European Journal of Orthodontics 22 (2000) 257–269

Statistical methods for the comparison of measurements derived from orthodontic imaging

Frank Krummenauer* and Gerhard Doll**
Departments of *Medical Statistics and Documentation and **Orthodontics, University of Mainz, Germany

SUMMARY Biometric comparison procedures for dental imaging methods with continuous outcome were reviewed, mainly concentrating on assessment and comparison of accuracy and precision according to the study design. Univariate graphical and numerical representations of corresponding deviations were summarized to derive a 'check list' of minimum information necessary to compare the measurement methods.

The methods reviewed in this investigation are illustrated by the comparison of conventional (radiographic) cephalometry versus assessment using the DigiGraph® in 50 female children. A paired *t*–test and the corresponding confidence interval approach were used to assess deviations in location of two imaging methods; the test procedures of Maloney/Rastogi, Hahn/Nelson, and Grubbs were surveyed as proposals for the comparison of precisions in paired data. The Krippendorff coefficient was used as an aggregate measure for method concordance.

Since these methods can be performed by simple modification of standard options available in most statistical software packages, this review intends to enable dental researchers to choose the correct methods, and perform adequate data analysis and representation.

Introduction

The comparison of clinical measurement methods is of broad interest for any medical discipline, which is based on continuous diagnostic methods such as orthodontics (cranial and in particular mandibular imaging), cardiology (cardiac imaging and two-dimensional cardial volumetry), radiology (signal intensity measurement in tissue subcuts), and pathology (antibody concentration in prepared tissue). Comparison problems will arise, for example, when an established, but expensive or invasive reference method is to be compared with a new and more attractive substitute measurement method, where the latter must not yield less accurate or precise measurements than the standard. Therefore, statistical comparison procedures are invoked to provide information on possible significant deviations in accuracy and/or precision, as well as on the order of agreement between the methods under consideration.

Despite the necessity of using adequate statistical procedures for data aggregation and representation, there are tendencies to oversimplify the latter by simply providing mean values and standard deviations for each of the imaging methods under consideration separately, or by using a correlation coefficient as a surrogate measure for agreement between the two measurements. Measurement comparison is usually based on paired data, that is the intra-individual information contained will be ignored by only presenting separate means and variances for each imaging method. The latter will result in crucial errors due to liberal decisions based on the suboptimal representation, that is significant results may only be significant due to the kind of data analysis, but not because of existing clinical differences between the imaging methods under consideration. This effect will be illustrated in the setting of the comparison of (paired) measurement precisions. Therefore, this review summarizes elementary procedures for the

biometric evaluation of paired data. These methods, however, are rarely available in standard software packages. Therefore standard programs such as Excel®, SPSS®, or Winstat® will be illustrated to modify available options to implement the subsequent methods. The corresponding formulae may appear tedious, but readers should concentrate on the interpretation of the results of the methods and on pitfalls in their application, rather than their derivation and mathematical representation. The latter are included merely as a basis for self-implementation and not as a short-cut tutorial on mathematical statistics.

Many applications of these procedures are devoted to the comparison of diagnostic methods, whereas the introduction and evaluation of new drugs or therapies often concentrates on mean or median effect estimates. The comparison of diagnostic procedures requires an additional comparison of scale. If a new measurement method is to be compared with a gold standard, two questions are of interest. However, only the first is frequently asked when designing method comparison studies: 'Is the new method valid?', i.e. are there significant mean or median deviations? Nevertheless, there is a second indication for substitution of an established method by a new one. For example, imaging methods are frequently used for the determination of cardiac functional parameters, which are clinically relevant indicators for decisions on further invasive diagnostic or immediate interventional therapy. Assessment of such endpoints, however, may become biased, when being performed by students of medicine or physicians with limited imaging experience. However, if the new method promises easier application, maybe due to additional features assisting assessment of the clinical endpoints of interest, there will also be a focus on the measurements' variation. If the new method turns out to be valid and easier to apply, a significant decrease in variability and thus additional gain in diagnostic quality as based on the clinical endpoints under consideration will be expected. Simultaneous application of both measurement instruments (new and established) will allow for the simultaneous intra-individual comparison of location and scale. Bearing in mind that this is mainly focused on tests for the detection of differences in paired variances, corresponding methods to establish 'diagnostic equivalence' of measurement methods will not be considered in the following. This will only consider univariate measurements; extensions to multivariate analysis are straightforward with the methods described below by applying multiple test procedures on the several clinical parameters of interest (see, e.g. Altman, 1991).

Next normality of the data under consideration will be assumed: If X_1 and X_2 denote the corresponding measurement results with methods 1 and 2, respectively, both will be assumed to follow a normal (Gaussian) distribution with respective population means μ_1, μ_2 and standard deviations σ_1, σ_2. This normal assumption can, for example, be enforced by taking replicate measurements of each patient with each measurement method of interest and using the respective replications' mean as a basis for analysis. Further let ρ denote the Pearson correlation between X_1 and X_2. The normal assumption allows for a quite obvious interpretation of *accuracy* and *precision* of the measurement methods represented by X_1 and X_2: different location (accuracy) can be described by the paired mean difference $\mu_1 - \mu_2 \neq 0$, deviation in scale (precision) is established for $\sigma_1^2 - \sigma_2^2 \neq 0$.

