Journal of Biopharmaceutical Statistics, 19: 150–173, 2009

ISSN: 1054-3406 print/1520-5711 online
DOI: 10.1080/10543400802535141

Taylor & Francis
Taylor & Francis Group

AN APPLICATION OF LINEAR MIXED EFFECTS MODEL TO ASSESS THE AGREEMENT BETWEEN TWO METHODS WITH REPLICATED OBSERVATIONS

Anuradha Roy
Department of Management Science and Statistics,
The University of Texas at San Antonio, San Antonio, Texas, USA

We study the problem of assessing the agreement between two methods with any number of replicated observations using linear mixed effects (LME) model with Kronecker product covariance structure in a doubly multivariate set-up. This method can also be used in the case of unbalanced designs when number of replications on each patient is unequal, as well as when the number of replications on each patient by respective methods is unequal. The model is implemented using the MIXED procedure of SAS. We demonstrate our proposed method with three real datasets.

Key Words: Assessment of agreement; Kronecker product covariance structure; Linear mixed effects model; *Proc Mixed*; Replicated observations.

1. INTRODUCTION

It is often necessary to compare a new measurement technique with an established one, measuring some quantity, such as carbon dioxide production, blood pressure, body fat, child's weight, or even measuring bone mineral density in children. The simple and relatively inexpensive methods for gathering quantitative data as compared to the expensive gold standard methods are always valued. It is often needed to see whether they agree so that they both can be used interchangeably. The question to be answered in this paper is, "Do the two methods of measurement agree statistically?" so that one can switch them if needed. The problem has been discussed by many authors (Argall et al., 2003; Barnhart et al., 2005, 2007; Bartko, 1994; Bland and Altman, 1983, 1986, 1990, 1999; Choudhary and Nagaraja, 2005, 2007; Haber et al., 2005; Lee et al., 1989; Lin, 1989; Lin et al., 2002; Quiroz, 2005; St. Laurent, 1998). However, all these authors except Bland and Altman (1986, 1999, 2007), Barnhart et al. (2005, 2007), Haber et al. (2005), and Quiroz (2005) used only a single measurement on each subject for each method. Most of these approaches used various factors, such as a systematic bias, a difference in variabilities (random errors), and a low correlation for measures of disagreement. Choudhary and Nagaraja (2005) used all three factors for

Received January 16, 2007; Accepted May 11, 2008
Address correspondence to Anuradha Roy, Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, Texas 78249, USA; E-mail: Anuradha.Roy@utsa.edu

assessment of agreement between two methods, and they used the intersection-union principle to deal with the multiple testing problem. Choudhary and Nagaraja (2007) extended their approach to incorporate a continuous covariate (Choudhary, 2007; Choudhary and Ng, 2006) and to deal with data with replications and longitudinal measurements (Choudhary, 2008). Lin et al. (2002) reviewed the literature and presented methodologies in measurements of agreement in terms of coverage probability. Barnhart et al. (2007) mentioned that when there is a disagreement between two methods, one must know whether the source of disagreement arose from systematic bias or random error, as a systematic bias can be fixed with ease through calibration, while a random error is often a more cumbersome exercise of variation reduction.

