

Contents

0.1 Bland-Altman methodology

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of paired sample t -test, correlation coefficients or simple linear regression. Simple linear regression is unsuitable for method comparison studies because of the required assumption that one variable is measured without error. In comparing two methods, both methods are assumed to have attendant random error.

Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which they are unsuitable for comparing two methods of measurement (?). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be ‘explained to non-statisticians’.

Notwithstanding previous remarks about linear regression, the first step recommended, which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality should also be shown, as it is necessary

to give the correct interpretation of how both methods compare. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.



Figure 1: Scatter plot For Fotobalk and Counter Methods.

? notes that scatter plots were very seldom presented in the Annals of Clinical Biochemistry. This apparently results from the fact that the ‘Instructions for Authors’ dissuade the use of regression analysis, which conventionally is accompanied by a scatter plot.

0.1.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, \dots, n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \dots, n$).

? proposes a scatterplot of the case-wise averages and differences of two methods of measurement. This scatterplot has since become widely known as the Bland-Altman plot. ? express the motivation for this plot thusly:

“From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

The case wise-averages capture several aspects of the data, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. ? cautions that it would be the difference against either measurement value instead of their average, as the difference relates to both value. This methodology has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical methodology for making a visual assessment of the data.

The magnitude of the inter-method bias between the two methods is simply the average of the differences \bar{d} . This inter-method bias is represented with a line on the Bland-Altman plot. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are also particularly

relevant. The variances around this bias is estimated by the standard deviation of these differences S_d .

0.1.2 Bland-Altman plots for the Grubbs data

In the case of the Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

The Bland-Altman plot for comparing the ‘Fotobalk’ and ‘Counter’ methods, which shall henceforth be referred to as the ‘F vs C’ comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.8	793.2	0.6	793.5
2	793.1	793.3	-0.2	793.2
3	792.4	792.6	-0.2	792.5
4	794.0	793.8	0.2	793.9
5	791.4	791.6	-0.2	791.5
6	792.4	791.6	0.8	792.0
7	791.7	791.6	0.1	791.6
8	792.3	792.4	-0.1	792.3
9	789.6	788.5	1.1	789.0
10	794.4	794.7	-0.3	794.5
11	790.9	791.3	-0.4	791.1
12	793.5	793.5	0.0	793.5

Table 2: Fotobalk and Terma methods: differences and averages.

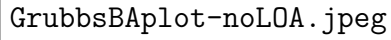


Figure 2: Bland-Altman plot For Fotobalk and Counter methods.

In Figure 1.3 Bland-Altman plots for the ‘F vs C’ and ‘F vs T’ comparisons are shown, where ‘F vs T’ refers to the comparison of the ‘Fotobalk’ and ‘Terma’ methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the ‘F vs C’ comparison than in the ‘F vs T’ comparison. Conversely there appears to be less precision in ‘F vs T’ comparison, as indicated by the greater dispersion of covariates.

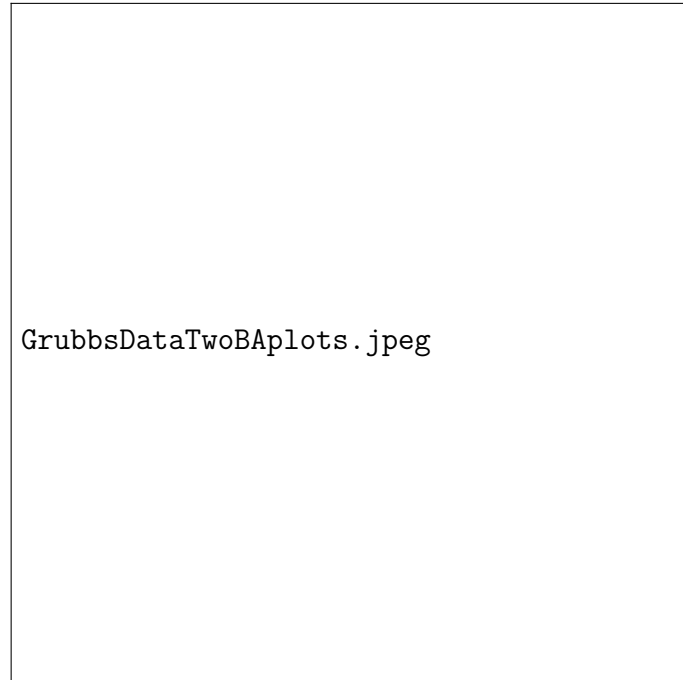


Figure 3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

0.1.3 Prevalence of the Bland-Altman plot

?, which further develops the Bland-Altman methodology, was found to be the sixth most cited paper of all time by the ?. ? describes the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. ? reviewed the use of Bland-Altman plots by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001. This study concluded that use of the BlandAltman plot increased over the years, from 8% in 1995 to 14% in 1996, and 3136% in 2002.

The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (?). Furthermore ? recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

0.1.4 Adverse features

Estimates for inter-method bias and variance of differences are only meaningful if there is uniform inter-bias and variability throughout the range of measurements. Fulfilment of these assumptions can be checked by visual inspection of the plot. The prototype Bland-Altman plots depicted in Figures 1.4, 1.5 and 1.6 are derived from simulated data, for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias, and is defined by ? as meaning that ‘one method gives values that are higher (or lower) than those from the other by an amount that is proportional to the level of the measured variable’. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (?) test, should be also be used.

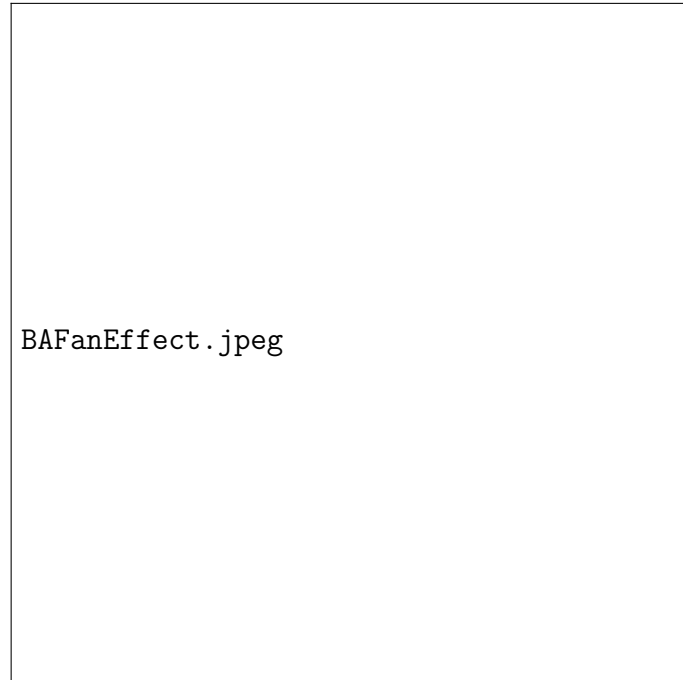


Figure 4: Bland-Altman plot demonstrating the increase of variance over the range.



Figure 5: Bland-Altman plot indicating the presence of proportional bias.

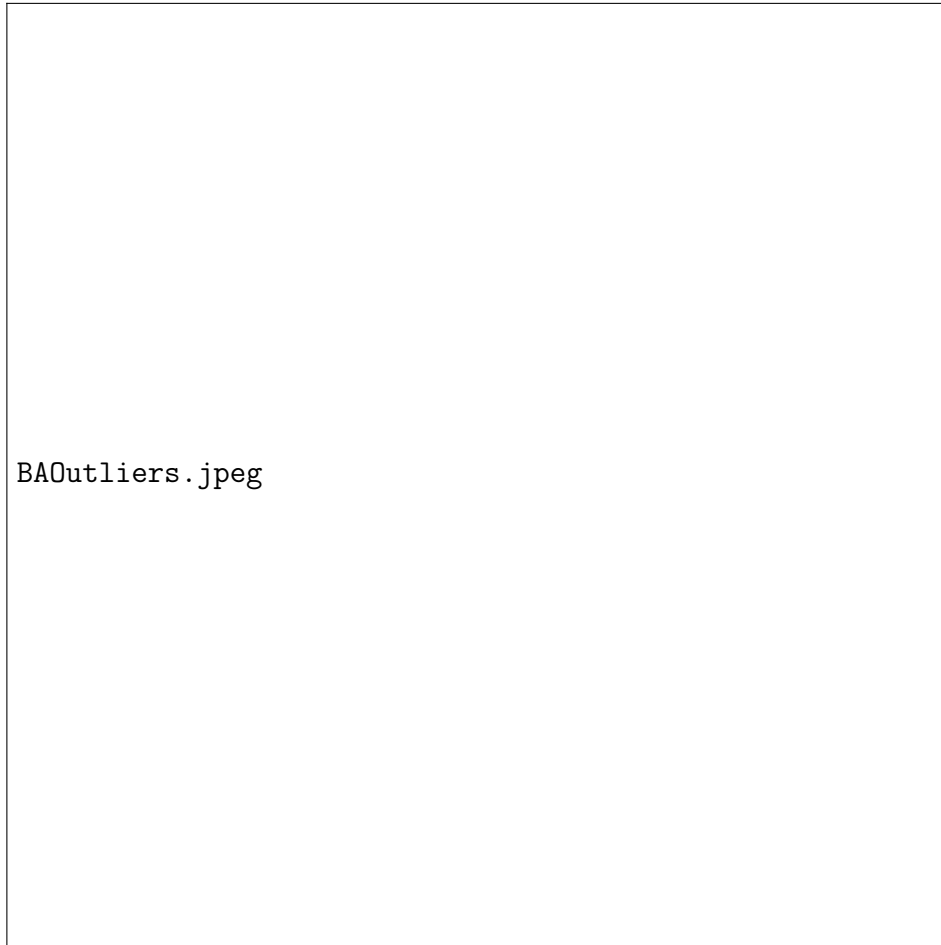


Figure 6: Bland-Altman plot indicating the presence of potential outliers.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. ? do not recommend excluding outliers from analyzes, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that ‘we usually find that this method of analysis is not too sensitive to one or two large outlying differences’. Figure 1.6 demonstrates how the Bland-Altman plot can be used to visually inspect the presence of potential outliers.