The normal assumption provides simple significance tests for the null hypotheses H_0: $\mu_1 = \mu_2$ (no difference in mean measurement location) and K_0: $\sigma_1^2 = \sigma_2^2$ (no difference in measurement scales versus corresponding alternatives H_1: $\mu_1 \neq \mu_2$ and K_1: $\sigma_1^2 \neq \sigma_2^2$.

Methods

The paired t-test

For the sake of completeness this section reviews the well-known *t*-test; its corresponding confidence interval procedure, however, is much less used, although remarkably more informative. The paired *t*-test can be applied to H_0 to test for statistically significant deviation in accuracy; to assess this deviation's possible clinical relevance

the corresponding $(1 - \alpha)$ confidence interval for the mean deviation in accuracy $\mu_1 - \mu_2$ should be computed. This interval provides a range, which contains the 'true' (but unknown) value of the mean difference in the underlying population with a probability of at least $1 - \alpha$. It therefore enables a decision as to whether the observed deviation in the measurement locations is not only of statistical significance, but also of clinical relevance.

If $\bar{d}$ denotes the arithmetic mean and s_d the empirical standard deviation of the differences $d_j = x_{1j} - x_{2j}$ for subject $j = 1, \ldots, n$, then the paired Student's t-test is based on

$$t = \frac{\bar{d}}{s_d} \cdot \sqrt{n} \ ,$$

where H_0 is rejected as soon as the absolute value of t becomes larger than the $(1 - \frac{\alpha}{2})$ quantile of Student's t distribution with $n - 1$ degrees of freedom, thus for $|t| \geq t_{n-1;1-\frac{\alpha}{2}}$. The latter quantile can be found in standard statistical tables.

Alternatively, a P value for this test can be provided by standard statistical software, which allows rejection of H_0 and thus acceptance of H_1 (significant differences in accuracy) as soon as this P value is less than or equal to the significance level α, that is for $P \leq \alpha$. The larger the value of t (indicating large mean differences between the methods), the smaller will be the corresponding P value.

The $(1 - \alpha)$ confidence interval for $\mu_1 - \mu_2$ can be computed similarily via

$$\bar{d} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$$

with the $(1 - \frac{\alpha}{2})$ quantile of t_{n-1} as above. It also provides a level α test for H_1: $\mu_1 \neq \mu_2$. It will reject H_0, if the value 0 is not contained in the $(1 - \alpha)$ confidence interval, the latter indicating significant differences between the population mean values μ_1 and μ_2. The confidence interval additionally provides information on the order of the mean difference, which has to be expected for the underlying population, from where the study subjects have been sampled. Thus, it illustrates the magnitude of clinical relevance contained in the observed deviation. Assuming a mean difference of $\bar{d} = 0.1$ has been observed for $n = 400$ patients with a standard deviation $s_d = 0.2$ to test H_0 at the significance level $\alpha = 5$ per cent, the necessary quantile then becomes $t_{400-1,1-\frac{0.05}{2}} = t_{399,0.025} = 1.96$ from standard tables. The limits for the 95 per cent confidence interval of $\mu_1 - \mu_2$ are thus

$$d_{\text{lower}} = 0.1 - 1.96 \cdot \frac{0.2}{\sqrt{400}} = 0.0804 \approx 0.08$$

$$d_{\text{upper}} = 0.1 - 1.96 \cdot \frac{0.2}{\sqrt{400}} = 0.1196 \approx 0.12.$$

Therefore, the population mean difference between the method accuracies lies between a minimum of 0.08 and a maximum of 0.12 (at a confidence level of 95 per cent). It is obviously context dependent as to whether this numerically small (although statistically significant) difference in the population means is also clinically relevant.

The Maloney and Rastogi (1970) test

To intra-individually compare measurement precisions the overall variance, as usually provided in the literature is not an indication of variability (and thus 'measurement error'), but rather a superposition of inter- and intra-individual data variation.

In the previous notation we thus have $\sigma_1^2 = \sigma_s^2 + \sigma_{e1}^2$ for method 1 and $\sigma_2^2 = \sigma_s^2 + \sigma_{e2}^2$ for method 2, with the inter-individual variation σ_2^2 and the measurement errors σ_{e1}^2 and σ_{e2}^2 of primary interest. This modified representation therefore distinguishes two different *sources* of variation and thus 'error' in the measurement realizations, that is the population variability and the 'intrinsic measurement error' representing the instrumental precision. The basic idea of the following tests is to exploit the above separation of subject and measurement variabilities σ_s^2 versus σ_{e1}^2, σ_{e2}^2 to obtain tests for the direct comparison of the intrinsic measurement components in terms of $\sigma_{e1}^2 - \sigma_{e2}^2$. Tests for different precisions are tests for K_0: $\sigma_{e1}^2 = \sigma_{e2}^2$.

However, the latter should not be accomplished by applying the usual F-ratio test, since paired variances are being considered: The standard F-ratio

$$F = \frac{\sigma_1^2}{\sigma_2^2} = \frac{\sigma_s^2 + \sigma_{e1}^2}{\sigma_s^2 + \sigma_{e2}^2}$$

will be smaller than the intended variance ratio

$$\frac{\sigma_{e1}^2}{\sigma_{e2}^2}$$

as soon as $\sigma_s^2 > 0$. Therefore, increasing population variance would cause increasing conservativeness of this *F*-ratio and a corresponding loss of indications of a (significant) difference in precisions. On the other hand, this inter-individual variance will often be required to be large to ensure sufficiently representative study groups.

