Contents

[Dunn’s Criticism of BA 1](#_Toc326062760)

[Which method of measurement is more precise? 1](#_Toc326062761)

[Bland –Altman 1999 2](#_Toc326062762)

[What are Tolerance Intervals? 3](#_Toc326062763)

[Schluter’s Bayesian Approach: 3](#_Toc326062764)

[Escaramis et TDI approach 3](#_Toc326062765)

[Reporting of MCS 5](#_Toc326062766)

[Intra-method agreements/repeatability 6](#_Toc326062767)

[Deming Regression 6](#_Toc326062768)

[Probability Based Approaches 6](#_Toc326062769)

[Roy 2009 paper 7](#_Toc326062770)

[Using ML or REML Fitting 9](#_Toc326062771)

# What is Agreement

The matter of how well two methods of measurement are said to be “in agreement” is known as a **method comparison study**, and is prevalent in medical statistics literature. **Barnhart et al** give examples of the diversity of applications of method comparison studies; blood plasma data, body fat, oximetry data etc.

Often the questions extends to whether or not the two methods of measurements have sufficiently levels of agreement that they could be used **interchangeably**. However, as pointed out by **Mantha et al**, interchangeability is not the goal of many studies.

A useful, and broadly consistent, set of definitions of what this “agreement” entail is put forth by **Barnhart et al** and **Roy (2009)**, and shall be discussed in detail.

Shared with previous contributions **(Bland and Altman, Carstensen)** is the condition that there should no systematic tendency for one of the methods to consistently provide a value higher that than of the other method. If such a tendency did exist, we would refer to it as an **inter-method bias**.

In earlier literature, the emphasis was placed up on single measurements simultaneously by each of the methods of measurement. Several different approaches, such as the **Bland-Altman plot**, and **Deming Regression** have been proposed.

# Bland and Altman’s Approach

**Altman and Bland (1983)** developed a simple graphical approach for assessing agreement, whereby the case-wise means are plotted against the case-wise differences. The mean of the case-wise differences, i.e. the inter-method bias is then drawn as a horizontal line across the plot.

**Bland and Altman (1986)** extended this graphical plot with what has become the most influential approach to MCS ; the **Limits of Agreement**. The Limits of Agreement are a tolerance interval for the case-wise differences defined as the **average difference ±2Sd**. Two method of measurement are said to be in agreement if the limits of agreement are not “clinically important”.

Because the results obtained are summary statistics, confidence intervals are required for generalization to the population of case-wise differences. Formulations of confidence intervals for derived values are provided by **Bland-Altman (1999)**.

# Dunn’s Criticism of BA

In trying to prevent widespread misuse of statistical methods, Bland and Altman have, in the view of Dunn, gone too far.

Dunn criticizes the use of the Bland-Altman methodology on the basis that it fails to identify the underlying model, and furthermore, usage of this approach requires assumptions that are "subject of serious and valid criticism"

Importantly, Dunn suggests the Bland-Altman approach as exploratory tool.

# Which method of measurement is more precise?

Dunn contends that the notion of equivalence used by Bland and Altman is too restrictive, and that the Bland-Altman approach fails to address the issue of determining which of the two methods is more precise. Indeed, as many methods comparison studies are motivated by the need to compare a new method of measurement against an established one, the question of which of the two methods is more precise is definitely of interest.

Formal analysis and estimation of relative precision should be based on fitting the appropriate statistical model to the data as a whole. An advantage of which is the assumptions used are open to scrutiny.

***Dunn’s Remarks on other matters***

* Dunn contends that "treating method comparison studies and repeatability as separate issues is inviting trouble".
* Grubbs (1948) was interested with the burning times of fuses, later he discussed the velocity of artilerry shells (Grubbs 1973).In both cases, it is only possible to get one measurement .
* Dunn 2000, and Carroll & Ruppert do not recommend the use of orthogonal regression, on the basis that variance ratio must be guessed. In the case of replicated measurements, the variance ratio can be determined. However, it is an inefficient approach, not using the available data to the fullest.

