Contents

Chapter 1

Method Comparison Studies

1.1 What is a method comparison study?

The question of properly assessing "agreement" between two or more methods of measurement is ubiquitous, and is commonly referred to as a 'method comparison study'. 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).

Historically comparison of two methods of measurement was carried out by use of paired sample $t-$test, correlation coefficients or simple linear regression. Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these approaches for comparing two methods of measurement, and proposed their own framework with this application in mind. Although the authors acknowledge the opportunity to apply other, more complex methodologies, they argue that simpler approaches are preferable, especially when the results must be 'explained to non-statisticians'.

A method of measurement should ideally be both accurate and precise. 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. 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'.

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.

1.1.1 Purpose of Method Comparison Studies

Different authors focus on different aspects of comparison problem. 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).
>
> "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. In the view of Dunn (2002), a question relevant to many practitioners is which of the two methods is more precise. 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.

In many cases the purpose of the study is to calibrate a new method of measurement against a "Gold Standard" method, a known method that is considered most precise in its measurement. For example, 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 approaches have been proposed for this purpose in recent years. It must be noted that absence of measurement error should not be assumed for gold standard methods.

1.1.2 Grubbs' Artillery Round Data

To illustrate the characteristics of a typical method comparison study consider the data in Table 1.1 (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm artillery piece 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 simple estimate of the inter-method bias is given by the differences between pairs of measurements, for example, in Table 1.2 shows possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method.

The absence of inter-method bias is, by itself, not sufficient to establish that 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. Hene, method comparison studies are required to take account of both inter-method bias and difference in the precision of measurements.

Round	Fotobalk (F)	Counter (C)	Difference (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.

1.1.3 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).

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.

Altman and Bland (1983) 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.

1.2 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.

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.

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), Bland and Altman (2003) and Giavarina (2015) as examples.

1.3 Replicate Measurements and Repeatability

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Repeated measurements on several subjects can be used to quantify measurement error; the variation between measurements of the same quantity on the same individual. Measurements taken in quick succession, so that no real systemic changes can take place on each item, by the same observer using the same instrument on the same item can be considered true replicates (Bland and Altman, 1999). Roy (2009) accords with Bland and Altman's definition, but notes that some measurements may not be 'true' replicates. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity posed by replicate measurements . Bland and Altman (1986) address the additional complexity 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 properly estimate the inter-method bias. However, Carstensen et al. (2008) is critical of both approaches, offering an alternative approach that shall be introduced later.

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.

1.3.1 Exchangeable and Linked measurements

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.

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.

1.3.2 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 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.

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. A similar set of criteria is described in the *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*.

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). The repeatability of two methods influence the amount of agreement which is possible between those methods. 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 (Bland and Altman, 1999; Roy, 2009). 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.

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.*

Statistical procedures on within-item 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 chapter three.

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.

1.3.3 Types of Comparison Studies

Lewis et al. (1991) 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' (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 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 (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 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.

Lewis et al. (1991) deals specifically with this issue. In the context of this thesis, it is the least relevant of the three cases.

Roy et al. (2015) 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.

1.4 Outline of Thesis

Thus, the basic concepts of, and need for method comparison are introduced. The intention of this thesis is to develop the theory of method comparison studies using Linear Mixed Effects models. Chapter two will provide a review of the prevalent methods, highlighting particular flaws where relevant. Chapter three shall describe Linear Mixed effects models, and how the use of the linear mixed effects models can be extended to method comparison studies. Implementations of important existing work is presented using the `R` programming language.

Chapter three shall describe linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented again, using the `R` programming language.

Model diagnostics are an integral component of a complete statistical analysis. In chapter four model diagnostics are described in depth, with particular emphasis on linear mixed effects models.

In the fifth chapter, important linear mixed effects model diagnostic methods shall be extended to method comparison studies, and proposed methods are demonstrated on data sets that have become well known in literature on method comparison. The purpose is to both calibrate these methods and to demonstrate applications for them. The last chapter deals with robust measures of important parameters such as agreement.

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.

Bland, J. M. and D. G. Altman (2003). Applying the right statistics: analyses of measurement studies. *Ultrasound in obstetrics & gynecology 22*(1), 85–93.

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).

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.

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

Giavarina, D. (2015). Understanding bland altman analysis. *Biochemia medica 25*(2), 141–151.

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. Tillotson (1991). The problem of conversion in method comparison studies. *Applied Statistics*, 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.

Roy, A., C. D. Fuller, D. I. Rosenthal, and C. R. Thomas Jr (2015). Comparison of measurement methods with a mixed effects procedure accounting for replicated evaluations (com 3 pare): method comparison algorithm implementation for head and neck igrt positional verification. *BMC medical imaging 15*(1), 35.

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