Maloney and Rastogi (1970) suggested testing K_0 by using the empirical Pearson correlation *r* of $X_1 + X_2$ and $X_1 - X_2$. To understand the basic idea of this test, one has to observe, that the population analogue of this correlation is virtually proportional to

$$\frac{\sigma_{e1}^2 - \sigma_{e2}^2}{\sigma_s^2}.$$

The latter is zero, if and only if $\sigma_{e1}^2 = \sigma_{e2}^2$, i.e. K_0 is true. Therefore testing of K_0 becomes equivalent to testing for zero correlation between the intra-individual sum and the difference of the two measurements. This test, based on the empirical Pearson correlation, is available in most standard software packages.

The Maloney/Rastogi test can thus be performed by simply correlating the intra-individual sum and the difference of the measurement observations. Its corresponding *P* value can be obtained as a *P* value for the test of zero correlation available in any standard software package.

However the above test depends on the population variance being proportional to $1/\sigma_s^2$. Therefore, the statistical power of the Maloney/Rastogi test (that is the probability of detecting differences in the method precisions based on a fixed number of individuals) will dramatically decrease with increasing population variance. Therefore, the simple study design, when only one measurement from each method on each of the subjects is used, appears more insufficient due to confounding population and measurement variation than would be intuitively expected.

Studies with the primary goal of precision comparison therefore call for a homogeneous study population (i.e. a small σ_s^2) or, with the intention of deconfounding population and measurement error, replicate measurements on each subject under investigation, the latter being the better choice.

The Hahn and Nelson (1970) test

Hahn and Nelson (1970) proposed two different study designs for comparing the unknown value σ_{e1}^2 of a new measurement method's precision to the known value σ_{e2}^2 of a reference method. They suggested taking replicate measurements with the standard method on each of the study subjects whenever possible. As an alternative design for settings, where this is ethically unacceptable, too expensive, or simply excluded due to the destructive nature of the tests, they invoked a simultaneous observation with a second reference method X_3 with known precision σ_{e3}^2. The replication design is a special case of the latter due to $\sigma_{e2}^2 = \sigma_{e3}^2$. In either case, Hahn and Nelson (1970) considered the modified null hypothesis of identical measurement precisions

$$\tilde{K}_0 : \sigma_{e1}^2 = \frac{\sigma_{e2}^2 + \sigma_{e3}^2}{2},$$

i.e. the second reference replication or the second reference method were used to calibrate the first reference measurement. A statistical test for $\tilde{K}_0$ can be derived quite similarily to the Maloney/Rastogi statistic by computing the intra-individual difference $W = X_2 - X_3$ and the weighted sum

$$U = \frac{\sigma_{e3}^2}{\sigma_{e2}^2 + \sigma_{e3}^2} \cdot (X_1 - X_2) + \frac{\sigma_{e2}^2}{\sigma_{e2}^2 + \sigma_{e3}^2} \cdot (X_1 - X_3).$$

The usual *F*-ratio test can now be applied to these weighted differences (Grubbs, 1973), i.e. one first computes the values of *U* and *W* for each study subject, the corresponding empirical variances S_u^2 and S_w^2, and then $\tilde{K}_0$ can be tested with their *F*-ratio (as available in most standard software packages) for either of the Hahn/Nelson designs:

If two different reference methods are involved, then $\tilde{K}_0$ can be tested via

$$N_1 = \frac{S_u^2}{S_w^2} \cdot \left\{ \frac{1}{2} + \left(\frac{\sigma_{e2} \cdot \sigma_{e3}}{\sigma_{e2}^2 + \sigma_{e3}^2} \right)^2 \right\}^{-1}.$$

If two replicate reference measurements are taken, then $\tilde{K}_0$ can be tested via

$$N_2 = \frac{4 \cdot S_u^2}{3 \cdot S_w^2} \cdot$$

In either design, the Hahn/Nelson test is easy to perform by computing the weighted intra-individual differences U and W of the original data and then applying the usual F-test to U and W, deriving P values, etc., as an output of standard software packages.

Note that the Hahn/Nelson design is asymmetric concerning reference and new method in both cases: Although the new method will often be less invasive or expensive, both Hahn/Nelson designs call for two *reference* measurements. The inverse design (taking two measurements with the new method and only one with the standard) might be more intuitively attractive to clinical researchers. However, this study design may result in dramatic errors when computing the precisions of the measurement methods. Dunn (1992) has provided an instructive example, where the 'inverse Hahn/Nelson design' made a new measurement method look significantly more reliable than the gold standard, although it in fact could be shown to be significantly less precise. The reader is referred to Dunn's (1992) overview for further details.

The Grubbs (1973) test

A further disadvantage of the Hahn/Nelson test is the assumption of the exact knowledge of the reference method's precision(s). As an extension of the Hahn/Nelson procedure, Grubbs (1973) proposed a test for $\tilde{K}_0$, which involved additional estimation of the reference precision(s): compute $W = X_2 - X_3$, where X_3 is a second reference measurement obtained by replication or an additional calibrating reference method. Now consider the weighted intra-individual difference

$$V = X_1 - \frac{1}{2} \cdot (X_2 + X_3)$$

and let r denote the empirical Pearson correlation of W and V. Then $\tilde{K}_0$ can be tested at level α via

$$G = \frac{\left(\frac{S_v^2}{S_w^2} - \frac{3}{4} \right) \cdot \sqrt{n-2}}{\sqrt{3(1-r^2) \cdot \left(\frac{S_v^2}{S_w^2} \right)}},$$

where again S_v^2 and S_w^2 denote the empirical variances of V and W, respectively. $\tilde{K}_0$ is rejected at significance level α as soon as $|G| \geq t_{n-2,1-\frac{\alpha}{2}}$.

