement[34] particularly when between more than two observers/methods.

'Limits of agreement' techniques

In 1983, Bland and Altman[35] published their seminal article on agreement analysis. The 'limits of agreement' technique has become an increasingly popular technique in agreement studies, and has been adopted by many clinical scientists due to it being simple to execute and easy to comprehend, using simple graphics and elementary statistics. This technique involves firstly calculating the differences for each pair of values, and then plotting the differences against the corresponding means for each pair. The values of the differences (A–B) should be normally distributed and should be equally scattered for all levels of the corresponding mean.[26] This graphical method also reveals extreme outliers affecting the data sample. The upper and lower 'limits of agreement' correspond to the mean difference (A–B)±1.96 standard deviations (SDs). Inspection of the graph will illustrate the upper and lower 'limits of agreement', which represents the interval within which 95% of differences between measurements/measurers are expected to lie. The decision as to whether good agreement is demonstrated is a matter of clinical judgement. Three hypothetical Bland–Altman plots (Figures 2–4) illustrate how bias can be identified by inspection of the plot. In interpreting Bland and Altman plots, it is important to consider if variability is comparable over the full range of measurements (Figure 3). Often, the variability increases as the

Table 1 Comparison of two different techniques for measuring IOP (Goldmann tonometry *vs* tonopen)

Study 1				*Study 2*			
Subject	*(A) Goldmann*	*(B) Tonopen*	*A–B*	*Subject*	*(A) Goldmann*	*(B) Tonopen*	*A–B*
1	20	22	−2	1	25	27	−2
2	22	20	2	2	22	20	2
3	20	20	0	3	20	20	0
4	22	22	0	4	18	18	0
5	20	20	0	5	15	15	0

All measurements are in mmHg.

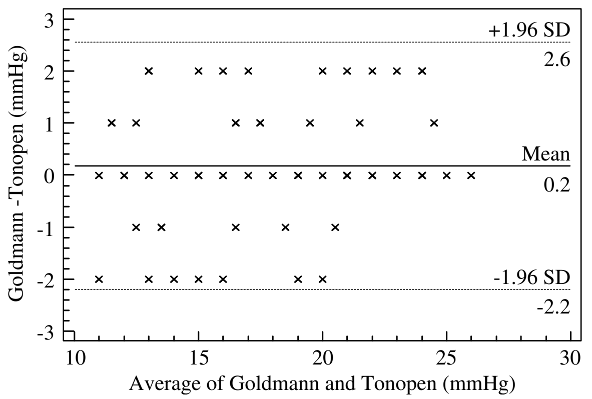

Figure 2 Hypothetical Bland–Altman plot of IOP recorded by Goldmann tonometry and tonopen. The solid line represents the mean difference (0.2 mmHg), and the dotted lines represent the upper (+2.6 mmHg) and lower (−2.2 mmHg) limits of agreement. This shows a mean difference between both measurements close to zero and no change in the magnitude of difference as the mean IOP increases.

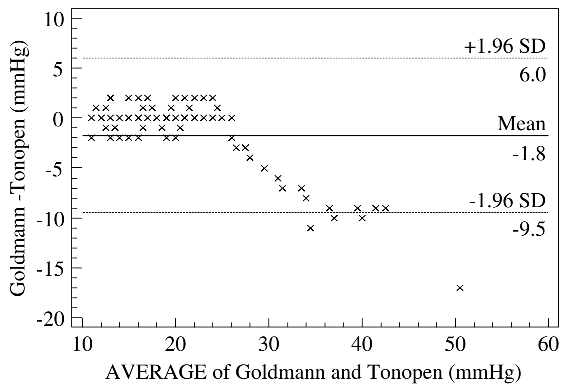

Figure 3 Hypothetical Bland–Altman plot of IOP recorded by Goldmann tonometry and tonopen. The solid line represents the mean difference (−1.8 mmHg), and the dotted lines represent the upper (+6 mmHg) and lower (−9.5 mmHg) limits of agreement. This hypothetical example illustrates good agreement between both methods of measurement for the range of IOP <25 mmHg, but beyond this range the relationship breaks down and the Tonopen measures much higher IOPs than the Goldmann tonometer.

