Understanding Bland Altman analysis

Davide Giavarina

Clinical Chemistry and Hematology Laboratory, San Bortolo Hospital, Vicenza, Italy

Corresponding author: davide.giavarina@ulssvicenza.it

Abstract

In a contemporary clinical laboratory it is very common to have to assess the agreement between two quantitative methods of measurement. The correct statistical approach to assess this degree of agreement is not obvious. Correlation and regression studies are frequently proposed. However, correlation studies the relationship between one variable and another, not the differences, and it is not recommended as a method for assessing the comparability between methods.

In 1983 Altman and Bland (B&A) proposed an alternative analysis, based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement.

The B&A plot analysis is a simple way to evaluate a bias between the mean differences, and to estimate an agreement interval, within which 95% of the differences of the second method, compared to the first one, fall. Data can be analyzed both as unit differences plot and as percentage differences plot.

The B&A plot method only defines the intervals of agreements, it does not say whether those limits are acceptable or not. Acceptable limits must be defined a priori, based on clinical necessity, biological considerations or other goals.

The aim of this article is to provide guidance on the use and interpretation of Bland Altman analysis in method comparison studies.

Key words: Bland-Altman; agreement analysis; laboratory research; method comparison; correlation of data

Received: February 23, 2015 *Accepted: April 30, 2015*

Introduction

Medical laboratories often need to assess the agreement between two measurement methods. Every time we have to change one method for another one, or evaluate a new or alternative method, or quite simply we have an alignment problem between two instruments, we need some tools to measure and appraise the differences as well as the cause of these differences.

Validation of a clinical measurement should include all of the procedures that demonstrate that a particular method used for the quantitative measurement of the variable concerned is both reliable and reproducible for the intended use.

The measurement of variables always implies some degree of error. When two methods are compared, neither provides an unequivocally correct measurement, so it could be interesting trying to assess the degree of agreement.

To assess this degree of agreement, the correct statistical approach is not obvious. Many studies give the product–moment correlation coefficient (*r*) between the results of two measurement methods as an indicator of agreement. However, correlation studies the relationship between one variable and another, not the differences, and it is not recommended as a method for assessing the comparability between methods.

http://dx.doi.org/10.11613/BM.2015.015 *Biochemia Medica 2015;25(2):141–51*

In 1983 Altman and Bland re-proposed an alternative analysis, firstly presented by Eksborg in 1981 (1), based on the quantification of the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement (2).

Correlation and linear regression

Correlation is a statistical technique that can show whether, and how strongly, pairs of variables are related. There are several different correlation techniques, including the Pearson or product-moment correlation, probably the most common one. The main result of a correlation is called the correlation coefficient (or "*r*"). It is computed as the ratio of covariance between the variables to the product of their standard deviations. The numerical value of *r* ranges from -1.0 to +1.0. This enables us to get an idea of the strength of relationship - or rather the strength of linear relationship between the variables. The closer the coefficients are to +1.0 or -1.0, the greater the strength of the linear relationship is. Usually, a linear regression study is performed together with correlation measurement. Actually, linear regression can be calculated only if the correlation exists and correlation coefficient can be interpreted only if the P value is significant. However, P is significant and regression can be calculated for most cases of method comparison. Linear regression finds the best line that predicts one variable from the other one. Linear regression quantifies goodness of fit with r^2, the coefficient of determination. Correlation describes linear relationship between two sets of data but not their agreement (3). Moreover, frequently a null hypothesis is used to verify if the two methods are not linearly related. With even a minimal trend, the probability of null hypothesis is very small and it can be safely, but sometimes erroneously, concluded that the two measurement methods are indeed related.

However, the two methods that are designed to measure the same variable should have good correlation when a set of samples are chosen in such manner that the property to be determined varies considerably. In the case of method comparison, this means that samples should cover a wide concentration range. A high correlation for any two methods designed to measure the same property could thus, in itself just be a sign that one has chosen a widespread sample.

Correlation quantifies the degree to which two variables are related. But a high correlation does not automatically imply that there is good agreement between the two methods. The correlation coefficient and regression technique are sometimes inadequate and can be misleading when assessing agreement, because they evaluate only the linear association of two sets of observations. The *r* measures the strength of a relation between two variables, not the agreement between them. Similarly, r^2, named the coefficient of determination, only tells us the proportion of variance that the two variables have in common. Finally, the test of significance may show that the two methods are related, but it is obvious that two methods designed to measure the same variable are related. Moreover, the test of significance could be misleading; the significance of the correlation depends on the values of the correlation coefficient. If the correlation coefficient is statistically significant with respect to the set limit ($P < 0.05$) only then we can interpret its value; which means that if we get for example $r = 0.22$ and $P = 0.027$ we should not conclude that there is a "significant relationship", but we can claim that there is no relationship between the variables, because, calculated coefficient of variation, which indicates the absence of correlation, is statistically significant.

The proposed Passing and Bablok regression analysis to overcome some limits of correlation analysis partially solves problems related with data distribution and with the detection of a constant or proportional difference between two methods. Compared with the other frequently proposed method, the Deming regression (4), the Passing and Bablok regression could be preferred for comparing clinical methods, because it does not assume measurement error is normally distributed, and is robust against outliers. However, it needs the residuals analysis, the distribution of difference around fitted regression line, for a complete interpretation of regression results (5). This is quite sim-

ilar, but more complicated than the analysis of differences, described below.