Grubbs' test can therefore be performed using tests for zero Pearson correlation, which are available in standard statistical software. Contrary to the Maloney/Rastogi proposal, this test is based on the above weighted intra-individual differences V and W to be computed in advance.

The Bland/Altman (1986) plot

Bland and Altman (1986) pointed out the misleading applications and interpretations of Pearson's correlation coefficient ρ in the method comparison setting. Whereas maximum correlation can occur in the presence of a severe deviation in scale and/or location, corresponding tests for deviation in accuracy and/or precision may tend to indicate significantly different measurement methods despite fine reproducibility, but very small (although systematic) shifts in location or scale. Similarily, the t-test will fail to reject the null hypothesis of identical means as soon as only the *mean* measurements are equal despite any unreproducibility or even discordance in the underlying data series. Bland and Altman (1986) further argued against the correlation pendant 'scattergram of X_1 versus X_2', which is often (mis)used in the same way as Pearson's correlation. Instead of using the common, but moderately informative scattergram, they proposed plotting the mean $\frac{1}{2}(X_1+X_2)$ versus the difference $X_1 - X_2$, and additionally providing two horizontal lines indicating $\bar{d} \pm 2 \cdot s_d$ (mean ± twice the standard deviation of the intra-individual difference). By further providing the 95 per cent confidence interval for the

expected intra-individual difference by additional parallel horizontal lines, three points of interpretation are provided by this plot:

1. If the horizontal line through 0 lies outside the 95 per cent confidence bounds, shifted location of the observations in the Bland/Altman plot indicates significantly different measurement accuracies at the 5 per cent significance level. The latter can be used to assess the order of deviation in accuracy and thus consider its clinical relevance. Therefore, the 'confidence interval limits' provide information on bias between the two measurement methods under consideration.
2. The 'twice standard deviation limits' provide indirect information on the agreement between the imaging methods at hand: according to the normal distribution theory, approximately 95 per cent of the individual differences should be contained within these limits, referred to as the 'agreement limits' (Bland and Altman, 1986). Provided differences within these limits would not be clinically important, the two methods could be interchanged.
3. Bearing in mind the Maloney/Rastogi proposal, the Bland/Altman plot also provides graphical information on differences in precision: If the Maloney/Rastogi test rejects the hypothesis K_0 and thus establishes significantly different measurement precisions, the Pearson correlation between the intra-individual sum and difference is significantly non-zero. Therefore the Bland/Altman plot (as the plot of intra-individual mean versus difference) will show a positive ($\sigma_{e1} > \sigma_{e2}$) or negative ($\sigma_{e1} < \sigma_{e2}$) direction. Structure and trend in the Bland/Altman plot therefore indicate different measurement precisions. The visual strength of the plot's direction provides an impression for the magnitude of deviation in precisions: the more direction in the Bland/Altman plot, the more evidence for deviation in measurement precisions. If there is no obvious relationship between difference and mean (i.e. no trend in the Bland/Altman plot), the lack of agreement between the methods can be summarized by computing their (constant) bias in terms of $\bar{d}$ and adjusted for it by subtracting $\bar{d}$ from the new method to reproduce the values of the reference.

Points 1 and 2 do not contradict each other, since two measurement methods may show very small bias despite poor agreement and, on the other hand, may show fine reproducibility of measurement despite systematic, but clinically irrelevant bias.

In the above sense, the Bland/Altman plot provides aggregate information necessary for an efficient method comparison. Nevertheless, method comparison could also *additionally* include the usual scattergram of X_1 versus X_2; the latter is more convenient to medical researchers and also allows for recognition of reproducibility between the methods' measurement results. However, this plot and its numerical pendant, the Pearson correlation coefficient, should not be the *only* information to be provided.

Figures 1 and 2 present the scattergram and the Bland/Altman plot for the graphical representation of 2000 simulated data points (x_1, x_2) to illustrate the information contained in both diagrams: the Bland/Altman representation immediately provides the information $\sigma_{e1} \ll \sigma_{e2}$ (note the negative direction in the intra-individual difference/mean regression plot), and the absence of bias between the two fictious measurement methods (the 0 line lies between the horizontal confidence limits, i.e. there is no significant difference in the location of the methods). The scattergram provides an indication of at least moderate correlation (estimated Pearson correlation $r = 0.66$). However, the difference in precision as recognised in the Bland/Altman plot is rather difficult to detect (despite the fact that one method has a range of approx –2 to +2 and the other of –4 to +4). It should be noted that the 95 per cent agreement limits in the Bland/Altman plot are approximately –2 to +2 and cover virtually the entire range of the data. Therefore, the agreement between these two fictious measurement methods should be considered rather poor, in contrast to the rather optimistic (but worthless) Pearson correlation of 0.66. This illustrates that method comparison should be based on intra-individual comparison of location (means)

