Contents

Chapter 1

Appendices

1.1 Appendix A : Improper Method Comparison Techniques

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 paired sample $t-$test, simple linear regression, or correlation coefficients.

Paired sample *t*-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_d = 0$. The test statistic is distributed as a t random variable with $n - 1$ degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{1.1}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. This method can be potentially misused for method comparison studies. Paired $t-$tests test only whether the mean responses are the same, and so provides a useful test for inter-method bias. However, no insight can be obtained about the variability of the case-wise differences by the paired $t-$test, critically undermining it as a stand-alone procedure. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

The Correlation Coefficient

Correlation is inadequate to assess agreement because it only evaluates only the linear association of two sets of observations. Nonetheless linear association is not the same as agreement. It is possible for two methods to be highly correlated, yet have poor agreement due to any combination of constant and proportional bias. Arguments against its usage have been made repeatedly in the relevant literature, with Altman and Bland (1983), Bland and Altman (1986), **?** and **?** as examples.

Regression Methods

On account of the fact that one set of measurements are linearly related to another, one could surmise that simple linear Regression is the most suitable approach to analyzing comparisons. However simple linear regression is considered by many authors to be wholly unsuitable for method comparison studies (Altman and Bland, 1983; Cornbleet and Cochrane, 1979; Ludbrook, 1997). Simple linear regression is defined as such with the name 'Model I regression' by Cornbleet and Cochrane (1979), in contrast to 'Model II regression' models, which shall be discussed later on.

A key assumptions of simple linear regression is that the independent variable values are without random error. For method comparison studies, both sets of measurement must be assumed to be measured with imprecision and neither case can be taken to be a reference method. Arbitrarily selecting either method as the reference (i.e. the independent variable) will yield conflicting outcomes: a regression of X on Y would yield an entirely different model from fitting Y on X.

Further criticisms of linear regression exist. Firstly regression methods are uninformative about the variability of the differences. Secondly regression models are unduly influenced by outliers. Lastly, regression models can not be used to effectively analyze repeated measurements.

The Identity Plot

Altman and Bland (1983) states that regression analysis can offer useful insights, and recommending an 'Identity Plot', a simple graphical approach that yields a cursory examination of how well the measurement methods agree. In the case of good agreement, the co-variates of the Identity plot accord closely with the $X = Y$ line. This plot is not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation. An identity plot shall complement demonstrations of commonly used approaches in the next chapter.

Decomposition of Inter-Method Bias

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing inter-method bias to be decomposed into constant bias and proportional bias. Regression methods can determine the presence of inter-method bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002).

Constant bias describes the case where one method gives values that are consistently different to the other across the whole range. Using a naive estimation of bias, such as the mean of differences, it may incorrectly indicate absence of bias, by yielding a mean difference close to zero. This would be caused by positive differences in the measurements at one end of the range of measurements being canceled out by negative differences at the other end of the scale. Proportional Bias exists when two methods agree on average, but exhibit differences over a range of measurements, i.e. the differences are proportional to the scale of the measurement. A measurement method may be subject to any combination of fixed bias or proportional bias, or both (Ludbrook, 2002).

Constant or proportional bias using linear regression can be detected by an individual test on the intercept or the slope of the line regressed from the results of the two methods to be compared. If there is no constant bias, the intercept is equal to zero and, similarly, if there is no proportional bias, the slope is equal to one. Thus, carrying out hypothesis tests on these coefficients (where the null hypotheses are $\beta_0 = 0$ and $\beta_1 = 1$) allow us to test for the presence of both types of bias.

If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined.

1.2 Appendix 2 : Variations of the Bland-Altman Plot

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As Bland and Altman (1986) point out this may not be the case. Importantly Bland and Altman (1999) 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 importance 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.

Due to limitations of the conventional difference plot, a series of alternative formulations for the Bland-Altman approach have been proposed. Referring to the assumption that bias and variability are constant across the range of measurements, Bland and Altman (1999) 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 be 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.

Bland and Altman (1999) 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 case-wise 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 case-wise ratios as percentage of averages, removing the need for logarithmic transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. Eksborg (1981) proposed such a ratio plot, independently of Bland and Altman. Dewitte et al. (2002) commented on the reception of this article by saying '*Strange to say, this report has been overlooked*'.

1.3 Appendix 3 : The Coefficient of Repeatability

Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement by collecting replicated data. Roy (2009) notes the lack of convenience in such calculations. The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999).

