The International Journal of Biostatistics

Volume 7, Issue 1

2011

Article 9

An Improved Bland-Altman Method for Concordance Assessment

Jason J. Z. Liao, *Teva Pharmaceuticals USA*
Robert Capen, *Merck Research Laboratories*

Recommended Citation:

Liao, Jason J. Z. and Capen, Robert (2011) "An Improved Bland-Altman Method for Concordance Assessment," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 9.
DOI: 10.2202/1557-4679.1295

An Improved Bland-Altman Method for Concordance Assessment

Jason J. Z. Liao and Robert Capen

Abstract

It is often necessary to compare two measurement methods in medicine and other experimental sciences. This problem covers a broad range of data with applications arising from many different fields. The Bland-Altman method has been a favorite method for concordance assessment. However, the Bland-Altman approach creates a problem of interpretation for many applications when a mixture of fixed bias, proportional bias and/or proportional error occurs. In this paper, an improved Bland-Altman method is proposed to handle more complicated scenarios in practice. This new approach includes Bland-Altman's approach as its special case. We evaluate concordance by defining an agreement interval for each individual paired observation and assessing the overall concordance. The proposed interval approach is very informative and offers many advantages over existing approaches. Data sets are used to demonstrate the advantages of the new method.

KEYWORDS: concordance, Bland-Altman Method, measurement error model, agreement interval, bias

Author Notes: Please send correspondence to: Jason J. Z. Liao, Teva Branded Pharmaceutical Products R&D, Inc., 425 Privet Rd, Horsham, PA 19044; phone: 215-293-7259; email: jason.liao@tevausa.com. We thank the editor and two referees for their comments that improved the presentation of the paper.

1. Introduction

It is often necessary in medicine and other experimental sciences to assess the agreement between two measurement methods. Before a new method or a new instrument is adopted for use in measuring a variable of interest, its agreement relative to other similar evaluators needs to be assessed. Measurements of agreement are needed to assess the reliability of multiple raters (or the same rater over time) in a randomized clinical trial (RCT). As pointed out by Fleiss (1986), any elegant clinical trial design cannot overcome the damage by unreliable measurement. A good agreement among raters is very important and crucial for a clinical trialist. The agreement problem covers a broad range of data with applications arising from many different fields. Various questions regarding agreement can be posed such as the listings by Bartko (1994). Can the measurements from two methods be used interchangeably? How does one define and measure agreement? What is the overall level of agreement? How much bias and variance is there between methods? In short, there are two objectives for an agreement study. The first and the most important one is to determine if the measurements from the two methods agree; and the second one is that if they do not, then provide the sources of disagreement such as the level of fixed and/or proportional bias and the unacceptable residual error in order to improve the agreement either by calibrating the two measurement methods and/or reducing the residual error via variance reduction procedure, if necessary, where the fixed bias indicates the systematic deviations as the average differences between the two measurement methods and the proportional bias is usually related to the range of measurements but can be also related to the quality of measurements. (Magari, 2002)

The agreement problem has a long history and can be traced back over 100 years to Pearson, who proposed the correlation coefficient which was later adopted to measure agreement. Following Liao and Capen (2008), the existing approaches can be classified into three categories. The first type of approach is the hypothesis testing approach, which tests the departure from the perfect agreement (i.e., intercept equal to 0 and slope equal to 1) using, for example, a functional or structural regression analysis. This type of approach heavily depends on the residual variance and can lead to rejecting a reasonably good agreement when this variance is small, but fail to reject a poor agreement when this variance is large (Lin, 1989). This type approach is therefore not appropriate for assessing agreement (Lin, 1989; Bland and Altman, 1999). The second category is an index approach such as the intraclass correlation coefficient (ICC), the concordance correlation coefficient (CCC) (Lin, 1989), and the improved concordance correlation coefficient (Liao, 2003). There are many critiques in the literature about using an index in assessing agreement (Bland and Altman (1986 & 1999),

Deyo, *et al.* (1991), Atkinson and Nevill (1997)). Major concerns about using an index approach are 1) it is very sensitive to the range of the measurements available in the sample and sensitive to sample heterogeneity – the greater this range, the higher the index; 2) it is not related to the actual scale of measurement or to the size of error which might be scientifically allowable; 3) the same index value has different meanings in different experiments.