The analysis of differences: the Bland and Altman method

Bland and Altman introduced the Bland-Altman (B&A) plot to describe agreement between two quantitative measurements (6). They established a method to quantify agreement between two quantitative measurements by constructing limits of agreement. These statistical limits are calculated by using the mean and the standard deviation (*s*) of the differences between two measurements. To check the assumptions of normality of differences and other characteristics, they used a graphical approach.

The resulting graph is a scatter plot XY, in which the Y axis shows the difference between the two paired measurements (A-B) and the X axis represents the average of these measures ((A+B)/2). In other words, the difference of the two paired measurements is plotted against the mean of the two measurements. B&A recommended that 95% of the data points should lie within ± 2*s* of the mean difference. This is the most common way to plot the B&A method, but it is also possible to plot the differences as percentages or ratios, and one can use the first method or the second one, instead of the mean of both methods.

The following example could help in familiarizing with the B&A graph plot. Table 1 shows a hypothetical series of paired data, from which it is possible to construct the B&A plot and to evaluate the agreement. In the first column a series of hypothetical variable measurements is shown, obtained by a method, named method A. The data is sorted from smallest to largest. The second column shows the measurements obtained for the same specimens but with a second, different method, B. Therefore, each line shows paired data. Figure 1 indicates the regression line between the two methods; correlation coefficient between the two methods is r = 0.996 (95% confidence interval, CI = 0.991-0.998, P < 0.001), and the regression equation is y = 7.08 (-0.30 to 19.84) + 1.06 (1.02 to 1.09) x; that could be evaluated as a very good agreement.

Table 1. Hypothetical data of an agreement between two methods (Method A and B).

	Method A (units)	Method B (units)	Mean (A+B)/2 (units)	(A – B) (units)	(A – B)/ Mean (%)
	1.0	8.0	4.5	-7.0	-155.6%
	5.0	16.0	10.5	-11.0	-104.8%
	10.0	30.0	20.0	-20.0	-100.0%
	20.0	24.0	22.0	-4.0	-18.2%
	50.0	39.0	44.5	11.0	24.7%
	40.0	54.0	47.0	-14.0	-29.8%
	50.0	40.0	45.0	10.0	22.2%
	60.0	68.0	64.0	-8.0	-12.5%
	70.0	72.0	71.0	-2.0	-2.8%
	80.0	62.0	71.0	18.0	25.4%
	90.0	122.0	106.0	-32.0	-30.2%
	100.0	80.0	90.0	20.0	22.2%
	150.0	181.0	165.5	-31.0	-18.7%
	200.0	259.0	229.5	-59.0	-25.7%
	250.0	275.0	262.5	-25.0	-9.5%
	300.0	380.0	340.0	-80.0	-23.5%
	350.0	320.0	335.0	30.0	9.0%
	400.0	434.0	417.0	-34.0	-8.2%
	450.0	479.0	464.5	-29.0	-6.2%
	500.0	587.0	543.5	-87.0	-16.0%
	550.0	626.0	588.0	-76.0	-12.9%
	600.0	648.0	624.0	-48.0	-7.7%
	650.0	738.0	694.0	-88.0	-12.7%
	700.0	766.0	733.0	-66.0	-9.0%
	750.0	793.0	771.5	-43.0	-5.6%
	800.0	851.0	825.5	-51.0	-6.2%
	850.0	871.0	860.5	-21.0	-2.4%
	900.0	957.0	928.5	-57.0	-6.1%
	950.0	1001.0	975.5	-51.0	-5.2%
	1000.0	960.0	980.0	40.0	4.1%
mean ($\bar{d}$)				-27.17	-17.40%
standard deviation (s)				34.81	-12.64%

Mean differences ($\bar{d}$) and standard deviation (s) are shown.

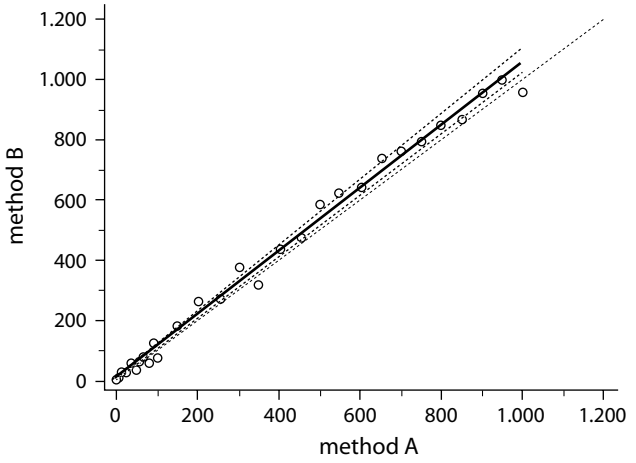

Figure 1. The regression line between hypothetical measurements done by method A and method B.
Regression equation is expressed as: y = a (95% CI) + b (95% CI) x (Passing & Bablok regression) (21). Regression line has a slope of 1.06 (1.02 to 1.09) and an intercept of 7.08 (-0.30 to 19.84). Correlation coefficient between the two methods is r = 0.996 95% confidence interval, CI = 0.991-0.998, P < 0.001.

