Use of the Limits of Agreement Approach in Periodontology

Diego Garcia Bassani[a]/Letícia Algarves Miranda[bc]/Anders Gustafsson[c]

Purpose: To discuss the statistical approaches that have been traditionally used to compare measures in periodontal research, highlighting its strengths and weaknesses and, finally, to suggest the use of the limits of agreement method of Altman and Bland (1983) as an alternative method to address this question.

Materials and Methods: Using a sample dataset of clinical periodontal measures as a background, the different possible approaches for agreement assessment are discussed and statistical and clinical points are considered. Eight hundred and forty repeated measures, belonging to the training phase of a clinical study, were performed in five individuals presenting different severities of periodontal conditions. The use of correlation coefficient, comparison of means, linear regression technique, Kappa coefficient, intra-class correlation coefficient and means versus differences plot is demonstrated.

Results: Most of the methods are applied without the appropriate care, resulting in misleading interpretations. The information that arises from some of the methods used so far is poorly informative and adds little understanding to the operational characteristics of the raters or instruments. Some of the resulting information from the correlation coefficient and kappa coefficient may even be false or not applicable for the entire range of possible values.

Conclusions: The graphical approach that plots differences against means, including the 95% limits of agreement estimated by the mean difference ± 1.96 standard deviation of the differences is the most informative approach and its application should be considered for continuous clinical periodontal measures.

Key words: agreement, periodontal attachment loss, probing depth, reproducibility

Oral Health Prev Dent 2007; 2: 119-124. *Submitted for publication: 22.03.06; accepted for publication: 08.08.06.*

Methods for assessing agreement between measures in dentistry, especially in periodontology, have been applied in scientific research for a long time. Comparing measurements is a task commonly performed in periodontal research, especially in studies measuring clinical attachment level (CAL) and probing depth (PD).

[a] *Centre for Addiction and Mental Health, Canada*

[b] *(formerly) Brazilian Lutheran University and (currently) Tissue Biology Laboratory, Pontifícia Universidade Católica – RS, Brazil*

[c] *Karolinska Institutet, Huddinge, Sweden*

Reprint requests: *Diego Garcia Bassani, Centre for Addiction and Mental Health, Health Systems Research & Consulting Unit, 33 Russell Street, Tower - T308, Toronto, ON - M5S 2S1, Canada. Tel: +01 (416) 535 8501 # 6465. Fax: +01 (416) 979 4703. Email: diego_bassani@camh.net*

Proper knowledge of the level of intra- and inter-examiner measurement error is a pivotal element of periodontal epidemiological investigation (Fleiss et al, 1991). Knowing these errors is important since they can influence the estimates of disease prevalence and severity as well as the magnitude of differences between subpopulations. Therefore, it is important that errors are documented whenever possible and also taken into account when interpreting the results of the studies (Kingman and Albandar, 2002).

Data reflecting agreement rates have been presented in a range of different ways in periodontal research, and how appropriate and meaningful they are also varies (Badersten et al, 1984; Best et al, 1990; Boushka et al, 1990; Lopez et al, 2003).

The aim of the present paper is to discuss the statistical approaches that have been traditionally used to

compare measures, highlighting their benefits and weaknesses and, finally, to suggest the use of the limits of agreement method of Altman and Bland (1983) as an alternative method to address this question in periodontal research.

MATERIALS AND METHODS

The illustrative example consists of 840 repeated measures of PD and CAL performed in five female individuals (mean age 53.7 years old) with different severities of periodontal condition. Clinical measurements were collected, by the same examiner, from six sites of all teeth present with a William type probe (Neumar, Sao Paulo, Brazil). Measurements were made in mm and were rounded to the lower whole mm. The interval between the repeated measures was 7 to 10 days. These data were obtained for the training phase of a clinical study. To these data, different methods used to assess agreement in periodontal research were applied and are further discussed. These include Pearson correlation coefficient, statistical comparison of means, linear regression, Kappa coefficient, intraclass correlation coefficient (ICC) and the alternative method of limits of agreement. All calculations and analyses were performed using Stata version 7.0 (Stata corporation, TX, USA).