The third category is an interval approach. Assume the difference $D_i = Y_i - X_i$ is approximately normally distributed, Bland and Altman (1986, 1999) proposed using the 95% interval of the observed differences

$$[\bar{D} - 2S_D, \bar{D} + 2S_D] \quad (1)$$

which they called the “limits of agreement” (LOA), to measure agreement, where

$$\bar{D} = \frac{1}{n} \sum_{i=1}^n D_i \text{ and } S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2.$$

These limits are compared to scientifically acceptable boundaries. The provided interval will ensure that 95% of all differences will fall into it if the differences are normally distributed and n is large. This method does not depend on the range of the sample, and the limits of agreement give some indication whether the discrepancy is acceptable in practice by comparing the limits to the acceptable boundary. As a supplement, they also proposed a mean-difference graphic that plots the difference D_i against the mean of the two measurements $M_i = (X_i + Y_i)/2$ along with the 95% limits of the difference, i.e., the limits of agreement.

Bland-Altman approach has been a favorite of medical researchers. It evaluates the agreement at each individual level and it is simple and intuitive to implement. The limits are directly linked to practitioners' subject knowledge. However, this approach does not make a conclusion about agreement, which is the first goal of an agreement study and it has two rather more interesting hidden, but often violated in practice, assumptions (Carstensen, *et al.*, 2008): 1) The variation of the differences is constant over the range of measurements. 2) The difference between the methods is constant over the range of measurements. Thus, the Bland-Altman approach creates a problem of interpretation when a mixture of fixed bias, proportional bias and/or proportional error occurs (Ludbrook, 1997). In cases where the additive agreement (i.e., fixed bias) is expected, for example, the test-retest situation, Bland-Altman's approach is applicable (Rousson, *et al.*, 2002). However, the most common cases in the biomedical/health science literature are the absolute agreement where no difference (i.e., no bias) is allowed (Stine (1989), Haber, *et al.* (2006)). When there is a fixed bias, the limits of agreement in equation (1) will cover all the differences. When there is a fixed and/or proportional bias between the two

measurements, the mean of the two measurements (M_i) is not a good metric for the true value. When there is a proportional bias between the two measurements, the mean of the difference for the two measurements ($\bar{D}$) is not a good metric for the bias and it cannot specify the magnitude of the fixed and proportional biases. Also S_D^2 is not a good metric for the residual variance. In addition, the mean of the two measurements (M_i) used in this approach is always a random variable even if one of the two measurements is a “gold” standard. Plus, the variance of the mean (M_i) from the two measurements can be larger than the variance of the difference (D_i) from these two measurements with a moderately large correlation. In addition, the mean-difference plot gives artifactual bias (i.e., the mean-difference plots showing a bias trend) information in measurements differing only in the extent of random error (Hopkins, 2004) because of the correlation between the difference and the mean. Furthermore, this artifactual bias can also be observed when the two measurement methods have substantially different locations and/or spread. The simple 95% interval of the difference generally does not give sufficient information about the agreement of the two methods because the limits will cover 95% of all the differences if the differences are normally distributed even when a fixed bias exists. When a poor agreement conclusion is reached, the level and type of bias between the two sets of measurements cannot be fully assessed, which is the second goal of an agreement study. As mentioned in the agreement literature, covariates involved in an agreement study cannot be incorporated into this approach.

Based on the above mentioned arguments, any new method developed for evaluating the concordance of two measurement methods should be able to fulfil the two goals of an agreement study and be very informative. When the agreement is poor, it should easily indicate what type and degree of bias exists so that appropriate calibration can be implemented, if needed. It is also desirable that the concordance assessment be able to check and adjust for covariates or factors that could potentially affect the conclusions from the study. In Section 2, an improved Bland-Altman interval approach with a supplemental statistical quality-control type graphic for assessing overall concordance is described. The improved Bland-Altman approach covers more complicated scenarios in practice. This approach is very informative and offers many advantages over existing approaches and includes the Bland-Altman approach as special case. Two data sets are used in Section 3 to set the agreement interval for each pair of measurement and to demonstrate the advantages of the new approach. Summary and discussion follow in Section 4.

2. An Improved Bland-Altman Interval Approach

In practice, a simple measure of agreement for each individual pair (i.e., within-individual between methods) is preferred. An obvious starting point is the difference between measurements for each pair. That is, we can judge the agreement of two measurement methods by deriving an agreement interval and then showing that the differences of the paired measurements fall within the specified interval. In other words, an agreement interval, Δ , is defined and a pair of measurements is claimed to be “in agreement” at a specified level if their difference is within the interval. The two methods are considered agree with each other if no more than k observed paired difference fall outside the agreement interval Δ . In practice, $k=0$ is usually chosen. This means that two methods agree only if all of the paired differences fall within the agreement interval Δ (i.e., every pair of measurements is in agreement). Note that k must be pre-specified. A non-zero k can be chosen to accommodate the tolerance for accepting pairs of measurements not in agreement.