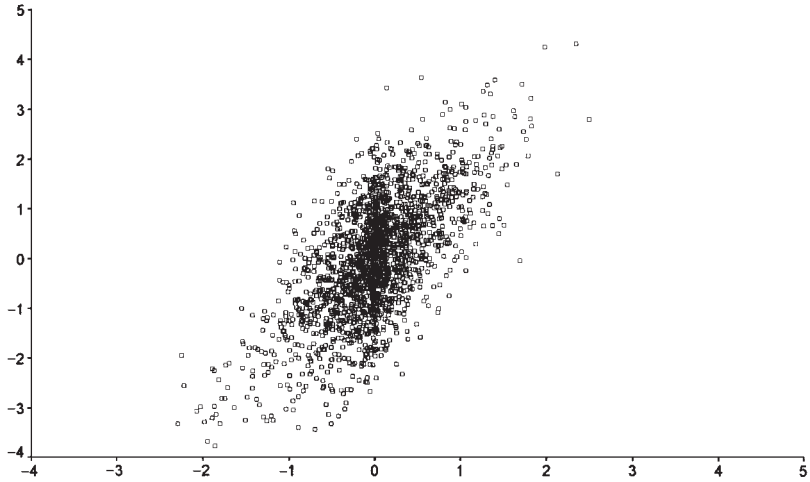

Figure 1 Scattergram for 2000 simulated data points (comparison of two fictious variates X_1 and X_2).

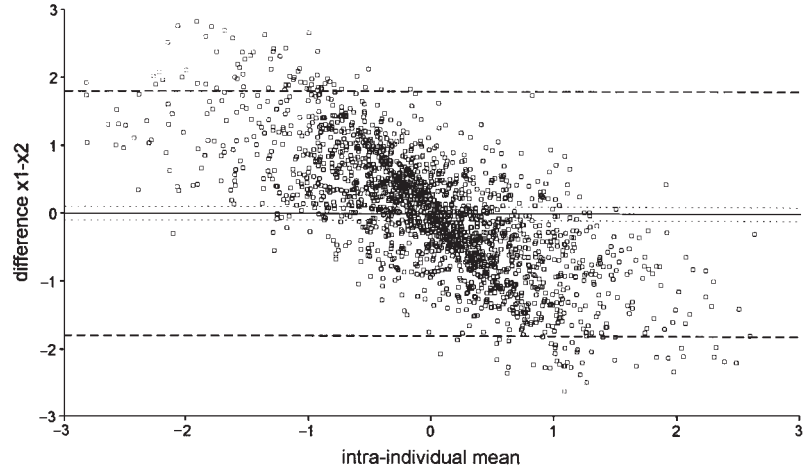

Figure 2 Bland/Altman plot for 2000 simulated data points (comparison of two fictious variates X_1 and X_2); exterior hatched lines indicate 'agreement limits' ($\bar{d} \pm 2_d^s$) and interior hatched lines the expected intra-individual difference's 95 per cent 'confidence interval limits'.

and precision (variances) instead of the more common, but often misleading, correlation.

The Krippendorf (1970) coefficient

Neither maximum correlation ρ nor agreement in accuracy and precision alone will suffice to prove concordance and thus sufficient reproducibility among methods; this requires $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$ and $\rho = +1$ simultaneously for the characterization of perfect measurement concordance. To obtain a coefficient with maximum range between –1 and +1, (such as established for the Pearson correlation), Lin

(1989) derived a coefficient K with a range between –1 (perfect discordance) to +1 (perfect concordance between the methods):

$$K = \frac{2\sigma_1\sigma_2 \cdot \rho}{2\sigma_1\sigma_2 + (\sigma_1 - \sigma_2)^2 + (\mu_1 - \mu_2)^2}$$

is +1 if and only if $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$ and $\rho = +1$ simultaneously. Therefore, K corresponds to the Bland/Altman plot in a similar way as ρ corresponds to the simple scattergram of X_1 versus X_2. It represents an analogue of Krippendorff's (1970) weighted kappa for ordinal data and can be estimated as

$$\hat{K} = \frac{2 \cdot r \cdot S_1 \cdot S_2}{S_1^2 + S_2^2 + \left(\overline{X}_1 - \overline{X}_2\right)^2},$$

where X_1, X_2, S_1^2, S_2^2, r denote the arithmetic means, the empirical variances and the observed Pearson correlation of the measurement methods 1 and 2, i.e. X_1, X_2, S_1^2, S_1^2 are the sample analogues of μ_1, μ_2, σ_1^2, σ_2^2, respectively. A $(1 - \alpha)$ confidence interval for K has been provided by Lin (1989).

The DigiGraph® data

The DigiGraph® (Nanda *et al.*, 1995) is a recent method for measuring the value of several orthodontic parameters in children by avoiding radiographic exposure. To assess its concordance with the usual cephalometric analysis of radiographic images, the results of twice repeated DigiGraph® measurements and a twice repeated planimetry analysis of 32 orthodontic parameters in 50 female patients aged 13 years were compared. Due to its non-invasive character and the time-efficient possibility of taking replicate measurements on children, it was hoped that the DigiGraph® would substitute X-ray based diagnostics in orthodontics as soon as sufficient concordance with the planimetry reference could be ensured.

It should be mentioned that the following results are only representative for one of 32 parameters under consideration. It general, quite satisfying reliability of the DigiGraph® and concordance with the reference cephalometry results were observed except for this one parameter. Nevertheless, this parameter, the posterior face height (Jarabak), has been chosen to illustrate the performance of the previous estimation techniques for situations of discordant measurement methods.