RESULTS AND DISCUSSION

Pearson correlation coefficient

The Pearson correlation coefficient between two measures is the first choice approach for agreement assessment discussed here. In the presented example, high values of r for CAL (0.98) and PD (0.92) would be obtained if the method was applied. However, the product-moment correlation coefficient, r, is not suited for agreement for a number of reasons. First, it depends on the variation of the true values and the measurement error (variation within raters) and reflects how close to a straight line is the relation between the two measures. Therefore, the value of the correlation coefficient is linked to the sampling strategy in a way that if the subject's measures are highly variable compared to the measurement error, the r-value will be high, and if the variation is low, the r-value will be low. As long as the values grow or reduce at the same pace for the two variables, systematic errors will allow r to be high even if there is not one similar value between the two observations. In a practical situation, if one rater is recording 2 mm in a CAL measure, and another rater, or even the same rater in a second observation, is recording 5 mm from the same site, as long as the difference is constantly close to ± 3 mm the r value will be very close or equal to one.

High correlation values, therefore, do not imply necessarily that two raters or two measures agree or reflect the same event consistently under similar situations. It only means that they vary, even when reporting incorrect measures, in a similar way, the two or more measures being associated, and nothing more.

The other issue to be considered is that Pearson correlation assumes that at least one of the variables or measures in question is normally distributed, which is not usually true for CAL and PD, known for distributions closer to a Poisson curve, specially for population-representative samples.

Finally, it is simple to rule out the Pearson coefficient for agreement assessment purposes when one understands that the coefficient is a measure of association or strength of the relationship, but not of agreement.

Comparison of means

Another simple but misleading method that is often used to assess agreement is the statistical comparison of means. This method consists of comparing the two measurements based on the means and applying a statistical test. In our example the resulting values for CAL would be 1.14 ± 0.07 mm for the first examination and 1.46 ± 0.12 mm for the second. The value for a *t* test with equal variances, comparing both means is 0.018, showing that the means differ significantly. For the PD example, the means and standard deviations would be 2.38 ± 0.05 and 2.59 ± 0.09 mm respectively, with a p value for the *t* test of comparison of the means of 0.039.

In this example, the similarity of the means, both for CAL and PD, is not evident and due to reduced standard errors, the *t* tests show a significant difference between the two measures. If one compared these values to the product-moment correlation coefficient presented above, conclusions concerning agreement may become hard to handle. In general, when agreement is reported through this method, authors say the measures are statistically similar, or the same, between the raters, based on the results of a not significantly different test. One point to consider whenever facing this type of approach is that depending on the algorithm applied to the test for difference between means, the greater the measurement error (standard error), the

greater the chance of not finding any significant difference. It does not even consider association between measures, only similarity between the means, as a function of the variances.

Linear regression

The use of linear regression to compare periodontal measurements between raters is used less often, but sometimes one may come across this procedure. In general, the linear regression line slope is compared to zero. This method results in exactly the same issues that were discussed when the product-moment method (Pearson correlation coefficient) was presented. Fig 1 represents the fitted values for the linear regression line for the CAL measures between the two raters.

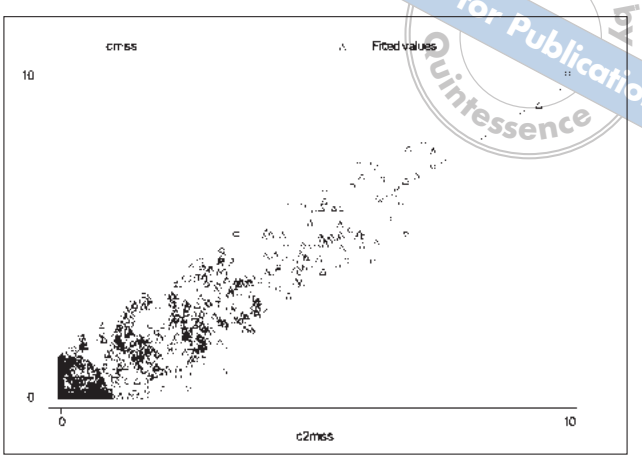

Fig 1 Linear regression line (fitted values) for the CAL measures between the two raters.

