Chapter 1

Review of MCS Methodologies

1.1 Bland-Altman methodology

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.

Dewitte et al. (2002) 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.

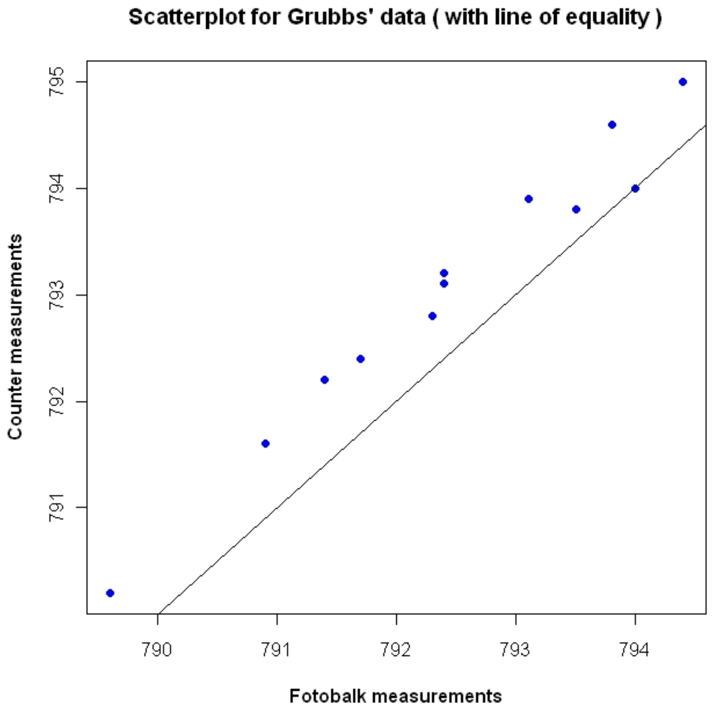

Figure 1.1: Scatter plot For Fotobalk and Counter Methods.

1.1.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, \ldots, n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \ldots, n$).

Altman and Bland (1983) 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. Altman and Bland (1983) 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. Bland and Altman (1986) 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.

1.1.2 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 1.1: 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 1.2: Fotobalk and Terma methods: differences and averages.

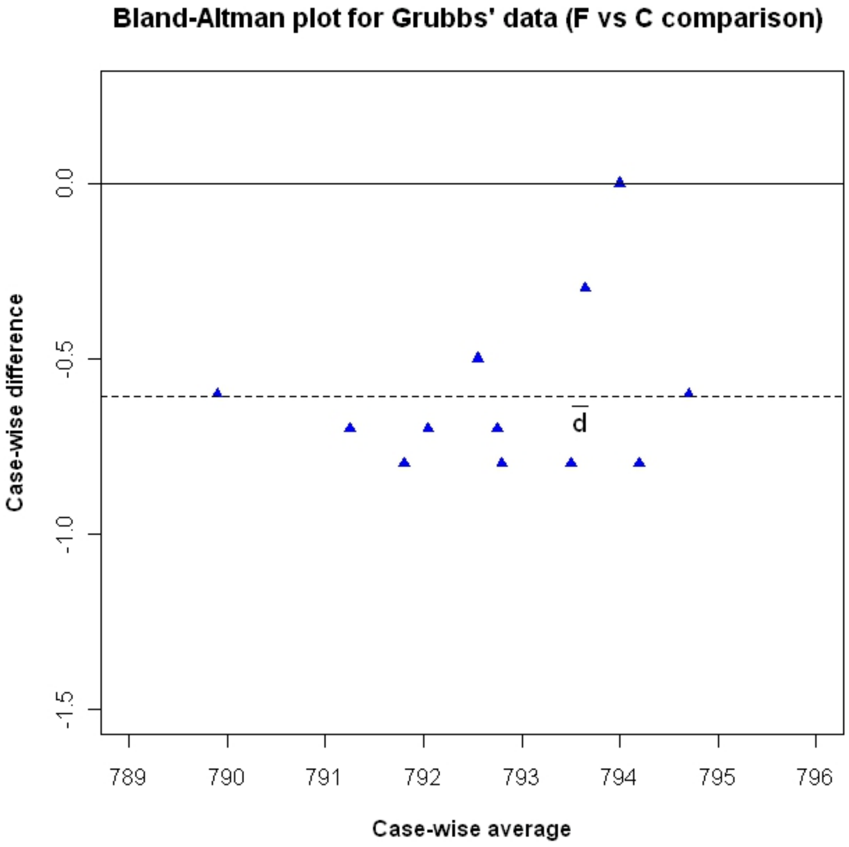

Figure 1.2: Bland-Altman plot For Fotobalk and Counter methods.