measurement increases.[28,36] If so, then the variation in the percent difference may be fixed, and the plot may be redrawn on a logarithmic scale, for example, plotting the ratio or the per cent difference, rather than the absolute difference between the two variables. If the variability is relatively constant, then one looks for any systematic trend in the mean difference (see Figure 4). The presence of bias may in itself not be a problem, provided it is known and can be adjusted for. An illustration of the advantage of the 'limits of agreement' technique over correlation in assessing agreement is provided by Murray and Miller.[37]

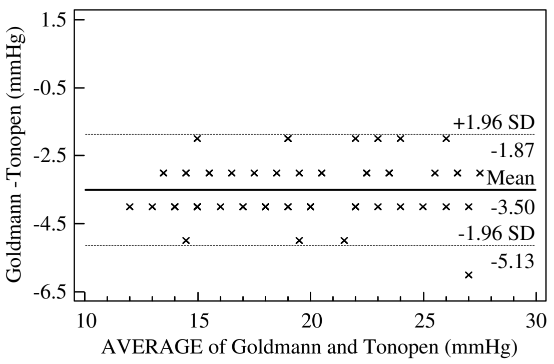

Figure 4 Hypothetical Bland–Altman plot of IOP recorded by Goldmann tonometry and tonopen. The mean difference was −3.5 mmHg; the upper and lower limits of agreement were −1.9 and −5.1 mmHg, respectively. This shows a fixed systematic bias (the Tonopen was consistently recording a higher IOP than the Goldmann tonometer, but the size of the difference did not change with increasing IOP).

When there are repeated measures (replicate measurements) performed by two methods on the same subjects, calculating the mean of the replicate measurements by each method and then using those pairs of means to compare the two methods can be performed using the 'limits of agreement' method.

Cotter *et al*[38] (in assessing test–retest reliability) and Beck *et al*[39] (assessing two methods and their test–retest reliability) used a combination of both the ICC and Bland–Altman plots in their analyses. Their analyses are explicit and comprehensive and allow the reader to accept their conclusions with confidence.

Linear regression techniques

Linear regression techniques can be used to assess agreement, but there are many models of linear regression and it is important to choose the correct regression model for the agreement study.[25,30] Models such as standardised principal component analysis,[30] Deming regression,[40] and the nonparametric Passing–Bablok model[41–43] may be used, but ordinary least squares regression (OLS) is inappropriate as an assumption of OLS regression is that the values of the y variable are random, whereas the x variable is fixed, without random error. This is rarely the case when examining agreement between measurers/methods.[30]

Coefficients of repeatability

The repeatability coefficient is a useful statistic when dealing with repeat measurements by the same method (test–retest reliability).[26,44,45] When there are only two measurements per subject, the repeatability coefficient is 2 × (SD of the differences) between the repeated measures. This is the repeatability coefficient adopted by the British Standards Institution (BSI).[6] As the mean difference between two measurements using the same method should be zero, we expect 95% of differences to be <2 SD. Repeatability coefficients can often be used in conjunction with other tests of test–retest reliability, for example, ICCs. As the 'repeatability coefficient' is measured in the same units as the variable being measured, it should not strictly be termed a 'coefficient' as coefficients are by definition dimensionless. However, the term 'coefficient of repeatability' has been adopted to describe this statistical technique. Ruamviboonsuk *et al*[46] use the coefficient of repeatability in comparing test–retest reliability between two visual acuity tests. However, no mention was made regarding whether the SD was unrelated to the magnitude of the score (a necessary assumption to use coefficient of repeatability).