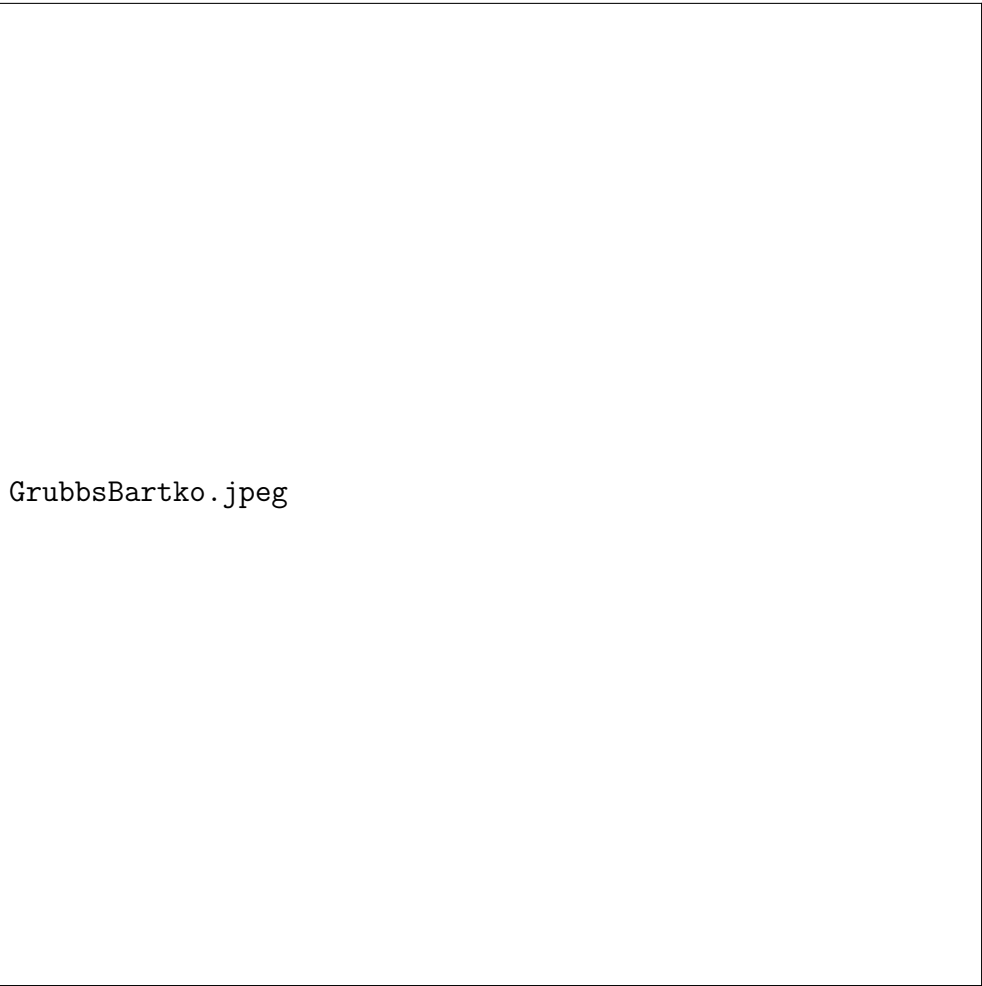
As a complement to the Bland-Altman plot, ? proposes the use of a bivariate confidence ellipse, constructed for a predetermined level. ? provides the relevant calculations for the ellipse. This ellipse is intended as a visual guidelines for the scatter plot, for detecting outliers and to assess the within- and between-subject variances.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other. Consequently Bartko’s ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a)$ is greater than $\text{var}(d)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a)$ is less than $\text{var}(d)$, the orientation of the ellipse is vertical.

The Bland-Altman plot for the Grubbs data, complemented by Bartko’s ellipse, is depicted in Figure 1.7. The fourth observation is shown to be outside the bounds of the ellipse, indicating that it is a potential outlier.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko’s ellipse. A covariate is added to the ‘F vs C’ comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this manipulated data set. By inspection of the confidence interval, a conclusion would be reached that this extra covariate is an outlier, in spite of the fact that this observation is wholly consistent with the conclusion of the Bland-Altman plot.

Importantly, outlier classification must be informed by the logic of the data’s formulation. In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra covariate. Conversely, the fourth observation, from




GrubbsBartko.jpeg

Figure 7: Bartko's Ellipse For Grubbs' Data.

the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

In classifying whether a observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the Grubbs test procedure is the absence of any outliers in the data set. Conversely, the alternative hypotheses is that there is at least one outlier present.

The test statistic for the Grubbs test (G) is the largest absolute deviation from the



GrubbsBartko2.jpeg

Figure 8: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

sample mean divided by the standard deviation of the differences,

$$G = \max_{i=1,\dots,n} \frac{|d_i - \bar{d}|}{S_d}.$$

For the 'F vs C' comparison it is the fourth observation gives rise to the test statistic, $G = 3.64$. The critical value is calculated using Student's t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n), n-2}^2}{n-2 + t_{\alpha/(2n), n-2}^2}}.$$

For this test $U = 0.75$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with p -value = 0.003, according with the previous result using Bartko's ellipse.

0.1.5 Inferences on Bland-Altman estimates

? advises on how to calculate confidence intervals for the inter-method bias and limits of agreement. For the inter-method bias, the confidence interval is a simply that of a mean: $\bar{d} \pm t_{(0.5\alpha, n-1)} S_d / \sqrt{n}$. The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

$$\text{Var}(LoA) = \left(\frac{1}{n} + \frac{1.96^2}{2(n-1)} \right) s_d^2.$$

If n is sufficiently large this can be following approximation can be used

$$\text{Var}(LoA) \approx 1.71^2 \frac{s_d^2}{n}.$$

Consequently the standard errors of both limits can be approximated as 1.71 times the standard error of the differences.

A 95% confidence interval can be determined, by means of the t distribution with $n-1$ degrees of freedom. However ? comment that such calculations may be ‘somewhat optimistic’ on account of the associated assumptions not being realized.

0.1.6 Formal definition of limits of agreement

? note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as ‘being like a reference interval’.

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the Shewhart limits are not based on any sample used for an analysis, but on the process’s historical values, a key difference with Bland-Altman limits of agreement.

? regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. ? offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975, n-1)} s_d \sqrt{1 + \frac{1}{n}}$$

where n is the number of subjects. Carstensen is careful to consider the effect of the sample size on the interval width, adding that only for 61 or more subjects is there a quantile less than 2.

? offers an alternative description of limits of agreement, this time as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence. ? describes them as a probability interval, and offers a clear description of how they should be used; 'if the absolute limit is less than an acceptable difference d_0 , then the agreement between the two methods is deemed satisfactory'.

The prevalence of contradictory definitions of what limits of agreement strictly are will inevitably attenuate the poor standard of reporting using limits of agreement, as mentioned by ?.

0.1.7 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as 'replicate

measurements'. ? recommends the use of replicate measurements, but acknowledges the additional computational complexity.

? address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. ? propose a correction for this.

? takes issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation. ? demonstrates how the limits of agreement calculated using the mean of replicates are 'much too narrow as prediction limits for differences between future single measurements'. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the 'mean of replicates' approach.

Chapter 1

The Bland Altman Plot

1.1 Bland Altman Plots In Literature

? contains a study the use of Bland Altman plots of 44 articles in several named journals over a two year period. 42 articles used Bland Altman's limits of agreement, wit the other two used correlation and regression analyses. ? remarks that 3 papers, from 42 mention predefined maximum width for limits of agreement which would not impair medical care.

The conclusion of ? is that there are several inadequacies and inconsistencies in the reporting of results ,and that more standardization in the use of Bland Altman plots is required. The authors recommend the prior determination of limits of agreement before the study is carried out. This contention is endorsed by ?, which makes a similar recommendation for the sample size, noting that *sample sizes required either was not mentioned or no rationale for its choice was given.*

In order to avoid the appearance of "data dredging", both the sample size and the (limits of agreement) should be specified and justified before the actual conduct of the trial. (?)

? remarks that the limits of agreement should be compared to a clinically acceptable difference in measurements.

1.1.1 Gold Standard

This is considered to be the most accurate measurement of a particular parameter.

The Gold Standard may not be financially feasible for general use, and therefore more economical methods, of suitable levels of precisions, must be devised. Method Comparison studies is used to ascertain the levels of precision of such methods.

1.2 Bland Altman Plots

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (?).

Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge that there are other valid, but complex, methodologies, and argue that a simple approach is preferable to this complex approaches, *especially when the results must be explained to non-statisticians* (?).

Notwithstanding previous remarks about regression, the first step recommended ,which the authors argue should be mandatory,is construction of a simple scatter plot of the data. The line of equality ($X = Y$) should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion

that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

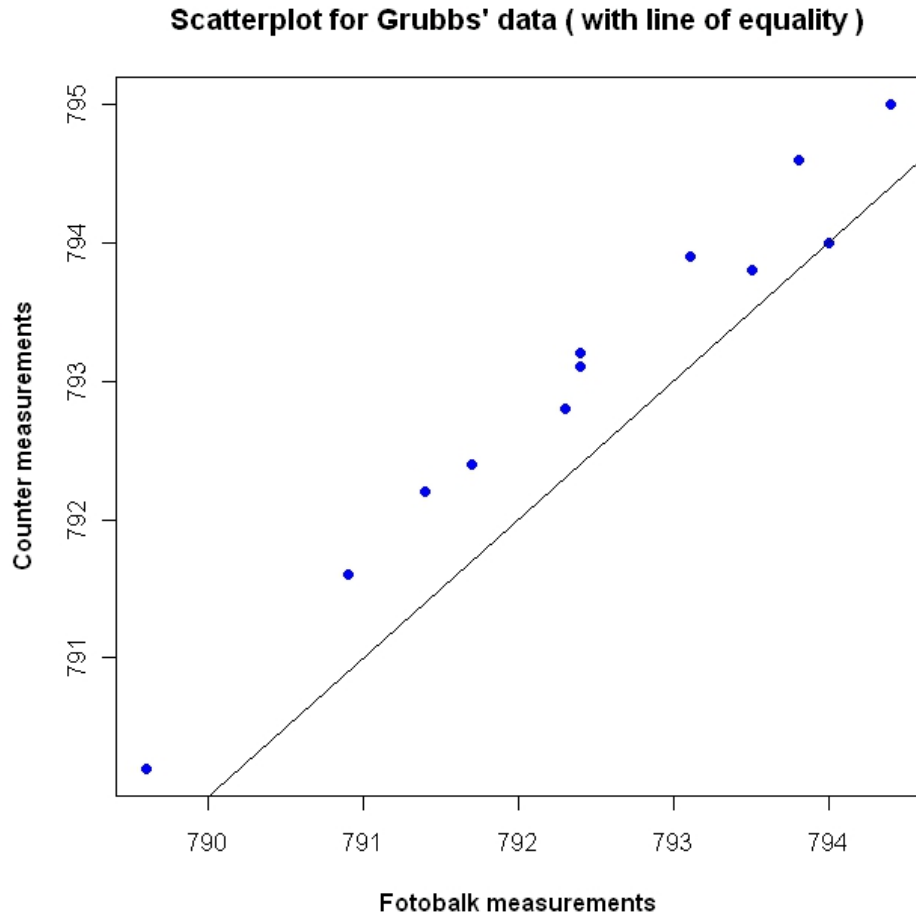


Figure 1.1: Scatter plot For Fotobalk and Counter Methods

In light of shortcomings associated with scatterplots, I recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 2.1). These differences and averages are then plotted (Figure 2.2).

The dashed line in Figure 2.2 alludes to the inter method bias between the two methods, as mentioned previously. Bland and Altman recommend the estimation of inter method bias by calculating the average of the differences. In the case of Grubbs

data the inter method bias is -0.6083 metres per second.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages $[(F+C)/2]$
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.1: Fotobalk and Counter Methods: Differences and Averages

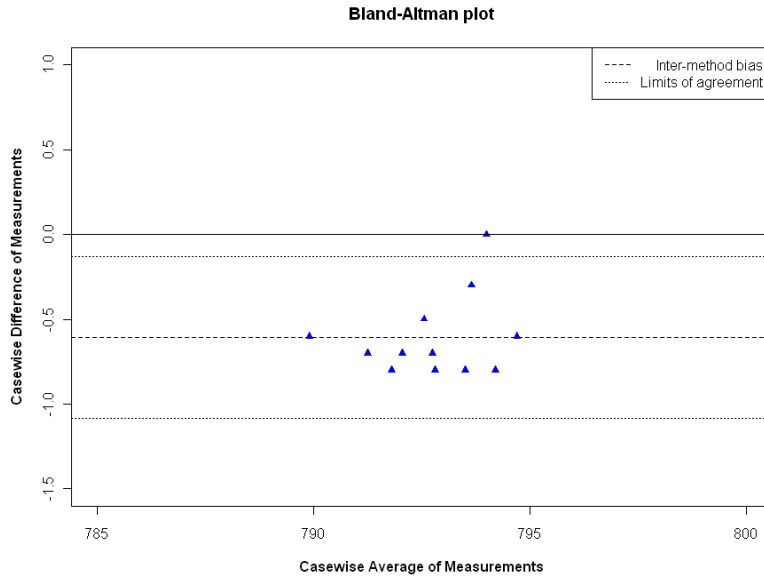


Figure 1.2: Bland Altman Plot For Fotobalk and Counter Methods

By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

1.2.1 Inspecting the Data

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. ? express the motivation for this plot thusly:

”From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

Figures 1.3 1.4 and 1.5 are three Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of trends

that would adversely affect use of the recommended methodology. Figure 1.3 demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range. Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (?).