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.

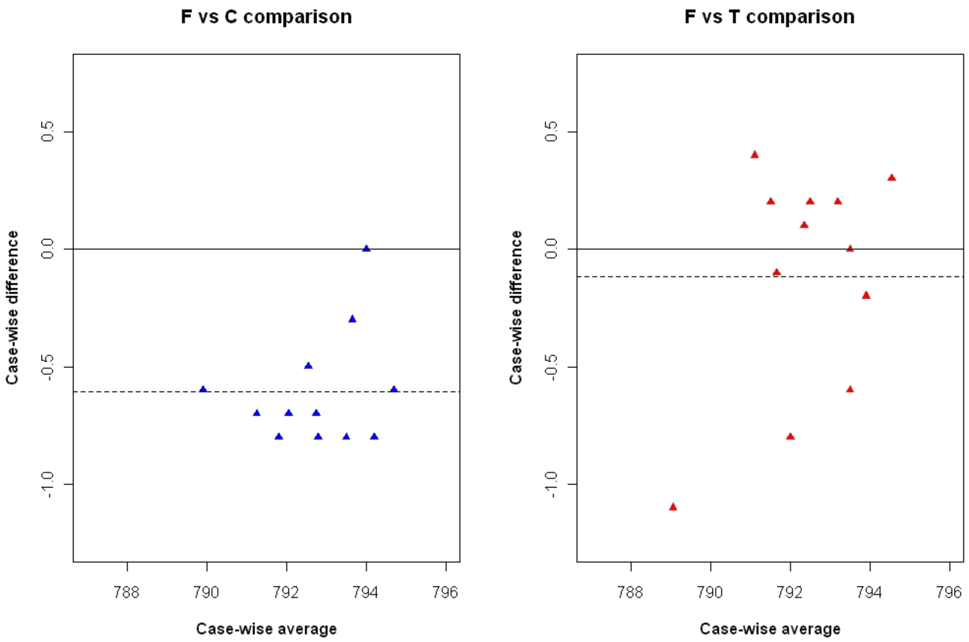

Figure 1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

1.1.3 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 Ludbrook (1997) 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 (Benjamini and Hochberg, 1995) test, should be also be used.

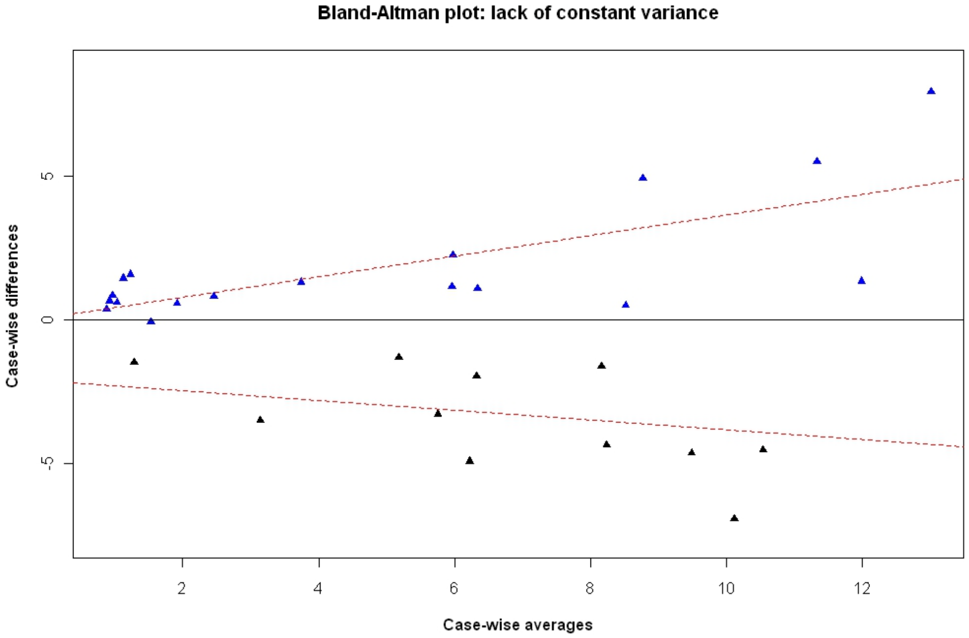

Figure 1.4: Bland-Altman plot demonstrating the increase of variance over the range.