The regression coefficient of the respective model (beta) is 0.97, which is considered a high value. This coefficient means that for each millimeter the first rater measures, the second rater, or second measure of the same rater, identifies 0.97 mm, with an intercept very close to zero, which is the expected situation. The main issue when applying linear regression to assess agreement is that the statistical test, the Wald test in this case, usually compares the slope of the least squares against zero, leading to an interpretation very similar to the one obtained from the Pearson correlation, when the r-value is tested to verify the difference from zero. Since the significance of a test of this nature is a function of the variance/measurement error, the same complications mentioned above are valid. Hence, finding a statistically significant association between the measures is fairly a function of a true agreement. Additionally, observing no differences also does not demonstrate agreement since, as in the comparison of means, measurement error plays a major role in the power of the test. Using linear regression models to predict CAL or PD measurements from a first rater compared to a second rater and obtaining the standard error for the predictive values may be an option, especially in cases where one is not directly interested in the comparability, but in a calibration parameter. In that sense, the coefficient is the amount of millimeters that a rater (perhaps less experienced) 'registers' as a function of each millimeter the reference rater (more experienced) identifies. The issue of linear regression not taking into account the measurement error (precision of the estimates) nor the range of the observations, added to the fact that the model is forcing the values to fit into a straight regression line, may be a factor to be considered in the interpretation of the parameters/coefficients. In addition, samples for calibration usually are not random samples, and there is also no random sample with which to compare the results. Consequently, the range of observations influencing the model will result in underestimation of the slope. Although there are strategies to correct the slope attenuation, other issues are still to be considered. Complex strategies to include error estimates and variance corrections in the model do not seem justifiable due to its complexity when the issue is just part of the methods session of a study that is not evaluating agreement as a primary objective.

Kappa coefficient

The Kappa coefficient, or Cohen's Kappa, is often used for agreement assessment in periodontal investigations. The Kappa statistic takes into account agreement by chance, avoiding the situations exemplified above where chance is not taken into account, and this is the main argument that has been used to support its use. This coefficient has advantages over simple agreement rates.

However, the Kappa coefficient is not suited for agreement assessment of continuous measures. Thus periodontal measures, in millimeters, fall under this rule. Even though manual periodontal measures may be considered discrete, including only integer values, the behaviour of the resulting scores (millimetres) might impair the Kappa calculations because some values might be absent for one of the raters or for the whole resulting matrix of data. Also, when the sample is selected the behaviour of the resulting scores

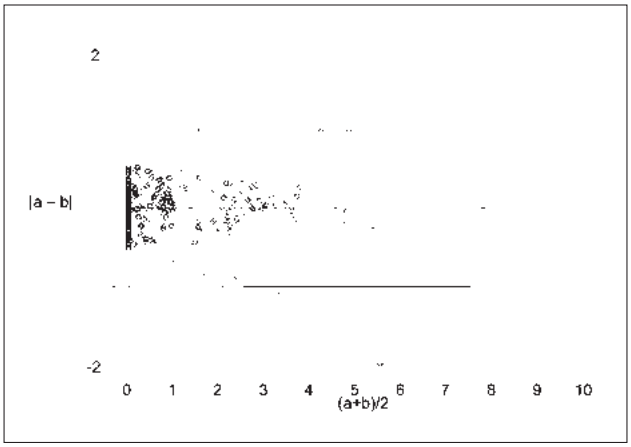

Fig 2 Standard deviation of the pairs of observations plotted against the mean of the same pairs for CAL. Note: since CAL and PD were measured manually and registered only in a categorical way (integers), a jitter approach was applied to the graphic representations throughout the paper to avoid having hundreds of superimposed observations.