If the aim is to evaluate the agreement between the two measurements, it could be interesting to statistically study the behaviors of the differences between one measurement and the other. Column 4 shows these differences. An ideal model would claim that the measurements obtained by one method or another gave exactly the same results. So, all the differences would be equal to zero. But any measurement of variables always implies some degree of error. Even the mere analytical imprecision for method A and method B generates a variability of the differences. However, if the variability of the differences were only linked to analytical imprecision of each of the two methods, the average of these differences should be zero. This is the first point required to evaluate the agreement between the two methods: look at the average of the differences between the paired data.

From our example, the average of the differences is -27.17 units (bottom line of table 1). This mean difference ($\bar{d}$) is not zero, and this means that on average the second method (B) measures 27.17 units more than the first one. This bias could be a constant or an average result arising from problems for specific concentrations or values. It is important to evaluate the differences at different magnitudes of the measured variable. If neither of the two methods is a "reference", the differences could be compared with the mean of the two paired values. The average can be seen in column 3. The B&A graph plot simply represents every difference between two paired methods against the average of the measurement, as shown in Figure 2. The differences between method A and method B are plotted against the mean of the two measurements. Plotting difference against mean also allows us to investigate any possible relationship between measurement error and the true value. But since we do not know the true value, the mean of the two measurements is the best estimate we have (7). If the first method is a standard or reference method, we can use these values instead of the mean of the two measurements (8), although this is controversial, because a plot of the difference against a "standard measurement" will always appear to show a relation between difference and magnitude when there is none (9).

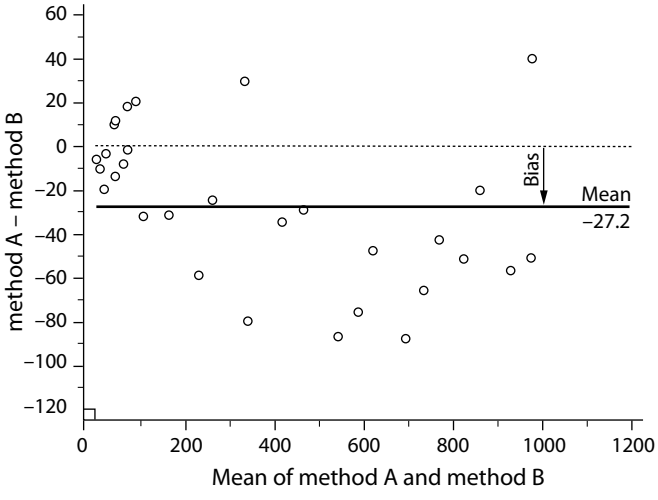

Figure 2. Plot of differences between method A and method B *vs.* the mean of the two measurements (data from table 1). The bias of -27.2 units is represented by the gap between the X axis, corresponding to a zero differences, and the parallel line to the X axis at -27.2 units.

The bias of -27.2 units is represented by the gap between the X axis, corresponding to zero differences, and the parallel line to the X axis at -27.2 units. This negative bias seems to be due to measurements over 200 units, while for lower concentrations data are closer to each other. A negative trend seems to be evident along the graph, as better shown in Figure 3. Drawing a regression line of the differences could help in detecting a proportional difference (10-12). The visual examination of the plot allows us to evaluate the global agreement between the two measurements. In our example, we can summarize the lack of agreement by calculating the bias, estimated by the mean difference ($\overline{d}$) and the standard deviation of the differences (*s*). We would expect most of the differences to lie between $\overline{d}$ -2*s* and $\overline{d}$ +2*s*, or more precisely, 95% of differences will be between $\overline{d}$-1.96*s* and $\overline{d}$ +1.96*s*, if the differences are normally distributed (Gaussian). Normal distribution of the differences must always be verified, for example by drawing a histogram. If this is skewed or has very long tails the assumption of normality may not be valid. From the example of table 1, the measurements of the two methods are not distributed normally, but on the other hand the differences

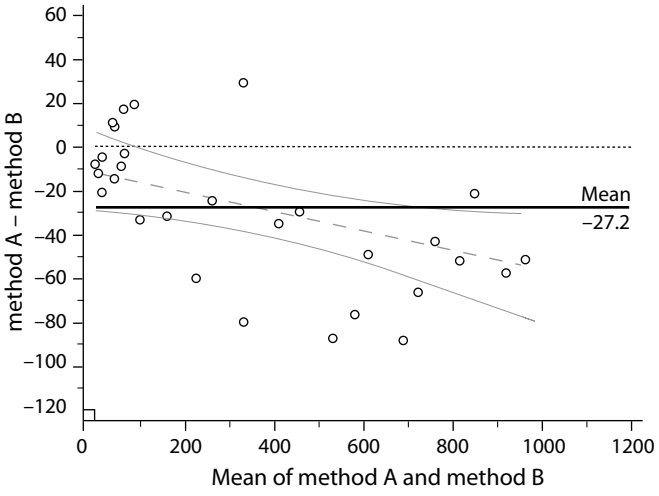

Figure 3. The same plot as Figure 1 including regression line and confidence interval limits.
Dotted line represents the regression line (y = -0.05 (-0.08 to -0.01)x – 10.15 (-28.07 to 7.77) confidence interval limits are presented as continuous line.