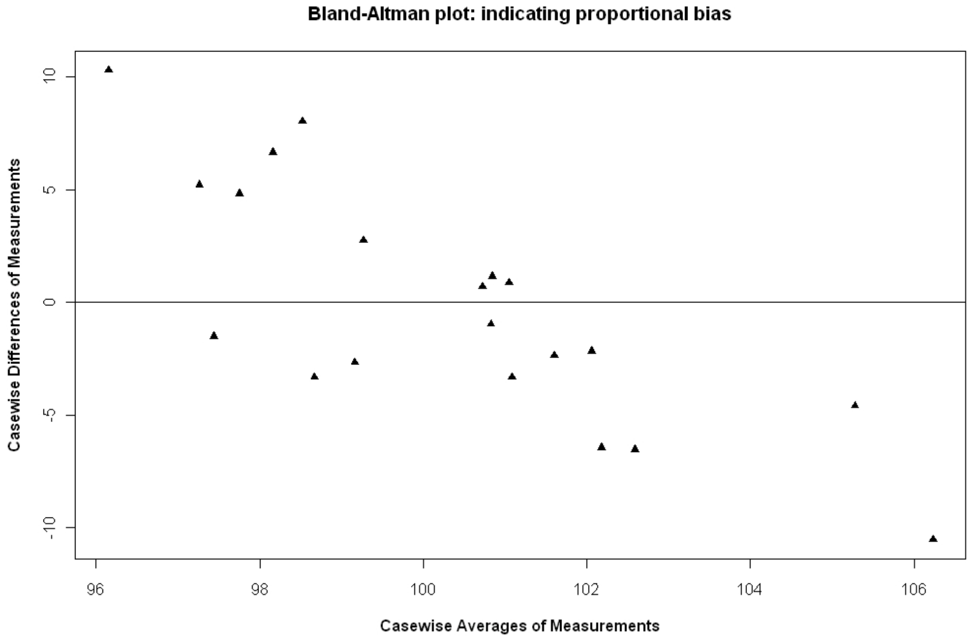

Figure 1.5: Bland-Altman plot indicating the presence of proportional bias.

1.1.4 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'. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity.

Bland and Altman (1986) 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. Bland and Altman (1986) propose a correction for this.

Carstensen et al. (2008) 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. Carstensen et al. (2008) 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.

1.1.5 Sampling Protocols

Dunn discusses the sampling protocols in depth. Consider a random sample of N specimens. A simple design is a set of measurements on each specimen using each of the two methods, yield 2N measurments. Dunn remarks that such a design would not yield much in the way of information. The criticism projected at the correlation coefficient is only valid if one is specifically interested in assessing agreement. However, it should be used as an exploratory tool in the first instance. Exchangealility encompasses the qualities of similar precision.

1.1.6 MCS Research Notes

The problem of comparing two methods of measurement is ubiquitous in scientific literature. The use of well-established methodologies, such as the paired t-test, correlation and regression approaches is criticised in Altman and Bland(1983). In the Bland-

Altman papers, the British Standards Institute emerge as the key authority on the definition of the Limits of agreement. It is assumed that, in the absence of a specified probability, that the level is 95%.

Bland and Altman proposed a simple graphical technique, plotting the case-wise differences against the case-wise means of the respective measurements. The benefit of such an approach is the plot makes it easier to assess the magnitude of the disagreement (both error and bias), spot outliers, and see whether there is any trend.

1.1.7 Success of Bland-Altmans plot

The success of the Bland-Altman approach is perhaps due to the fact that only a visual inspection of the plot is required. Bland and Altmans paper was later reported to be the sixth most widely cited statistical paper ever (Hollis 1996, for example). Hollis, S (1996), Annals of clinical biochemistry (Annals of Biochemistry 33,1-4) Ryan, T and Woodall W (2005). The most cited statistical papers Journal of applied Statistics 32(5), 461-474. Bland and Altman emphasis the clinical importance of the range of between the limits of agreement, and use this range as a basis for evaluating agreement. The question arises as to whether or not it is statistically valid to arrive at a decision about the population probability from an observed coverage range in a sample.

Altman and Bland (1983) show that their graphical approach can be supplemented by a test of significance on the Pearson product correlation coefficient of the plotted quantities. This test is equivalent to the test of the hypothesis that the method variances are equal (Pitman 1939) Bland and Altman recommend a test of significance of Spearmans rank correlation coefficient of the absolute differences and the case-wise means. Hayes et al (2006) examines the pitfalls that arises when an outlier is assesses using an informal criterion based on a fixed number of standard deviations rather than a more formal standard approach.

