Contents

Chapter 1

Method Comparison Studies

1.1 What is a method comparison study?

The problem of compare the results of two different measurement techniques, or in other words, assessing the *agreement* between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Ludbrook (1997) states that the purpose of comparing two measurements "of a continuous biological variable" is to uncover systematic differences, not to point to similarities". The need to compare the results of two different measurement techniques is common in medical statistics.

In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and either a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years.

Published examples of method comparison studies can be found in disciplines as diverse as pharmacology (Ludbrook, 1997), anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000).

The approach proposed by Roy deals with the question of agreement, and indeed

interchangeability, as developed by Bland and Altman's corpus of work. In the view of Dunn, a question relevant to many practitioners is which of the two methods is more precise.

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm gun and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels 'Fotobalk', 'Counter' and 'Terma'.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 1.1.1: Velocity measurement from the three chronographs (Grubbs 1973).

An important aspect of the these data is that all three methods of measurement are assumed to have an attended measurement error, and the velocities reported in Table 1.1 can not be assumed to be 'true values' in any absolute sense.

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those

qualities. An accurate measurement method will give results close to the unknown 'true value'. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy. The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The smaller the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than those of the other, they can not be considered to be in agreement. With this in mind a methodology is required that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.1.2: Difference between Fotobalk and Counter measurements.

1.2 Agreement

Bland and Altman (1986) defined perfect agreement as the case where all of the pairs of measurement data, when plotted on a conventional scatter-plot, 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 line $X = Y$ on the Cartesian plane).

To carry their idea a step further, we define a specific numerical measure of agreement as twice the expected squared perpendicular distance of the pair of random variables (X_1, X_2) to the line of equality or agreement in the (X_1, X_2)-plane, that is, $E(X_1 - X_2)/2$, where X_1 and X_2 denote the continuous measurements of rater 1 and rater 2, respectively.

Obviously, other L_p norms may be considered for the purpose of numerically measuring agreement and warrant future consideration.

Agreement is the extent to which the measure of the variable of interest, under a constant set of experimental conditions, yields the same result on repeated trials (Sanchez and Binkowitz, 1999). The more consistent the results, the more reliable the measuring procedure.

1.3 Purpose of Method Comparison Studies

Carstensen (2010) provides a review of many descriptions of the purpose of Method Comparison studies, several of which are reproduced here.

> "The question being answered is not always clear, but is usually epxressed as an attempt to quantify the agreement between two methods" (Bland and Altman, 1995).
>
> "Some lack of agreement between different methods of measurement is inevitable. What matters is the amount by which they disagree. We want to know by how much the new method is likely to differ from the old, so that it is not enough to cause problems in the mathematical interpretation we can preplace the old method by the new, or even use the two interchangeably" (Bland and Altman, 1999).
>
> "It often happens that the same physical and chemical property can be measured in different ways. For example, one can determine For example, one can determine sodium in serum by flame atomic emission spectroscopy or by isotope dilution mass spectroscopy. The question arises as to which method is better" (Mandel, 1991).
>
> "In areas of inter-laboratory quality control, method comparisons, assay validations and individual bio-equivalence, etc, the agreement between observations and target (reference) values is of interest" (Lin et al., 2002).
>
> "The purpose of comparing two methods of measurement of a continuous

biological variable is to uncover systematic differences, not to point to similarities‘" (Ludbrook, 1997).

"In the pharmaceutical industry, measurement methods that measure the quantity of prdocuts are regulated. The FDA (U.S. Food and Drug Administration) requires that the manufacturer show equivalency prior to approving the new or alternatice method in quality control" (Tan & Inglewicz, 1999).

While several major commonalities are present in each definitions, there is a different emphasis for each, which will inevitably give rise to confusion. Carstensen (2010) seems to endorse a simple phrasing of the research question that is proposed by Altman and Bland (1983), i.e. "*do the two methods of measurement agree sufficiently closely?*" with Carstensen (2010) expressing the view that other considerations (for example, the "equivalence" of two methods) to be treated as separate research questions. As such, we will revert to other research questions, such as "equivalence of methods" later, focussing on agreement and repeatability of methods.

1.4 Bias as a source of Lack Of Agreement

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

Chapter 2

Improper MCS Techniques

2.1 Methods of assessing agreement

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 assume to have attendant random error.

2.2 Paired sample *t* test

This method can be applied to test for statistically significant deviations in bias. This method can be potentially misused for method comparison studies. Paired t tests test only whether the mean responses are the same. Certainly, we want the means to be the same, but this is only a small part of the story. The means can be equal while the (random) differences between measurements can be huge.

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 a t random variable

with $n-1$ degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{2.1}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

2.3 Inappropriate use of the 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.

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.

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 coeffiecient 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.

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.

The correlation coefficient measures linear agreement–whether the measurements go up-and-down together. Certainly, we want the measures to go up-and-down together, but the correlation coefficient itself is deficient in at least three ways as a measure of agreement.

The correlation coefficient can be close to 1 even when there is considerable bias between the two methods. For example, if one method gives measurements that are always 10 units higher than the other method, the correlation will be 1 exactly, but the measurements will always be 10 units apart.

