How to use difference plots in quantitative method comparison studies

Patrick J Twomey

Address
Department of Clinical Biochemistry,
The Ipswich Hospital, Ipswich IP4 2SU, UK

Correspondence:
Dr Patrick J Twomey
E-mail: Taptwomey@aol.com

Abstract

Quantitative method comparison studies are fundamental to clinical biochemistry. The interpretation of quantitative method comparison studies relied heavily on correlation and regression methods until Bland and Altman first described the concept of absolute difference plots. Since then, many clinical biochemistry journals advocate the use of difference plots; however, there is a lot of ignorance about the validity as well as the pros and cons of the various difference plots. The most important issue in quantitative method comparisons studies is to determine limits of agreement that are valid across the whole range of values in the study so that correct data interpretation and conclusions occur. This article discusses validity as well as the pros and cons of difference plots and provides means to determine limits of agreement that are valid across the whole range of values in method comparison studies. Accordingly, correct data interpretation will be more likely and better conclusions should be arrived as a result.

Ann Clin Biochem 2006; **43:** 124–129

Until the mid-1980s, the interpretation of quantitative method comparison studies relied heavily on correlation and regression methods. In 1986, Bland and Altman first described in the medical literature the concept of absolute difference plots,[1] having already done so in a statistical journal in 1983.[2] The mean and absolute difference between each pair of readings from the two different methods is determined, and then all the absolute differences are plotted on the *y*-axis, against their corresponding means on the *x*-axis (Figure 1). Bland and Altman provided another means of examining data obtained from quantitative method comparison studies, and started the trend against the sole use of correlation/regression methods. As a result the absolute difference plot is frequently employed in quantitative method comparison studies, especially as many journals advocate its use.[3,4] However, there is a lot of ignorance about the validity as well as the pros and cons of the absolute difference plot and, accordingly, many users appear to employ the technique solely because it is expected by many journals.

As health-care professionals treat individual patients, the real issue in method comparison studies is not *whether* the two methods agree, but *how well* the two methods agree from the point of individual specimens.[5] Correlation coefficients assess the association between the two methods and linear regression methods assess whether the points lie on a straight line. While the line of equality, the regression line and other data can be added to scatter (*x* versus *y*) plots to improve such an assessment, such methods are not ideal, especially because of the poor resolution of such plots. The major thrust put forward by Bland and Altman was that the absolute difference plot could be used to provide a way of comparing the agreement between two methods and more importantly to determine whether the degree of agreement is acceptable from a clinical context. They proposed that the mean and standard deviation (SD) can be calculated from the absolute difference between the methods. They quite correctly pointed out that it is essential that the differences approximate to a Gaussian distribution (the *t* distribution when the sample number is small) before such procedures are attempted (Figure 1a and b), and if this is not the case, that the raw data be transformed such that the differences then assume a Gaussian distribution. Logarithmic transformation is often but not always sufficient to meet this criterion. Unfortunately, many authors inappropriately apply the difference plot technique[5,6] and therefore do not determine limits of agreement that are valid across the whole range of values. As a result, incorrect data interpretation and conclusions may arise.