# Bland –Altman 1999

***How Bland and Altman have extended the Difference Plot (with Limits of agreement) to the case of replicate measurements.***

When replicate measurements are taken with each method on each item, more complex methodologies are required. Bland & Altman(1999) describe a number of approaches to adopt in such a case. The first approach is to take the average measurement over the replicates measurements on each subject by each method.

The second approach is to simply treat these replicate measurements as independent measurements in their own right, and use the pre-existing Bland-Altman approach for single measurements.

The proposed approach shares the same overall structure as their earlier work. However complex calculations are now required to compute the variance of case-wise differences. This is compounded by the fact that different approaches are required depending on whether or not there are equal or unequal numbers of replicates.

***What have they missed? Carstensen et al discuss in depth the shortfalls in this approach.***

Carstensen et al (2008) address the flaws in both approaches, demonstrating how the first approach would lead to an underestimation of the variance of case-wise differences. The second case would also lead to incorrect estimates, although it is pointed out that the second approach is much better than the first.

# What are Tolerance Intervals?

A tolerance interval is a statistical interval within which, with some confidence level, a specified proportion of a population falls.

The ***Engineering Statistics Handbook***describes the difference: *Confidence limits are limits within which we expect a given population parameter, such as the mean, to lie. Statistical tolerance limits are limits within which we expect a stated proportion of the population to lie.*

It is useful to make the distinction between tolerance intervals and confidence intervals clear. The confidence interval describes a single-valued population parameter, commonly the mean, with a specified confidence level. The tolerance interval, on the other hand, describes the range of data values that includes a specific proportion of the population.

As discussed in *Vardeman (1992),* the tolerance interval is not as widely used as the confidence interval and prediction interval, largely because of the emphasis placed on these in undergraduate teaching. Furthermore, *Vardeman(1992)* argues this lack of awareness can lead to misuse of confidence intervals where other types of intervals are more appropriate.

Curiously Carstensen et al (2008) describe the Limits of agreement as a prediction interval, although stating that it is formulated in correctly for that purpose.

***Why Tolerance Intervals are appropriate?***

It is clear from the definition of Tolerance intervals that they function precisely as Bland-Altman intend.

# Schluter’s Bayesian Approach:

***consistent with philosophy of hypothesis testing free stats, but it is unclear what Bland and Altman had in mind.***

*Schluter(2009)* develops a multivariate hierarchical Bayesian approach to Bland-Altman’s methodology, presenting two methods of analysis that complement pre-existing literature.

# Escaramis et TDI approach

Escaramis et al propose to estimate the TDI by constructing a probability interval of the difference in paired measurements between methods, demonstrating how the statistical tolerance interval (TI) procedure as a natural way to make inferences about interval estimates.

***Alternative Approaches***

**Kelly (1985)** advocates the use of a generalization of regression analysis, known as **Structural Equation Modelling** to assess both intra-method agreement and inter-method agreement. This approach was criticised by Bland and Altman on the basis that it did not sufficiently address the issue of precision.

***Criticism of the Bland-Altman approach.***

Arguably, for the single replicate case, the established methodologies are sufficient for assessing agreement between two methods.

The Bland Altman approach has come in for criticism from several authors.  **Graham** **Dunn** criticizes the lack of formal testing. There is little guidance on how to deal with **outliers**.

# Reporting of MCS

Indeed a precise set of instructions on how to correctly use the methodology has not been set out. In a survey of published papers, **Mantha et al** criticizes the absence of pre-specified threshold values for limits of agreement.

The use of structural equation modelling as an alternative to the Bland-Altman methodology was posited by several authors (**Kelly 1985, Dunn, Hopkins**) . Indeed Kelly’s approach generated an exchange of letters.

An early contribution to MCS literature was made by **Grubbs 1948,** who proposed ANOVA models to determine the appropriate estimates. Grubb’s approach was later extended by **Carstensen et al 2008** to account for **replicate measurements**, and to allow for the computation of Limits of Agreement.