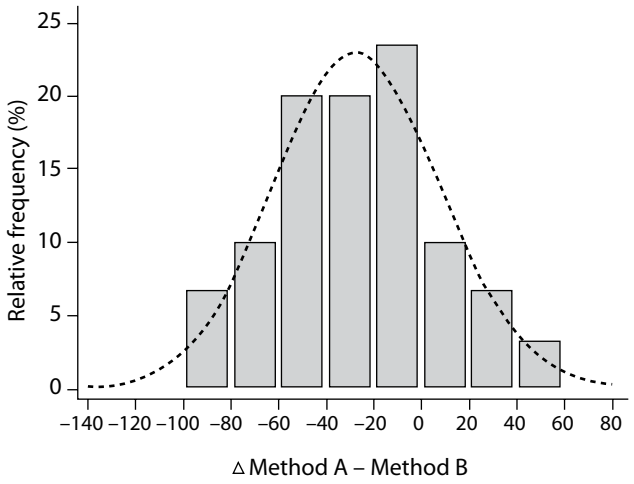

Figure 4. Distribution plot of differences between measurement by methods A and B.
The dotted line represents Normal distribution. Shapiro-Wilk test for normal distribution accepted normality (P = 0.814).

do seem to be (Figure 4). Statistical tests should always be used to determine if the distribution is normal, since in some cases normality cannot be determined simply by observing the histogram plot. If any statistical software is available, a test for normal distribution (such as Shapiro-Wilk test (13), D'Agostino-Pearson test (14), Kolmogorov-Smirnov test (15)) can be done, for the hypothesis that the distribution of the observations in the sample is normal (if P < 0.05 then reject normality). If differences are not normally distributed, a logarithmic transformation of original data can be tried.

After ensuring that our differences are normally distributed, we can use the *s* to define the limits of agreement. From data of table 1, *s* = 34.8, so 95% of differences will be

$\overline{d}$-1.96*s* = -27.2 – (1.96 x 34.8) = -95.4

$\overline{d}$ +1.96*s* = -27.2 + (1.96 x 34.8) = 41.1

So, results measured by method A may be 95 units below or 41 above method B (Figure 5).

Bias and agreement limits

The B&A plot system does not say if the agreement is sufficient or suitable to use a method or

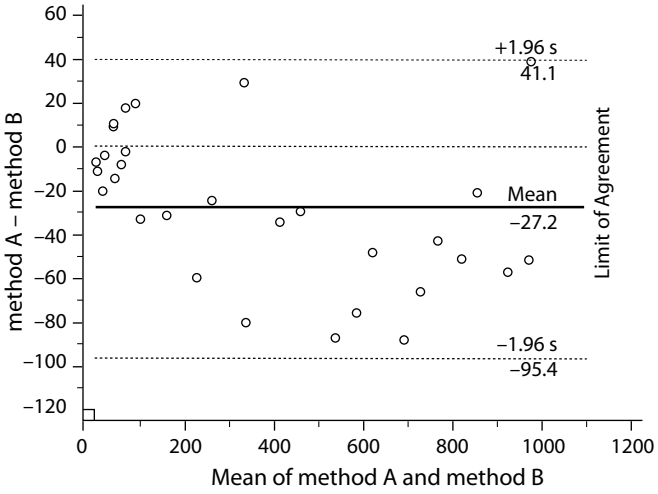

Figure 5. Bland and Altman plot for data from the table 1, with the representation of the limits of agreement (doted line), from -1.96*s* to +1.96*s*.

the other indifferently. It simply quantifies the bias and a range of agreement, within which 95% of the differences between one measurement and the other are included. It is possible to say that the bias is significant, because the line of equality is not within the confidence interval of the mean difference (Figure 6, see over), but only analytical, biological or clinical goals could define whether the agreement interval is too wide or sufficiently narrow for our purpose. The best way to use the B&A plot system would be to define *a priori* the limits of maximum acceptable differences (limits of agreement expected), based on biologically and analytically relevant criteria, and then to obtain the statistics to see if these limits are exceeded, or not.

Precision of estimated limits of agreement

As with any statistical evaluation, we only estimate a value which applies to whole population. Our estimating precision depends on the amount of observed data, i.e. on the sample size. It would be opportune to calculate the confidence interval (CI) in order to see how precise our estimates are. In particular, the 95% CI of the mean difference illustrates the magnitude of the systematic difference. If the line of equality is not in the interval, there is a significant systematic difference, i.e. the second method constantly under- or over- estimates compared to the first one.

The 95% CI of agreement limits allows for the estimate of the size of the possible sampling error. It can be measured by using standard error provided the differences follow a distribution which is approximately normal (16). Standard error of $\bar{d}$ is $\sqrt{s^2/n}$ and standard error of $\bar{d}$-2*s* and $\bar{d}$ +2*s* is about $\sqrt{3s^2/n}$. 95% CI corresponds to the observed value minus *t* standard errors to the observed value plus *t* standard errors, where *t* is the value of *t* distribution (17) with *n*-1 degrees of freedom. Table 2 shows all the B&A plot statistics, including CIs. But usually simple statistic programs can perform all these calculations and what matters is to understand the significance

Table 2. Bland and Altman plot statistics from data of table 1, including the elements to calculate confidence intervals.