The magnitude of the correlation coefficient is affected by the range of subjects/units studied.

The correlation coefficient can be made smaller by measuring samples that are similar to each other and larger by measuring samples that are very different from each other. The magnitude of the correlation says nothing about the magnitude of the differences between the paired measurements which, when you get right down to it, is all that really matters.

The usual significance test involving a correlation coefficient– whether the population value is 0–is irrelevant to the comparability problem. What is important is not merely that the correlation coefficient be different from 0. Rather, it should be close to (ideally, equal to) 1!

2.4 Regression Methods

Scatterplots are recommended by Altman and Bland (1983) 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. O'Brien et al. (1990) 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 does not 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.

2.4.1 Simple Linear Regression

Regression methods are uninformative about the variability of the differences.

Another inappropriate approach is the regressing one set of measurements against the other. According to this methodology the measurement methods could considered equivalent if the confidence interval for the regression coefficient included 1. Analysts sometimes use least squares (referred to by Ludbrook as Model I) regression analysis to calibrate one method of measurement against another.

In this technique, the sum of the squares of the vertical deviations of y values from the line is minimized. This approach is invalid, because both y and x values are attended by random error.

2.4.2 Useful Insights

Bland and Altman have stated that regression analysis offers insights into MCS problems. The Identity Plot is a simple graphical approach, advocated by Bland and Altman (1986), that yields a cursory examination of how well the measurement methods agree. In the case of good agreement, the co-variates of the plot accord closely with the $X = Y$ line.

Ludbrook (1997, 2002) criticizes Bland-Altman 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 agreeement, but offer no guidance on how to correct for those effects.

2.4.3 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 fixed bias and proportional bias. Fixed 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.

Regression analysis is typically misused by regressing one measurement on the other and declare them equivalent if and only if the confidence interval for the regression coefficient includes 1. Some simple mathematics shows that if the measurements are comparable, the population value of the regression coefficient will be equal to the correlation coefficient between the two methods.

Regression methods can determine the presence of bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002).

Constant Bias is a form of systematic deviations estimated as the average difference between the test and the reference method.

Constant or proportional bias in method comparison studies 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.

Proportional Bias exists when two methods agree on average, but exhibit differences over a range of measurements. Proportional Bias is a difference in the two measures

which is proportional to the scale of the measurement. Proportional bias describes the difference in measurements getting progressively greater, or smaller, across the range of measurements. A measurement method may have either an attendant fixed bias or proportional bias, or both (Ludbrook, 2002).

If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined. Outliers are a source of error in regression estimates.

Proportional bias describes the difference in measurements getting progressively greater, or smaller, across the range of measurements. A measurement method may have either an attendant fixed bias or proportional bias, or both. (**?**). Determination of these biases shall be discussed in due course.

Inference Procedures

A 95% confidence interval for the intercept estimate can be used to test the intercept, and hence fixed bias, is equal to zero. This hypothesis is accepted if the confidence interval for the estimate contains the value 0 in its range.

Should this be, it can be concluded that fixed bias is not present. Conversely, if the hypothesis is rejected, then it is concluded that the intercept is non zero, and that fixed bias is present.

Testing for proportional bias is a very similar procedure. The 95% confidence interval for the slope estimate can be used to test the hypothesis that the slope is equal to 1.

This hypothesis is accepted if the confidence interval for the estimate contains the value 1 in its range. If the hypothesis is rejected, then it is concluded that the slope is significant different from 1 and that a proportional bias exists.

Intra-class correlation coefficient

The ICC, which takes on values between 0 and 1, is based on analysis of variance techniques. It is close to 1 when the differences between paired measurements is very small compared to the differences between subjects. Of these three procedures–t test, correlation coefficient, intra-class correlation coefficient–the ICC is best because it can be large only if there is no bias and the paired measurements are in good agreement, but it suffers from the same faults ii and iii as ordinary correlation coefficients. The magnitude of the ICC can be manipulated by the choice of samples to split and says nothing about the magnitude of the paired differences.

Chapter 3

Repeated Measurements and Repeatability

3.1 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.

Bland and Altman attend to the issue of repeated measures in 1996.

Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual.

Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation, and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the '$t-$' limits of agreement (the outer pair of dashed lines) centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

Roy (2009) accords with Bland and Altman's definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

Repeated measurements are said to be linked if a direct correspondence exists between successive measurements across measurements, i.e. pairing. Such measurements are commonly made with a time interval between them, but simultaneously for both methods. Paired measurements are exchangeable, but individual measurements are not.

If the paired measurements are taken in a short period of time so that no real systemic changes can take place on each item, they can be considered true replicates. Should enough time elapse for systemic changes, linked repeated measurements can not be treated as true replicates.

In this model, the variances of the random effects must depend on m, since the different methods do not necessarily measure on the same scale, and different methods naturally must be assumed to have different variances. Carstensen (2004) attends to the issue of comparative variances.

Bland and Altman (1999) also remark that an important feature of replicate observations is that they should be independent of each other. This issue is addressed by Carstensen (2010), in terms of exchangeability and linkage. Carstenen advises that repeated measurements come in two *substantially different* forms, depending on the circumstances of their measurement: exchangable and linked.