The measurement difference interval is a very practical and reasonable approach. Let $(X_i, Y_i), i=1, \dots, n$, be the n pairs of observations, which might represent reportable values or transformed values, such as the log-transformed values. Both X and Y are allowed to contain measurement error. Consider a linear measurement error model (Fuller (1987), Casella & Berger (1990)) as follows:

$$Y_i = Y_i^0 + \varepsilon_i = a_0 + b_0 \times X_i^0 + \varepsilon_i \quad (2)$$

$$X_i = X_i^0 + \delta_i \quad (3)$$

where $(X_i^0, Y_i^0), i=1, \dots, n$ are the unobserved (fixed) true values, ε_i and δ_i are independently normally distributed with zero means and equal variance, σ^2 .

Remarks:

1. When one measurement is a “gold” standard, i.e., there is no measurement error, the above defined model in (2) and (3) is still valid and it just reduces to the classical regression.
2. When there is no information about which measurement is a reference, the above defined model is ready to be used because of the symmetric position of two measurements.
3. Note in theory, there are possible many disagreement sources. However, it is reasonable to assume three possible disagreement sources (fixed and/or proportional bias, unacceptable residual variance, i.e., measurement imprecision) for practical use, and these three sources will be identified by the measurement error model. The imprecision needs to be assessed, and usually is done in practice, before any agreement testing. (Magari, 2002)

In general, $Y-X$ is normally distributed with mean $a_0 + (b_0 - 1) \times X^0$ and variance $2\sigma^2$. Under the assumption of perfect agreement ($a_0 = 0, b_0 = 1$), $Y-X$ will have a normal distribution with mean zero and variance $2\sigma^2$. Therefore, the agreement interval Δ is defined such that $1 - \alpha = P(Y - X \in \Delta)$, and can be constructed as follows.

$$\Delta = \left(-t_{1-\alpha/2, n-1} \sqrt{2}\hat{\sigma}, +t_{1-\alpha/2, n-1} \sqrt{2}\hat{\sigma} \right) \quad (4)$$

where $t_{1-\alpha/2, n-1}$ is the $100 \times (1 - \alpha/2)^{th}$ quantile of a t -distribution with degrees of freedom $n-1$ and $\hat{\sigma}^2 = \frac{1}{n} \left(S_{xx} - \frac{S_{xy}^2}{\hat{b}_0} \right)$ is obtained from the model defined by

equations (2) and (3), $\hat{b}_0 = \frac{-(S_{xx} - S_{yy}) + \sqrt{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}}{2S_{xy}}$, $S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2$,

$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$, $S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2$, $S_{xy} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$, $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. Thus, if the

difference of any paired measurement falls within the interval defined by equation (4), it is claimed to be in agreement; and the two measurement methods agree if no more than the pre-specified k number of differences fall outside of this interval.

Remarks:

1. The value $t_{1-\alpha/2, n-1} \sqrt{2}\hat{\sigma}$ represents the minimum difference between two methods that needs to be exceeded to be fairly certain that a real “change” has occurred (Eliaszew, *et al.*, 1994). Thus, the agreement interval defined in (4) indicates the possible range for the difference if there is no real “changes” occurred.
2. If there is no bias or only fixed bias exists, then $\sqrt{2}\hat{\sigma}$ in the new approach is equivalent to S_D in equation (1). Thus, if there is no bias between the two methods, then the interval defined in equation (4) from this new approach is the same as the limit of agreement interval in equation (1) from Bland-Altman’s approach except the choice of the critical value.
3. Equation (4) is used to assess the absolute agreement where only zero difference (i.e., no bias) is allowed. It is the most common case in biomedical/health science literature.

4. If one of the measurements is a “gold” standard measurement, then $\sqrt{2}\hat{\sigma}$ should be replaced by $\hat{\sigma}$ when constructing the interval in equation (4) and the following equations (7), (8) and (9) correspondingly.

