NIWA
Taihoro Nukurangi

A Proposal for Strength-of-Agreement Criteria for Lin.s.Concordance Correlation Coefficient

NIWA Client Report: HAM2005-062
May 2005

NIWA Project: MOH05201

NIWA
Taihoro Nukurangi

A Proposal for Strength-of-Agreement Criteria for Lin.s.Concordance Correlation Coefficient

G. B. McBride

Prepared for

Ministry of Health

NIWA Client Report: HAM2005-062
May 2005

NIWA Project: MOH05201

National Institute of Water & Atmospheric Research Ltd
Gate 10, Silverdale Road, Hamilton
P O Box 11115, Hamilton, New Zealand
Phone +64-7-856 7026, Fax +64-7-856 0151
www.niwa.co.nz

Contents

Executive Summary		iv
1.	Introduction	1
2.	Modelling	2
3.	Conclusions	6
4.	Acknowledgement	6
5.	References	7
APPENDIX Lin's Concordance Correlation Coefficient		8

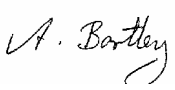
Reviewed by:

Dr R. Craggs

Approved for release by:

Dr R.J. Davies-Colley

Formatting checked



Executive Summary

Strength-of-agreement criteria for Lin's concordance correlation coefficient are proposed to assess the degree of equivalence between a new laboratory method and a "gold standard" method, as follows:

Strength-of-agreement	Continuous variables	QuantiTray methods
Almost perfect	>0.99	>0.90
Substantial	0.95..0.99	0.8..0.9
Moderate	0.90..0.95	0.65..0.8
Poor	<0.90	<0.65

To assess the degree of agreement for a given set of data it is proposed that the lower one-sided 95% confidence limit for the calculated concordance correlation coefficient should be compared to the values in this table. For example, consider a QuantiTray method, using a single set of 51 wells or a 49x48 sheet of larger and small wells. If the lower one-sided 95% confidence limit on its coefficient returns a value of 0.81, then one can confidently conclude that the degree of agreement is "Substantial".

It is desirable that the assessment be performed on at least 25 samples, preferably 50.

Introduction

The Ministry of Health wishes to have a statistically defensible measure of concordance when considering the performance of a new analytical test as compared to that of a "gold standard" method. Lin's concordance correlation coefficient (Lin 1989, 2000; Zar 1996; McBride 2003a; McBride 2005) has emerged as the best measure of agreement for two methods of measuring the same continuous variable (e.g., chemical concentration). Perfect agreement would be signified by all data lying exactly on the 1:1 line, and Lin's coefficient is unique in basing its formulae on exactly that notion.

Here we use statistical modelling to examine how the coefficient relates to several random variations in the responses of laboratory tests to known true concentrations, for both continuous measures (such as for a chemical concentration) and discrete measures (i.e., MPN results, using multi-well procedures, such as Colilert^{TM}). The results have been used by the author and competent (ESR) microbiologists to derive the strength-of-agreement table, with separate columns for the continuous and discrete cases. This separation is most important, because MPN results are inherently "noisy".

The criteria are broadly similar to those developed by Landis & Koch (1977), as already adopted by the Ministry of Health (see McBride 2003a), and a Ministry of Health/NIWA website (http://www.niwa.co.nz/services/statistical/kappa).

Only the broad outlines of this modelling are given here. Full details are given in an accompanying @RISK workbook.[15] The full set of equations is listed in this report's Appendix.

[15] This is the Excel workbook "Criteria for Lin concordance correlation coefficient.xls". @RISK (Palisade Corporation 2000) is a plug-in to Excel. If it is not installed on the machine used to read the workbook one can still see all the comments and formulae, but the Monte Carlo calculation results and graphics cannot be displayed. The first sheet of the workbook (named README) summarises the calculation procedures, which are further elaborated in individual sheets.