Coefficient of variation

The coefficient of variation provides a relative measure of data dispersion compared to the mean, expressed as either a quotient or as a percentage of the within-subject SD divided by the mean. As the coefficient of variation is dimensionless, it can be used to assess repeatability between two methods of measurement recorded on different scales. However, to be used correctly, the coefficient of variation should be independent of the mean.[47–50]

Categorical data

The following techniques can be used to compare agreement for categorical data.

A cross tabulation (row × column) table

A cross tabulation with rater 1's category frequencies attributed to the row, and rater 2's attributed to the column (see Table 2) provides almost all relevant information for assessing agreement. The diagonal of the table represents where rater 1 agrees with rater 2. For good agreement, one would expect on inspection the diagonal of the tabulation to have the greatest number (see Table 2). The data can then be summarised further by calculating kappa or weighted kappa statistics.

Table 2 Hypothetical $r \times c$ array of observed frequencies between two raters on a scale (A, B, C in increasing order)

Rater 2	*Rater 1*			
	A	*B*	*C*	*Row total*
A	**7**	0	0	7
B	1	**9**	2	12
C	0	2	**4**	6
Column total	8	11	6	

Bold values occur on the agreement diagonal. All other values of the $r \times c$ are termed off-diagonal entries. K value = 0.69 (if A, B, C are a nominal scale). $K_w = 0.81$ (if A, B, C are an ordinal scale).

Cohen's Kappa coefficient[51,52]

The original purpose of the kappa statistic (the unweighted kappa (K)) is to compare two measurers who use the same nominal scale.[53] The kappa statistic gives a value that is an indication of the amount of agreement present, corrected for that which would have occurred by chance.[29,53–55] The values of K can range from -1 to $+1$ (zero translates as agreement no better than that which would have occurred by chance). For example, Azuara-Blanco *et al*[56] use the unweighted kappa coefficient (in this case, for bivariate data) to analyse agreement between intra/interobserver reliability for glaucoma experts in the detection of glaucomatous changes of the optic disk.

Weighted kappa statistic (K_w)

This is intended for ordinal categorical data (eg none, mild, moderate, severe). A weighting system is incorporated into the K statistic, so that greater degrees of disagreement (eg none pairing with severe) are given greater penalty. The commonest weighting system is a quadratic weighting system in which the weights for proportional disagreement progress geometrically.

For the K_w, the value depends solely on the values that are not on the diagonal line of agreement (ie all off-diagonal entries). The cells on the diagonal line of agreement are given a value of 0. A simple method for interpreting the K or K_w value is the empirical approach proposed by Landis and Koch,[57] whereby $0.81 \leq K \leq 1.00$ represents almost perfect agreement, $0.61 \leq K \leq 0.80$ represents substantial agreement, and so on until $0.00 \leq K \leq 0.20$ represents slight agreement.

As K and K_w statistics are correlative statistics, they are dependent on the prevalence of the characteristic being studied.[54,58] This makes it difficult to compare two or more kappa values when the true prevalence for the groups or characteristics differs.

Percentage agreement

This is a value that relates the number of measurements that agree to the total number of comparisons (expressed as a percentage). It is a crude assessment that does not tell us a great deal about agreement and does not incorporate any adjustment for agreement by chance.[29] Hence, it is of little use in agreement studies, especially