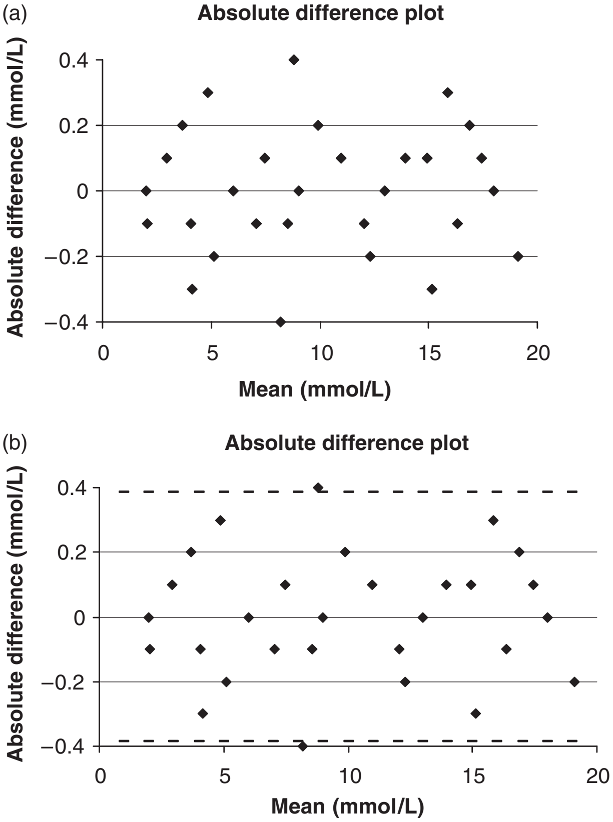

Figure 1 *(a) Absolute difference plot with the mean degree of spread around the central tendency being independent of the mean concentration. The slope, r, mean difference, SD, SEM and median difference for the difference plot data are 0.002, 0.043, 0.000, 0.1927, 0.0358 and 0.000 mmol/L, respectively. (b) Absolute difference plot as per (a) with the 95% limits of agreement, using the Gaussian distribution; the t-distribution would result in slightly wider 95% limits of agreement*

To overcome such problems, I suggest that the following procedure is followed:

1. The absolute difference plot data are plotted before calculating the mean and the SD and the data are visually inspected to see if they are 'well-behaved' as was originally intended.[5] The latter includes a uniform data distribution throughout the plot. In the first instance, it is important to check for outliers – data points 'whose discordancy from the majority of the sample is excessive in relation to the assumed distributional model for the sample, thereby leading to the suspicion that it is not generated by this model'[7] by applying Dixon's test (when the sample size is between 3 and 25)[8] or similar.[9] Gaussian data should be scattered at random; any pattern to the data is unlikely to have such a distribution including a funnel effect (increase in the degree of the differences with increasing mean values due to an increasing SD of the differences – Figure 2a) and slopes in the data (Figure 3a) that differ significantly from zero when there is a calibration error.[10] However, owing to the usually smaller y scale relative to the x scale in absolute difference plots, the slope of such data is often not sensitive.
2. The correlation coefficient for the difference plot data should be determined – it should approximate to zero.
3. If the data passes these conditions, I then suggest that the mean and 95% limits of agreement for the differences be added to the absolute difference plot and visually inspected to verify that the data is symmetrical about the mean (by counting the number of points above and below the mean, between the mean ± 1SD and between the mean ± 2SD) and that the 95% limits of agreement are valid across the whole of the range of values. The mean should be roughly equal to the median.
4. The number of points with values between the mean ± 1SD and between + 1SD and + 2SD/−1SD and −2SD should each approximate 34% and 14%, respectively, as expected in a Gaussian distribution.

Failure to adhere to these four points means that the data are unlikely to have a Gaussian distribution. If the data does not have a Gaussian distribution, the relative difference plot[11] may be appropriate (Figure 3a and b) and, if not, the raw data must be transformed, for example by log transformation, such that the differences assume a Gaussian distribution after the transformation if the benefits of the parametric statistics are to be realized. Such transformations allow extreme scores to be kept in the data-set while maintaining the relative ranking of scores, yet the error variance and skew present in the variable(s) can be reduced.[12] However, log-transformation may not be appropriate for the model being tested, or may affect its interpretation in undesirable ways by altering the relationship with the original variable, and the transformed data can be difficult to interpret.[13,14] Furthermore, you must be very careful transforming such data back to the original scale to avoid substantial downward 'transformation bias'.[15]