Modelling

In essence, a synthetic random series of true concentrations have been generated and a laboratory test's results have been generated to compare with them. In the case of the continuous variables, these latter results have been generated by perturbing the series true series using normal distributions, with drift and shift parameters (as shown on Figure 1), and also using an uneven distribution of values (also shown on Figure 1). ("Shift" denotes a constant upward bias, "drift" denotes an ever-increasing increasing bias.) Example sets of outputs are given in Figures 1–4. The values on the Y-axis on these figures represent the laboratory method's simulated response to the true values shown on the X-axis.

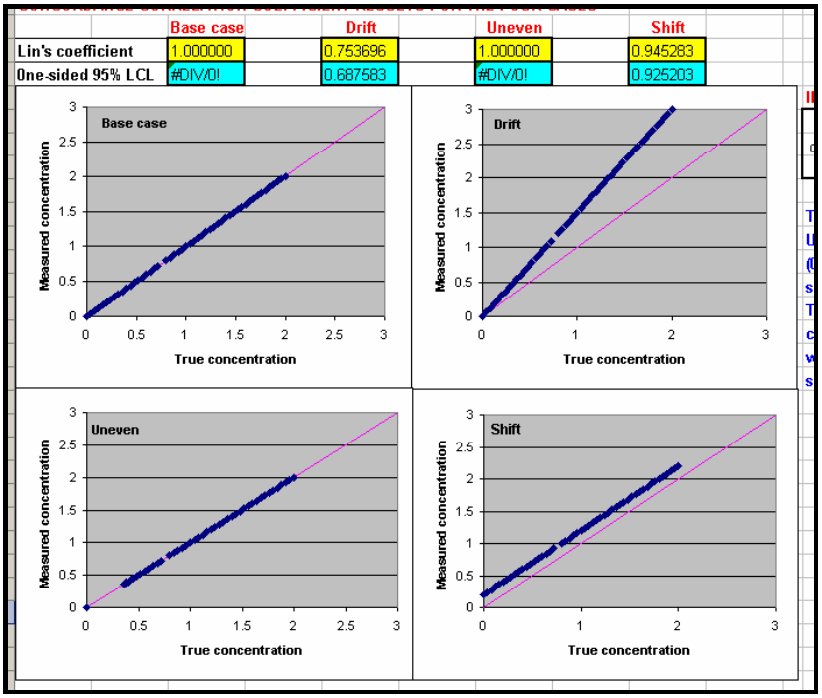


Figure 1: Base case for continuous variables (continuous variable).[16]

[16] The "#DIV/0!" entry for the lower one-sided 95% confidence interval arises only when there is no simulated noise and all data lie exactly on the 1:1 line.

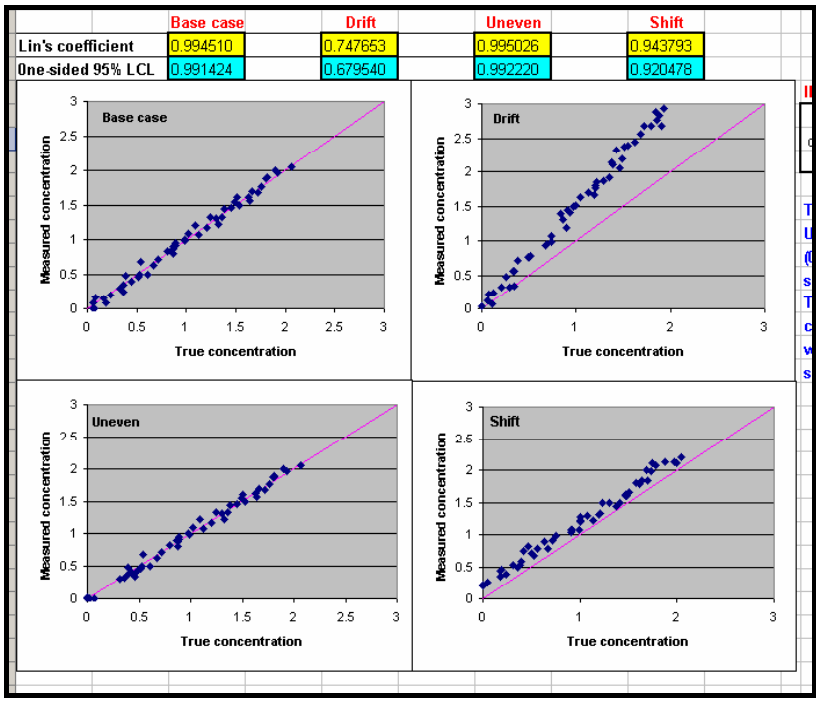


Figure 2: Small random variations (continuous variable).