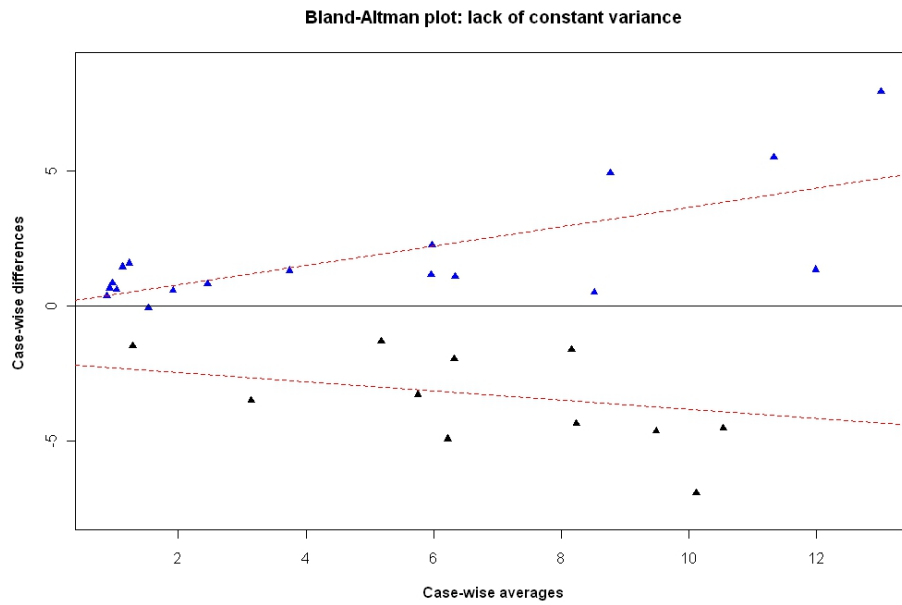


Figure 1.3: Bland-Altman Plot demonstrating the increase of variance over the range

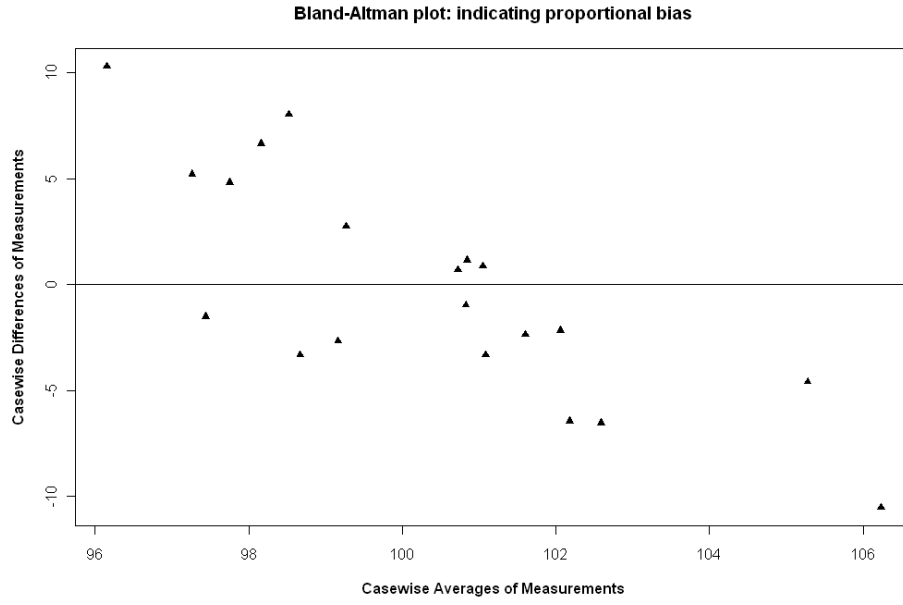


Figure 1.4: Bland-Altman Plot indicating the presence of proportional bias

Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as *proportional bias* (Ludbrook, 1997). Both of these cases violate the assumptions necessary for further analysis using limits of agreement, which shall be discussed later. The plot also can be used to identify outliers. An outlier is an observation that is numerically distant from the rest of the data. Classification thereof is a subjective decision in any analysis, but must be informed by the logic of the formulation. Figure 1.5 is a Bland Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively.

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Hence any observation, such as the one on the extreme right of figure 1.5, should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster. The one on the extreme left should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

? do not recommend excluding outliers from analyses. However recalculation of the inter-method bias estimate , and further calculations based upon that estimate, are useful for assessing the influence of outliers.(?) states that "*We usually find that this method of analysis is not too sensitive to one or two large outlying differences.*"

1.2.2 Limits of Agreement

? introduces an elaboration of the plot, adding to the plot ‘limits of agreement’ to the plot. These limits are based upon the standard deviation of the differences. The discussion shall be reverted to these limits of agreement in due course.

1.2.3 Variations of the Bland Altman Plot

? remarks that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. ? acknowledge that this is not easy to interpret, and that it is not suitable in all cases.

? offers two variations of the Bland -Altman plot that are intended to overcome potential problems that the conventional plot would inappropriate for.

The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages.

1.2.4 Agreement

Bland and Altman (1986) defined perfect agreement as the case where all of the pairs of rater data lie along the line of equality, where the line of equality is defined as the 45 degree line passing through the origin(i.e. the $X = Y$ line).

Bland and Altman (1986)expressed this in the terms *we want to know by how much the new method is likely to differ from the old; if this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably. How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparisonand to choose the sample size .*

1.2.5 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

1.2.6 Inappropriate assessment of Agreement

Paired T tests

This method can be applied to test for statistically significant deviations in bias. This method can be potentially misused for method comparison studies.

It is a poor measure of agreement when the rater's measurements are perpendicular to the line of equality[Hutson et al]. In this context, an average difference of zero between the two raters, yet the scatter plot displays strong negative correlation.

Inappropriate Methodologies

Use of the Pearson Correlation Coefficient, although seemingly intuitive, is not appropriate approach to assessing agreement of two methods. Arguments against its usage have been made repeatedly in the relevant literature. It is possible for two analytical methods to be highly correlated, yet have a poor level of agreement.

Pearson's Correlation Coefficient

It is well known that Pearson's correlation coefficient is a measure of the linear association between two variables, not the agreement between two variables (e.g., see Bland and Altman 1986). This is a well known as a measure of linear association between two variables. Nonetheless this is not necessarily the same as Agreement. This method is considered wholly inadequate to assess agreement because it only evaluates only the association of two sets of observations.

1.2.7 Inappropriate use of the Correlation Coefficient

It is intuitive when dealing with two sets of related data, i.e the results of the two raters, to calculate the correlation coefficient (r). Bland and Altman attend to this in their 1999 paper.

They present a data set from two sets of meters, and an accompanying scatterplot. An hypothesis test on the data set leads us to conclude that there is a relationship between both sets of meter measurements. The correlation coefficient is determined to be $r = 0.94$. However, this high correlation does not mean that the two methods agree. It is possible to determine from the scatterplot that the intercept is not zero, a requirement for stating both methods have high agreement. Essentially, should two methods have highly correlated results, it does not follow that they have high agreement.

1.2.8 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.

Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

1.2.9 The Bland Altman Plot

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. It has proved highly popular ever since. This is a simple, and widely used , plot of the differences of each data pair, and the corresponding average value. An important requirement is that the two measurement methods use the same scale of measurement.

scatter plots

The authors advise the use of scatter plots to identify outliers, and to determine if there is curvilinearity present. In the region of linearity ,simple linear regression may yield results of interest.

1.2.10 Effect of Outliers

Another argument against the use of model I regression is based on outliers. Outliers can adversely influence the fitting of a regression model. Cornbleet and Cochrane compare a regression model influenced by an outlier with a model for the same data set, with the outlier excluded from the data set. A demonstration of the effect of outliers

was made in Bland Altman's 1986 paper. However they discourage the exclusion of outliers.

1.2.11 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

The bias is computed as the average of the difference of paired assays.

If one method is sometimes higher, and sometimes the other method is higher, the average of the differences will be close to zero. If it is not close to zero, this indicates that the two assay methods are producing different results systematically.

Precision of Limits of Agreement

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. A different sample would give different limits of agreement. ? advance a formulation for confidence intervals of the inter-method bias and the limits of agreement. These calculations employ quantiles of the 't' distribution with $n - 1$ degrees of freedom.

1.2.12 Appropriate Use of Limits of Agreement

Importantly ? makes the following point:

These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

Carstensen attends to the issue of repeated data, using the expression replicate to express a repeated measurement on a subject by the same methods. Carstensen formulates the data as follows Repeated measurement - Arrangement of data into groups, based on the series of results of each subject.

1.2.13 The Bland Altman Plot - Variations

Variations of the Bland Altman plot is the use of ratios, in the place of differences.

$$D_i = X_i - Y_i \tag{1.1}$$

Altman and Bland suggest plotting the within subject differences $D = X_1 - X_2$ on the ordinate versus the average of x_1 and x_2 on the abscissa.

measurements

1.3 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.

Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

1.3.1 Bland Altman plots using 'Gold Standard' raters

According to Bland and Altman, one should use the methodology previous outlined, even when one of the raters is a Gold Standard.

1.3.2 Bias Detection

further to this method, the presence of constant bias may be indicated if the average value differences is not equal to zero. Bland and Altman does, however, indicate the indication of absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

1.3.3 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A

Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis

A third element of the Bland-Altman methodology, an interval known as ‘limits of agreement’ is introduced in ?, (sometimes referred to in literature as 95% limits of agreement). Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. ? refer to this as the ‘equivalence’ of two measurement methods. It must be established clearly the specific purpose of the limits of agreement. ? comment that the limits of agreement “how far apart measurements by the two methods were likely to be for most individuals”, a definition echoed in their 1999 paper:

“We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits of agreement. These values define the range within which most differences between measurements by the two methods will lie.”

The limits of agreement (LoA) are computed by the following formula:

$$LoA = \bar{d} \pm 1.96S(d) \quad (1.2)$$

with \bar{d} as the estimate of the inter method bias, $S(d)$ as the standard deviation of the differences and 1.96 is the 95% quantile for the standard normal distribution. (However, in some literature, 2 standard deviations are used instead for simplicity.) For the Grubbs ‘F vs C’ comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure 1.9 shows the resultant Bland-Altman plot, with the limits of agreement shown in dashed lines.

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As ? point out this may not be the case. Bland and

Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion.

‘How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size.’(?)

? note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as “being like a reference interval.”

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the limits, the mean and standard deviation, are not based on any sample used for an analysis, but on the process’s historical values, a key difference with Bland-Altman limits of agreement.

? regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. ? offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975, n-1)} S_d \sqrt{1 + \frac{1}{n}} \quad (1.3)$$

where n is the number of subjects. Only for 61 or more subjects is there a quantile less than 2.

? describes limits of agreement as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence.