Two replicates per child and each method were each aggregated as mean values providing normally distributed measurement data. In the following two characteristic scenarios will be illustrated: First, the mean values to obtain information on the concordance between the DigiGraph® and the planimetry reference will be compared (Table 1 and Figure 3), and secondly the relative performance of the two intra-individual DigiGraph® replicates to obtain an impression of its reproducibility (Table 2 and Figure 4). It should be noted that the latter represents a different kind of intra-individual data dependency, which should undergo further investigation using analysis of variance, e.g. to derive an intra-class correlation. The following, however, is used to illustrate the previous statistical methods rather than to report standard techniques for repeated measurement analysis. Nevertheless, taking the average of repeated measurements may influence the univariate procedures reviewed in this article as already indicated for the Maloney/Rastogi test. In general, the repeated measurements should not be immediately aggregated without exploiting their explicit information.

Tables 1 and 2 each present the 95 per cent confidence interval (CI) for the mean difference, the P values of the paired t-test, the Maloney/

Table 1 Numerical information for the comparison of planimetry (X_1) and DigiGraph® (X_2) assessment based on 50 children.

$\bar{x}_1$ (mean)	1.79
$\bar{x}_2$ (mean)	1.41
s_1 (standard deviation)	0.40
s_2 (standard deviation)	0.52
r (Pearson correlation)	0.70
95% CI for mean difference	(0.27;0.48)
t-test (P value)	$P < 0.001$
Maloney/Rastogi test	$P = 0.02$
Grubbs test (P value)	$P = 0.009$
K (Krippendorff coefficient)	0.51
95% CI for K	(0.35;0.67)

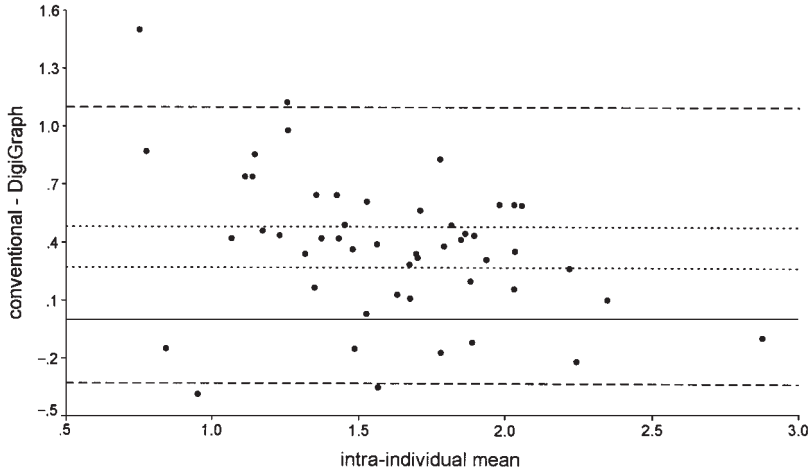

Figure 3 Bland/Altman plot for the comparison of planimetry (X_1) and DigiGraph® (X_2) assessment based on 50 children; exterior hatched lines indicate 'agreement limits' ($\bar{d} \pm 2s_d$) and interior hatched lines the expected intra-individual difference's 95 per cent 'confidence interval limits'.

Table 2 Numerical information for the comparison of first (X_1) and second (X_2) DigiGraph® replication based on 50 children.

$\bar{x}_1$ (mean)	1.40
$\bar{x}_2$ (mean)	1.41
s_1 (standard deviation)	0.52
s_2 (standard deviation)	0.53
r (Pearson correlation)	0.99
95% CI for mean difference	(–0.03;0.01)
t-test (P value)	$P = 0.19$
Maloney/Rastogi test	$P = 0.80$
Grubbs test (P value)	$P = 0.69$
K (Krippendorff coefficient)	0.99
95% CI for K	(0.99;1.00)

Rastogi and the Grubbs' test, the empirical Pearson correlation r and the Krippendorff coefficient K, respectively. The 95 per cent confidence intervals for K are also shown.

Figures 3 and 4 present the Bland/Altman plot (the dotted lines limit the 95 per cent confidence interval for the mean difference) for the mean replicate DigiGraph® and planimetry assessments (Figure 3) and the original DigiGraph® replications (Figure 4). Whereas reproducibility amongst the two DigiGraph® replicates is fine, reproducibility between the planimetry and the DigiGraph® measurements is unacceptably low as illustrated by the corresponding Bland/Altman plot. Figure 3 indicates the existence of a significant absolute bias between the methods, since the horizontal line through 0 is not contained in the 95 per cent confidence interval limits. Next, a negative trend in this plot can be observed, which indicates a clinically relevant difference in measurement precisions; since this trend is negative, it can be concluded that the measurement variance component of the DigiGraph® was larger than that of the gold standard. Therefore, the DigiGraph® is seen to be less precise than conventional cephalometry. Finally, one can interpret the range –0.35 to 1.18 of the agreement limits, which indicates an underestimation of more than 1 unit by the DigiGraph® (at least for some of the study subjects). This order of deviation makes the two measurement methods look much less concordant than their Pearson correlation 0.70 (Table 1) would have suggested. All three characteristics (accuracy, precision and correlation) involved into estimation of K can be recognized in these plots; thus, to a certain extent, $\hat{K}$ corresponds to the Bland/Altman plot like the empirical Pearson correlation r corresponds to the linear regression scattergram.