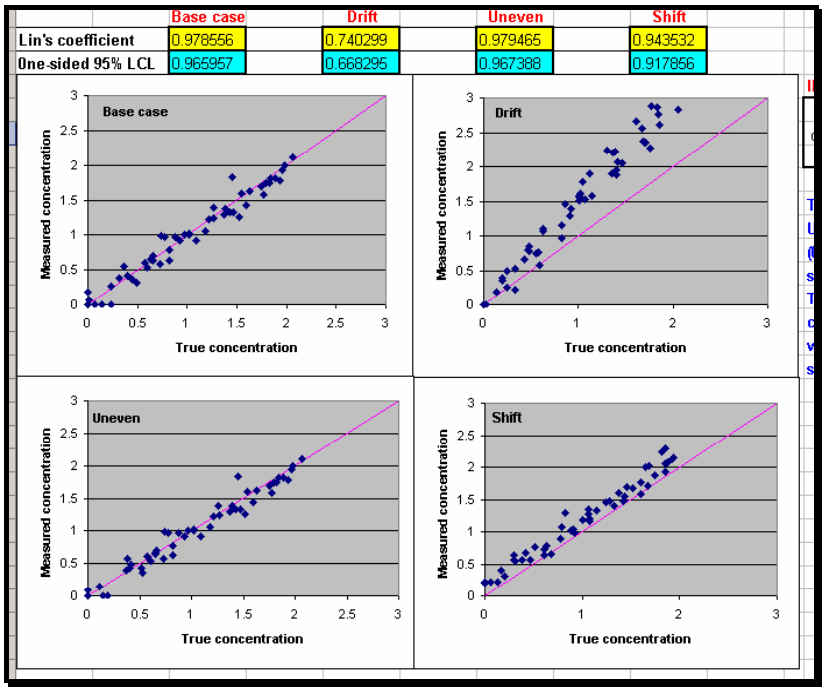


Figure 3: Medium random variations (continuous variable).

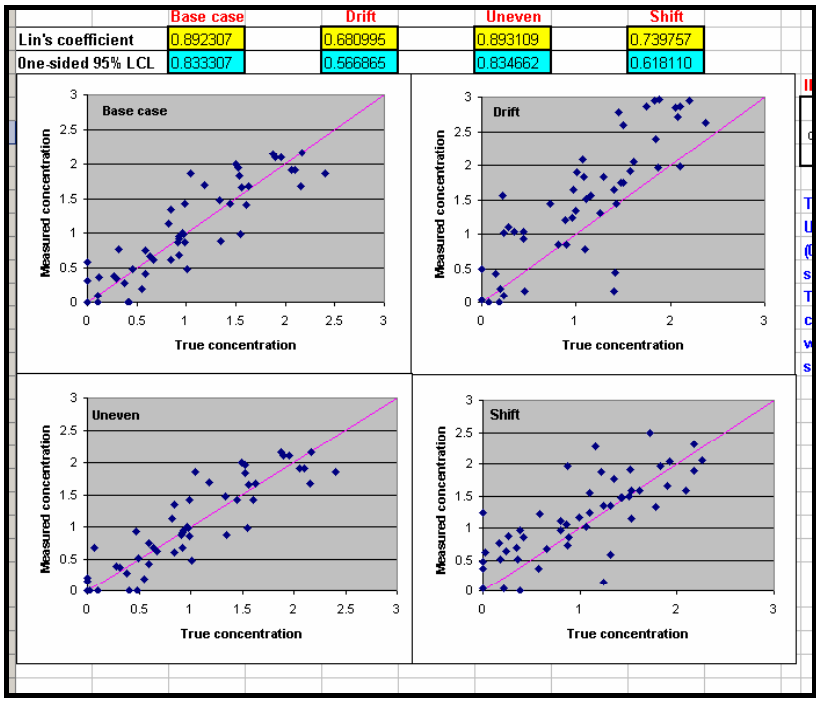


Figure 4: Large random variations (continuous variable).