Table 1 Weighted and unweighted Kappa coefficients for the comparison of the two measures of clinical attachment level (CAL) and probing depth (PD) and p value for the Ho: Kappa coefficient = 0

	Kappa coefficient			
Weight (|diff|=1) (%)	CAL	p	PD	p
0	0.81	<0.0001	0.82	<0.0001
80	0.86	<0.0001	0.95	<0.0001
100	0.88	<0.0001	0.99	<0.0001

(millimetres) night impair the Kappa calculations because some values night be absent for one of the raters or for the whole resulting matrix of data. Also, when the sample is selected from a healthy population, where the range of possible CAL and PD values is narrower, or in a sample where, by chance, some values are not observed, comparing Kappa statistics is not suitable since the agreement by chance is affected by the range of possible observations and this is different for the two samples.

It is important to note that in general, Kappa values are interpreted and compared according to Landis and Koch interpretation tables, not taking into consideration the number of possible codes for the measurement under evaluation.

The Kappa calculations for the fictional data used as an example in this report are presented in Table 1. The Kappa coefficient can be weighted, but even then, the information the measure provides is poor in a broader sense.

The explanation for the effect of weighting the error differently on the Kappa value is clear when one considers the range of possible values of PD (1–8) and CAL (0–10). Therefore, the weighting strategy must be carefully considered as a function of the range of the scale (possible values) and the future interpretations and comparability of Kappa coefficient values if the coefficient is used for agreement assessment, despite the above presented arguments. As a final appeal, the Kappa statistic does not show where the rater is deficient in the examination and does not allow corrections in the training phase of a clinical study, for example.

Intra-class correlation coefficient

The use of the ICC for agreement assessment should be carefully considered for this purpose. First of all, ICC is a general name for a series of different measures and the researcher should be aware of which formula is being used or which methods are applied by the chosen software. The following example and brief discussion uses the one-way coefficient, derived from the analysis of variance. The definition of ICC might help to understand how a researcher should handle the coefficient. The ICC is a restriction of the usual correlation concept (Pearson correlation) to the case of interchangeable measures. While for two repeated measures from the same rater at different points of time interchangeability of the measures could be discussed, for different raters, when comparing observers for example, the exclusion of this measure from the list of possibilities is clear. The values for ICC for CAL in our example were 0.98 (95% CI 0.94–1.0), and for PD 0.89 (95% CI 0.73–1.0), both similar to the r-values obtained from the Pearson coefficient. Once the ICC is a ratio of the variance between measures over the total variance, if the conditions under which the measures are obtained are exactly the same and if the researcher is not interested in comparing the evolution of a rater or the maintenance of a pattern of reproducibility over time or over the range of possible values, this measure might be useful. The careful researcher should consider, however, that the same issues previously discussed about reliability of the agreement measure through the whole

range of possible values is still relevant for the ICC case, including the negative points raised for the Pearson correlation coefficient.

Limits of agreement or difference versus mean

Factors that affect single periodontal measurements, such as time of day, position of subject, light, type of probe and the observer himself, have a direct effect on the repeatability and reproducibility of repeated measures. Once all the other effects are fixed, except for the observer error, the method described by Altman and Bland (1983) for assessment of repeatability of measures is relatively simple and presents many advantages over the previously described strategies.

When evaluating repeated measures, it is important to certify that the pattern of agreement does not vary along the range of measurements; in other words, that the agreement does not vary, or vary minimally, as the measures get higher or lower. The method consists of plotting the difference of the pairs of observations against the mean of the same pairs (in the case of two repeated measures per subject/site). The plotting of the differences against only one of the measures can be misleading, is not recommended and has been discussed previously (Bland and Altman, 1995a). When there are two replicated measures a and b, the plot will reflect the |a-b| against (a+b)/2. The resulting plot for the CAL of the example we have been using is showed in Fig 2.

It can be visualised from a plot of this kind whether there is any trend for higher or lower agreement as the measures change in magnitude. From the example above, it can be seen that for high values of CAL in mm, there is less difference between the raters a and b, while for low values, differences of around 1.5 mm are observed. In this case, calculating a product-moment coefficient (Pearson coefficient) between the mean and the difference is informative.