The agreement interval Δ should be compared to a scientifically acceptable boundary, if such exists. The purpose for this is to identify the unacceptable residual variance as the disagreement source. If the agreement interval is too wide to be considered scientifically acceptable, then actions should be taken to reduce the residual variance σ^2 , which can and usually in practice be done in the design stage. The data can be graphically displayed, as in Figure 1, to allow for better visual assessment and diagnosis. In the figure, the differences are plotted against sample numbers. They can also be plotted against the sorted samples based on X values if a proportional bias is found. Note that this figure is essentially a statistical process control chart (Montgomery, 1996) where the “control” limits are the agreement interval calculated in equation (4). Note that this control chart type plot is mainly used to draw a conclusion about the agreement of two measurement methods, not to assess the bias information. The bias is obtained from the model in equations (2) and (3) and can be evaluated as follows.

Figure 1: Graphical description of the proposed interval approach using a statistical process control-type technique, where the dotted line is the perfect agreement line and the two solid lines define the agreement window.

An approximate $100 \times (1 - \gamma)\%$ confidence interval for the fixed bias is

$$\hat{a}_0 - \frac{z_{\gamma/2} \hat{\sigma}_a}{\sqrt{n}} \leq a \leq \hat{a}_0 + \frac{z_{\gamma/2} \hat{\sigma}_a}{\sqrt{n}} \quad (5)$$

where $z_{\gamma/2}$ is the $100 \times (\gamma/2)^{th}$ quantile of a standard normal distribution,

$\hat{a}_0 = \bar{Y} - \hat{b}_0 \times \bar{X}$, $\hat{\sigma}_a^2 = \frac{\hat{\sigma}^2}{n} + \bar{X}^2 \hat{\sigma}_b^2$, and $\hat{\sigma}_b^2 = \frac{(1 + \hat{b}_0^2)(S_{xx}S_{yy} - S_{xy}^2)}{(S_{xx} - S_{yy})^2 + 4S_{xy}^2}$. Similarly, an

approximate $100 \times (1 - \gamma)\%$ confidence interval for the proportional bias at X_i is

$$\left[(b_0^L - 1) \times X_i, (b_0^U - 1) \times X_i \right] \quad (6)$$

where $b_0^L = \hat{b}_0 - \frac{z_{\gamma/2} \hat{\sigma}_b}{\sqrt{n}}$, $b_0^U = \hat{b}_0 + \frac{z_{\gamma/2} \hat{\sigma}_b}{\sqrt{n}}$ and X_i is the observation from the

reference measurement (Casella and Berger, 1990).

This new approach improves the Bland-Altman's approach and includes it as a special case. However, the new approach is different from Bland-Altman's approach in a number of ways. First, generally speaking, the limit of agreement (LOA) in Bland-Altman's approach is the process (i.e., the difference) capability limits, not necessary for the "expected" process (i.e., agreement) behavior. In contrast, the proposed agreement interval is the process capability limits for the "expected" process behavior. Thus, a more appropriate name for LOA should be the limits of capability (LOC). Second, this new approach uses a statistical process control tool to make an overall agreement conclusion by pre-choosing a k to accommodate the tolerance in accepting an agreement conclusion. Third, if the two methods are discordant and it is desirable to calibrate them, then bias information will be needed. This is easy to obtain with the new approach. The total bias can be assessed by $\hat{a}_0 + (\hat{b}_0 - 1) \times X_i^0$, where $\hat{a}_0 = \bar{Y} - \hat{b}_0 \bar{X}$ is the fixed bias and $(\hat{b}_0 - 1) \times X_i^0$ is the proportional bias.

In some cases, it is known a priori that a fixed and/or proportional bias exists between the two methods. For example, in the context of test versus re-test, there is generally an improvement in the re-test scores because of the learning gained from the first test. Thus, perfect agreement could not be achievable without first accounting for the expected fixed bias, which should not be counted as discordant (Rousson, *et al.*, 2002). Another example occurs in the context of two measurement scales for the same subject when there is a one-to-one monotonic relation between the two scales. Existing methods cannot handle this situation. However, the proposed approach can by simply adjusting the agreement interval. The agreement interval Δ can be modified as

$$\Delta = \left(a_0 - t_{1-\alpha/2, n-1} \sqrt{2} \hat{\sigma}, a_0 + t_{1-\alpha/2, n-1} \sqrt{2} \hat{\sigma} \right), \quad (7)$$

for a fixed bias a_0 ($a_0 \neq 0, b_0 = 1$), or

$$\Delta_i = ((b_0 - 1) \times X_i - t_{1-\alpha/2, n-1} \sqrt{2\hat{\sigma}}, (b_0 - 1) \times X_i + t_{1-\alpha/2, n-1} \sqrt{2\hat{\sigma}}), \quad (8)$$