Summarizing the following facts can be recognized: Planimetry and DigiGraph® assessments differ in accuracy (note the quite large

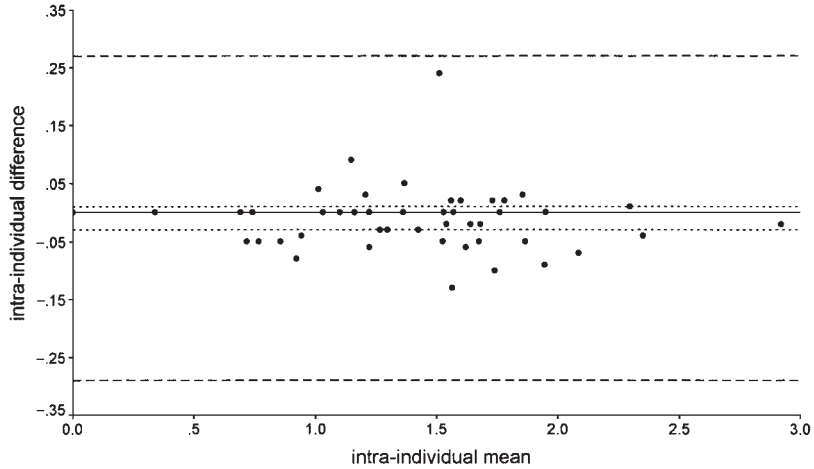

Figure 4 Bland/Altman plot for the comparison of first (X_1) and second (X_2) DigiGraph® replication based on 50 children; exterior hatched lines indicate 'agreement limits' ($\bar{d} \pm 2s_d$) and interior hatched lines the expected intra-individual difference's 95 per cent 'confidence interval limits'.

mean difference with 95 per cent confidence interval (0.27; 0.48), which indicates a substantial underestimation of the posterior face height (Jarabak) by the DigiGraph®. In addition, the DigiGraph® appears significantly less precise than planimetry (Grubbs $P = 0.009$) as indicated by the obvious trend in the corresponding Bland/Altman plot. The Pearson correlation $r = 0.70$ indicates quite moderate reproducibility which is also mirrored in the even smaller value of the Krippendorff coefficient $K = 0.53$. Therefore, regarding this certain parameter, the DigiGraph® can hardly be established as a diagnostic substitute for the usual X-ray based cephalometry. The cause of the observed discordance between the measurement methods might be the posterior face height's dependence on the exact fixing of the sella coordinates.

Regarding the relative performance of the two DigiGraph® replications, there is almost perfect reproducibility between them, which is confirmed by the maximum possible range of the 95 per cent confidence interval for the Krippendorff coefficient (Table 2); neither differences in accuracy nor in precision are found (paired t-test $P = 0.19$, Grubbs $P = 0.69$). The correlation between the intra-individual replications is also substantial indicating perfect reliability of the DigiGraph®. The same is indicated in the corresponding Bland/Altman plot (Figure 4), which does not show any trend (and thus absence of deviations in precisions) or any deviation concerning accuracy, since the value 0 is obviously contained in the difference's 95 per cent confidence interval. Nevertheless, the DigiGraph® assessment of this parameter seems to be confounded by other influental factors concerning orthodontic geometry which will be investigated in subsequent studies.

Discussion

This paper has tried to survey some standard methods for the comparison of clinical imaging procedures and provide elementary estimates and test procedures for assessing differences in accuracy, precision and reliability of the methods under consideration. Such an overview cannot be complete and was never intended to be. It should rather provide simple suggestions for approaches to method comparison. Of course, there are many more efficient and elaborate multivariate statistical methods available (e.g. Jaech, 1985; Fleiss, 1986; Dunn, 1992) to analyse, for example, reliability data as implicitly contained in the DigiGraph® data. Nevertheless, many practical data applications will require simple and thus robust analysis due to the possibly uncontrollable sources of bias and errors inevitable in clinical reality. In such

situations, if even the simple models reviewed in this article cannot be justified for describing the data, more complex approaches hardly will. Therefore, in agreement with Bland and Altman (1986), it is suggested that simple, but clinically interpretable statistical methods should be applied, rather than more efficient, but also more complicated and restrictive approaches.

It should be mentioned, that the choice of only one method for analysis may not be appropriate in practice, but efforts should be made to obtain a sufficient descriptive overview of the data. Results of tests such as presented in this paper should be regarded as descriptive *P* values and sources of additional 'trial and error approaches' rather than as the results of static, but confirmatory, hypotheses testing (since the latter may not be flexible enough to detect the real reasons of disagreement between measurement methods). A further possible source of confusion can be an inadequate study design: whereas a precision comparison study calls for an optimum homogeneous population to reduce the loss in efficiency shown for the Maloney/Rastogi test, a simple replication design will be employed to include a representatively wide range of individuals to keep the reliability estimates derived from this study representative for a maximum general population. This relates to the population dependence of the results: dental imaging methods may be highly efficient and reliable in patients with severe dental disorders, but they undergo dramatic loss in accuracy or precision as soon as applied to healthy probands or patients with moderate or different sources of dental disorder. Therefore, planning of method comparison studies should sensitively incorporate the main goal of the investigation intention, that is either providing a representative range of observers and instruments and a maximum homogeneous study population (if measurement errors are focused on), or a representative range of study subjects to also allow for estimation of subject-specific properties. These study designs are comparable to diagnostic score evaluation studies (which are primarily intended to assess rater variation as a surrogate measure of reliability of the proposed diagnostic score) and diagnostic studies (which are primarily interested in a subject's optimum reliable diagnostic findings and adjust the latter for rater variation).