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009) notes that some measurements may not be 'true' replicates.

Roy's methodology assumes the use of 'true replicates'. However data may not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one $AR(1)$ structure. However determining MLEs with such a structure would be computational intense, if possible at all.

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice (Bland and Altman, 1999).

3.2 Definition of Replicate measurements

Further to Bland and Altman (1999), a formal definition is required of what exactly replicate measurements are

By replicates we mean two or more measurements on the same individual taken in identical conditions. In general this requirement means that the measurements are taken in quick succession.

Roy accords with Bland and Altman's definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

3.3 Exchangeable and Linked measurements

Repeated measurements are said to be exchangeable if no relationship exists between successive measurements across measurements. If the condition of exchangeability exists, a group of measurement of the same item determined by the same method can be re-arranged in any permutation without prejudice to proper analysis. There is no reason to believe that the true value of the underlying variable has changed over the course of the measurements.

Exchangeable repeated measurements can be treated as true replicates. For the purposes of method comparison studies the following remarks can be made. The $r-$th measurement made by method 1 has no special correspondence to the $r-$th measurement made by method 2, and consequently any pairing of repeated measurements are as good as each other.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

Repeatability is the ability of a measurement method to give consistent results for a particular subject, i.e. a measurement will agree with prior and subsequent measurements of the same subject. Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study, a view endorsed by Carstensen et al. (2008). Before there can be good agreement between two methods, a method must have good agreement with itself. If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Roy, 2009).

3.3.1 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

3.4 Repeatability

Repeatability describes the variation in measurements taken by a single method of measurement on the same item and under the same conditions.

A measurement may be said to be repeatable when this variation is smaller than some agreed limit.

Repeatability is practically used, for example, in medical monitoring of conditions. In these situations, there is often a predetermined "critical difference", and for differences in monitored values that are smaller than this critical difference, the possibility of pre-test variability as a sole cause of the difference may be considered in addition to, for examples, changes in diseases or treatments.

A measurement is said to be repeatable when this variation is smaller than some prespecified limit. In these situations, there is often a predetermined "critical difference",

and for differences in monitored values that are smaller than this critical difference, the possibility of pre-test variability as a sole cause of the difference may be considered in addition to, for examples, changes in diseases or treatments.

The assessment of method-specific repeatability and reproducibility is of interest in its own right.

Repeatability and reproducibility can only be assessed when replicate measurements by each method are available.

If replicate measurements by a method are available, it is simple to estimate the measurement error for a method, using a model with fixed effects for item, then taking the residual standard deviation as measurement error standard deviation. However, if replicates are linked, this may produce an estimate that biased upwards.

The quality of repeatability is the ability of a measurement method to give consistent results for a particular subject. That is to say that a measurement will agree with prior and subsequent measurements of the same subject.

A measurement method can be said to have a good level of repeatability if there is consistency in repeated measurements on the same subject using that method. Conversely, a method has poor repeatability if there is considerable variation in repeated measurements.

Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study.

Barnhart et al. (2007) and Roy (2009) highlight the importance of reporting repeatability in method comparison, because it measures the purest random error not influenced by any external factors. Importantly, before there can be good agreement between two methods, a method must have good agreement with itself. Statistical procedures on within-subject variances of two methods are equivalent to tests on their respective repeatability coefficients. A formal test is introduced by Roy (2009), which will discussed in due course.

Repeatability is important in the context of method comparison because the repeatability of two methods influence the amount of agreement which is possible be-

tween those methods. If one method has poor repeatability, the agreement is bound to be poor. If both methods have poor repeatability, agreement is even worse. If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Roy, 2009).

Barnhart et al. (2007) remarks that it is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors, while further remarking '*curiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked*. Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. However Roy (2009) notes the lack of convenience in such calculations.

Repeatability is defined by the IUPAC (2009) as '*the closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time)*' and is determined by taking multiple measurements on a series of subjects.

According to the *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, the following conditions need to be fulfilled in the establishment of repeatability:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time.
- same objectives

3.5 Carstensen Move to Chapter 2

, 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 (2) 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 $2.8\sqrt{\omega^2 + \sigma_m^2}$.

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 $2.8\sigma_m$.

3.6 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 Dunn (2002), 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.

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.

? discusses the relevance of gold Standards in the context of MCS.

It is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors (Barnhart et al., 2007).

As noted by Bland and Altman 1999, the repeatability of two methods of measurement can potentially limit importance of repeatability' curiously replicate measure-

ments are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked.

3.7 Other Types of Studies / Gold Standards

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.

?, 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' (**?**). 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 **?**. 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.

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

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 (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet 346*, 1085–87.

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

Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics 5*(3), 399–413.

Carstensen, B. (2010). *Comparing Clinical Measurement Methods : A Practical Guide*. Wiley.

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).

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

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. `http://goldbook.iupac.org/R05293.html`.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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., 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.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

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.

Sanchez, M. M. and B. S. Binkowitz (1999). Guidelines for measurement validation in clinical trial design. *Journal of biopharmaceutical statistics 9*(3), 417–438.