1.3.4 Appropriate Use of Limits of Agreement

Importantly ? makes the following point:

These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

1.3.5 Problems with Limits of Agreement

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values. The formulation is also heavily influenced by outliers. An Example in ? demonstrates the effect of recalculating without a particular outlier. Referring to the VCF data set in the same paper, there is more than one outlier.

1.4 The Bland Altman Plot

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. It has proved highly popular ever since. This is a simple, and widely used, plot of the differences of each data pair, and the corresponding average value. An important requirement is that the two measurement methods use the same scale of measurement.

Variations of the Bland Altman plot is the use of ratios, in the place of differences.

$$D_i = X_i - Y_i \tag{1.4}$$

Altman and Bland suggest plotting the within subject differences $D = X_1 - X_2$ on the ordinate versus the average of x_1 and x_2 on the abscissa.

1.5 Bland-Altman Plots

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of paired sample t-test, correlation coefficients or simple linear regression. Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (?). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be ‘explained to non-statisticians’.

Notwithstanding previous remarks about regression, the first step recommended,

which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

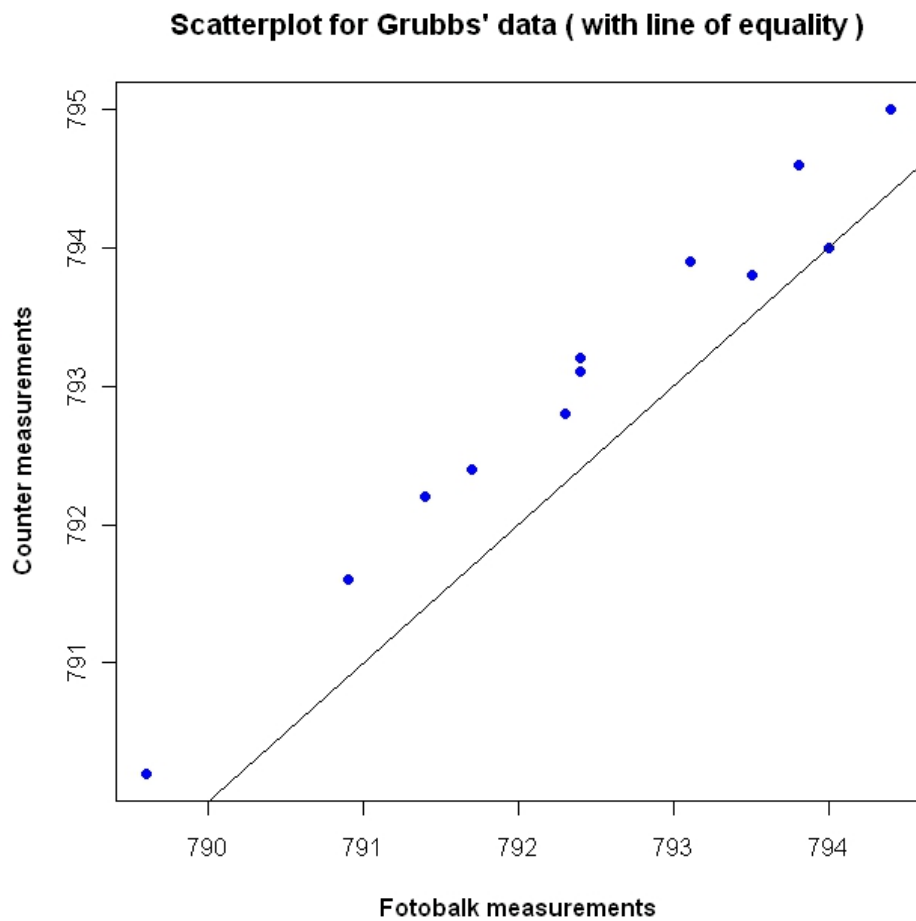


Figure 1.5: Scatter plot For Fotobalk and Counter Methods.

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated, and then the

average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \dots, n$). These differences and averages are then plotted. This methodology, now commonly known as the ‘Bland-Altman Plot’, has proved very successful. [1], which further develops the methodology, was found to be the sixth most cited paper of all time by the ISI. [2] also commented on the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals ([3]). Furthermore [4] recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

The magnitude of the inter-method bias between the two methods is simply the average of the differences \bar{d} . The variances around this bias is estimated by the standard deviation of the differences $S(d)$. This inter-method bias is represented with a line on the Bland-Altman plot. These estimates are only meaningful if there is uniform inter-bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot. In the case of Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.2: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.80	793.20	0.60	793.50
2	793.10	793.30	-0.20	793.20
3	792.40	792.60	-0.20	792.50
4	794.00	793.80	0.20	793.90
5	791.40	791.60	-0.20	791.50
6	792.40	791.60	0.80	792.00
7	791.70	791.60	0.10	791.65
8	792.30	792.40	-0.10	792.35
9	789.60	788.50	1.10	789.05
10	794.40	794.70	-0.30	794.55
11	790.90	791.30	-0.40	791.10
12	793.50	793.50	0.00	793.50

Table 1.3: Fotobalk and Terma methods: differences and averages.

1.5.1 Using Bland-Altman Plots

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. ? express the motivation for this plot thusly:

“From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

The Bland-Altman plot is simply a scatterplot of the case-wise averages and differences of two methods of measurement. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are particularly. Later it will be shown that case-wise differences are the sole component of the next part of the methodology, the limits of agreement.

For creating plots, the case wise-averages fulfil several functions, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. ? cautions that it would be the difference against either measurement value instead of their average , as the difference relates to both value.

The Bland-Altman plot for comparing the ‘Fotobalk’ and ‘Counter’ methods, which shall henceforth be referred to as the ‘F vs C’ comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

In Figure 1.3 Bland-Altman plots for the ‘F vs C’ and ‘F vs T’ comparisons are shown, where ‘F vs T’ refers to the comparison of the ‘Fotobalk’ and ‘Terma’ methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons.

By inspection, there exists a larger inter-method bias in the ‘F vs C’ comparison than in the ‘F vs T’ comparison. Conversely there appears to be less precision in ‘F vs

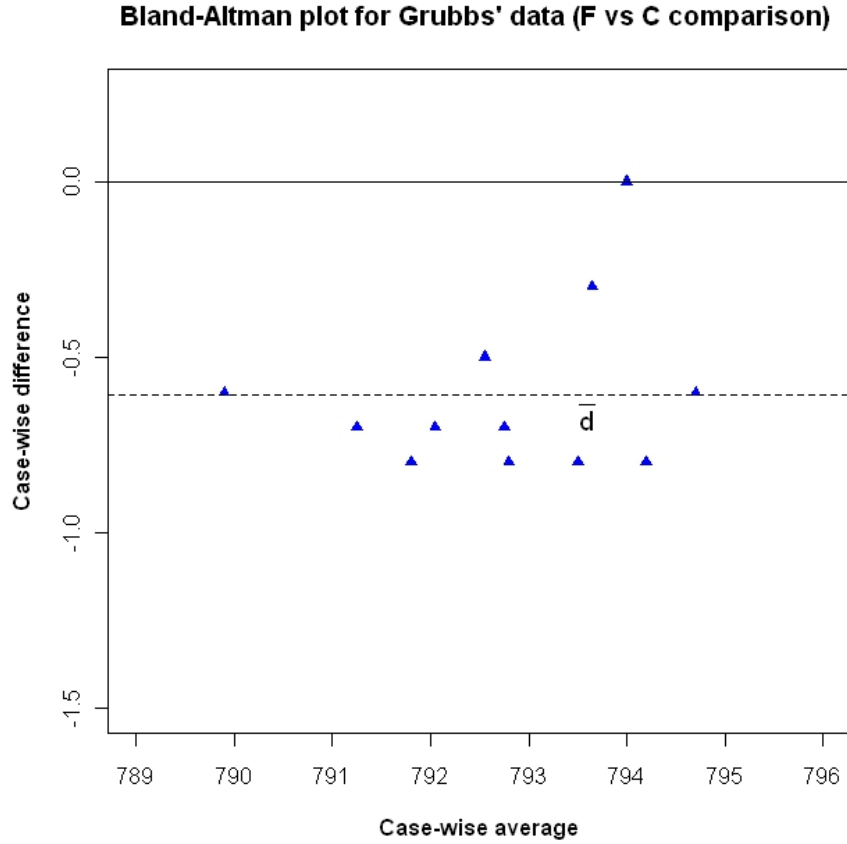


Figure 1.6: Bland-Altman plot For Fotobalk and Counter methods.

T' comparison, as indicated by the greater dispersion of co-variates.

Figures 1.4, 1.5 and 1.6 are three prototype Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

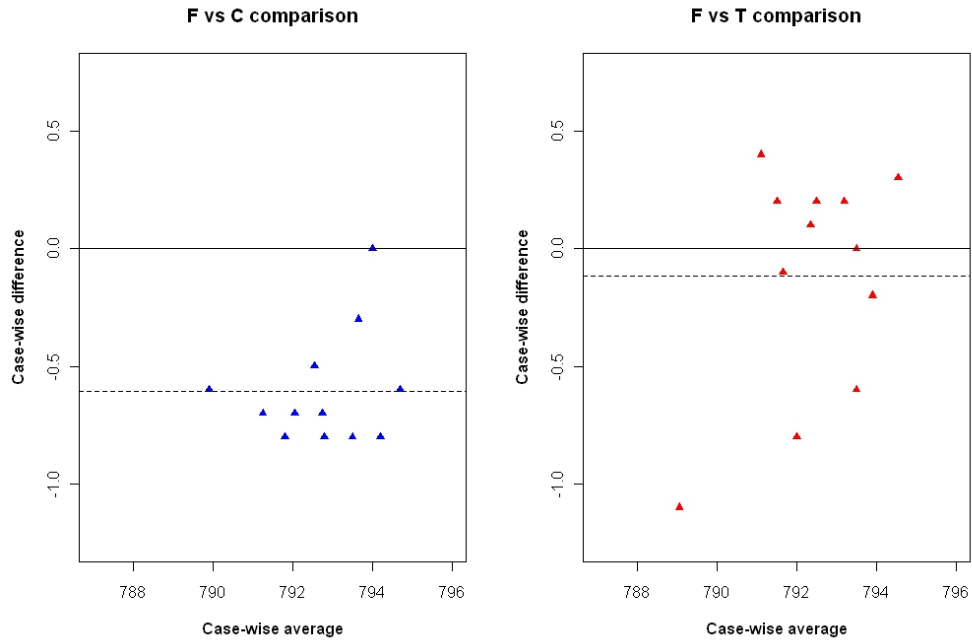


Figure 1.7: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (?) test, should be also be used.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. Classification of outliers can be determined with numerous established approaches, such as the Grubb's test, but always classification must be informed by the logic of the data's formulation. Figure 1.6 is a Bland-Altman plot

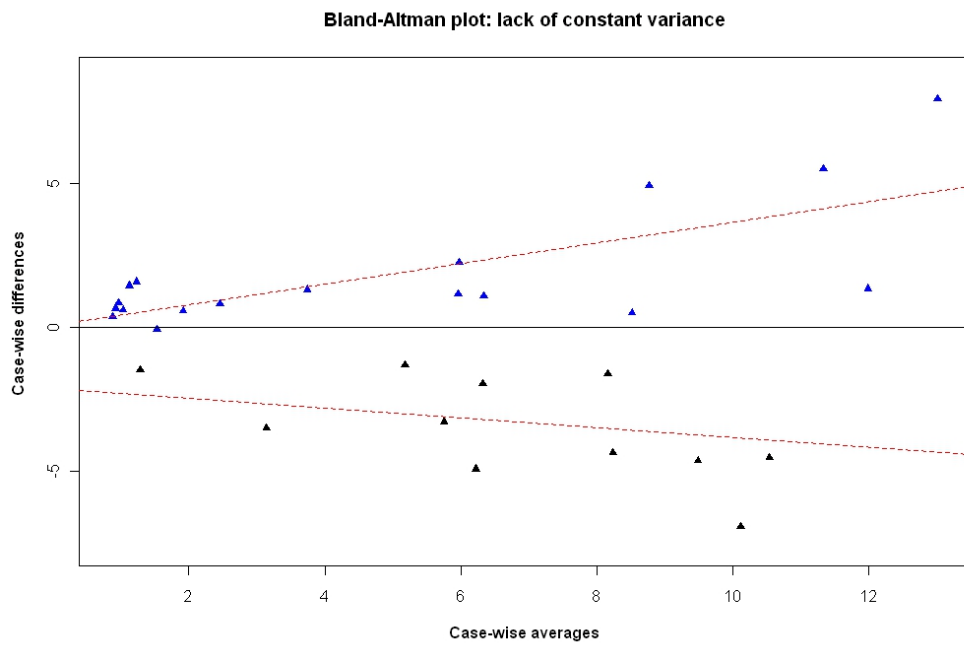


Figure 1.8: Bland-Altman plot demonstrating the increase of variance over the range.