Parameter	Unit	Standard error formula	Standard error (*se*)	*t* value for 29 degrees of freedom	Confidence (*se* * *t*)	Confidence intervals		
						from	–	to
number (*n*)	30							
degrees of freedom (*n*-1)	29							
difference mean ($\bar{d}$)	-27.17	$\sqrt{s^2/n}$	6.35	2.05	13.00	-40.16		-14.17
standard deviation (*s*)	34.81							
$\bar{d}$ –1.96*s*	-95.39	$\sqrt{3s^2/n}$	11.01	2.05	22.51	-117.90		-72.88
$\bar{d}$ +1.96*s*	41.05	$\sqrt{3s^2/n}$	11.01	2.05	22.51	18.54		63.56

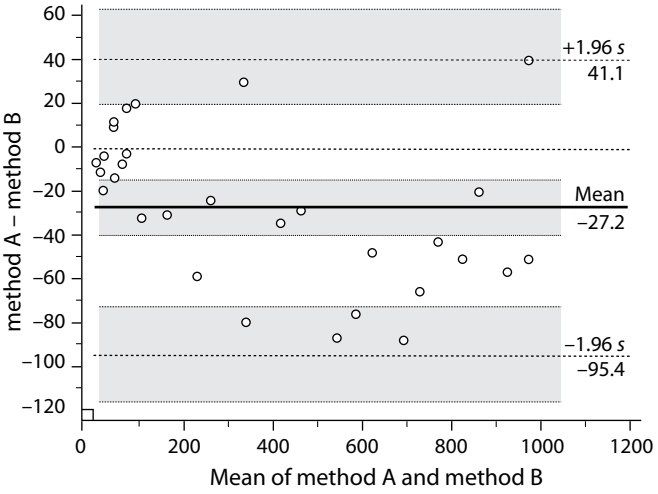

Figure 6. Same plot as Figure 2, with the representation of confidence interval limits for mean and agreement limits (shaded areas, data from table 2).

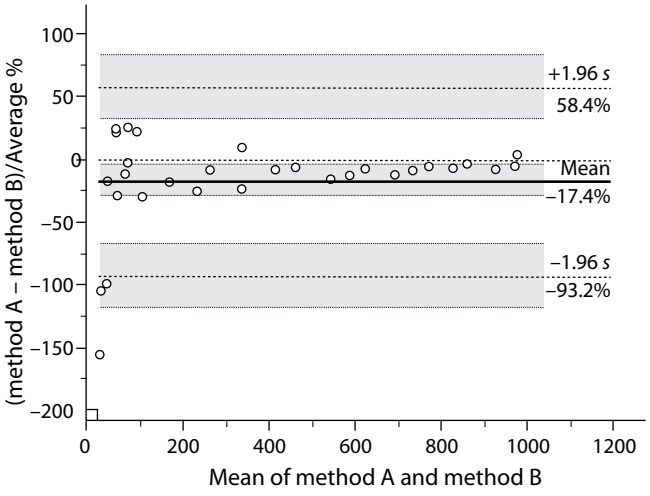

Figure 7. Plot of differences between method A and method B, expressed as percentages of the values on the axis [(method A – Method B)/mean%)], *vs.* the mean of the two measurements (data from table 1). Shaded areas present confidence interval limits for mean and agreement limits.

of the areas of confidence around the mean difference and the agreement limits, as shown in Figure 6. In summary, the CIs of mean difference and of the agreement limits simply describe a possible error in the estimate, due to a sampling error. The greater the number of samples used for the evaluation of the difference between the methods, the narrower will be the CIs, both for the mean difference and for the agreement limits.

Bland and Altman method: plot difference as percentage

In a B&A plot system the differences can be also expressed as percentages of the values on the axis (i.e. proportionally to the magnitude of measurements [(method A – Method B)/mean %)]. This option is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. Figure 7 represents the same data as Figure 6, plotted as percentage of differences. The bias (mean difference) is -17.4%, almost constant for all the measured concentrations, with the exception of very low values. As for the plot of unit values, this bias is significant, since the line of equality is not in the CI. The agreement limits are from -93.2% to 58.4%.

Common instances in laboratory diagnostics

Proposed in 1983 (2), the B&A plot method is now widespread. Their paper in the Lancet "*Statistical methods for assessing agreement between two methods of clinical measurement*" (17) has been cited more than 30,000 times by a large number of peer reviewed scientific papers (18). Many examples are available in scientific literature, usually as supplements to regression analysis and the scatter plot (19), a practice that is also recommended by the Clinical and Laboratory Standards Institute (CLSI) (20).

In Figure 8 some common models, which could represent general behaviors of agreement analysis are reported. Five cases are proposed, one for each line, each one analyzed by regression analysis and B&A plot, in unit (second column) and percentage values (third column) *versus* the mean of the two methods.