For discrete data the base case already contains some variability—because MPN data can only take on a fixed number of numbers; most numbers are actually impossible in the MPN system. Accordingly, if the true concentration does not accord with one of the few possible MPN numbers, the laboratory result cannot lie on the 1:1 agreement line. For example, if a MPN table has adjacent entries of 14 and 18 MPN per 1200 mL, and the true value is 16 per 100 mL, "perfect" concordance is formally impossible.

For MPN calculations reported herein, the random selections are made using "occurrence probabilities", as described in McBride (2003b), as is depicted in Figure 5. The Figure also indicates what happens when some random variations ("noise") and "drift" are included—this is for the IDEXX 51-well setup (all wells with constant volume).

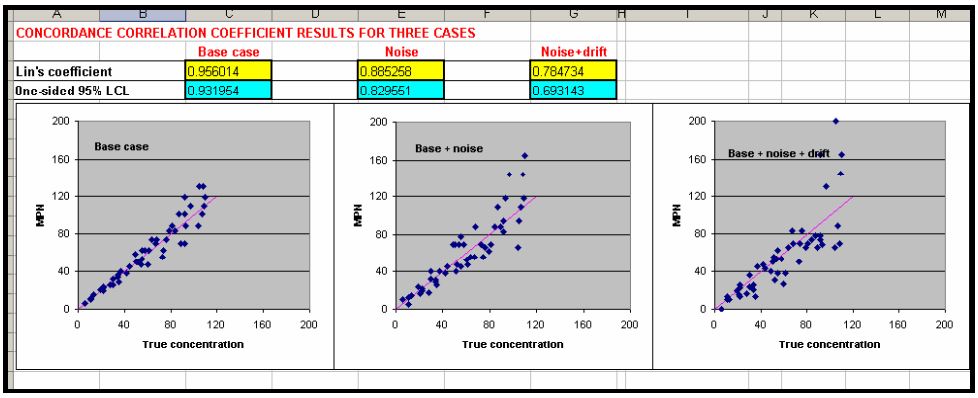


Figure 5: Cases studies for discrete MPN data, using IDEXX 51-well setup.

Monte Carlo simulations have also been performed on simulated 3x3x3 decimal dilution series data. Results are very variable—the base case can show extremely low concordance correlation coefficients. In other words, there were too few tubes to make any meaningful concordance statements about such tests.

Note that we have not analysed the IDEXX QuantiTray/2000 (49x48 wells) in this manner, because the computational effort would seem to be immense.[17] Instead, we appeal to a similarity with the QuantiTray 51-well results. The top two MPN values of the QuantiTray/2000 table jump by about 20% (from 1986 to 2419 per 100 mL), in very similar fashion to the top values of the 51-well table (whose top two values are 165.2 and 200.5 per 100 mL). That is, we are assuming that the basic outlier patterns obtained with the 51-well table would be similar to those obtained using the QuantiTray/2000 table. In contrast, the 3x3x3 table has enormous increases in predicted MPN values at the top end (i.e., 3-3-0, 3-3-1, 3-3-2 patterns of positive tubes give MPNs of 23.7, 45.9 and 110 per 100 mL—more than a doubling in the highest two values).[18]

Many simulations have been performed using these models. The results are summarised in the conclusions.

[17] It would require many random samples to be taken from a 48^2 array for about 1000 values of true concentrations.
[18] This is why its base case can have such low Lin coefficient values.