for a proportional bias $(b_0 - 1)X_i$ ($a_0 = 0, b_0 \neq 1$), or

$$\Delta_i = (a_0 + (b_0 - 1) \times X_i - t_{1-\alpha/2, n-1} \sqrt{2\hat{\sigma}}, a_0 + (b_0 - 1) \times X_i + t_{1-\alpha/2, n-1} \sqrt{2\hat{\sigma}}), \quad (9)$$

for both fixed bias a_0 and proportional bias $(b_0 - 1)X_i$ ($a_0 \neq 0, b_0 \neq 1$), where X_i is the observation from the reference measurement.

Remarks:

1. Equation (7) is used to assess additive agreement where a fixed difference (bias) is allowed. When there is only fixed bias, the interval defined in equation (7) from this new approach is the same as the interval in equation (1) from Bland-Altman's approach except the choice of the critical value. Thus, the limits of agreement in Bland-Altman approach is the bias adjusted agreement interval in the new approach when there is only fixed bias.
2. Equations (8) and (9) are used to assess multiplicative and linear agreement, respectively (Stine, 1989). Examples of the additive, multiplicative and linear agreement can be found in Stine (1989).
3. Usually the fixed bias a_0 and the proportional bias factor b_0 are unknown and they need to be estimated. However, if a_0 and b_0 are known, one can also simply transfer the measurement Y first and then construct the agreement interval.
4. Bland and Altman (1999) used the regression of the difference D_i against the mean M_i to find the limit of agreement when there are biases. However, the results from this regression can be misleading. First, the variance of the mean M_i could be much bigger than the variance of the difference D_i . Second, as mentioned early, regression of the difference D_i against the mean M_i will give artifactual bias information and is highly misleading when the difference D_i and the mean M_i are correlated. Third, on other hand, this regression may mis-present the true bias information. For example, the regression of the difference D_i against the mean M_i indicates no proportional bias for data (1, 2), (1.5, 1.5), (2, 1) and (3, 0) which however has a true proportional bias and this true bias can be detected by the new approach.

The proposed approach can also be used in the case when the two measurement error variances are not assumed equal (i.e., proportional error), for example, $\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2} > 1$, where $\sigma_\delta^2 = \text{Var}(\delta)$ and $\sigma_\varepsilon^2 = \text{Var}(\varepsilon)$. In practice, the reliability ratio λ is usually around 1 and the unequal variance situation usually can be avoided in the design stage through replication. For example, let m_x and m_y be the number of replicate determinations made on X and Y, respectively, for each subject i . Then m_x and m_y would be chosen so that $\text{Var}(X_i) = \text{Var}(Y_i)$. When the reliability ratio λ is known, the estimates in the linear measurement error model defined by equations (2) and (3) are unique (Fuller (1987), Casella & Berger (1990)). In this case, one possibility for the agreement interval could be chosen as

$$\Delta = \left(-t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta, +t_{1-\alpha/2, n-1} \sqrt{1 + \lambda} \hat{\sigma}_\delta \right). \quad (10)$$

In many real life examples, an agreement study often includes covariates or “nuisance” factors, that should be evaluated and adjusted. The proposed approach can handle this issue by adding covariate terms in the linear measurement error model.

3. Two Examples

3.1. Plasma Volume Data Set

Consider the data reported in Table 2 in Bland and Altman (1999). The data include the measurements of plasma volume expressed as a percentage of the expected value in 99 healthy individuals using two alternative sets of normal values due to Nadler and Hurley. Per the analysis done in Bland and Altman (1999), a logarithm transformation of the data is needed to assess the agreement of the measurements of plasma volume for Nadler and Hurley normal values. Figure 2 shows the scatterplot of the log-transformed data.

Figure 2: Agreement scatterplot of $\log(\text{Hurley})$ and $\log(\text{Nadler})$ for the plasma volume data. Solid line: perfect agreement line $\log(\text{Hurley}) = \log(\text{Nadler})$.

Figure 3: Concordance Assessment for the plasma volume data. Dotted line: perfect agreement line or estimated bias; Solid line: interval boundary. Left panel: agreement interval for the new approach; Middle panel: agreement interval with bias adjusted for the new approach; Right panel: limit of agreement for Bland-Altman approach.