Bland and Altman (1999) in their renowned article pointed out correctly that a single measurement on each subject cannot judge which method is more precise; lack of preciseness can certainly interfere with the comparison of two methods. Thus, they strongly recommended simultaneous estimation of repeatability and agreement by collecting replicated data, even though for more than two replicated measurements their calculations were not totally convenient, especially for the case with unequal replications. If a single measurement is taken by each method, we can only examine the overall agreement between two methods. If there is substantial disagreement between two methods, we would not be able to find out if it is due to the replication error (intra-method) if we do not have replications. A method cannot be useful unless it gives the same or similar results when operated repeatedly to the same individual within a time period too short for any change to take place. Thus, replications help us to segregate out the sources of disagreement, and assess total, inter-method, and intra-method agreement (Barnhart et al., 2005) separately, where inter-method agreement requires measurement of between-subject variation and intra-method agreement requires measurement of within-subject variation. If one method has poor repeatability in the sense of considerable variation, the agreement between the two methods is bound to be poor. Even if the old method has more variation, a new method which is perfect will not agree with it. Thus, it is important to report repeatability when assessing measurement, because it measures the purest random error (Barnhart et al., 2007) that is not influenced by any other factors. Sometimes genuine replicates cannot be obtained, in those cases we have a loose sense of repeatability based on some assumptions. Like Bland and Altman (1999), by replicates we also mean two or more measurements on the same individual taken under identical conditions. In general this means that the measurements are taken in a short period of time so that no real biological changes can take place. We can thus assume that the variances and covariances of these replicated measurements are homogeneous, in other words, we can assume that these replicated measurements are equicorrelated, and we must take this equicorrelated correlation structure of the replicates into account when assessing the agreement between two methods. Bland and Altman (1999) used a method of moment approach to estimate the bias and the repeatability coefficient of each method. They calculated the repeatability coefficient of each method using within-subject variance, but regrettably they did not test the agreement between them formally. They also did not test the bias between two methods in a formal way, however, their method may be used for testing the presence of bias using the confidence interval approach. These two authors explored the agreement between two measurement methods by asking the question, "Do the

two methods of measurement agree sufficiently closely?" and they answered this question by estimating two limits of agreement. Because of its easiness and intuitive appeal, limits of agreement are used extensively in medical literature for assessing agreement between two methods. Nevertheless, we study the problem of assessing agreement between two methods by fitting linear mixed effects (LME) model with Kronecker product covariance structure in a doubly multivariate set-up, instead of straightforward graphical techniques and tedious statistical calculations. This LME model is very easy to implement using *PROC MIXED* of *SAS* and the results are also easy to interpret.

As mentioned in Barnhart et al. (2007), sources of disagreement may arise from differing population means, differing between-subject variances, differing within-subject variances between two methods, and poor correlation between measurements of the two methods. Many authors used correlation coefficient as one factor of method comparison, nonetheless many authors did not, as it is inappropriate for assessing random error. (Correlation coefficient-type approaches based on a bivariate normal distribution of the data are given in Bartko, 1994; Lee et al., 1989; Lin, 1989; St. Laurent, 1998.) Argall et al. (2003) used Pearson correlation coefficient to compare two methods of weight estimation and described that the correlation coefficient 0.82 is good in comparing two methods. Lee et al. (1989) claimed that intraclass correlation coefficient (ICC) is the correct statistic for measuring agreement between two methods instead of Pearson correlation coefficient. Lin (1989) objected to the use of ICC and developed an alternative called the concordance correlation coefficient (CCC). There are vast literatures on repeated measurements that deal with different kinds of correlation coefficients due to repeated measurements. Lam et al. (1999) estimated the correlation coefficient between two variables with repeated observations on each variable. Hamlett et al. (2003, 2004) and lately Roy (2006) estimated it by using LME model. Roy modeled the true overall correlation coefficient between two variables by calculating it in two parts: the within-subject correlation coefficient between two variables added to the subject effect to it. The ICC and CCC for replicated observations are discussed in Barnhart et al. (2005). Haber et al. (2005) used ICC and CCC, and found that both of these values are equal to 0.997 for their replicated coronary artery calcium data; confirming that there is a perfect agreement between the two radiologists. However, their new coefficient of interobserver variability (CIV) suggests that the agreement between the two radiologists is less than perfect. Haber et al. then claimed that ICC and CCC are unable to reflect the observer disagreement when the between-subject variability is substantially higher than the within-subject variability that is present in the coronary artery calcium data. Our method also concludes that there is a perfect disagreement between the two radiologists due to random error (p-value $= 6.5095E$-9), although the overall correlation coefficient between the two radiologists is 0.9957. These observations suggest that correlation coefficient type index may indeed be misleading, and thus, may not be used for all kinds of datasets, especially for datasets with replications. If it is used at all, it should be used with great caution. At this point we must note (Roy, 2006) that for datasets with replicated observations, there are three other overall correlation coefficients besides the overall correlation coefficient between the two methods, such as the overall correlation coefficient between two methods for two different replicates and the overall correlation coefficient between any two different replicates of