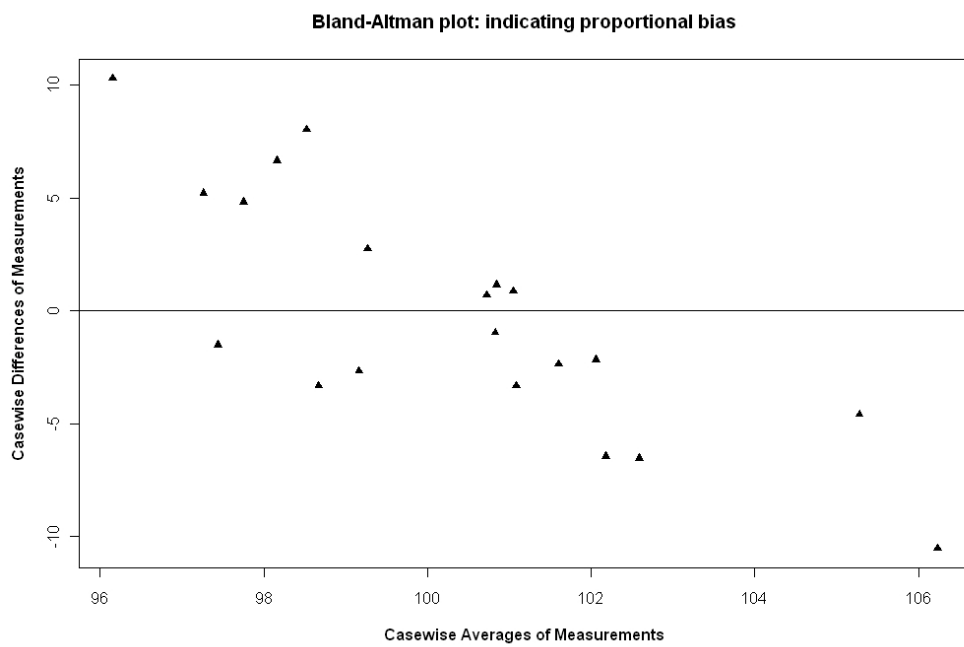


Figure 1.9: Bland-Altman plot indicating the presence of proportional bias.

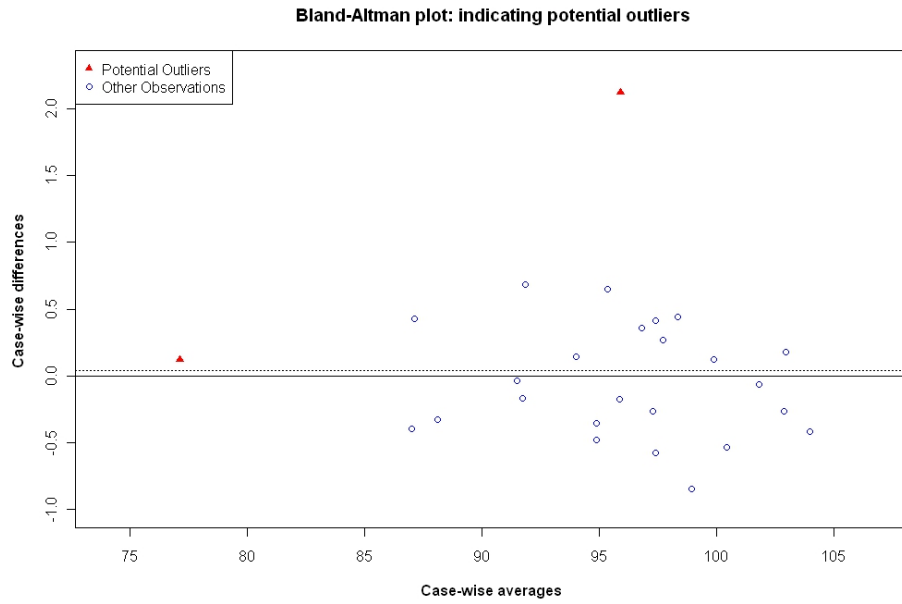


Figure 1.10: Bland-Altman plot indicating the presence of potential outliers.

with two potential outliers.

? do not recommend excluding outliers from analyses, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that ‘we usually find that this method of analysis is not too sensitive to one or two large outlying differences’.

In classifying whether a observation from a univariate data set is an outlier, Grubbs’ outlier test is widely used. In assessing whether a co-variate in a Bland-Altman plot is an outlier, this test is useful when applied to the difference values treated as a univariate data set. For Grubbs’ data, this outlier test is carried out on the differences, yielding the following results.

The null and alternative hypotheses is the absence and presence of at least one outlier respectively. Grubbs’ outlier test statistic G is the largest absolute deviation from the sample mean divided by the standard deviation of the differences. For the

‘F vs C’ comparison, $G = 3.6403$. The critical value is calculated using Student’s t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n),n-2}^2}{n-2+t_{\alpha/(2n),n-2}^2}}. \quad (1.5)$$

For this test $U = 0.7501$. The conclusion of this test is that the fourth observation in the ‘F vs C’ comparison is an outlier, with $p - value = 0.002799$.

As a complement to the Bland-Altman plot, ? proposes the use of a bivariate confidence ellipse, constructed for a predetermined level.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other. ? provides the relevant calculations for the ellipse. Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers (furthermore ? proposes formal testing procedures, that shall be discussed in due course). Inspection of Figure 1.7 shows that the fourth observation is outside the bounds of the ellipse, concurring with the conclusion that it is an outlier.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko’s ellipse. A co-variate is added to the ‘F vs C’ comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this enhanced data set. By inspection of the confidence interval, a conclusion would be reached that this extra co-variate is an outlier, in spite of the fact that this observation is consistent with the intended conclusion of the Bland-Altman plot.

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main clus-

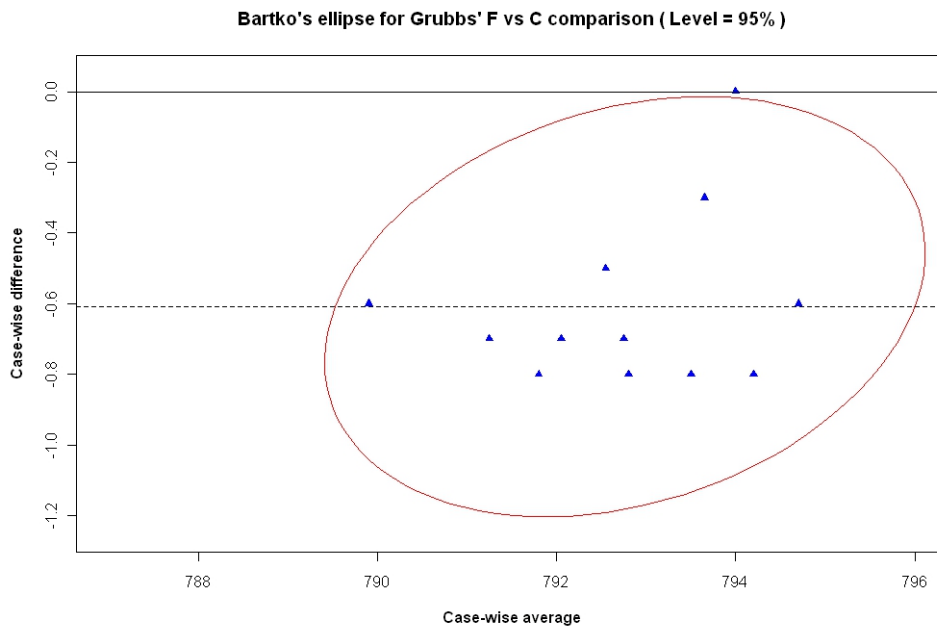


Figure 1.11: Bartko's Ellipse For Grubbs' Data.

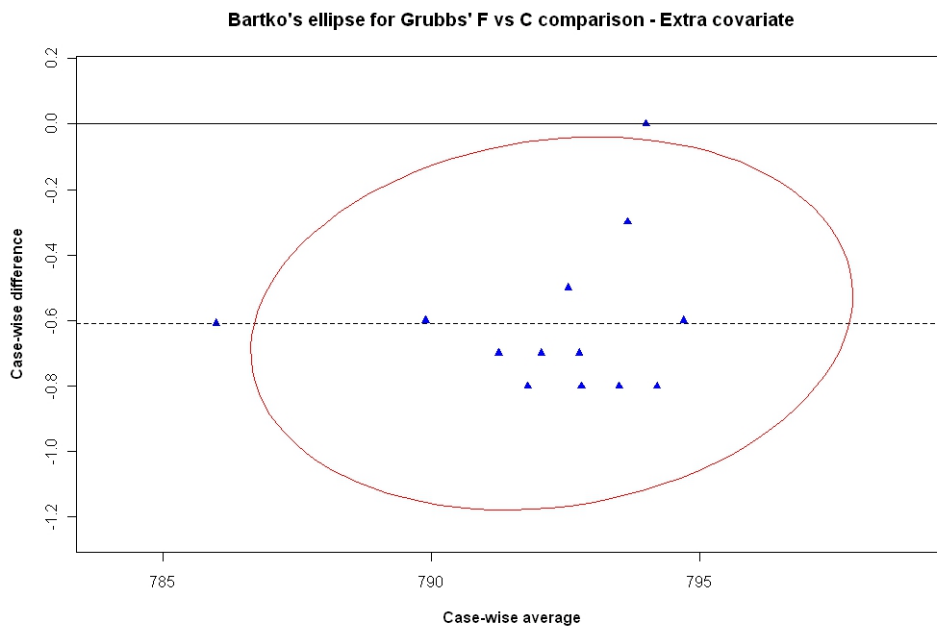


Figure 1.12: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

ter, as in the case with the extra co-variate. Conversely, the fourth observation, from the original data set, should be considered an outlier, as it has a noticeable vertical

displacement from the rest of the observations.

Bartko's ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a_i)$ is greater than $\text{var}(d_i)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a_i)$ is less than $\text{var}(d_i)$, the orientation of the ellipse is vertical.

1.5.2 Using Bland-Altman Plots

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. ? express the motivation for this plot thusly:

“From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

The Bland-Altman plot is simply a scatterplot of the case-wise averages and differences of two methods of measurement. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are particularly. Later it will be shown that case-wise differences are the sole component of the next part of the methodology, the limits of agreement.