Consider the linear measurement error model,

$$\begin{aligned}\log(\text{Nadler}) &= a_0 + b_0 \times \log(\text{Hurley})^0 + \varepsilon \\ \log(\text{Hurley}) &= \log(\text{Hurley})^0 + \delta\end{aligned}$$

where ε and δ are independently and identically distributed normal random variables with mean 0 and variance σ^2 , and $\log(\text{Hurley})^0$ is the unobserved true value. The results are $\hat{a}_0=0.113$, $\hat{b}_0=0.997$, $\hat{\sigma}=0.015$. Based on 98 degrees of freedom, the agreement interval in (4) was $\Delta = (-0.043, 0.043)$ in the log-scale and it was $(-4.193, 4.377)$ in terms of the % difference in the measurement method raw scale. To visually interpret the results, the arithmetic differences of $\log(\text{Nadler})-\log(\text{Hurley})$ were plotted against the observation numbers in the left

panel of Figure 3. The left panel indicates that almost all the differences are above the upper limit of the agreement interval. Thus, some real “changes” have occurred. To investigate what caused these “changes”, we look at the fixed-bias and the proportional bias within the observed plasma volume range due to Hurley. The log-total bias (fixed and proportional bias) ranges from 0.099 to 0.101, and the log-proportional bias ranges from -0.015 to -0.013. Thus, there is a constant fixed-bias $\hat{a}_0 = 0.113$ with an ignorable proportional bias. If this fixed bias is not considered as a discordant factor, then the bias-adjusted agreement interval using (7) is $\Delta = (0.070, 0.156)$ in the log-scale and the plot is shown in the middle panel of Figure 3. In comparison, the Bland-Altman approach for the data gave $\bar{D} = 0.099$ and $S_D = 0.022$ and the limits of agreement (LOA) in equation (1) were $(0.056, 0.141)$ in the log-scale. The mean-difference plot is shown in the right panel of Figure 3, which is very similar to the graph in the middle panel. Thus, the Bland-Altman’s limit of agreement is the bias-adjusted agreement interval defined in this paper but with different bias-adjustment.

Figure 4: Agreement scatterplot of Baseline SIP and two-week Follow-up SIP scores for the low back pain data. Solid line: perfect agreement line $Baseline = Follow-up$.

3.2. Low Back Pain Data Set

Consider a data set reported in Deyo, *et al.* (1991). The data were from a randomized trial of treatments for chronic low back pain. All the patients reported having low back pain for at least three months and completed at least two weeks in the trial. Patients completed a series of questionnaires at baseline, two weeks, four weeks, and three months. The Sickness Impact Profile (SIP) was used to assess the sickness-related dysfunction in 12 different categories and to produce a score for each category. At each visit, both patients and clinicians were asked to rate subject improvement on a six-point ordinal scale. If both patient and clinician indicated any degree of improvement on this scale, the patient was classified as improved. The classification was based on clinician and patient assessment without SIP scores available. This consensus judgment was used as an external standard for examining the association of SIP score changes with clinical changes in patient status. Figure 4 shows the SIP scores at the baseline and the two-week for those 34 patients classified as improved.

Consider the linear measurement error model,

$$\begin{aligned} \text{Follow-up} &= a_0 + b_0 \times \text{Baseline}^0 + \varepsilon \\ \text{Baseline} &= \text{Baseline}^0 + \delta \end{aligned}$$

where ε and δ are independently and identically distributed normal random variables with mean 0 and variance σ^2 , and Baseline^0 is the unobserved true value. The results are $\hat{a}_0 = -21.126$, $\hat{b}_0 = 1.924$, $\hat{\sigma} = 5.653$. Based on 33 degrees of freedom, the agreement interval in (4) was $\Delta = (-16.264, 16.264)$. To visually interpret the results, the arithmetic SIP differences of $(\text{Follow-up} - \text{Baseline})$ were plotted against the observation numbers in the upper left panel of Figure 5. The upper left panel indicates that most of the differences are inside the agreement interval. Thus, not many real “changes” have occurred. In comparison, the Bland-Altman approach for the data gave $\bar{D} = -5.397$ and $S_D = 8.488$ and the limits of agreement (LOA) in equation (1) were $(-22.372, 11.578)$. The mean-difference plot is shown in the upper right panel of Figure 5, which shows an increasing trend, i.e., proportional bias. Note that when there is a proportional bias, the bias estimate $\bar{D}$ and residual variance estimate S_D^2 from Bland-Altman approach are not valid. To investigate this increasing trend, the regression of the difference D_i against the mean M_i was used as suggested in Bland and Altman (1999). The difference against the mean regression gives $D = -9.791 + 0.307 \times M$. If using original SIP scores from baseline and follow-up, this will lead to $\text{Follow-up} = -11.564 + 1.362 \times \text{Baseline}$. When comparing to the bias