the two methods. Overall correlation coefficient between two methods and the overall correlation coefficient between two methods for two different replicates provide measures of association between two methods, and high values of these two suggest a good association between the two methods. The overall correlation coefficients between any two different replicates of the two methods offer measures of association among the replicates of the two methods, and high values of these two suggest good associations among the replicates of the two methods. Even though high values of these four correlation coefficients appear to provide some indication of assessing the agreement between two methods, in reality they do not, and we will see that later in Section 5. To see how correlation coefficient may depend on the internal structure of the data with replicated observations, further study using extensive simulation is needed. Correlation is related to, but not synonymous with, agreement. Bland and Altman (1986) effectively demonstrated an example where correlation was excellent, but agreement was not. The differences in correlation and agreement between the male and female volunteers in Bowling et al. (1993) further strengthen this argument. Thus, we strongly feel that correlation coefficient type of measure is inappropriate for assessing random errors, and hence in this paper we do not use correlation coefficient as one of the main factors for the assessment of agreement, but we do report all four above-mentioned correlation coefficients, and use them as additional features or information.

In this article we propose a novel method for assessing the agreement between a new method (often easier or cheaper) and an established method (often delicate or harmful), with unbalanced data and with unequal replications for different subjects. We do this by fitting LME model with Kronecker product covariance structure in a doubly multivariate set-up, and properly testing the three factors, the bias, the between-subject variabilities, and the within-subject variabilities (agreement between the repeatability coefficients) of the two methods. Between-subject variability is needed when one is interested in the true or real difference between the two methods giving different measurements on the same subject, while within-subject variability is needed to calculate the random error among the replications taken by the same method on the same subject. By doubly multivariate set-up we mean the information in each patient is multivariate in two levels, the number of methods and the number of replicated measurements. Several authors (Boik, 1991; Chaganty and Naik, 2002; Galecki, 1994; Naik and Rao, 2001; Shults and Morrow, 2002; Shults et al., 2004) have observed many advantages of using Kronecker product structure over usual unstructured variance covariance matrix for analyzing doubly multivariate data. Naik and Rao (2001) used the structure $\boldsymbol{V} \otimes \boldsymbol{\Sigma}$, where $\boldsymbol{V}$ is compound symmetry structure to represent the correlation matrix between the repeated measures in their analysis of the data using a MANOVA model. Then Roy and Khattree (2007) used this same structure for discriminant analysis and the related hypotheses testing problems.

We approach the assessment of the agreement between two methods with replicates by using the maximum likelihood estimation where the replicated observations are linked over time. We can easily extend the method to situations where the replicated observations are not linked. To the best of the author's knowledge this is the first time that hypotheses testings on the bias, the between-subject variabilities, as well as the within-subject variabilities (repeatability coefficients) of two methods are accomplished in a formal way with any number

of replicated measurements. We propose the following three conditions, using the three factors as mentioned before, to verify whether two methods for measuring a quantitative variable can be considered interchangeable.

1. No significant bias, i.e., no difference between the means of the two methods.
2. No difference in the between-subject variabilities of the two methods.
3. No difference in the within-subject variabilities of the two methods, i.e., no difference between the *repeatability coefficients* (defined in Section 4.3) of the two methods.

Instead of proposing the last two conditions separately, one can integrate them together and propose one condition, such as no difference in the overall variabilities of the two methods, where overall variability is the sum of between-subject variability and within-subject variability. However, we cannot then separate out the two types of variabilities. Between-subject variabilities are crucial when one is concerned about the true difference between the two methods. Sometimes there can be significant difference in between-subject variabilities of the two methods, but not in within-subject variabilities. Testing of hypothesis on the difference in overall variabilities of two methods, along with other testings, is described in Section 4.

Albeit we are not taking correlation coefficient as a factor for interchangeability, we report all four correlation coefficients and use the overall correlation coefficient (Roy, 2006) as an added feature, along with the bias, the between-subject variabilities, and the within-subject variabilities, to assess the agreement between two methods. We maintain the value 0.82, like Argall et al. (2003), as the edge of the overall correlation coefficients while comparing two methods, but one can always change it according to one's requirements.