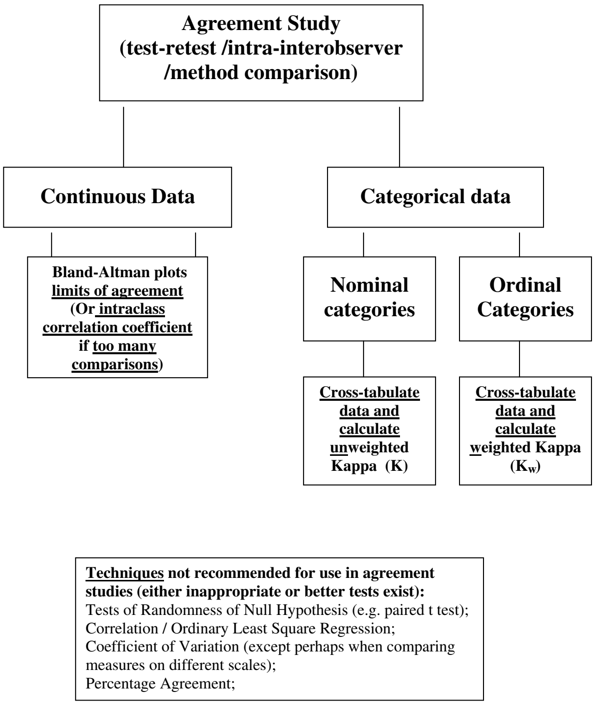
Agreement Study
(test-retest /intra-interobserver
/method comparison)
Continuous Data
Categorical data
Bland-Altman plots
limits of agreement
(Or intraclass
correlation coefficient
if too many
comparisons)
Nominal
categories
Ordinal
Categories
Cross-tabulate
data and
calculate
unweighted
Kappa (K)
Cross-tabulate
data and
calculate
weighted Kappa
(Kw)
Techniques not recommended for use in agreement
studies (either inappropriate or better tests exist):
Tests of Randomness of Null Hypothesis (e.g. paired t test);
Correlation / Ordinary Least Square Regression;
Coefficient of Variation (except perhaps when comparing
measures on different scales);
Percentage Agreement;

Figure 5 Flow chart to choose an appropriate statistical approach for a particular agreement study.

as there are superior techniques to compare categorical agreement.

Summary

We describe simple and appropriate statistical strategies that can be used in the analysis of agreement studies. In addition, some approaches that are inappropriate have been highlighted, to discourage their future use in ophthalmic journals. Below, we provide a flow chart (Figure 5) to help choose an appropriate method when analysing agreement of both continuous and categorical data. The reader should note that this serves as a flexible guide and not a regimented structure. It should also be noted that the technique to use should if at all possible be decided *a priori*, at the study design stage.

In a recent review article, Altman emphasised the importance of the misuse of statistics in medical journals.[59] It is important for all those practising evidence-based medicine to be familiar with appropriate statistical techniques for agreement analysis, so that they can judge the relative merits of those publications claiming to show reliability for a particular test.

References

1 Barman SA, Hollick EJ, Boyce JF, Spalton DJ, Uyyanonvara B, Sanguinetti G *et al*. Quantification of posterior capsular opacification in digital images after cataract surgery. *Invest Ophthalmol Vis Sci* 2000; **41**: 3882–3892.

2 Aslam TM, Dhillon B. Neodymium:YAG laser capsulotomy: a clinical morphological analysis. *Graefes Arch Clin Exp Ophthalmol* 2002; **240**: 972–976.

3 Findl O, Buehl W, Menapace R, Georgopoulos M, Rainer G, Siegl H *et al*. Comparison of 4 methods for quantifying posterior capsule opacification. *J Cataract Refract Surg* 2003; **29**: 106–111.

4 Cioffi GA, Alm A. Measurement of ocular blood flow. *J Glaucoma* 2001; **10**: S62–S64.

5 Aydin A, Wollstein G, Price LL, Schuman JS. Evaluating pulsatile ocular blood flow analysis in normal and treated glaucomatous eyes. *Am J Ophthalmol* 2003; **136**: 448–453.

6 British Standards Institution. *Precision of Test Methods I: Guide for the Determination and Reproducibility for a Standard Test Method (BS 5497, Part I)*. BSI: London, 1979.

7 Burns C. Parallels between research and diagnosis: the reliability and validity issues of clinical practice. *Nurse Pract* 1991; **16**: 42, 45, 49–50.

8 Koran LM. The reliability of clinical methods, data and judgments (second of two parts). *N Engl J Med* 1975; **293**: 695–701.