In the first example, case A, two highly correlated measurements are compared. Notwithstanding a determination coefficient of 0.9992, differences between the two measurements can be seen better in the B&A plot, that defines a bias of -7.1 units and an agreement range from -60.5 and 46.4 units. A difference plot allows us to evaluate a moderate

negative trend of differences, proportional to the magnitude of the measurement. The bias seems to change with concentration, becoming lower when the concentration is higher. Moreover, the differences seem to be constant, with a slight enlargement of the agreement limits, correlating with the concentration levels (absolute values, A2). However, a difference of plus 46 or less 60 units would be important for a measurement of 100 or 200 or 300 units, while they would not be significant for 1000 or 2000 or 3000 unit measurements. This information is better represented when the differences are plotted as percentage of the concentration (A3). The bias is -0.5% and the 2*s* agreement range is ± 11% (from -11.5% to 10.5%), principally caused by the lower measurements; above 500 units, the 2*s* agreement range seems to be less than 5%.

A model of this behavior in the differences comparison is case B, where a constant $s = \pm 50$ units was hypothesized. If the variability of the differences between the two measurements procedures is constant, the two plots will appear as they do in case B; the spread of the differences remain consistent across the range of concentration on the reported units difference plot (B2), but it increases significantly with decreasing concentration on the percentage difference plot (B3).

In the case of proportional difference variability between measurements, i.e. constant coefficient of variation across the range of concentration, the effect on the B&A plot in reported unit difference is a widening trend of the agreement range with increasing concentrations (C2). Intuitively, in the percentage difference plot, the trends remain parallel to the x axis (C3).

For constant differences across the intervals of concentrations, the reporting unit difference provides a better representation of the difference between the two measurements, while percentage difference plot is preferable for proportional difference variability (constant coefficient of variation).

If other errors overlap these sources of variability, they add their effects to the previous one. For instance, in case D we hypothesized a constant error of plus 15 units in method B, given the same proportional variability (CV%) of 5%, as in case C. An example of a constant systematic error could be an error in the blank reagent, or a matrix effect interfering with one method but not with the other. This constant error is immediately returned as a bias of -15 units in the unit difference plot. The percentage difference plot shows how this error affected more measurements of low concentrations, while the percentage bias verges to 0% for higher ones.

The last case, E, hypothesizes a proportional constant error, overlapped with the same proportional variability (CV%) of 5%, as in case C. An example could be a calibration error in one method, or a problem in some constants in an equation when computing the final results. The effect is that the magnitude of difference (bias) changes in a linear fashion. The widening trend of data with increasing concentrations is due to the constant CV% = 5%. If a proportional constant error was overlapped with a constant variability, the variability of the differences will be consistent across the measuring interval, but the bias will show a linear slope. Case E could be a model for data from Table 2, plotted in Figures 3 and 7. Case E is the only case in which the linear regression provides clear information about a problem of agreement between the two measurements, with a significant change in the slope of the regression line. On the contrary, when the agreement analysis is conducted on a wide range of concentrations, correlation and linear regression are not particularly informative, and could also be misunderstanding. Cases A to D are quite similar if only correlation is taken into account.

Summary and highlights

If you want to evaluate whether the differences between two measurements of the same substance are significant, study the differences, not the agreement. The correlation between methods is always misleading and should not be used for assessing the method comparability. The B&A plot analysis is a simple way to evaluate a bias between the mean differences, and to estimate an agree-