***MCS with Replicate measurements***

In subsequent contributions, the matter of assessing agreement in the presence of replicate measurements was addressed. Some approaches extended already established approaches **(Bland-Altman 1999).** Other contributions were based on methodologies not seen previously in Method comparison Study Literature(for example, **Carstensen et al 2008** and **Roy 2009,** using LME models).

A review of recent literature demonstrates how useful and effective the use of LME models are.

More recent contributions to method comparison literature are **Total Deviation Index** and the **Coverage Probability**, introduced by **Lin 2000**, and extended **by Lin 2002** and **Choudhary &Nagaraja 2007**

# Intra-method agreements/repeatability

The matter of intra-method agreement, also known as **repeatability**, has received greater attention in recent contributions. Intra-method agreement describes a method of measurement giving a consistent set of measurements for the same item under identical conditions, such that variability of these measurements can be assumed to be caused by random error only. Many authors argue that the

**Bland and Altman 1999** propose the **coefficient of repeatability (CR)** as a measure of intra-method agreement. Recent literature has set out as a one of the conditions for agreement that two methods must have comparable coefficients of repeatability, or an equivalent condition thereof **(Roy 2009).**

# Deming Regression

**Deming Regression** describes a type of regression model where error is assumed for both variables, rather than just one of them. One of the key drawbacks is that **the variance ratio**, the ratio of the residual variances must be specified( equivalently the ratio of the respective coefficient of variability).

**Orthogonal regression** is a special case of Deming Regression whereby the residual variances are assumed to be equal. In the case of single measurements, a practitioner would be forced to make this assumption. If it was found the coefficients of repeatability were known for both methods, the variance ratio could be computed accordingly as the ratio of CRs.

This begets the question as to how should a Deming regression be performed, given that there are replicates measurements for each item. **Linear Mixed Effects models** are the most plausible solution to this.

# Probability Based Approaches

More recently, agreement between two measurement systems can be determined using the **total deviation index (TDI)** or the **coverage probability (CP)** criteria, as proposed by **Lin (2000)** and **Lin et al. (2002)**. Based on Lin et al's approach, an LME model was proposed by **Choudhary (2007)**.

# Roy 2009 paper

Roy (2009) proposes a suite of hypothesis tests for assessing the agreement of two methods of measurement, when replicate measurements are obtained for each item, using a LME approach. (An item would commonly be a patient).

Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects. Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other.

Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods.

Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal.

***Testing for Inter-method Bias***

Firstly, a practitioner would investigate whether a significant inter-method bias is present between the methods. This bias is specified as a fixed effect in the LME model. For a practitioner who has a reasonable level of competency in R and undergraduate statistics (in particular simple linear regression model) this is a straight-forward procedure.

***Reference Model (Ref.Fit)***

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

|  |
| --- |
| > Ref.Fit = lme(y ~ meth-1, data = dat, *#Symm , Symm#*  + random = list(item=pdSymm(~ meth-1)),  + weights=varIdent(form=~1|meth),  + correlation = corSymm(form=~1 | item/repl),  + method="ML") |

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model.

***Nested Model (Between-Item Variability)***

|  |
| --- |
| > NMB.fit = lme(y ~ meth-1, data = dat, *#CS , Symm#*  + random = list(item=pdCompSymm(~ meth-1)),  + correlation = corSymm(form=~1 | item/repl),  + method="ML") |

***Nested Model (Within –item Variability)***

|  |
| --- |
| > NMW.fit = lme(y ~ meth-1, data = dat, *#Symm , CS#*  + random = list(item=pdSymm(~ meth-1)),  + weights=varIdent(form=~1|meth),  + correlation = corCompSymm(form=~1 | item/repl),  + method="ML") |

***Nested Model (Overall Variability)***

Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.

|  |
| --- |
| > NMO.fit = lme(y ~ meth-1, data = dat, *#CS , CS#*  + random = list(item=pdCompSymm(~ meth-1)),  + correlation = corCompSymm(form=~1 | item/repl),  + method="ML") |