For creating plots, the case wise-averages fulfil several functions, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. ? cautions that it would be the difference against either measurement value instead of their average , as the difference relates to both value.

The Bland-Altman plot for comparing the ‘Fotobalk’ and ‘Counter’ methods, which shall henceforth be referred to as the ‘F vs C’ comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

In Figure 1.3 Bland-Altman plots for the ‘F vs C’ and ‘F vs T’ comparisons are shown, where ‘F vs T’ refers to the comparison of the ‘Fotobalk’ and ‘Terma’ methods.

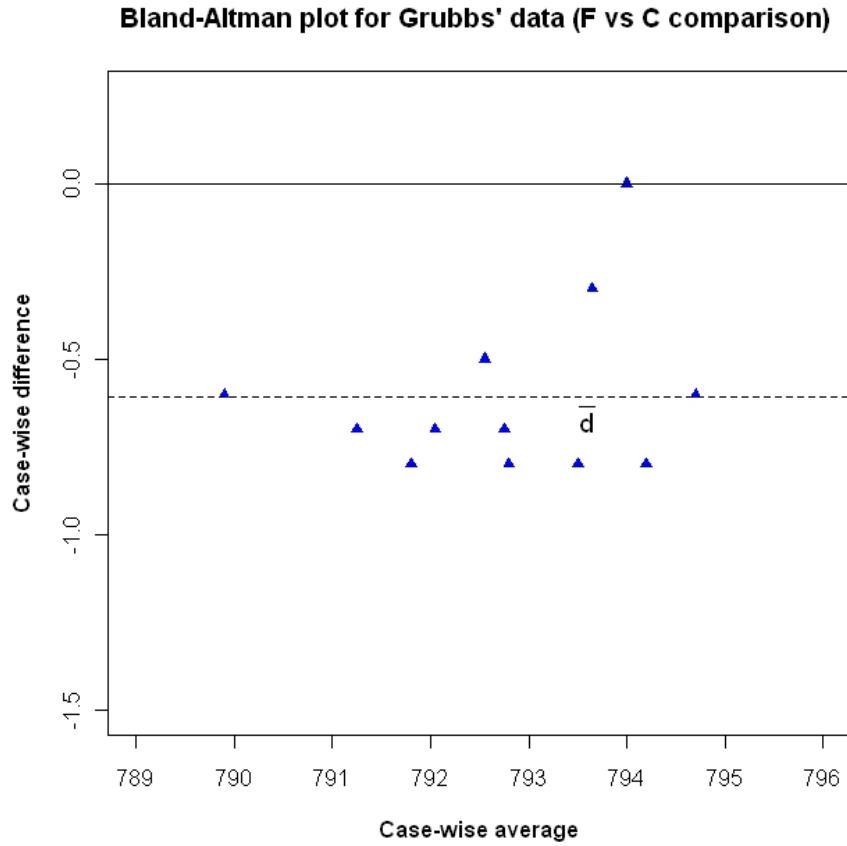


Figure 1.13: Bland-Altman plot For Fotobalk and Counter methods.

Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons.

By inspection, there exists a larger inter-method bias in the ‘F vs C’ comparison than in the ‘F vs T’ comparison. Conversely there appears to be less precision in ‘F vs T’ comparison, as indicated by the greater dispersion of co-variates.

Figures 1.4, 1.5 and 1.6 are three prototype Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the

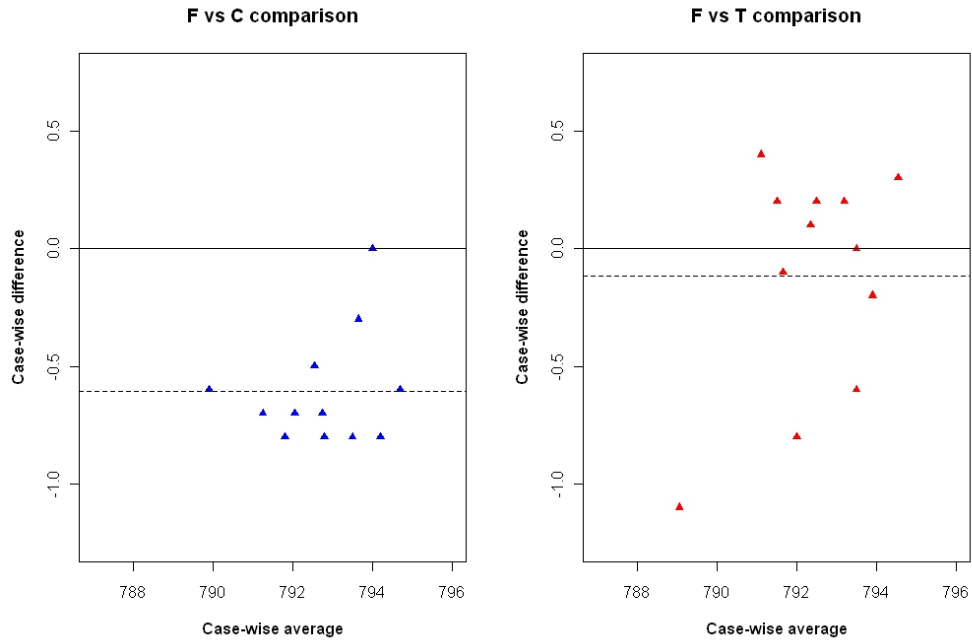


Figure 1.14: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (?) test, should be also be used.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses

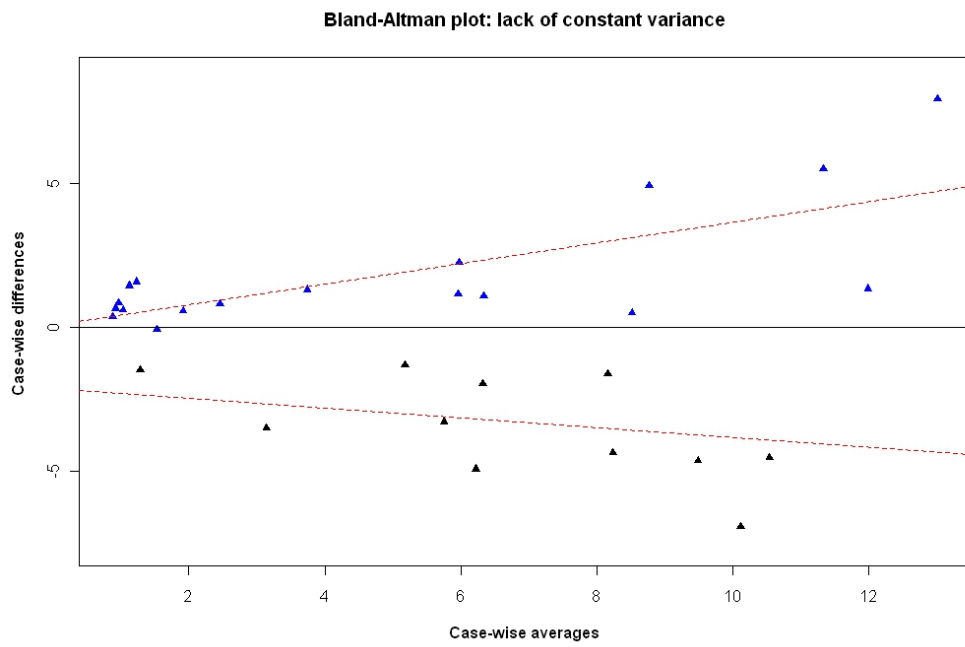


Figure 1.15: Bland-Altman plot demonstrating the increase of variance over the range.

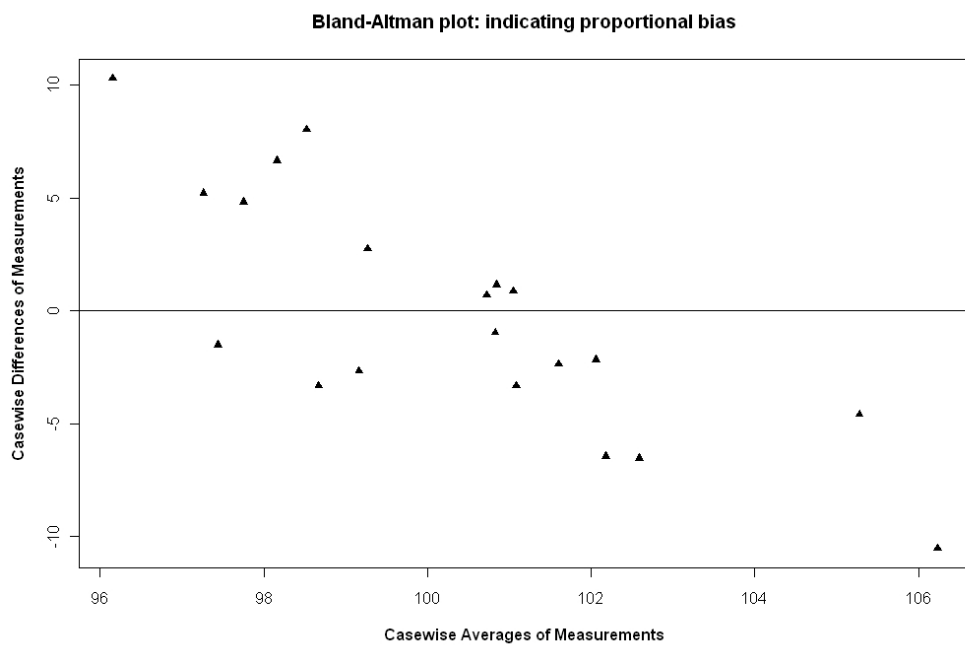


Figure 1.16: Bland-Altman plot indicating the presence of proportional bias.

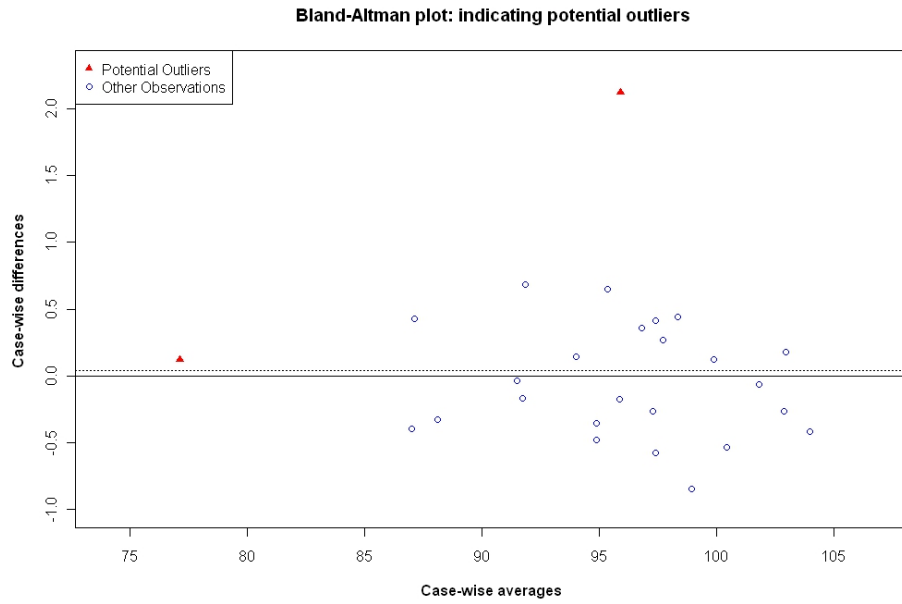


Figure 1.17: Bland-Altman plot indicating the presence of potential outliers.

suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. Classification of outliers can be determined with numerous established approaches, such as the Grubb’s test, but always classification must be informed by the logic of the data’s formulation. Figure 1.6 is a Bland-Altman plot with two potential outliers.