The next step is to examine the implications of the mean and the SD from a clinical perspective.[1,16] Unfortunately, few authors do so[6] or do so appropriately.[5] However, this step is crucial to the usefulness of any new test and is one of the key points that separate clinical biochemistry from pure biochemistry. If no systematic bias exists between the two methods, the mean difference should be approximately zero. The further the mean is from zero, the more likely that a systematic bias exists. By using the standard error of the mean, the 95% confidence interval for the mean can be calculated and if this interval does not include zero, then there is a statistically significant degree of systematic bias between the two methods. However, the 95% confidence interval for the mean can and should also be compared with acceptable clinical limits for systematic bias for the analyte in question to see if the difference is clinically acceptable or not. With regards to the SD, the important point to note is that it can be used as a comparative tool, as was originally intended by Bland and Altman.[1] As the data has a Gaussian distribution, one can reasonably expect approximately 95% of the differences between the two methods to lay between the mean ± 1.96SDs. These limits of agreement can be employed to decide whether the agreement for the individual data from the two methods is clinically acceptable: that is, does this interval represent acceptable agreement or not from a clinical point of view? However, many of those carrying out method comparative studies do not put the 95% limits of agreement into a clinical context. There may actually be no significant bias between the two methods, but the difference between the 95% limits of agreement may be so large from a clinical perspective that the two methods do not clinically agree. For example, imagine two glucose methods that have the relationship $y=x$ and a 95% limit of agreement of 1 mmol/L; there is agreement about the central tendency for the data, but a specimen with a value of 6.3 mmol/L from a fasting patient by one method could be consistent with euglycaemia, impaired fasting glycaemia or diabetes mellitus when analysed by the alternative method. Such a difference is clinically unacceptable and thus there is poor agreement between the methods despite minimal proportional and constant bias. Stöckl *et al.*[17] have logically added to the concept of the 95% limits of agreement by utilizing the flexibility of the Gaussian/*t*-distribution to add confidence intervals to the limits of agreement. Thus, the significance of the differences found when using small samples can be appreciated. If the difference is clinically but not statistically significant, a decision can be made to analyse more specimens if deemed appropriate; however, it is best to assess the power of the study (that is, the minimum sample size that should be used) prior to analysis.[18]

If the data cannot be made to assume a Gaussian distribution, then non-parametric statistical methods can potentially be employed to evaluate the spread of the individual differences from a clinical point of view as per the parametric equivalents. Non-parametric