The value of r (-0.08) shows that both measures are poorly correlated, but the sample size and the low variance of the measures makes the test of the hypothesis that H_0: r = 0 obtain a p value of 0.0498. If the variation (determined by |a-b|) was found to be independent of the size of the measurement (CAL), then the residual standard deviation obtained from a one-way analysis of variance would represent the measure of repeatability/agreement for the rates in the overall sample.

Since the value of r was different from zero, further calculations are necessary to determine the repeatability/agreement measure. Testing the parameters

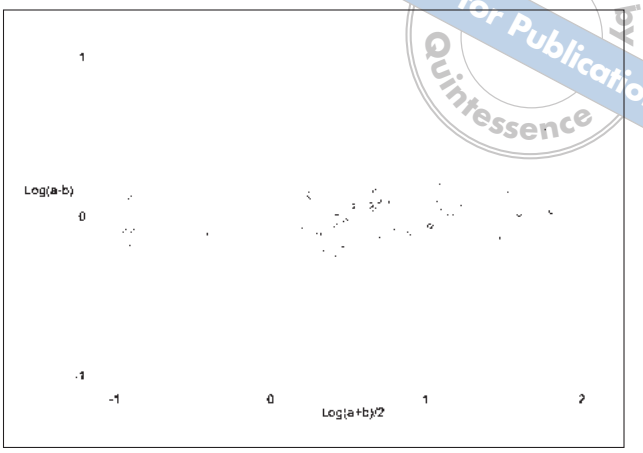

Fig 3 Log-transformed values for CAL.

(|a-b| and (a+b)/2) for normality helps to determine whether to use a simple approach (analysis of variance) or a slightly more complex procedure. In this example, using a one-way analysis of variance would lead to misleading results, due to the fact that there is no normal distribution of one of the components. A logarithmic transformation of the components being handled is usually useful to remove the association and allow the calculation of the residual standard deviation through the procedure described above (analysis of variance). Back-transformation of the results will return the agreement rate in the original unit. Fig 3 shows the log-transformed values for the previous example.

As can be seen, the relationship between the difference and the mean is more constant along the range of values. The reduced number of observations in this plot is a consequence of the exclusion of the observations where difference is null (|a-b| = 0).

After logarithmic transformation, the one-way analysis of variance can be calculated and the resulting residual standard deviation back-transformed to an agreement measure in millimetres. Back-transformation, however, is not always possible or meaningful. In such situations, sub-cohorts of data or categorisation of the scale to be analysed in separate categories may be useful. When using manual CAL or PD measures, if there is any agreement, the resulting (|a-b|) will be equal to zero for some cases. These cases are not included in the plot and will result in observation losses under-powering the test.

The visual/graphical evaluation of the plot is useful, and alternative parameters, such as the mean of the differences ($\sum$ |a-b|/n), standard deviations and 95% limits of agreement can be displayed in the graphical representation to achieve a more meaningful inter-

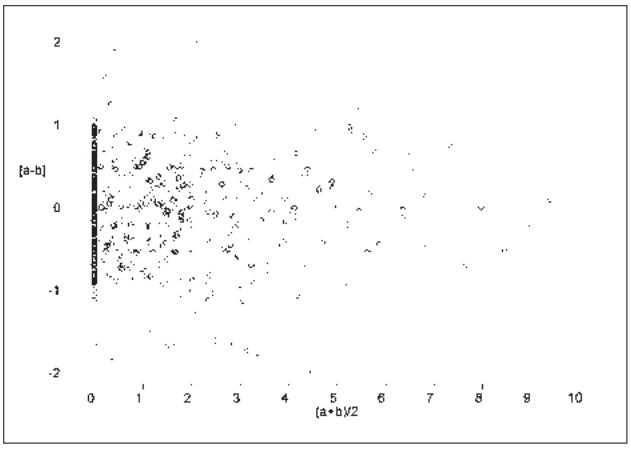

Fig 4 95% confidence interval and the mean of the difference for the CAL values.

pretation. Personal standards of acceptable agreement should be established in advance. Fig 4 represents the mean versus difference plot including the 95% confidence interval and the mean of the difference for the CAL values.