1.1.8 Underlying Model

The model underlying the Bland-Altman approach can be expressed as an LME model with heterogeneous variances.

$$y_{ij} = \beta_j + b_i + \varepsilon_{ij}$$

The case-wise differences and case-wise means follow a bivariate normal distribution, with expected values and variances specified as [input equations].

1.1.9 Outlier detection

Additionally, there is no clear guidance in any of the Bland-Altman papers on the treatment of outliers that may arise in a plot. An example used in Bland-Altman 1986 identifies a clear outlier, where it is advised by the authors that in practice, one could omit this subject. Bland and Altman 1999 recommend the computationally intensive approach of calculating the limits of agreement with, and then without, suspected outliers, in order to assess the impact on the results. However, they are clear that they do not recommend excluding outliers from analyses.

1.2 Westgard et Al

Westgard et al. (1)(2)(3) outlined the basic principles for method comparison in a clear, easy to follow manual. They also introduced the concept of allowable analytical error and gave an overview of published performance criteria. They recommended that the estimated analytical imprecision and bias be compared with these performance criteria in method evaluation as well as in method comparison. Their approach made use of a scatter-plot and calculations based on regression lines, but with confidence limits and judgment of acceptability based on the criteria for allowable analytical error.

These principles of comparing analytical performance with performance criteria, however, have not been universally accepted, and recent publications have criticized the misuse of correlation coefficients (4) and overinterpretation of regression lines in method comparison (5)(6)(7). Bland and Altman (4) recommended the difference plot (or bias plot or residual plot) as an alternative approach for method comparison. On the abscissa they used the mean value of the methods to be compared, to avoid regression towards the mean, and on the ordinate they plotted the calculated difference between measurements by the two methods. They further estimated the mean and standard deviation of differences and displayed horizontal lines for the mean and for 2 the standard deviation. However, they missed the concept of a more objective criterion for acceptability. Recently, Hollis (5) has recommended difference plots as the only acceptable method for method comparison studies for publication in Annals of Clinical Biochemistry, but without specifying criteria for acceptability.

However, a few difference plots with evaluation of acceptability according to defined criteria have been published, e.g., in evaluation of estimated biological variation compared with analytical imprecision (8), and in external quality assessment of plasma proteins for the possibilities of sharing common reference intervals (9).

Maybe the scarcity of such publications is more a question of interpretation of the data by plotting than a strict choice between scatter-plot and difference plot, as discussed by Stckl (10) recently. Investigators seem to rely too much on regression lines and r-values, without doing the equally important interpretation of the data points of the plot. This is becoming more and more disadvantageous with the increasing number of Reference Methods available for comparison with field methods, because in these cases, it is not a question of finding some relationships, but simply of judging the field method to be acceptable or not.

NCCLS has recently published guidelines for method comparison and bias estimation by using patients samples (11), where both scatter-plots and bias plots are advised. The document also recommends plotting of single determinations as mean values and stresses the need of visual inspection of data. Further, comparison with performance criteria is recommended, but these criteria are not specified and they are not used in the graphical interpretation. Recently, Houbouyan et al. (12) used ratio plots in their validation protocol of analytical hemostasis systems, where they used a preset, but arbitrarily chosen, acceptance limit of inaccuracy of 15

In the following, we will use the difference plot (or bias plot) in combination with simple statistics for the principal judgment of the identity or acceptability of a field method. The difference plot makes it easier to apply the concept; in principle, however, the same evaluations could be performed for a scatter-plot in relation to the line of identity (y = x).

The aim of this contribution is to pay attention to the hypothesis of identity and the concept of acceptable analytical quality in method comparison, especially when one of the methods is a Reference Method.

1.3 Other Types of Studies

Lewis et al. (1991) categorize method comparison studies into three different types. The key difference between the first two is whether or not a 'gold standard' method is used. In situations where one instrument or method is known to be 'accurate and precise', it is considered as the'gold standard' (Lewis et al., 1991). A method that is not considered to be a gold standard is referred to as an 'approximate method'. In calibration studies they are referred to a criterion methods and test methods respectively.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). (In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively.) Altman and Bland (1983) make clear that their methodology is not intended for calibration problems.

2. Comparison problems. When two approximate methods, that use the same units of measurement, are to be compared. This is the case which the Bland-Altman methodology is specfically intended for, and therefore it is the most relevant of the three.

3. Conversion problems. When two approximate methods, that use different units of measurement, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement. Lewis et al. (1991) deals specifically with this issue. In the context of this study, it is the least relevant of the three.