Finally, the necessity and number of measurement replications should be considered during the planning phase of the study. As long as costs and ethical considerations do not explicitly prevent replication, replicate assessments should always be taken with all clinical measurement methods of interest, but *at least* with the reference method as indicated for the Hahn/Nelson test(s). If tests are destructive or cannot be repeated due to ethical or economic reasons, a second reference method should be considered to calibrate the first one and then compare the reference methods' 'consensus assessment' to the non-reference result. A remarkable loss in power was illustrated for the Maloney/Rastogi test, if method comparison is based on single measurements without a second calibrating reference and without sufficiently reducing the population variance within the planning phase. In general, care should be taken when designing reliability studies including replicate measurements for some of the methods to be compared.

The results of measurement comparison studies not only depend on the data at hand, but are also influenced by both the study design and the underlying population, and last (but not least) by the clinical context of interest.

Summarizing tips for practice

The above items indicate the following recommendations to provide minimum univariate information necessary for the statistical representation of method comparison trials (bearing in mind that there are a number of comprehensive multivariate approaches for the analysis of method comparison studies such as reviewed by Dunn, 1992):

1. Numerical representation may be structured as indicated in Tables 1 and 2, where *K* and *r* mainly provide information on method agreement and reproducibility, whereas means, standard deviations and the corresponding paired significance tests rather refer to location and scale, i.e. to bias, accuracy, precision and thus validity.

2. Graphical representation may be performed as indicated in Figures 1 and 2, where the scattergram mainly provides information on reproducibility, the Bland/Altman plot mainly on location and scale, i.e. on bias, accuracy, precision and thus validity.
3. Confidence intervals on differences in locations should be provided whenever possible to illustrate the clinical and not only the statistical information contained in the data at hand.
4. The Krippendorff coefficient K corresponds to the Bland/Altman plot in a similar way as the Pearson correlation r corresponds to the scattergram of X_1 versus X_2.
5. The usual F-ratio test must not be applied to paired data, since this may result in overly liberal decisions, i.e. under-estimation of differences in measurement precisions. The tests of Maloney and Rastogi, Hahn and Nelson or Grubbs should be applied instead. The latter can be used, if replicate measurements based on reference methods are available. The Maloney/Rastogi test is an expedient for, e.g. destroying test scenarios, where replications are not possible, whereas the Hahn/Nelson test assumes and involves exact knowledge of both bias and precision of the reference method(s), Grubbs' test can be applied without this knowledge.
6. Studies with the primary goal of precision comparison call for an homogeneous study population (i.e. a small population variability) or, with the intention of deconfounding population and measurement error, even better replicate measurements on each subject under investigation.
7. Although a new measurement method will often be less invasive or expensive, both Hahn/Nelson designs call for two *reference* measurements. The inverse design (two replicates with the new method and only one with the reference), which might appear somewhat more attractive to clinical researchers, may result in dramatic errors when comparing the precisions of the measurement methods.
8. The Maloney/Rastogi test can be performed by simply correlating the intra-individual sum and difference of the measurement observations, a corresponding P value can be obtained as a P value for the test of zero correlation available in most standard software packages.
9. In either replication design, the Hahn/Nelson test is easy to perform by computing the weighted intra-individual differences U and W (Methods section) from the original data and applying the usual F-ratio test to U and W, providing a P value as an output of standard software packages.
10. The Grubbs' test for differences in instrumental precisions can be performed using corresponding tests for zero Pearson correlation as available in standard statistical software. Contrary to the Maloney/Rastogi proposal, this test is based on the weighted intra-individual differences V and W as indicated in the Methods section.

Address for correspondence

Dr F. Krummenauer
Department of Medical Statistics and Documentation
Obere Zahibacher Strasse 69
University of Mainz
D–55131 Mainz
Germany

Acknowledgements

The authors wish to thank Professor Jörg Michaelis (Department of Medical Statistics and Documentation, University Clinics of Mainz) for helpful comments during the preparation of this work, and Dr Heidi Gärtner (Department of Orthodontics, University Clinics of Mainz) for providing the DigiGraph® data.

References

Altman D G 1991 Practical statistics for medical research. Chapman & Hall, London

Bland M, Altman D G 1986 Statistical methods for assessing agreement between two methods of clinical measurement. Lancet i: 307–310

Dunn G 1992 Design and analysis of reliability studies. Statistical Methods in Medical Research 1: 123–157

Fleiss J L 1986 The design and analysis of clinical experiments. John Wiley & Sons, New York

Grubbs F E 1973 Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. Technometrics 15: 53–66

Hahn J H, Nelson W 1970 A problem in the statistical comparison of measuring devices. Technometrics 12: 95–102

Jaech J L 1985 Statistical analysis of measurement errors. John Wiley & Sons, New York

Krippendorff K 1970 Bivariate agreement coefficients for reliability of data. In: Borgatta F, Bohrnstedt G W (eds) Sociological methodology. Jossey-Bass, San Francisco, pp. 139–150

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

Maloney C J, Rastogi S C 1970 Significance tests for Grubbs' estimators. Biometrics 26: 671–676

Nanda R S, Ghosh J, Bazakidou E 1995 Three-dimensional facial analysis using a video imaging system. Angle Orthodontist 66: 181–188