Conclusions

To assess the degree of equivalence between a new laboratory method and a "gold standard" method, I propose strength-of-agreement criteria for Lin's concordance correlation coefficient as follows:

Strength-of-agreement	Continuous variables	QuantiTray methods
Almost perfect	>0.99	>0.90
Substantial	0.95..0.99	0.8..0.9
Moderate	0.90..0.95	0.65..0.8
Poor	<0.90	<0.65

To assess the degree of agreement for a given set of data the lower one-sided 95% confidence limit for the calculated concordance correlation coefficient should be compared to the values in this table (as is fully documented in McBride 2003a). For example, consider a QuantiTray method, using a single set of 51 wells or a 49x48 sheet of larger and small wells. If the lower one-sided 95% confidence limit on its coefficient returns a value of 0.81, then one can confidently conclude that the degree of agreement is "Substantial".

It is desirable that the assessment be performed on at least 25 samples, although 50 would be better still (for reasons given in McBride 2003a).

It must be stressed that these results have not been tested in the open science literature, by acceptance and publication in a reputable science journal. The task has been much more complex than was first envisaged, and a more complete analysis (including comparison of cases of MPN versus MPN results applied to the same dilution series setup) should be contemplated, with a view to publication of the results in the peer-reviewed scientific literature.

Acknowledgement

Mr Andrew Ball (ESR, Christchurch) provided detailed advice on the strength of agreement criteria reported herein.

References

Altman, D. G.; Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. Statistician 32: 307–317.

Bland, J. M.; Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, February 8: 307–310.

Landis, J.R.; Koch, G.G. (1977). The measurement of observer agreement for categorical data. Biometrics 33: 159–174.

Lin, L.I-K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics 45: 255–268.

Lin, L.I-K. (2000). A note on the concordance correlation coefficient. Biometrics 56: 324–325.

McBride, G.B. (2003a). Statistical validation criteria for drinking-water microbiological methods. NIWA Client Report: HAM2003-012. Report to Ministry of Health. 17 p.

McBride, G.B. (2003b). Preparing exact Most Probable Number (MPN) tables using occupancy theory, and accompanying measures of uncertainty. NIWA Technical Report No. 121 (ISSN 1174-2631). 62 pp.

McBride, G.B. (2005). Using Statistical Methods for Water Quality Management: Issues, Options and Solutions. Wiley, Hoboken, NJ.

Steichen, T. J.; Cox, N. J. (2002). A note on the concordance correlation coefficient. Stata Journal 2(2): 183–189.

Zar, J.H. (1996). Biostatistical Analysis. 3rd ed. Prentice-Hall, Upper Saddle River, NJ.

APPENDIX Lin.s.Concordance Correlation Coefficient

The following material is repeated from McBride (2003a), with minor corrections; it is included here for completeness. It is also discussed in some length in McBride (2005).

The sample concordance correlation coefficient (denoted as $\hat{\rho}_c$) was first proposed by Lin (1989) for assessment of concordance in continuous data. It represents a breakthrough in assessing concordance between alternative methods where the data are continuous (i.e., enumerative, not discrete),[19] in that it appears to avoid all the shortcomings associated with the panoply of usual procedures (Pearson correlation coefficient r, paired t-tests, least squares analysis for slope and intercept, coefficient of variation, intraclass correlation coefficient). It also appears to be superior to the previously proposed limits-of-agreement procedure (Altman & Bland 1983, Bland & Altman 1986), as discussed by Steichem & Cox (2002).

The concordance correlation coefficient can range from –1 to +1, as does Pearsons r, but it cannot exceed r in absolute value. It is robust on as few as 10 pairs of data (Lin 1989). It appears, with a worked example, in a recent edition of a popular biostatistical text (Zar 1996).[20] It is important to note that there are typographical errors in Lin's original paper (and therefore repeated in Zar's book). Corrections have recently been published (Lin 2000), along with the claim that they have negligible effect. However, it has been shown by Steichen & Cox (2002) that the errors can be problematic when the assessed relationship approaches strong concordance – highly relevant in studies of equivalence.