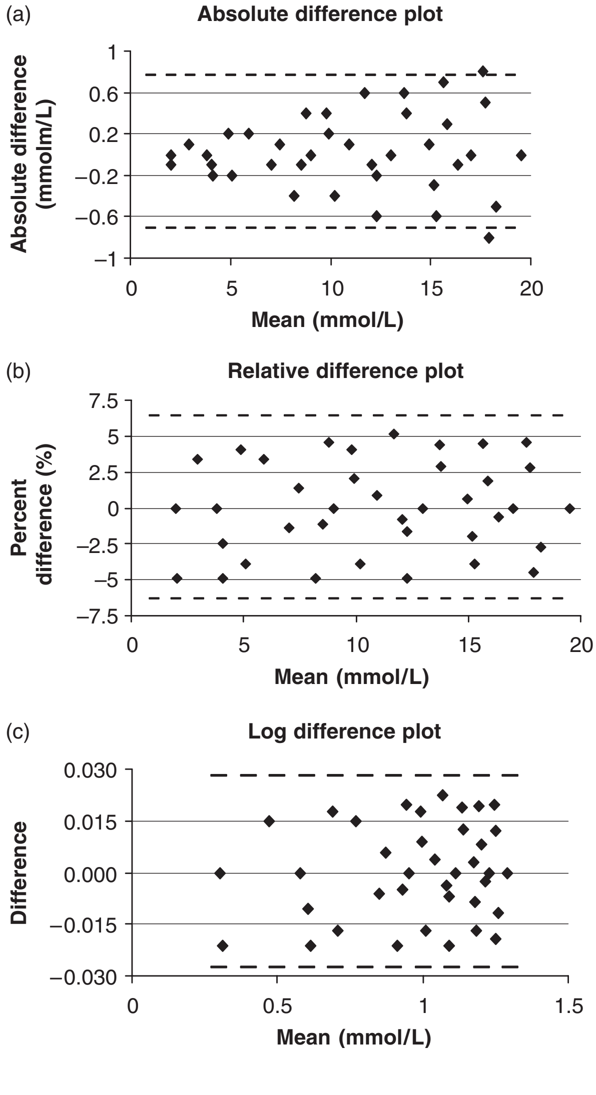

Figure 2 *(a) Absolute difference plot with a 'funnel' effect (that is, the spread around the central tendency increases with increasing mean concentration). Horizontal 95% limits of agreement are not valid across the whole ranges of values, whereas funnel-shaped limits of agreement would be. The slope, r, mean difference, SD, SEM and median difference for the difference plot data are 0.004, 0.051, 0.023, 0.3631, 0.0581 and 0.000 mmol/L, respectively. (b) Relative difference plot for the data in (a) with 95% limits of agreement calculated using the Gaussian distribution; the t-distribution would result in slightly wider 95% limits of agreement. The slope, r, mean difference, SD, SEM and median difference for the difference plot data are 0.070, 0.113, 0.051, 3.1456, 0.5037 and 0.000 mmol/L, respectively. (c) Log difference plot for the data in (a) with 95% limits of agreement calculated using the Gaussian distribution. The slope, r, mean difference, SD, SEM and median difference for the difference plot data are 0.007, 0.146, 0.000 log (mmol/L), 0.0137 log (mmol/L), 0.0022 log (mmol/L) and 0.000 log (mmol/L), respectively. Examination of (b) and (c) shows that all the points are within ± 2SDs of the mean – closer examination shows that 55% of the points fall within ± 1SD of the mean, and thus the data are not perfectly Gaussian in nature, producing an SD value that is too large for the data. Such data can occur due to a pathological process when a small group of specimens have the same mean but a larger SD than that of the majority of specimens*

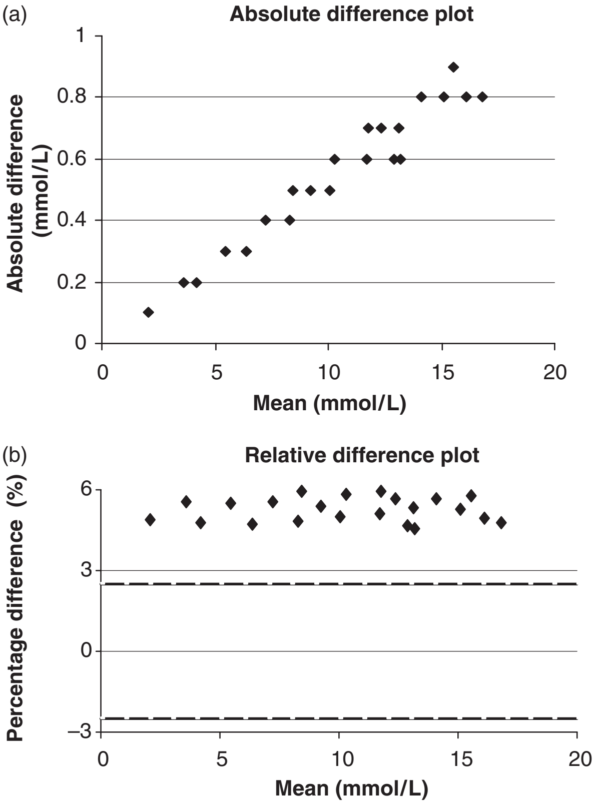

Figure 3 *(a) Absolute difference plot with a slope in the data that differs significantly from zero. The slope, r, mean difference, SD and median difference for the difference plot data are 0.052, 0.974, 0.5456, 0.2262 and 0.60 mmol/L, respectively. Horizontal 95% limits of agreement are not valid across the whole ranges of values, whereas funnel-shaped limits of agreement would be. Note that the median difference and the mean difference are in relative agreement despite the fact that horizontal limits of agreement are not valid. (b) Relative difference plot for the data from (a) with total error criteria of +2.5% and −2.5% plotted. The slope, r, mean difference, SD, SEM and median difference for the difference plot data are 0.004, 0.039, 5.258, 0.453, 0.097 and 5.31%, respectively. All of the data points fall outside the total error criteria and accordingly, the difference between the two methods is significant*