? do not recommend excluding outliers from analyses, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that ‘we usually find that this method of analysis is not too sensitive to one or two large outlying differences’.

In classifying whether a observation from a univariate data set is an outlier, Grubbs’ outlier test is widely used. In assessing whether a co-variate in a Bland-Altman plot is an outlier, this test is useful when applied to the difference values treated as a univariate data set. For Grubbs’ data, this outlier test is carried out on the differences, yielding

the following results.

The null and alternative hypotheses is the absence and presence of at least one outlier respectively. Grubbs' outlier test statistic G is the largest absolute deviation from the sample mean divided by the standard deviation of the differences. For the 'F vs C' comparison, $G = 3.6403$. The critical value is calculated using Student's t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_{\alpha/(2n), n-2}^2}{n-2 + t_{\alpha/(2n), n-2}^2}}. \quad (1.6)$$

For this test $U = 0.7501$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with $p - value = 0.002799$.

As a complement to the Bland-Altman plot, ? proposes the use of a bivariate confidence ellipse, constructed for a predetermined level.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other. ? provides the relevant calculations for the ellipse. Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers (furthermore ? proposes formal testing procedures, that shall be discussed in due course). Inspection of Figure 1.7 shows that the fourth observation is outside the bounds of the ellipse, concurring with the conclusion that it is an outlier.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko's ellipse. A co-variate is added to the 'F vs C' comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this enhanced data set. By inspection of the confidence interval, a conclusion would be reached that this extra co-variate is an outlier, in spite of the fact that this observation is consistent with

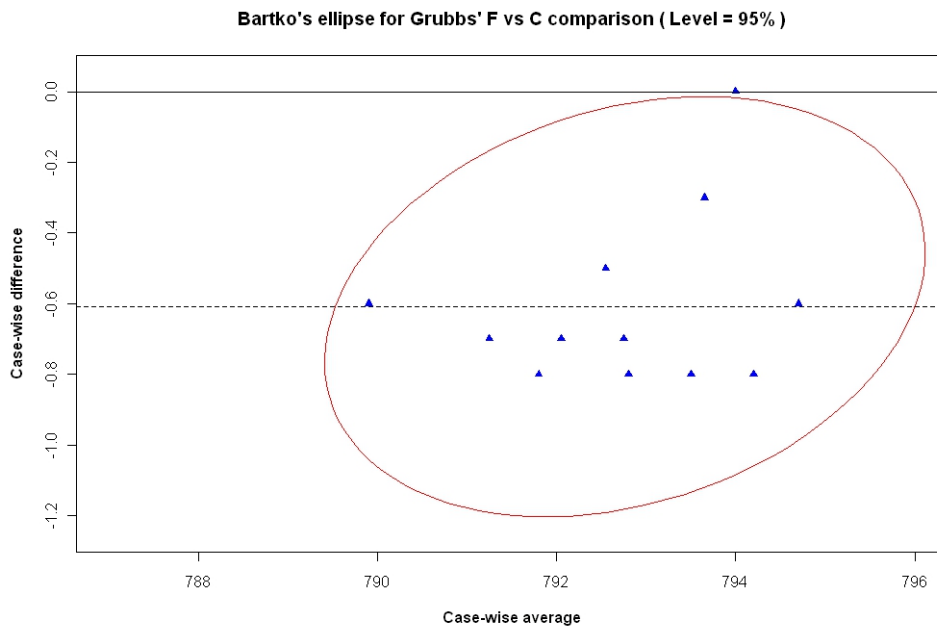


Figure 1.18: Bartko's Ellipse For Grubbs' Data.

the intended conclusion of the Bland-Altman plot.

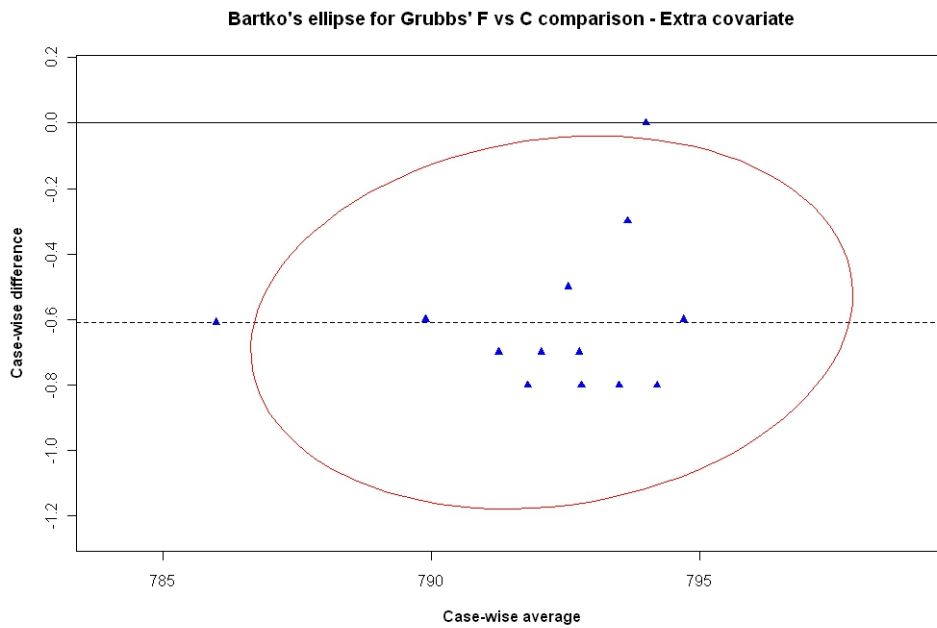


Figure 1.19: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

In the Bland-Altman plot, the horizontal displacement of any observation is sup-

ported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra co-variate. Conversely, the fourth observation, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

Bartko's ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a_i)$ is greater than $\text{var}(d_i)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a_i)$ is less than $\text{var}(d_i)$, the orientation of the ellipse is vertical.

1.5.3 Variations of the Bland-Altman Plot

Referring to the assumption that bias and variability are constant across the range of measurements, ? address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman acknowledge that this is not easy to interpret, and may not be suitable in all cases.

? offers two variations of the Bland-Altman plot that are intended to overcome potential problems that the conventional plot would inappropriate for. The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages. This will remove the need for log transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. ? proposed such a ratio plot, independently of Bland and Altman. ? commented on the reception of this article by saying ‘Strange to say, this report has been overlooked’.

1.5.4 Variations of the Bland-Altman Plot

Referring to the assumption that bias and variability are constant across the range of measurements, ? address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magni-

tude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman acknowledge that this is not easy to interpret, and may not be suitable in all cases.

? offers two variations of the Bland-Altman plot that are intended to overcome potential problems that the conventional plot would be inappropriate for. The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages. This will remove the need for log transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. ? proposed such a ratio plot, independently of Bland and Altman. ? commented on the reception of this article by saying ‘Strange to say, this report has been overlooked’.

1.5.5 Regression-based Limits of Agreement

Assuming that there will be no curvature in the scatter-plot, the methodology regresses the difference of methods (d) on the average of those methods (a) with a simple intercept slope model; $\hat{d} = b_0 + b_1 a$. Should the slope b_1 be found to be negligible, \hat{d} takes the value \bar{d} .

The next step to take in calculating the limits is also a regression, this time of the residuals as a function of the scale of the measurements, expressed by the averages a_i ;

$$\hat{R} = c_0 + c_1 a_i$$

With reference to absolute values following a half-normal distribution with mean

$\sigma\sqrt{\frac{2}{\pi}}$, ? formulate the regression based limits of agreement as follows

$$\hat{d} \pm 1.96\sqrt{\frac{\pi}{2}}\hat{R} = \hat{d} \pm 2.46\hat{R} \quad (1.7)$$

1.5.6 Regression-based Limits of Agreement

Assuming that there will be no curvature in the scatter-plot, the methodology regresses the difference of methods (d) on the average of those methods (a) with a simple intercept slope model; $\hat{d} = b_0 + b_1a$. Should the slope b_1 be found to be negligible, \hat{d} takes the value \bar{d} .

The next step to take in calculating the limits is also a regression, this time of the residuals as a function of the scale of the measurements, expressed by the averages a_i ;
 $\hat{R} = c_0 + c_1a_i$

With reference to absolute values following a half-normal distribution with mean $\sigma\sqrt{\frac{2}{\pi}}$, ? formulate the regression based limits of agreement as follows

$$\hat{d} \pm 1.96\sqrt{\frac{\pi}{2}}\hat{R} = \hat{d} \pm 2.46\hat{R} \quad (1.8)$$

1.5.7 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as ‘replicate measurements’. ? recommends the use of replicate measurements, but acknowledges that additional computational complexity.

? address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as

a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. ? propose a correction for this.

? takes issue with the limits of agreement based on mean values, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation. ? demonstrates how the limits of agreement calculated using the mean of replicates are ‘much too narrow as prediction limits for differences between future single measurements’. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the ‘mean of replicates’ approach.

The approach proposed by ? is a formal test on the Pearson correlation coefficient of case-wise differences and means (ρ_{ad}). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. For the Grubbs data, the correlation coefficient estimate (r_{ad}) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers ‘r to z’ transformation (?). The null hypothesis ($\rho_{ad}=0$) would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has been no further mention of this particular test in ?, although ? refers to Spearman’s rank correlation coefficient. ? comments ‘we do not see a place for methods of analysis based on hypothesis testing’. ? also states that consider structural equation models to be inappropriate.

? highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be

estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example α may take the value of the inter-method bias estimate from Bland-Altman methodology. Another assumption is that the precision ratio $\lambda = \frac{\sigma_s^2}{\sigma_b^2}$ may be known.

? considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

? recommends the following approach for analyzing method comparison data. Firstly he recommends conventional Bland-Altman methodology; plotting the scatter-plot and the Bland-Altman plot, complemented by estimate for the limits of agreement and the correlation coefficient between the difference and the mean. Additionally box-plots may be useful in considering the marginal distributions of the observations. The second step is the calculations of summary statistics; the means and variances of each set of measurements, and the covariances.

When both methods measure in the same scale (i.e. $\beta = 1$), ? recommends the use of Grubbs estimators to estimate error variances, and to test for their equality. A test of whether the intercept α may be also be appropriate.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.4: Fotobalk and Counter Methods: Differences and Averages

1.5.8 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as ‘replicate measurements’. ? recommends the use of replicate measurements, but acknowledges that additional computational complexity.

? address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as

a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. ? propose a correction for this.

? takes issue with the limits of agreement based on mean values, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation. ? demonstrates how the limits of agreement calculated using the mean of replicates are ‘much too narrow as prediction limits for differences between future single measurements’. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the ‘mean of replicates’ approach.

The approach proposed by ? is a formal test on the Pearson correlation coefficient of case-wise differences and means (ρ_{ad}). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. For the Grubbs data, the correlation coefficient estimate (r_{ad}) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers ‘r to z’ transformation (?). The null hypothesis ($\rho_{ad} = 0$) would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has been no further mention of this particular test in ?, although ? refers to Spearman’s rank correlation coefficient. ? comments ‘we do not see a place for methods of analysis based on hypothesis testing’. ? also states that consider structural equation models to be inappropriate.

? highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be

estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example α may take the value of the inter-method bias estimate from Bland-Altman methodology. Another assumption is that the precision ratio $\lambda = \frac{\sigma_s^2}{\sigma_b^2}$ may be known.

? considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

? recommends the following approach for analyzing method comparison data. Firstly he recommends conventional Bland-Altman methodology; plotting the scatter-plot and the Bland-Altman plot, complemented by estimate for the limits of agreement and the correlation coefficient between the difference and the mean. Additionally box-plots may be useful in considering the marginal distributions of the observations. The second step is the calculations of summary statistics; the means and variances of each set of measurements, and the covariances.

When both methods measure in the same scale (i.e. $\beta = 1$), ? recommends the use of Grubbs estimators to estimate error variances, and to test for their equality. A test of whether the intercept α may be also be appropriate.

1.5.9 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects, Bland Altman suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the

effect of repeated measurement error. Bland Altman propose a correction for this.

Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

1.6 Bland Altman Plots

The issue of whether two measurement methods are comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of matched pairs correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (?).

As an alternative they proposed a simple statistical methodology specifically appropriate for method comparison studies. They acknowledge that there are other valid methodologies, but argue that a simple approach is preferable to complex approaches, *”especially when the results must be explained to non-statisticians”* (?).

The first step recommended which the authors argue should be mandatory is construction of a simple scatter plot of the data. The line of equality ($X = Y$) should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1).

The averages of the two measurements is considered by Bland and Altman to the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2). ?express the motivation for this plot thusly:

”From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, ..n$). These differences and averages are then plotted. This methodology, now commonly known as the ‘Bland-Altman Plot’, has proved very successful. ?, which further develops the methodology, was found to be the sixth most cited paper of all time by the ?. ? also commented on the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (?). Furthermore ? recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

The magnitude of the inter-method bias between the two methods is simply the average of the differences \bar{d} . The variances around this bias is estimated by the standard deviation of the differences $S(d)$. This inter-method bias is represented with a line on the Bland-Altman plot. These estimates are only meaningful if there is uniform inter-

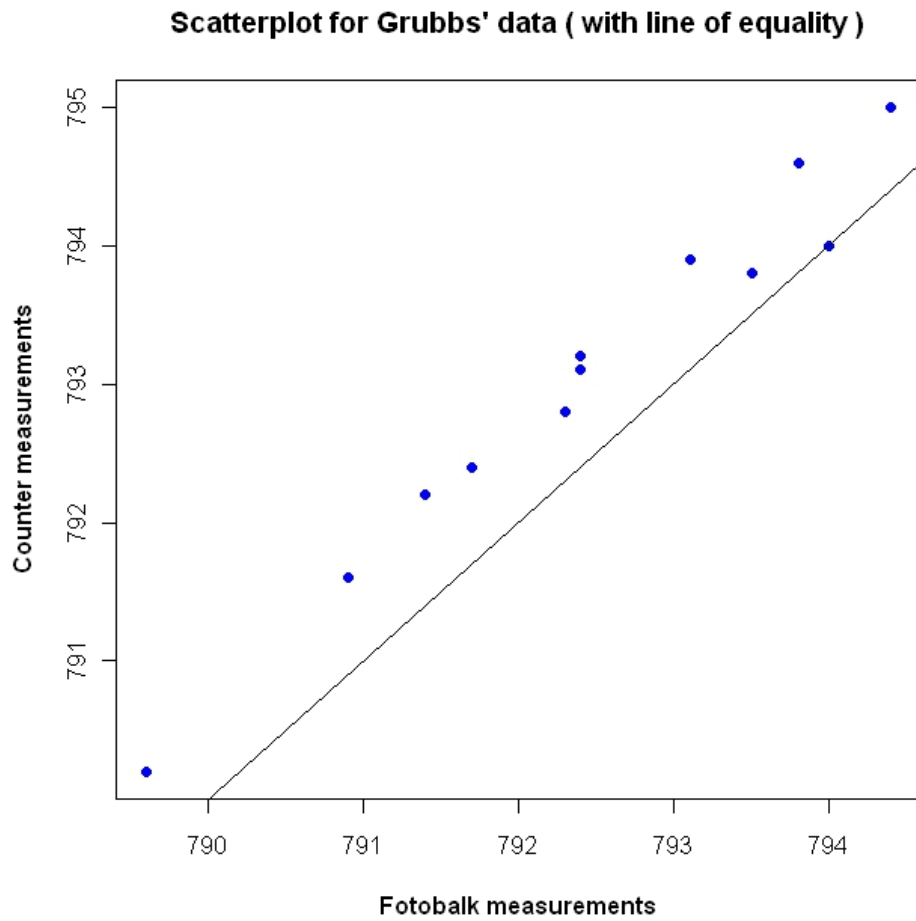


Figure 1.20: Scatter plot For Fotobalk and Counter Methods.

bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot. In the case of Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.5: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.80	793.20	0.60	793.50
2	793.10	793.30	-0.20	793.20
3	792.40	792.60	-0.20	792.50
4	794.00	793.80	0.20	793.90
5	791.40	791.60	-0.20	791.50
6	792.40	791.60	0.80	792.00
7	791.70	791.60	0.10	791.65
8	792.30	792.40	-0.10	792.35
9	789.60	788.50	1.10	789.05
10	794.40	794.70	-0.30	794.55
11	790.90	791.30	-0.40	791.10
12	793.50	793.50	0.00	793.50

Table 1.6: Fotobalk and Terma methods: differences and averages.

1.6.1 Introduction to Limits of Agreement

- Comparing two methods of measurement is normally done by computing limits of agreement (LoA), i.e. prediction limits for a future difference between mea-

surements with the two methods. When the difference is not constant it is not clear what this means, since the difference between the methods depends on the average; hence, unlike the case where the difference is constant, LoA cannot directly be translated into a prediction interval for a measurement by one method given that of another.

- The main point in the paper by Bland and Altman [1] is however different from the outlook in this paper; Bland and Altman mainly discuss whether two methods of measurement can be used interchangeably and how to assess this with the help of proper statistical methods to derive LoA, i.e. prediction limits for differences between two methods. This paper takes as starting point that the classical LoA can be converted to a prediction interval for one method given a measurement by the other (details in the next section). This sort of relationship can be shown in a plot as a line with slope 1 and prediction limits as lines also with slope 1; applicable for the prediction both from method 1 to method 2 and vice versa. In the case of non-constant difference it would be desirable to be able to produce a similar plot, usable both ways. Thus, the aim of this paper is to produce a conversion from one method to another that also applies in the case where the difference between methods is not constant.
- In this paper, I set up a proper model for data for method comparison studies which in the case of constant difference between methods leads to the classical LoA, and in the case of linear bias gives a simple formula for the prediction. The paper only addresses the situation where only one measurement by each method is available, although replicate measurements by each method are desirable whenever possible [2]. Moreover, the situation with non-constant variance over the range of measurements is not covered either.

1.6.2 Discussion

I have here proposed a simple twist to the results from regression of the differences on the sums in the case of a linear relationship between two methods of measurement. It is consistent with the obvious underlying model, and exploits the fact that although the parameters of the model cannot be estimated, those functions of the parameters that are needed for creating predictions can be estimated. The prediction limits provided have the attractive property that if the prediction line with limits is drawn in a coordinate system, the chart will apply in both ways; hence, both the line and the limits are symmetric. Precisely as the prediction intervals derived from the classical LoA are in the case where the difference between methods is constant. The drawback is that the regression of the differences on the means ignores that the averages are correlated with the residuals (i.e. the error terms), and therefore gives biased estimates if the slope linking the two methods is far from 1 or the residual variances are very different. However, both of these are rather uncommon in method comparison studies, so the method proposed here is widely applicable. When considering LoA, the only feasible transformation is the log-transform, which gives LoA for the ratio of measurements, which is immediately understandable. If, for example, the measurements are fractions where some are close to either 0 or 1 a logit transform may be adequate.

LoA would then be for (log) odds-ratios, not very easily understood. For other more arbitrarily chosen transformation the situation may be even worse. But if a plot with conversion lines and limits are constructed, then the plot is readily back-transformed to the original scale for practical use.

1.6.3 Distribution of Maxima

It is possible to use Order Statistics theory to assess conditional probabilities. With two random variables T_0 and T_1 , we define two variables Z and W such that they take the maximum and minimum values of the pair of T values.

1.6.4 Plot of the Maxima against the Minima

In Figure 1, The Maximas are plotted against their corresponding minima. The Critical values of the Maxima and Minima are displayed in the dotted lines. The Line of Equality depicts the obvious logical constraint of the each Maximum value being greater than its corresponding minimum value.

The scientific question at hand is the correct approach to assessing whether two methods can be used interchangeably. ? expresses this as follows:

We want to know by how much (one) method is likely to differ from the (other), so that if it not enough to cause problems in the mathematical interpretation we can ... use the two interchangeably.

Consequently, of the categories of method comparison study, comparison studies, the second category, is of particular importance, and the following discussion shall concentrate upon it. Less emphasis shall be place on the other three categories.

Further to ?, 'equivalence' of two methods expresses that both can be used interchangeably. ?, p.49 remarks that this is a very restrictive interpretation of equivalence, and that while agreement indicated equivalence, equivalence does not necessarily reflect agreement.

The main difference between Myers proposed method and the Bland Altman is that the random effects model is used to estimate the within-subject variance after adjusting

for known and unknown variables. The Bland Altman approach uses one way analysis of variance to estimate the within subject variance. In general, the random effects model is an extension of the analysis of the ANOVA method and it can adjust for many more covariates than the ANOVA method

1.7 Conclusions about Existing Methodologies

Scatterplots are recommended by ? for an initial examination of the data, facilitating an initial judgement and helping to identify potential outliers. They are not useful for a thorough examination of the data. ? notes that data points will tend to cluster around the line of equality, obscuring interpretation.

The Bland Altman methodology is well noted for its ease of use, and can be easily implemented with most software packages. Also it doesn't require the practitioner to have more than basic statistical training. The plot is quite informative about the variability of the differences over the range of measurements. For example, an inspection of the plot will indicate the 'fan effect'. They also can be used to detect the presence of an outlier.

??criticizes these plots on the basis that they presents no information on effect of constant bias or proportional bias. These plots are only practicable when both methods measure in the same units. Hence they are totally unsuitable for conversion problems. The limits of agreement are somewhat arbitrarily constructed. They may or may not be suitable for the data in question. It has been found that the limits given are too wide to be acceptable. There is no guidance on how to deal with outliers. Bland and Altman recognize effect they would have on the limits of agreement, but offer no guidance on how to correct for those effects.

There is no formal testing procedure provided. Rather, it is upon the practitioner opinion to judge the outcome of the methodology.

1.8 Treatment of Outliers

Bland and Altman attend to the issue of outliers in their 1986 paper, wherein they present a data set with an extreme outlier

1.9 Bland Altman Plots In Literature

? contains a study the use of Bland Altman plots of 44 articles in several named journals over a two year period. 42 articles used Bland Altman's limits of agreement, wit the other two used correlation and regression analyses. ? remarks that 3 papers, from 42 mention predefined maximum width for limits of agreement which would not impair medical care.

The conclusion of ? is that there are several inadequacies and inconsistencies in the reporting of results ,and that more standardization in the use of Bland Altman plots is required. The authors recommend the prior determination of limits of agreement before the study is carried out. This contention is endorsed by ?, which makes a similar recommendation for the sample size, noting that *sample sizes required either was not mentioned or no rationale for its choice was given.*

In order to avoid the appearance of "data dredging", both the sample size and the (limits of agreement) should be specified and justified before the actual conduct of the trial. (?)

? remarks that the limits of agreement should be compared to a clinically acceptable difference in measurements.

1.9.1 Gold Standard

This is considered to be the most accurate measurement of a particular parameter.