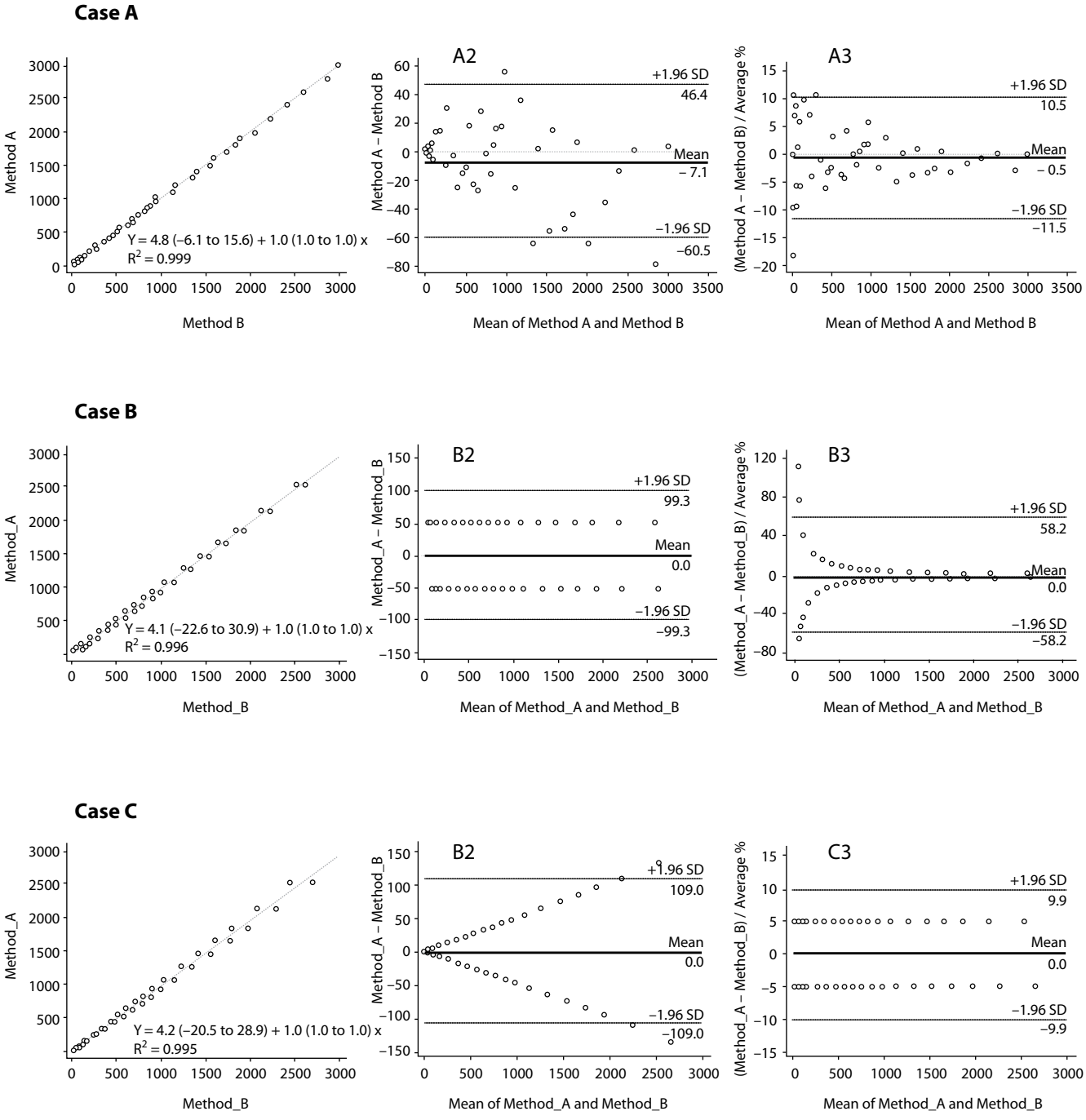

Figure 8. Method comparisons of two measurements in five different cases presented as regression analysis (column 1), Bland and Altman plot where differences are presented as units (column 2) and Bland and Altman plot where differences are presented as percentage (column 3).
Cases A, B, C, D and E represent hypothetical examples: A - random variability; B - constant variability, *s* = ± 50 units; C - constant coefficient of variation, CV% = 5%; D - constant error of plus 15 units in method B, given the same proportional variability (CV%) of 5%, as in case C; E - proportional constant error over CV% = 5%. Regression equation is expressed as: y= a (95% CI) + b (95% CI)x.
CI – confidence interval.