information in the improved new approach from the measurement error model which leads to $Follow-up = -21.126 + 1.924 \times Baseline$, the difference versus the mean regression underestimates the bias for this data set, which was also mentioned in the remark 4 in Section 2. If using the method described in Bland and Altman (1999) for constructing the limits of agreement for biased data, the lower limit of agreement is $D = -7.518 - 0.821 \times M$ and the upper limit of agreement is $D = -12.064 + 1.434 \times M$ and the mean difference plot is shown in the lower left panel of Figure 5.

Figure 5: Concordance Assessment for the low back pain data. Dotted line: perfect agreement line or estimated bias; Solid line: interval boundary. Upper left panel: agreement interval for the new approach; Upper right panel: limit of agreement for Bland-Altman approach; Lower left panel: limit of agreement adjusted for bias in Bland-Altman approach; Lower right panel: agreement interval adjusted for bias in the new approach for sorted samples.

As a comparison to Bland-Altman's approach, the lower agreement limit after bias adjustment using (9) in this paper is $Follow-up = -37.390 + 0.924 \times Baseline$

and the upper agreement limit using (9) in this new approach after bias adjustment is $Follow-up = -4.862 + 0.924 \times Baseline$. The arithmetic SIP differences of $(Follow-up - Baseline)$ were also plotted against the sorted observation samples in the lower right panel of Figure 5. Note the difference of the bias-adjusted boundaries from these two approaches. In addition to the center (i.e., the bias) changes, the length of the interval is getting wider for Bland-Altman's approach. However, the length of the interval of the new approach is constant but the center (i.e., the bias) is different. Note that the curve increases dramatically at the last sample point in the lower right panel of Figure 5. This is due to the dramatic big difference in terms of the baseline SIP value between the last sample point with a baseline SIP value 42.1 and its previous sample point with a baseline SIP value 28.8 as indicated clearly in Figure 4.

4. Summary and Discussions

In this paper, a very informative interval approach was proposed to assess the agreement of two measurement methods. This approach improves Bland-Altman's approach to cover more complicated practical scenarios, and it includes the Bland-Altman's approach as special case. It defines an agreement interval to evaluate each individual difference and can easily catch any existing bias (fixed, proportional or both) through a measurement error regression analysis. If a "not in agreement" conclusion is reached, the proposed approach clearly indicates what type of bias exists and provides a way to estimate it for calibration purposes, which is an important issue for practitioners. In this regard, the proposed approach is a disaggregate approach as classified and described in Barnhart, *et al.* (2007). When there are covariates, such as a subject's age and gender, involved in the experiment, then the concordance assessment can be adjusted for these covariates by adding the covariates in the linear measurement error model defined by equations (2) and (3). The details will be presented in a separate paper.

As seen in the agreement literature, the two sets of measurements can be generated from two gene sequences in disease mapping, two analysts, two laboratories, two instruments (old vs. new), two treatments, two clinical endpoints, two formulations (bioequivalence), two assays (Liao, *et al.*, 2006), two batches, two raters, test vs. re-test, and two models in the context of evaluating the goodness-of-fit of a generalized linear and/or non-linear mixed-effects model where agreement is assessed between the observed and predicted results (Vonesh, *et al.*, 1996). Thus, the proposed approach can also be used in these applications.

With the proposed approach, the agreement interval Δ can be constructed using equations (7), (8) and (9) for a fixed bias, proportional bias, or both, respectively. The proposed method can also be used when there is proportional error by using the Δ in equation (10). However, existing approaches cannot

handle these situations appropriately. Note that proportional error case is usually assessed before a formal agreement study, and it can usually be avoided in the design stage through appropriate replication to produce equal variances. For example, in comparing two analytical methods, if the current method has a precision (%RSD) of 5% and the new method has a precision of 3%, then defining a “result” as the average of triplicate determinations for the current method leads to a precision approximately equal to the precision of the new method. If the data driven interval Δ is too wide to be scientifically defensible, then actions should be taken to reduce the variance σ^2 or increase the sample size so that Δ is scientifically acceptable. When the variance is not constant within the range, then the agreement interval can be extended to an agreement curve over the range. If there is a statistically significant covariate effect on the residual variance, then different agreement intervals can be used for each covariate level.