***ANOVAs for Original Fits***

The likelihood Ratio test is very simple to implement in ***R***. All that is required it to specify the reference model and the relevant nested mode as arguments to the command *anova().*

The figure below displays the three tests described by Roy (2009).

|  |
| --- |
| > testB = anova(Ref.Fit,NMB.fit) # Between-Subject Variabilities  > testW = anova(Ref.Fit,NMW.fit) # Within-Subject Variabilities  > testO = anova(Ref.Fit,NMO.fit) # Overall Variabilities |

# Using ML or REML Fitting

Noticeably **Roy (2009)** uses ML estimation when specifying the LME models. No explanation is given, although plausibly it is due to the constraints of the computational environment being used.

Both **West et al (2010)** and **Pinheiro & Bates (2000)** compare ML and REML estimation, describing what types of tests are appropriate for each. When variance components are being tested, REML estimation is in fact the correct approach.

However, **Choudhary & Nagaraja(2007)** point out that for a joint test of Fixed and Random effects that ML estimation is the appropriate estimation method.

***Comparison of ML and REML fits***

|  |  |
| --- | --- |
| ***Fit 1 (ML)***  Dataset: ***Blood RS***  Fixed : ***127.3126 , 143.0275***  AIC: ***4075.594***  ***Between Subject Variability*** | ***Fit1r (REML)***  Dataset: ***Blood RS***  Fixed : ***127.3126 , 143.0275***  AIC: ***4068.172***  ***Between Subject Variability*** |