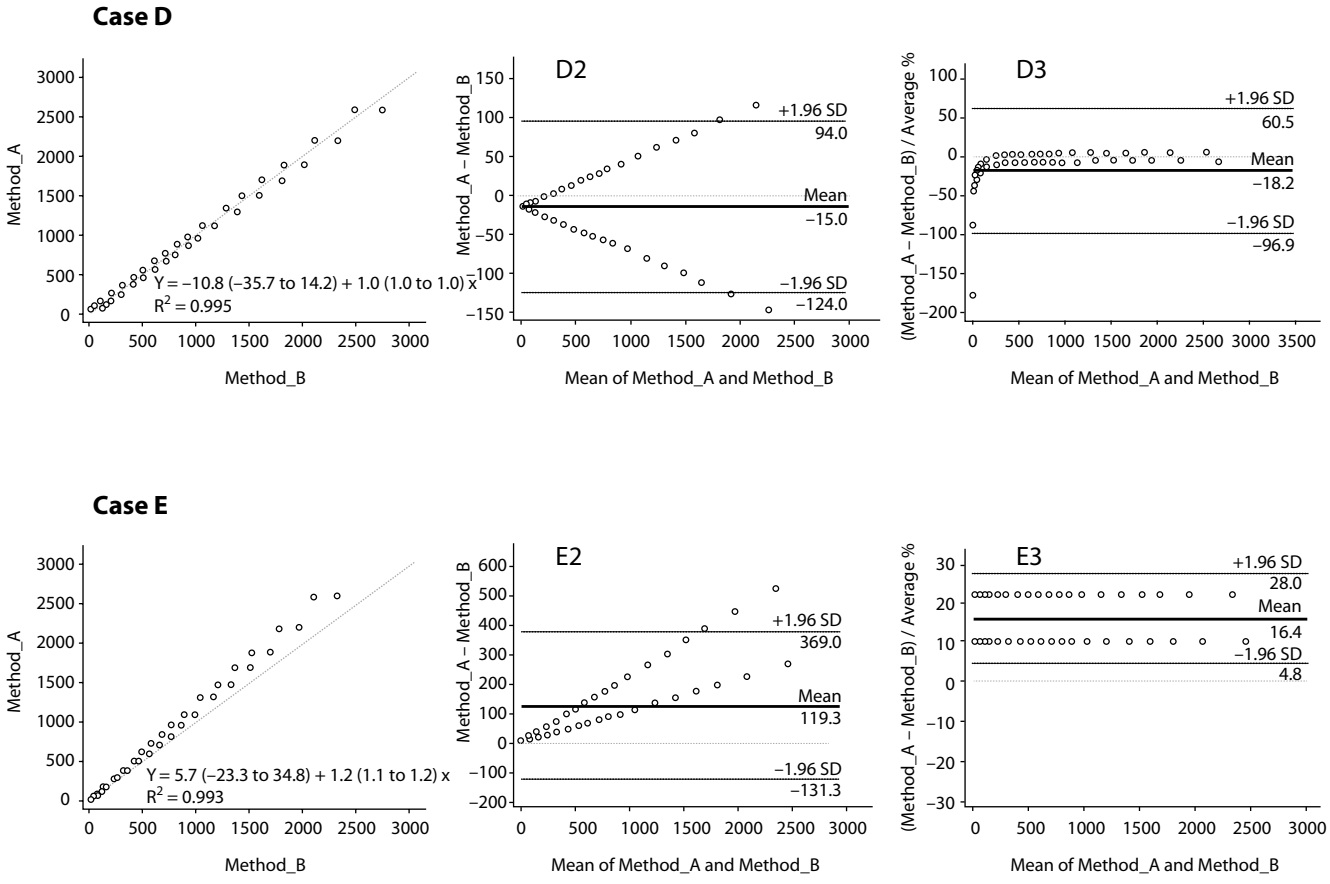

Figure 8. Method comparisons of two measurements in five different cases presented as regression analysis (column 1), Bland and Altman plot where differences are presented as units (column 2) and Bland and Altman plot where differences are presented as percentage (column 3).
Cases A, B, C, D and E represent hypothetical examples: A - random variability; B - constant variability, $s = \pm 50$ units; C - constant coefficient of variation, CV% = 5%; D - constant error of plus 15 units in method B, given the same proportional variability (CV%) of 5%, as in case C; E - proportional constant error over CV% = 5%. Regression equation is expressed as: y= a (95% CI) + b (95% CI)x.
CI – confidence interval.

ment interval, within which 95% of the differences of the second method, compared to the first one fall. Data can be logarithmically transformed, if differences seem not to be normally distributed. For bias and agreement limits, appropriate CIs can be computed, in order to consider the sampling error in relation to the dimension of the sample. Data can be analyzed as unit differences plot or as percentage differences plot. Both the plots may be considered, to allow the better evaluation. The B&A plot method only defines the intervals of agreements, it does not say whether those limits are acceptable or not. Acceptable limits must be defined *a priori*, based on clinical necessity, biological considerations or other goals.

Potential conflict of interest

None declared.

References

1. Eksborg S. Evaluation of method-comparison data. Clin Chem 1981;27:1311-2.
2. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. Statistician 1983;32:307–17. http://dx.doi.org/10.2307/2987937.
3. Udovičić M, Baždarić K, Bilić-Zulle L, Petrovečki M. What we need to know when calculating the coefficient of correlation? Biochem Med (Zagreb) 2007;17:10-5. http://dx.doi.org/10.11613/BM.2007.002.
4. Martin RF. General Deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clin Chem 2000;46:100-4.
5. Bilić-Zulle L. Comparison of methods: Passing and Bablok regression. Biochem Med (Zagreb) 2011;21:49-52. http://dx.doi.org/10.11613/BM.2011.010.
6. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-60. http://dx.doi.org/10.1191/096228099673819272.
7. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Int J Nurs Stud 2010;47:931-6. http://dx.doi.org/10.1016/j.ijnurstu.2009.10.001.
8. Krouwer JS. Why Bland-Altman plots should use X, not (Y+X)/2 when X is a reference method. Stat Med 2008; 27:778-80. http://dx.doi.org/10.1002/sim.3086.
9. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995;346:1085-87. http://dx.doi.org/10.1016/S0140-6736(95)91748-9.
10. Armitage P, Berry G, Matthews JNS eds. Statistical methods in medical research. 4th ed. Maiden, MA: Blackwell Science, 2002. http://dx.doi.org/10.1002/9780470773666.
11. Bland M. An introduction to medical statistics. 3rd ed. Oxford University Press, Oxford, 2000.
12. Eisenhauer JG. Regression through the origin. Teach Stat 2003;25:76-80. http://dx.doi.org/10.1111/1467-9639.00136.
13. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965;52:3-4. http://dx.doi.org/10.1093/biomet/52.3-4.591.
14. Sheskin DJ. Handbook of parametric and nonparametric statistical procedures. 5th ed. Chapman & Hall / CRC Press, Boca Raton, FL, 2011.
15. Neter J, Wasserman W, Whitmore GA eds. Applied statistics. 3rd ed. Allyn and Bacon, Boston, MA, 1998.
16. Bland JM, Altman DG. Statistical method for assessing agreement between two methods of clinical measurement. Lancet 1986;327:307-10. http://dx.doi.org/10.1016/S0140-6736(86)90837-8.
17. Medcalc manual. Available at: http://www.medcalc.org/manual/t-distibution.php. Accessed January 23rd, 2015.
18. Google Scholar search engine, Available at http://scholar.google.it/scholar?cites=7296362022254018043&as_sdt=2005&sciodt=0,5&hl=it. Accessed February 2nd, 2015.
19. Dewitte K, Fierens C, Stöckl D, LM Thienpont. Application of the Bland-Altman plot for interpretation of method - comparison studies: a critical investigation of its practice. Clin Chem 2002;48:799-801.
20. Clinical and Laboratory Standards Institute (CLSI): Measurement procedure comparison and bias estimation using patient samples. Approved guideline - Fifth Edition. CLSI document EP09-A3. Wayne, PA, USA, 2013.
21. 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 I. J Clin Chem Clin Biochem 1983;21:709-20. http://dx.doi.org/10.1515/cclm.1983.21.11.709.