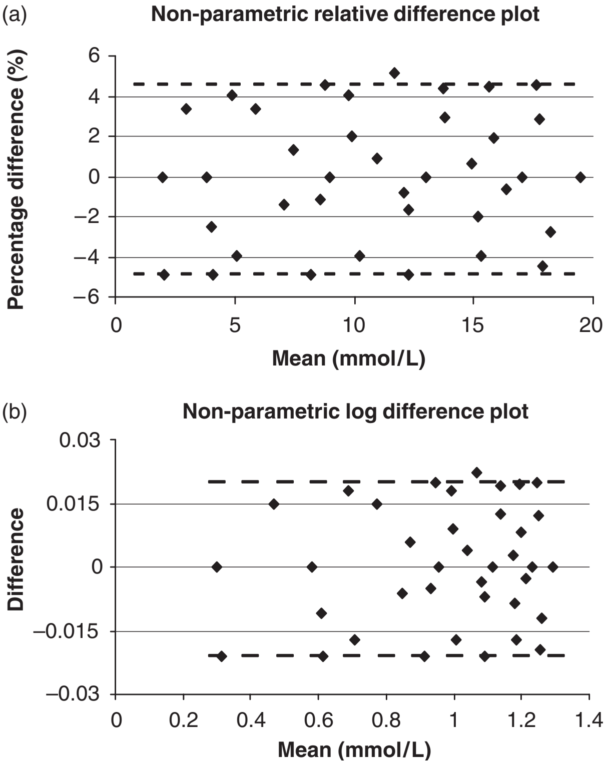

Figure 4 *(a) Relative difference plot for the data in Figure 2a with non-parametrically derived 95% limits of agreement, using MS Excel®[22] (median, 2.5th percentile and 97.5th percentile being 0.0000, −4.878 and 4.575%, respectively). (b) Logarithmic difference plot for the data in Figure 2a with non-parametrically derived 95% limits of agreement, using MS Excel®[21] (median, 2.5th percentile and 97.5th percentile being 0.0000 log (mmol/L), −0.0212 log (mmol/L) and 0.0199 log (mmol/L), respectively). Compared with Figures 2(b) and (c) respectively, the non-parametrically derived 95% limits of agreement are visually more appropriate and thus are more representative of the data; furthermore, the relative difference plot is easier to interpret*

methods are distribution-free estimators and are based on the order statistics from the sample. Therefore, the median can be employed instead of the mean, and the 2.5th and 97.5th percentiles can be employed instead of the ±1.96SD limits if and only if they can be reliably derived and are valid across the whole range of values. Such conditions may be difficult to meet for some data-sets (for example, Figures 2a and 3a) but may be easier for other data-sets (for example, Figure 2b and c) the non-parametrically derived 95% limits of agreement are demonstrated in Figure 4a and b. It must be noted that there is no generally accepted formula to estimate percentiles apart from the median, and as a result the statistical software package SAS® has a choice of five different versions to calculate percentiles non-parametrically.[19] Non-parametric statistics are equally valid but are inefficient compared with parametric statistics and as a result require more data points than parametric statistics: a minimum of 39 data points for a 95% limit of agreement and a minimum of 120 data points to derive the 90% confidence intervals for such limits of agreement. However, valid non-parametric data are more valid than invalid parametric data (compare Figure 2b with Figure 3a, and Figure 2c with Figure 3b). The most important issue is not how the data are derived (parametric or non-parametric methods) but that the derived limits of agreement are valid across the whole range of values investigated.