1.4 Fuzzy Gold Standards

The Gold Standard is considered to be the most accurate measurement of a particular parameter. But even gold standard raters must be assumed to have some level of measurement error. Fuzzy gold standard are considered by Phelps and Hutson (1994)

Dunn (2002, p.47) cautions that'gold standards' should not be assumed to be error free. 'It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement' (Dunn, 2002). Pizzi (1999) similarly addresses the issue of glod standards, 'well-established gold standard may itself be imprecise or even unreliable'.

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well-being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram, used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. This is reported as sensitivity of 95% and a specificity of 92% (ACR, 2008).

In literature they are, perhaps more accurately, referred to as 'fuzzy gold standards' (Phelps and Hutson, 1995). Consequently when one of the methods is essentially a fuzzy gold standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

Dunn (2002) makes two important points in relation to these categories. Firstly he remarks that there isn't clear cut differences between each category.

Secondly he comments on the clinician gold standard, the sphygmomanometer, *leaves considerable room for improvement*. Pizzi (1999) also attends to this issue: *well-established gold standard may itself be imprecise or even unreliable.*

The Magnetic resonance angiogram is considered to the gold standard for measuring aortic dissection, with a sensitivity of 95% and a specificity of 92% . (ACR, 2008)

In literature they are, perhaps more accurately, referred to as 'bronze standards'.

Consequently when one of the methods is essentially a bronze standard, as opposed to a true gold standard, the comparison procedure should be considered as being of the second category.

1.5 Fuzzball Agreement

Fuzzball agreement is a case where the correlation coefficient is close to zero. The sample values is restricted to a narrow range. but an examination of a relevant scatter-plot would indicate that there is agreement between the two methods.

Agreement - a numerical measure Hutson et al define a numerical measure for agreement.

For example, suppose the pairs of rater measurements are (1, 1), (1.1, 1), (1, 1.1), and (1.1, 1.1) then the sample Pearson correlation r = .0, yet the two raters or devices are considered to be in good agreement. We will refer to the instance where r is close to 0, yet there may be good agreement as ”fuzzball agreement.”

Fuzzball agreement occurs quite often in practice when the sample values have very narrow or restricted ranges. Fuzzball agreement is just one instance where the correlation coefficient is a poor measure of agreement.

Furthermore, note that the ICC is also a poor measure of agreement when there is fuzzball agreement. At the other extreme suppose the same raters given in the previous example had pairs of measurements (1, 101), (2, 102), (3, 103), and (4, 104) on the same relative scale as before. In this instance, r = 1.0, yet there is large disagreement between rater.

1.6 Repeatability and gold standards

Currently the phrase ‘gold standard’ describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the ‘gold standard’, yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a ’bronze standard’. Again, no formal definition of a ’bronze standard’ exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a ‘gold standard’. For example, by determining the ratio of CR to the sample mean $\bar{X}$. Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value of $\lambda = CR/\bar{X}$ with $\lambda < 0.1\%$. Similarly, a silver standard may be defined as the method with the lowest value of λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.

1.7 The Conversion Problem

In this section, we will reconsider the conversion problem, where by the methods of measurements are denominated in different units. Conversion problems arise when the comparison is between two approximate methods of measurement each of which measures the quantity in different units.

This situation can arise when the methods in question proceed by measuring different proxies for the underlying quantity of interest. (lewis 1991)

For the single measurement case, the author can not foresee any scope for insights that are not already offered by using a structural relation model, as proposed by lewis et 1991, or error-in-variables regression. In the case of orthonormal regression, it is not reasonable to assume that both methods have equal measurement variance, when they are denominated in different units. The analyst may attempt to mitigate the problem by scaling the variance of one method, but even still problems remain. Similarly for Deming regression, no further insights on how to properly estimate the variance ratio can be offered.

For the case of conversion problem with replicate measurements, a framework that incorporates the ideas offered by Roy (2009) can be proposed. Estimates for between-subject and within-subject variances may be sought. However Roy's tests on variability are no longer applicable, as one would not expect the method to have similar estimates. An estimate for the scaling factor β may be sought, where $Y_i \approx \beta X$.

$$X_i = \tau_i + \delta_i$$

$$Y_i = \alpha + \beta X \tau_i + \epsilon_i$$

We will simulate a data set based in lewis conversion problems, provide three replicates values for both measurements. To acheive this we add "jitter noise" to three copies of each original measurement.

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

Phelps, C. and A. Hutson (1995). Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical decision making 15*, 144–57.

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.