As mentioned previously, Barnhart et al. (2007) emphasize the importance of repeatability as part of an overall method comparison study. The coefficient of repeatability was proposed by Bland and Altman (1999), and is referenced in subsequent papers, such as Carstensen et al. (2008). **?** define a coefficient of repeatability as *the value below which the difference between two single test results....may be expected to lie within a specified probability.* Bland and Altman (1999) defines the repeatability coefficient as the upper limits of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances.

Once the within-item variability for both methods has been estimated, the relevant calculations for the coefficients of repeatability are straightforward. The coefficient is calculated from the within-item variability σ^2_m as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$. For 95% of subjects, two replicated measurement by the same method will be within this repeatability coefficient.

The coefficient of repeatability may provide the basis for the formulation a formal definition of a 'gold standard'. For example, by determining the ratio of the repeatability coefficient (CR) to the sample mean $\bar{X}$. Advisably the sample size should 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.

1.4 Appendix 4 : 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 (**?**). 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 framework 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.

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.

? discusses the importance of gold standards in the context of method comparison studies. Currently the phrase ‘gold standard’ describes the most accurate method of measurement available. No other criteria are set out. Further to Dunn (2002), various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer (i.e. a blood pressure measurement cuff), which is prone to measurement error. Consequently it can be said that a measurement method can be the ‘gold standard’, yet have poor repeatability. Dunn (2002) recognizes 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.

Dunn (2002, p.47) cautions that ‘gold standards’ should not be assumed to be error free and that ‘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’. Pizzi (1999) similarly addresses the issue of gold 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 tests based upon the angiogram are reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8% (ACR, 2008).

In literature gold standards are, perhaps more accurately, can be 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 both in the context of a comparison study and a calibration study.

According to Bland and Altman, one should use the approach previous outlined, even when one of the methods is a gold standard.

1.4.1 Similar Problems

? categorize method comparison studies into three different types, namely: calibration, comparison and conversion. 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' (**?**). 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 as 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 (**?**). 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 framework 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 for which Bland and Altman's Methodology is intended, and therefore it is the most relevant of the three for this thesis.

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.

? deals specifically with this issue. In the context of this thesis, it is the least relevant of the three cases.

? discusses the importance of gold Standards in the context of method comparison studies. Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to Dunn (2002), various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer (i.e. a blood pressure measurement cuff), which is prone to measurement error. Consequently it can be said that a measurement method can be the ‘gold standard’, yet have poor repeatability. Dunn (2002) recognizes 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.

Dunn (2002, p.47) cautions that ‘gold standards’ should not be assumed to be error free and that ‘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’. Pizzi (1999) similarly addresses the issue of gold 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 tests based upon the angiogram are reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8% (ACR, 2008).

In literature gold standards are, perhaps more accurately, can be 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 both in the context of a comparison study and a calibration study.

According to Bland and Altman, one should use the methodology previous outlined, even when one of the methods is a gold standard.

1.5 Appendix 5 : Indices and Graphical Techniques

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation in assessing agreement. The mean square deviation is defined as the expectation of the squared differences of two readings. The MSD is usually used for the case of two measurement methods X and Y, each making one measurement for the same subject, and is given by:

$$MSDxy = E[(x-y)^2] = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1-\rho_{xy}).$$

Barnhart et al. (2007) advises the use of a predetermined upper limit for the MSD value, MSD_{ul}, to define satisfactory agreement. However, a satisfactory upper limit may not be easily determinable, thus creating a drawback to this technique.

Alternative indices, proposed by Barnhart et al. (2007), are the square root of the MSD and the expected absolute difference (EAD).

$$EAD = E(|x-y|) = \frac{\sum |x_i - y_i|}{n},$$

Both of these indices can be interpreted intuitively, since their units are the same as that of the original measurements. They can also be compared to the maximum acceptable absolute difference between two methods of measurement d_0. For the sake of brevity, the EAD will be considered solely.

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions. A consequence of using absolute differences is that high variances would result in a higher EAD value.