As can be seen, the issue of higher agreement for higher CAL values is clear and the matter of how far from the confidence interval the variation may fall is also well represented for the lower values of CAL in millimetres. This advantage of the plot of differences versus magnitude (mean) is very helpful in a training phase of a study and for descriptive objectives, because trends towards larger or smaller disagreement along the range of possible values are easily identifiable. The magnitude of the agreement or disagreement and the error, variation and presence of outliers along the distribution of possible values is also visualised in a simple and meaningful way. For the researcher used to visualising regression diagnostic plots of residuals versus outcome, the similarity of the interpretation will be evident.

As presented by Altman and Bland (1983), when the association between the difference and the mean is present and the transformation is successfully employed, the 95% confidence interval will be asymmetrical with non-constant error. The plot will give a reasonable insight as to which transformation should be employed (i.e. exponential, logarithmic, square root, reciprocal cube). If none of the transformations is suitable or does not normalise the distribution, regressing the difference (a-b) on the mean (a+b)/2 is a suitable alternative that should be used. Further information on this subject can be obtained from the original papers from Bland and Altman (Altman and Bland, 1983, 1986; Bland and Altman, 1986, 1999). Finally, as described by Bland and Altman (1995b), the presented method is suitable either for repeatability or agreement, being a powerful and very informative tool in the description of reliability of measures in periodontal research.

The measurement of agreement and repeatability in periodontal research is of fundamental importance. The way it has been conducted and neglected so far increases the need for consistent methodologies and qualified approaches. A simple graphical approach may be a suitable and highly informative alternative to the agreement assessment approaches used previously. The quality of the information and the misleading results and interpretations that can arise from the use of inadequate evaluation methods should be considered carefully by the scientific journals. The simplicity of the method proposed by Bland and Altman in the early 1980s should be readily incorporated to enhance the quality of the research reports in dental journals and research meetings.

REFERENCES

1. Altman D, Bland BH. Measurements in medicine: the analysis of method comparison studies. Statistician 1983;32:307-317.
2. Altman DG, Bland JM. Comparison of methods of measuring blood pressure. J Epidemiol Community Health 1986;40:274-277.
3. Badersten A, Nilveus R, Egelberg J. Reproducibility of probing attachment level measurements. J Clin Periodontol 1984; 11:475-485.
4. Best AM, Burmeister JA, Gunsolley JC, Brooks CN, Schenkein HA. Reliability of attachment loss measurements in a longitudinal clinical trial. J Clin Periodontol 1990;17:564-569.
5. Bland JM, Altman DG. Comparing methods of measurement: why plotting difference against standard method is misleading. Lancet 1995a;346:1085-1087.
6. Bland JM, Altman DG. Comparing two methods of clinical measurement: a personal history. Int J Epidemiol 1995b;24(Suppl 1):S7-S14.
7. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res 1999;8:135-160.
8. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-310.
9. Boushka WM, Marinez YN, Prihoda TJ, Dunford R, Barnwell GM. A computer program for calculating kappa: application to interexaminer agreement in periodontal research. Comput Methods Programs Biomed 1990;33:35-41.
10. Fleiss JL, Mann J, Paik M, Goultchin J, Chilton NW. A study of inter- and intra-examiner reliability of pocket depth and attachment level. J Periodontal Res 1991;26:122-128.
11. Kingman A, Albandar JM. Methodological aspects of epidemiological studies of periodontal diseases. Periodontol 2000 2002;29:11-30.
12. Lopez R, Retamales C, Contreras C, Montes JL, Marin A, Vaeth M, Baelum V. Reliability of clinical attachment level recordings: effects on prevalence, extent, and severity estimates. J Periodontol 2003;74:512-520.