References:

- Atkinson, A. and Nevill, A. (1997). Comment on the Use of Concordance Correlation to Assess the Agreement between Two Variables. *Biometrics*, **53**, 775 – 777.
- Barnhart, H.X., Haber, M., and Lin, L.I. (2007), An Overview on Assessing Agreement with Continuous Measurement. *Journal of Biopharmaceutical Statistics*, **17**, 529 – 569.
- Bartko, J.J. (1994). Measurements of Agreement: A Single Procedure. *Statistics in Medicine*, **13**, 737 – 745.
- Bland, J.M. and Altman, D.G. (1986). Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *Lancet*, **2**, 307 – 310.
- Bland, J.M. and Altman, D.G. (1999). Measuring Agreement in Method Comparison Studies. *Statistical Methods in Medical Research*, **8**, 135 – 160.
- Carstensen, B., Simpson, J. and Gurrin, L.C. (2008), Statistical Models for Assessing Agreement in Method Comparison Studies with Replicates Measurements. *The International Journal of Biostatistics*, **4**(1), Article 16, 1 – 26.
- Casella, G. and Berger, R.L. (1990). *Statistical Inference*. Duxbury Press. Belmont, California.
- Deyo, R. A., Diehr, P. and Patrick, D. L. (1991). Reproducibility and Responsiveness of Health Status Measures: Statistics and Strategies for Evaluation. *Controlled Clinical Trials*, **12**, 142S – 158S.

- Eliasziw, M, *et al.* (1994). Statistical Methodology for the Concurrent Assessment of Interrater and Intrarater Reliability: Using Goniometric Measurements as an Example. *Physical Therapy*, **74**, 777 – 788.
- Fleiss, J.L. (1986), *The Design and Analysis of Clinical Experiments*. John Wiley & Sons, New York.
- Fuller, W.A. (1987). *Measurement Error Models*. John Wiley & Sons. New York.
- Haber, M. and Barnhart, H.X. (2006). Coefficients of Agreement for Fixed Observers. *Statistical Methods in Medical Research*, **15**, 255 – 271.
- Hopkins, W.G. (2004). Bias in Bland-Altman but not Regression Validity Analyses. *Sportscience*, **8**, 42 – 46.
- Liao, J.J.Z. (2003). An Improved Concordance Correlation Coefficient. *Pharmaceutical Statistics*, **2**(4), 253 – 261.
- Liao, J.J.Z. and Capen, R.C. (2008). Multiple Evaluators. In *Wiley Encyclopedia of Clinical Trials* (Ralph D'Agostino, ed.), Volume 3, pp. 186 – 194. Hoboken: John Wiley & Sons, Inc.
- Liao, J.J.Z., Capen, R.C. and Schofield, T.L. (2006), Assessing the Reproducibility of an Analytic Method, *J. of Chromatographic Science*, **44**, 119 – 122.
- Lin, L. I-K. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics*, **45**, 255 – 268.
- Lin, L.I-K. (2000). Total Deviation Index for Measuring Individual Agreement with Applications in Laboratory Performance and Bioequivalence. *Statistics in Medicine*, **19**, 255 – 270.
- Lin, L.I-K., Hedayat, A.S., Sinha, B. and Yang, M. (2002). Statistical Methods in Assessing Agreement: Models, Issues, and Tools. *J. of American Statistical Association*, **97**, 257 – 270.
- Ludbrook, J. (1997). Comparing Methods of Measurement. *Clinical and Experimental Pharmacology and Physiology*, **24**, 193 – 203.
- Magari, R. T. (2002), Statistics for Laboratory Method Comparison Studies. *BioPharm*, January, 28 – 32.
- Montgomery, D.C. (1996). *Introduction to Statistical Quality Control*. Third edition. Wiley, New York.
- Rousson, V., Gasser, T. and Seifert, B. (2002). Assessing Intrarater, Interrater and Test-retest Reliability of Continuous Measurements. *Statistics in Medicine*, **21**, 3431 – 3446.
- Stine, W.W. (1989). Interobserver Relational Agreement. *Psycho. Bull.* **106**, 341 – 347.
- Vonesh, E.F., Chinchilli, V.M., Pu, K. (1996). Goodness-of-fit in Generalized Nonlinear Mixed-effects Models. *Biometrics*, **52**, 572 – 587.