|  |
| --- |
| # Systolic blood pressure measurements made  # simultaneously by two observers (J and R)  # and an automatic blood pressure measuring  # machine (S), each making three observations  # in quick succession (supplied by Dr E O'Brien)  Blood = matrix(data=c(100, 106, 107, 98, 98, 111, 122, 128, 124, 108, 110, 108, 108, 112, 110, 121,  127, 128, 76, 84, 82, 76, 88, 82, 95, 94, 98,108, 104, 104, 110, 100, 106, 127, 127, 135,124, 112, 112, 128, 112, 114, 140, 131, 124,122, 140, 124, 124, 140, 126, 139, 142, 136,116, 108, 102, 118, 110, 102, 122, 112, 112,114, 110, 112, 112, 108, 112, 130, 129, 135,100, 108, 112, 100, 106, 112, 119, 122, 122,108, 92, 100, 108, 98, 100, 126, 113, 111,100, 106, 104, 102, 108, 106, 107, 113, 111,108, 112, 122, 108, 116, 120, 123, 125, 125,112, 112, 110, 114, 112, 110, 131, 129, 122,104, 108, 104, 104, 108, 104, 123, 126, 114,106, 108, 102, 104, 106, 102, 127, 119, 126,122, 122, 114, 118, 122, 114, 142, 133, 137,100, 102, 102, 102, 102, 100, 104, 116, 115,118, 118, 120, 116, 118, 118, 117, 113, 112,140, 134, 138, 138, 136, 134, 139, 127, 113,150, 148, 144, 148, 146, 144, 143, 155, 133,166, 154, 154, 164, 154, 148, 181, 170, 166,148, 156, 134, 136, 154, 132, 149, 156, 140,  174, 172, 166, 170, 170, 164, 173, 170, 154,174, 166, 150, 174, 166, 154, 160, 155, 170,140, 144, 144, 140, 144, 144, 158, 152, 154,128, 134, 130, 128, 134, 130, 139, 144, 141,  146, 138, 140, 146, 138, 138, 153, 150, 154,146, 152, 148, 146, 152, 148, 138, 144, 131,220, 218, 220, 220, 218, 220, 228, 228, 226,208, 200, 192, 204, 200, 190, 190, 183, 184,  94, 84, 86, 94, 84, 88, 103, 99, 106,114, 124, 116, 112, 126, 118, 131, 131, 124,126, 120, 122, 124, 120, 120, 131, 123, 124,124, 124, 132, 126, 126, 120, 126, 129, 125,110, 120, 128, 110, 122, 126, 121, 114, 125,90, 90, 94, 88, 88, 94, 97, 94, 96,106, 106, 110, 106, 108, 110, 116, 121, 127,218, 202, 208, 218, 200, 206, 215, 201, 207,130, 128, 130, 128, 126, 128, 141, 133, 146,136, 136, 130, 136, 138, 128, 153, 143, 138,100, 96, 88, 100, 96, 86, 113, 107, 102,100, 98, 88, 100, 98, 88, 109, 105, 97,  124, 116, 122, 126, 116, 122, 145, 102, 137,164, 168, 154, 164, 168, 154, 192, 178, 171,100, 102, 100, 100, 104, 102, 112, 116, 116,136, 126, 122, 136, 124, 122, 152, 144, 147,114, 108, 122, 114, 108, 122, 141, 141, 137,148, 120, 132, 146, 130, 132, 206, 188, 166,160, 150, 148, 160, 152, 146, 151, 147, 136,84, 92, 98, 86, 92, 98, 112, 125, 124,156, 162, 152, 156, 158, 152, 162, 165, 189,110, 98, 98, 108, 100, 98, 117, 118, 109,100, 106, 106, 100, 108, 108, 119, 131, 124,100, 102, 94, 100, 102, 96, 136, 116, 113,86, 74, 76, 88, 76, 76, 112, 115, 104,106, 100, 110, 106, 100, 108, 120, 118, 132,108, 110, 106, 106, 118, 106, 117, 118, 115,168, 188, 178, 170, 188, 182, 194, 191, 196,166, 150, 154, 164, 150, 154, 167, 160, 161,146, 142, 132, 144, 142, 130, 173, 161, 154,204, 198, 188, 206, 198, 188, 228, 218, 189,96, 94, 86, 96, 94, 84, 77, 89, 101,134, 126, 124, 132, 126, 124, 154, 156, 141,138, 144, 140, 140, 142, 138, 154, 155, 148,134, 136, 142, 136, 134, 140, 145, 154, 166,156, 160, 154, 156, 162, 156, 200, 180, 179,124, 138, 138, 122, 140, 136, 188, 147, 139,114, 110, 114, 112, 114, 114, 149, 217, 192,112, 116, 122, 112, 114, 124, 136, 132, 133,112, 116, 134, 114, 114, 136, 128, 125, 142,202, 220, 228, 200, 220, 226, 204, 222, 224,132, 136, 134, 134, 136, 132, 184, 187, 192,158, 162, 152, 158, 164, 150, 163, 160, 152,88, 76, 88, 90, 76, 86, 93, 88, 88,170, 174, 176, 172, 174, 178, 178, 181, 181,182, 176, 180, 184, 174, 178, 202, 199, 195,112, 114, 124, 112, 112, 126, 162, 166, 148,120, 118, 120, 118, 116, 120, 227, 227, 219,110, 108, 106, 110, 108, 106, 133, 127, 126,112, 112, 106, 112, 110, 106, 202, 190, 213,154, 134, 130, 156, 136, 132, 158, 121, 134,  116, 112, 94, 118, 114, 96, 124, 149, 137,108, 110, 114, 106, 110, 114, 114, 118, 126,106, 98, 100, 104, 100, 100, 137, 135, 134,122, 112, 112, 122, 114, 114, 121, 123, 128),  nrow = 85, ncol = 9, byrow = TRUE,  dimnames = list(NULL, c("J1","J2","J3","R1","R2","R3","S1","S2","S3")) )  #####################################################################  #Preparing the Blood Data  library(nlme)  blood = groupedData( y ~ meth | item ,  data = data.frame( y = c(Blood), item = c(row(Blood)),  meth = rep(c("J","R","S"), rep(nrow(Blood)\*3, 3)),  repl = rep(rep(c(1:3), rep(nrow(Blood), 3)), 3) ),  labels = list(BP = "Systolic Blood Pressure", method = "Measurement Device"),  order.groups = FALSE )    #pick out two of the three methods ( use J and S )    dat = subset(blood, subset = meth != "R") |