To illustrate the use of EAD, consider Table 1.5.1. The inter-method bias of 0.03, which is desirably close to zero in the context of agreement. However, an identity plot

	U	V	$U - V$	$\lvert U - V \rvert$
1	98.05	99.53	-1.49	1.49
2	99.17	96.53	2.64	2.64
3	100.31	97.55	2.75	2.75
4	100.35	96.03	4.32	4.32
5	99.51	99.00	0.51	0.51
6	98.50	100.76	-2.26	2.26
7	100.66	99.37	1.29	1.29
8	99.66	108.87	-9.21	9.21
9	99.70	105.16	-5.45	5.45
10	101.55	94.31	7.24	7.24

Table 1.5.1: Example data set

would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, which is a wide interval for this data. As with the identity plot, this would indicate lack of agreement. As with inter-method bias, an EAD value close to zero is desirable. However, from Table 1.5.1, the EAD can be computed as 3.71. The Bland-Altman plot remains a useful part of the analysis. In Figure 1.5.2, it is clear there is a systematic decrease in differences across the range of measurements.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD, using simulation studies, would be interesting, while further adding that '*It will be of interest to investigate the benefits of these possible new unscaled agreement indices*'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. The EAD

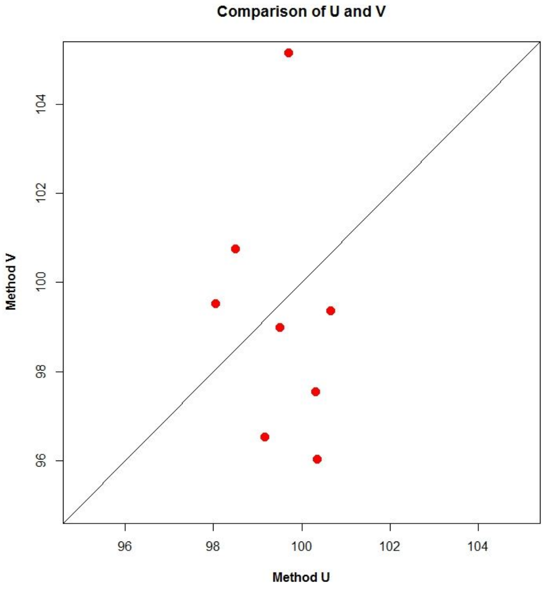

Figure 1.5.1: Identity Plot for example data

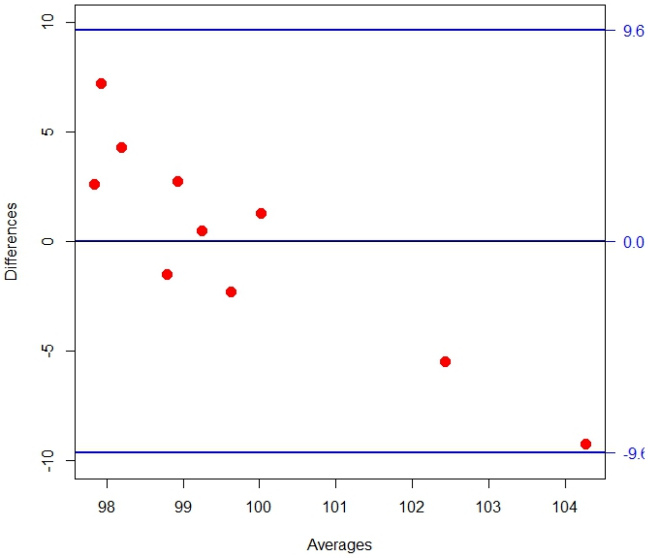

Figure 1.5.2: Bland-Altman Plot for UV comparison

values for both comparisons are therefore much closer.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two

	F vs C	F vs T
Inter-method bias	-0.61	0.12
Difference variance	0.06	0.22
Limits of agreement	(-1.08, -0.13)	(-0.81,1.04)
EAD	0.61	0.35

Table 1.5.2: Agreement indices for Grubbs' data comparisons.

measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{1.2}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. This boundary is known as the 'Total Deviation Index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

1.6 Carstensen Coefficient of Repeatability

The limits of agreement are not always the only issue of interest, the assessment of method specific repeatability and reproducibility are of interest in their own right. Repeatability can only be assessed when replicate measurements by each method are available.

Under the model for linked replicates, there are two possibilities depending on the circumstances. If the variation between replicates within item can be considered a part of the repeatability it will be $2.8\sqrt{\omega^2 + \sigma^2_m}$.

However, if replicates are taken under substantially different circumstances, the variance component ω^2 may be considered irrelevant in the repeatability and one would therefore base the repeatability on the measurement errors alone, i.e. use $2.8\sigma_m$.

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.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

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

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

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).

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

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.

Eksborg, S. (1981). Evaluation of method-comparison data [letter]. *Clinical Chemistry 27*, 1311–1312.

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

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

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

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

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

O'Brien, E., J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. Altman, M. Bland, A. Coats, and N. Atkins (1990). The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension 8*, 607–619.

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.

Roy, A. (2009). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.