Equations

In essence, Lin's coefficient measures vertical departures from the 1:1 (45^o) line of agreement between a gold standard method (on the X-axis) and a candidate method (on the Y-axis). The true value of the concordance correlation coefficient ρ_c is defined as

$$\rho_c = 1 - \frac{\delta}{\delta^*}, \qquad \text{(A.1)}$$

where

$$\delta = \text{expected squared perpendicular deviation from the } 45^o_0 \text{ line}, \qquad \text{(A.2)}$$

$$\delta^* = \delta \text{ for uncorrelated data,} \qquad \text{(A.3)}$$

[19] Enumerative data may at first be thought to be discrete, but the use of dilutions and reporting to standard volumes (e.g., 100 mL in the case of E. coli) effectively makes the data continuous.
[20] Zar denotes the coefficient as r_c.

where 45^{o}_{0} is the 1:1 concordance line passing through the origin.

To proceed, let us denote our n pairs of data as (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n). Also we define sample means and variances in the following manner, i.e.,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \,,\; S_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \,, \tag{A.4}$$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \,,\; S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 \,, \tag{A.5}$$

and the sample covariance

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right). \tag{A.6}$$[21]

Then the sample concordance correlation coefficient (see Lin 1989 for its derivation) is

$$\hat{\rho}_c = \frac{2S_{XY}}{S_X^2 + S_Y^2 + \left(\bar{X} - \bar{Y}\right)^2}. \tag{A.7}$$[22]

A standard error can be calculated for this estimate. In doing so it has been shown that the "Fisher transformation" is desirable to better meet the normal approximations to be invoked when confidence intervals are calculated or hypothesis tests are performed (Lin 1989). This transformation is

$$\hat{Z} = \tanh^{-1}\left(\hat{\rho}_c\right) = \frac{1}{2}\log_e\left\{\frac{\left(1 + \hat{\rho}_c\right)}{\left(1 - \hat{\rho}_c\right)}\right\}, \tag{A.8}$$

and we obtain the sample standard error of estimate (of $\hat{Z}$) as

$$S_{\hat{Z}} = \sqrt{\frac{\left(1 - r^2\right)\hat{\rho}_c^2}{\left(1 - \hat{\rho}_c^2\right)r^2} + \frac{2\hat{\rho}_c^3\left(1 - \hat{\rho}_c\right)u^2}{r\left(1 - \hat{\rho}_c^2\right)^2} - \frac{\hat{\rho}_c^4 u^4}{2r^2\left(1 - \hat{\rho}_c^2\right)^2}}{n - 2}}\,, \tag{A.9}$$[23]

[21] Note that the divisor for these S terms is n, whereas it is usual to use n–1 to ensure that the estimates are unbiased (unbiasedness is not always the most desirable property).
[22] The n–1 term in the denominator of Zar's version of this equation (eq. 18.76) is wrong: it should be just n.
[23] In Lin (1989) and in Zar (1996) the second group of terms under the radical had a coefficient of 4 on the numerator and the third group had a coefficient of 2 on the numerator. These were

where r is Pearson's correlation coefficient defined in the usual way [$r = S_{XY}/(S_X S_Y)$] and u is the "location shift relative to the scale" parameter, defined by

$$u = \frac{\left(\overline{X} - \overline{Y}\right)}{\sqrt{S_X S_Y}}. \qquad (A.10)^{24}$$

Then the lower one-sided confidence 95% confidence interval for Z is

$$L_Z = \text{lower one - sided 95\% confidence limit for } Z = \hat{Z} - 1.6449 S_{\hat{Z}}, \qquad (A.11)$$

and so, inverting the transformation in eq. (A.8) the 95% lower confidence limit for ρ_c is

$$L_{\rho_c} = \tanh\left(L_Z\right). \qquad (A.12)$$

corrected by Lin (2000), as shown in eq. (B.8), i.e., the first coefficient is 2 on the numerator and the second becomes a 2 on the denominator.

[24] In McBride (2003a) the right-hand-side included a multiplicative factor of (n–1)/n. This is not necessary.