9 Koran LM. The reliability of clinical methods, data and judgments (first of two parts). *N Engl J Med* 1975; **293**: 642–646.

10 Dunn G. Design and analysis of reliability studies. *Stat Methods Med Res* 1992; **1**: 123–157.

11 Shrout PE. Measurement reliability and agreement in psychiatry. *Stat Methods Med Res* 1998; **7**: 301–317.

12 van Saane N, Sluiter JK, Verbeek JH, Frings-Dresen MH. Reliability and validity of instruments measuring job satisfaction—a systematic review. *Occup Med (London)* 2003; **53**: 191–200.

13 Wittchen HU. Reliability and validity studies of the WHO—Composite International Diagnostic Interview (CIDI): a critical review. *J Psychiatr Res* 1994; **28**: 57–84.

14 Feinstein AR. A bibliography of publications on observer variability. *J Chronic Dis* 1985; **38**: 619–632.

15 Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). *J Clin Epidemiol* 1992; **45**: 567–580.

16 Margo CE, Harman LE, Mulla ZD. The reliability of clinical methods in ophthalmology. *Surv Ophthalmol* 2002; **47**: 375–386.

17 Yoshida A, Feke GT, Mori F, Nagaoka T, Fujio N, Ogasawara H *et al*. Reproducibility and clinical application of a newly developed stabilized retinal laser Doppler instrument. *Am J Ophthalmol* 2003; **135**: 356–361.

18 van Leeuwen R, Chakravarthy U, Vingerling JR, Brussee C, Hooghart AJ, Mulder PG *et al*. Grading of age-related maculopathy for epidemiological studies: is digital imaging as good as 35-mm film? *Ophthalmology* 2003; **110**: 1540–1544.

19 Henderer JD, Liu C, Kesen M, Altangerel U, Bayer A, Steinmann WC *et al*. Reliability of the disk damage likelihood scale. *Am J Ophthalmol* 2003; **135**: 44–48.

20 Bhan A, Bhargava J, Vernon SA, Armstrong S, Bhan K, Tong L *et al*. Repeatability of ocular blood flow pneumotonometry. *Ophthalmology* 2003; **110**: 1551–1554.

21 Carpineto P, Ciancaglini M, Zuppardi E, Falconio G, Doronzo E, Mastropasqua L. Reliability of nerve fiber layer thickness measurements using optical coherence

tomography in normal and glaucomatous eyes. *Ophthalmology* 2003; **110**: 190–195.

22 Holz FG, Jorzik J, Schutt F, Flach U, Unnebrink K. Agreement among ophthalmologists in evaluating fluorescein angiograms in patients with neovascular age-related macular degeneration for photodynamic therapy eligibility (FLAP-study). *Ophthalmology* 2003; **110**: 400–405.

23 Patton N, Aslam T. Reproducibility and clinical application of a newly developed stabilized retinal laser Doppler instrument. *Am J Ophthalmol* 2003; **136**: 578–579; author reply 579.

24 Delgado J, Fernandez-Jimenez MC. Inappropriate analysis of reproducibility. *Br J Haematol* 2003; **123**: 745; author reply 745–746.

25 Ludbrook J. Statistical techniques for comparing measurers and methods of measurement: a critical review. *Clin Exp Pharmacol Physiol* 2002; **29**: 527–536.

26 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **1**: 307–310.

27 Rousson V, Gasser T, Seifert B. Assessing intrarater, interrater and test–retest reliability of continuous measurements. *Stat Med* 2002; **21**: 3431–3446.

28 Musadiq M, Patsoura E, Hughes S, Yang Y. Measurements of linear dimensions on fundus photographs: comparison between photographic film and digital measurements. *Eye* 2003; **17**: 619–622.

29 Kramer MS, Feinstein AR. Clinical Biostatistics LIV: the biostatistics of concordance. *Clin Pharmacol Ther* 1981; **29**: 111–123.