Various methods exist for the determination of quality specifications for analytical methods, including those based on specific clinical decision-making, biological variation, professional recommendations and regulatory authorities such as Clinical Laboratory Improvement Amendments (CLIA).[20] The clinical perspective has the highest hierarchy and is independent of statistics. If a method comparison study is to be objective, such criteria should be set before analysis. Irrespective of how the total error is determined, it can be directly employed in difference plots (Figure 3b). The 95% limits of agreement (with their respective confidence intervals) as determined above should be within such total error limits for the differences between the methods to be acceptable. Such total error limits may be displayed on difference plots.

Individual difference plots have their own advantages and disadvantages. The absolute difference plot is ideal when there is minimal proportional error between the methods but there is a constant error, while the relative difference plot is ideal when there is a proportional error between the two methods but a minimal constant error. The relative difference plot has been derived from the absolute difference plot and accordingly similar rules apply – the data must exhibit a Gaussian distribution before the mean and SD can be calculated. Once this is the case, similar procedures to those employed with the absolute difference plot can be applied. Another reason for using the relative difference plot is because the total error criteria are often expressed in relative terms. If the relationship between the two methods exhibits a significant degree of proportional and constant error between the methods, then it may be best to transform the data to a Gaussian distribution before attempting to use parametric statistics. When the data is transformed to such a distribution, it is worth noting that although parametric statistical methods can be employed, most people do not understand what the data means because it has been transformed. However, the resulting mean and confidence intervals can be transformed back to aid data interpretation. Another influence on the choice of plot is the concentration range of the *x*-axis – the absolute plot tends to be favoured for a small data range, the percentage plot for a medium range and the logarithmic plot for a large range. Why else should you employ difference plots in quantitative method comparison studies? Compared with the scatter (*x* on *y*) plot, the correct difference plot has better resolution because of freely scalable *y*-axes, such that the data at critical clinical cut-offs can be examined in more detail. Similarly, the influence of subgroups can be easier to see by the use of subgroup-specific symbols for the data points. Another benefit of the better resolution is the increased detection of outliers. The detection of outliers is important for several reasons. Firstly, they allow us to determine more detailed differences between methods, for example possible interferences due to sample handling, lipaemia, drugs, disease and so forth. Secondly, if inappropriately included in the method comparison study they may affect the data interpretation.

If no significant systematic bias exists between the two methods but the difference exceeds clinical requirements, the precision of both methods at several different levels should be evaluated as the agreement between the methods will be reduced if one or both of the methods have poor precision.[21] Differences between different patient populations may also be contributing to the large limit of agreement and, accordingly, effects due to different subgroups should be investigated as described previously.

One final issue is the determination of the value for the *x*-axis in difference plots. This is dependent on the hierarchy of the methods examined. If one method is a reference method or gold standard, by definition no other method can have a higher hierarchy, and thus the values that this method produces are employed on the *x*-axis. This would be the case where a laboratory is comparing its cholesterol method to the Centres for Disease Control and Prevention (CDC) reference laboratory using the Liebermann–Burchard method. However, this rarely occurs in most method comparison studies. If the laboratory is comparing its routine cholesterol method to another routine cholesterol method, then these two methods have the same hierarchy. In such a situation, the correct answer is not known and thus the mean of the two values determined for each specimen is employed on the *x*-axis. Such a practice prevents regression towards the mean. Another scenario could be the evaluation of a point-of-care HbA_{1c} analyser in a clinic; in such scenarios values are often confirmed by the routine laboratory method if there is a perceived discrepancy between the result and the clinical impression and thus the routine laboratory method may be deemed to have a higher hierarchy.

Difference plots are very useful and powerful tools but, as with all such tools, they need to be used appropriately. They are very flexible and can be used to evaluate total error criteria: absolute total errors can best be examined using absolute difference plots and relative total errors by using relative difference plots. Should difference plots be employed appropriately, they will greatly enhance method comparison studies and in particular will help put the results into a clinical context, which is the main goal of such studies.