30 Ludbrook J. Comparing methods of measurement. *Clin Exp Pharmacol Physiol* 1997; **24**: 193–203.

31 Lee J, Koh D, Ong CN. Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput Biol Med* 1989; **19**: 61–70.

32 Fisher AR. *Statistical Methods for Research Workers*. Oliver and Boyd Ltd: Edinburgh, 1925.

33 Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput Biol Med* 1990; **20**: 337–340.

34 Muller R, Buttner P. A critical discussion of intraclass correlation coefficients. *Stat Med* 1994; **13**: 2465–2476.

35 Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *Statistician* 1983; **32**: 307–317.

36 Patton N, Aslam T. Statistical analysis of agreement in measurement comparison studies. *Eye* 2005; **19**: 363.

37 Murray G, Miller R. Statistical comparison of two methods of clinical measurement. *Br J Surg* 1990; **77**: 384–387.

38 Cotter SA, Chu RH, Chandler DL, Beck RW, Holmes JM, Rice ML *et al*. Reliability of the electronic early treatment diabetic retinopathy study testing protocol in children 7 to <13 years old. *Am J Ophthalmol* 2003; **136**: 655–661.

39 Beck RW, Moke PS, Turpin AH, Ferris III FL, SanGiovanni JP, Johnson CA *et al*. A computerized method of visual acuity testing: adaptation of the early treatment of diabetic retinopathy study testing protocol. *Am J Ophthalmol* 2003; **135**: 194–205.

40 Demming WE. *Statistical Adjustment of Data*. Wiley: New York, 1943.

41 Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, Part 1. *J Clin Chem Biochem* 1983; **21**: 709–720.

42 Passing H, Bablok W. Comparison of several regression procedures for method comparison studies and determination of sample sizes. Applicationof linear regression procedures for method comparison studies in clinical chemistry, Part II. *J Clin Chem Biochem* 1984; **22**: 431–445.

43 Payne RB. Method comparison. *Ann Clin Biochem* 1997; **34**: 319–320.

44 Bland JM, Altman DG. Measurement error. *BMJ* 1996; **312**: 1654.

45 Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999; **8**: 135–160.

46 Ruamviboonsuk P, Tiensuwan M, Kunawut C, Masayaanon P. Repeatability of an automated Landolt C test, compared with the early treatment of diabetic retinopathy study (ETDRS) chart testing. *Am J Ophthalmol* 2003; **136**: 662–669.

47 Chinn S. The assessment of methods of measurement. *Stat Med* 1990; **9**: 351–362.

48 Rankin G, Stokes M. Reliability of assessment tools in rehabilitation: an illustration of appropriate statistical analyses. *Clin Rehabil* 1998; **12**: 187–199.

49 Shechtman O. Is the coefficient of variation a valid measure for detecting sincerity of effort of grip strength? *Work* 1999; **13**: 163–169.

50 Allison DB. Limitations of coefficient of variation as index of measurement reliability. *Nutrition* 1993; **9**: 559–561.

51 Scott WA. Reliability of content analysis: the case of nominal scale coding. *Public Opin Quart* 1955; **75**: 321–325.

52 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; **20**: 37–46.

53 Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. *Stat Med* 2002; **21**: 2109–2129.

54 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987; **126**: 161–169.

55 Cyr L, Francis K. Measures of clinical agreement for nominal and categorical data: the kappa coefficient. *Comput Biol Med* 1992; **22**: 239–246.

56 Azuara-Blanco A, Katz LJ, Spaeth GL, Vernon SA, Spencer F, Lanzl IM. Clinical agreement among glaucoma experts in the detection of glaucomatous changes of the optic disk using simultaneous stereoscopic photographs. *Am J Ophthalmol* 2003; **136**: 949–950.

57 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.

58 Thompson WD, Walter SD. A reappraisal of the kappa coefficient. *J Clin Epidemiol* 1988; **41**: 949–958.

59 Altman DG. Statistics in medical journals: some recent trends. *Stat Med* 2000; **19**: 3275–3289.