References

1 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986; **i**: 307–10

2 Altman DG, Bland JM. Measurement in medicine: the analysis of comparative studies. *Statistician* 1983; **32**: 307–17

3 *Clinical Chemistry Information for Authors*. http://www.aacc.org/ccj/infoauth.stm (last accessed 4 August 2006)

4 *Annals of Clinical Biochemistry Statistical Guidelines*. http://www.rsmpress.co.uk/acbstats.pdf (last accessed 4 August 2006)

5 Altman DG, Bland, JM. Commentary on quantifying agreement between two methods of measurement. *Clin Chem* 2002; **48**: 801–2

6 Dewitte K, Fierens C, Stöckl D, Thienpont LM. Application of the Bland–Altman plot for interpretation of method-comparison studies: a critical investigation of its practice. *Clin Chem* 2003; **48**: 799–801

7 Hawkins D. Outliers. In: Johnson N, Kotz S, eds. *Encyclopedia of Statistical Sciences*. New York: John Wiley & Sons, 1985, pp. 539–43

8 Dixon W. Processing data for outliers. *Biometrics* 1953; **9**: 74–89

9 Barnett V, Lewis T. *Outliers in Statistical Data*. New York: John Wiley & Sons, 1984

10 Twomey PJ. Plasma glucose measurement with the Yellow Springs Glucose 2300 STAT and the Olympus AU640. *J Clin Pathol* 2004; **57**: 752–4

11 Pollock MA, Jefferson SG, Kane JW, Lomax K, MacKinnon G, Winnard CB. Method comparison – a different approach. *Ann Clin Biochem* 1992; **29**: 556–60

12 Hamilton LC. *Regressions with Graphics: A Second Course in Applied Statistics*. Monterey, CA: Brooks/Cole, 1992, ISBN 0534159001

13 Newton RR, Rudestam KE. *Your Statistical Consultant: Answers to Your Data Analysis Questions*. Thousand Oaks, CA: Sage, 1999, ISBN 0803958234

14 Osborne JW. Notes on the use of data transformations. *Practical Assessment, Research, and Evaluation* 2002; **8**. http://pareonline.net/getvn.asp?v=8&n=6 (last accessed 25 August 2005)

15 Parkhurst DF. Arithmetic versus geometric means for environmental concentration data. *Environ Sci Technol* 1998; **32**: 92A–8A

16 Petersen PH, Stöckl D, Blaabjerg O, *et al.* Graphical interpretation of analytical data from comparison of a filed method with a reference method by use of difference plots. *Clin Chem* 1997; **43**: 2039–46

17 Stöckl D, Cabaleiro DR, Van Uytfanghe K, Thienpont LM. Interpreting method comparison studies by use of the Bland–Altman plot: reflecting the importance of sample size by incorporating confidence limits and predefined error limits in the graphic. *Clin Chem* 2004; **40**: 2216–8

18 Linnet K. Necessary sample size for method comparison studies based n regression analysis. *Clin Chem* 1999; **45**: 882–94

19 SAS Version 8, SAS Institute GmbH, P.O. Box 105340 Neuenheimer Landstr. 28–30, D-69043, Heidelberg, Germany

20 Kenny D, Fraser CG, Hyltoft Petersen P, Kallner A. Consensus agreement. *Scan J Clin Lab Invest* 1999; **59**: 585

21 Twomey PJ, Don-Wauchope AC, McCullough D. Statistical interpretation of quantitative method comparison studies using serum osmolality as an example. *Chem Pathol* 2004; **1**: 3–7. www.clinchem.org.uk (last accessed 7 August 2005)

22 Microsoft Ireland, Atrium Building Block B, Carmenhall Road, Sandyford Industrial Estate, Dublin 18, Ireland

Accepted for publication 26 October 2005