2. LINEAR MIXED EFFECTS MODEL

As mentioned in the introduction the number of replicated measurements on each patient or subject may not be equal, and also the number of replications of the two methods on the same subject may not be equal. Let p_i^e and p_i^n be the number of replications on subject i by the established method (e), and a new method (n) respectively. Let $p_i = \max(p_i^e, p_i^n)$, and $n_i = 2p_i$. Therefore, the number of observations on the ith subject is n_i, under the assumption that the ith subject has $|p_i^e - p_i^n|$ missing values. The missing values are assumed to be missing at random (MAR), i.e., the probability that a value is missing depends only on the observed values of the subject, but not on the missing values of the subject. Even if MAR does not hold, a procedure based on MAR is less biased than naive methods (e.g., data deletion). Let y_i represent a response (any combination of method and replication) of the ith subject; additional multiple subscripts and superscripts will be used on y_i for further description. As we compare two methods, we choose the intercept and the two methods as the fixed effects, as well as choosing the two methods as the random effects. Thus, the LME model for the ith subject can be written as

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + b_{1i} z_{i1} + b_{2i} z_{i2} + \epsilon_i.$$

In this model every method replication combination response is denoted separately with the single subscript i. The terms x_{i1} and x_{i2} are used to indicate the method to

which the response y_i belongs and will take the value one or zero—if the response comes from the first method, x_{i1}, then it will equal one (corresponding to β_1, which represents the first method effect), x_{i2} will equal zero. The quantity β_o is the intercept, β_1 and β_2 are fixed effects parameters corresponding to two methods, and b_{1i} and b_{2i} are random effects parameters corresponding to two methods. The terms z_{i1} and z_{i2} are used to indicate the method to which the response y_i belongs and will take the values one or zero in exactly the same way as x_{i1} and x_{i2}. We arrange all responses (all method replication combinations) of the ith subject one below the other, as described below.

Let y_{it}^e and y_{it}^n be the responses by the established method and a new method of the ith subject at the tth replicate, $i = 1, 2, \ldots, N,\ t = 1, 2, \ldots, p_i$. Let $\boldsymbol{y}_{it} = (y_{it}^e, y_{it}^n)'$ be the 2×1 vector of measurements corresponding to the ith subject at the tth replicate. Let $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}', \boldsymbol{y}_{i2}', \ldots, \boldsymbol{y}_{ip_i}')'$ be the $(n_i \times 1)$-dimensional random vector corresponding to the ith subject. That is, the vector $\boldsymbol{y}_i$ is obtained by stacking the responses of the established method and a new method at the first replication, then stacking the responses of the established method and the new method at the second replication, and so on. For example, suppose we have two methods, each with two replications. Then, $\boldsymbol{y}_i = (y_{i1}^e, y_{i1}^n, y_{i2}^e, y_{i2}^n)'$ will be a (4×1)-dimensional random vector corresponding to the ith subject. Since all responses from the same subject are stochastically dependent on each other, we must analyze all responses from the same subject together. Thus, we write all responses ($\boldsymbol{y}_i$) of the ith subject in a matrix equation as described in Laird and Ware (1982) as

$$\begin{aligned}
\boldsymbol{y}_i &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,\\
\mathbf{b}_i &\sim N_m(\mathbf{0}, \boldsymbol{D}),\\
\boldsymbol{\epsilon}_i &\sim N_{n_i}(\mathbf{0}, \boldsymbol{R}_i),
\end{aligned}$$

where $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N$, and $\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N$ are independent, and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ are also all independent. LME model allows for the explicit analysis of between-subject ($\boldsymbol{D}$) and within-subject ($\boldsymbol{R}_i$) sources of variation of the two methods. We define the two methods by a vector variable M_var; M_var $= e$ for the established method and M_var $= n$ for a new method. As mentioned previously, we choose the intercept and the vector variable M_var as fixed effects, thus the design matrix $\boldsymbol{X}_i$ has three columns, and consequently $\boldsymbol{\beta} = (\beta_o, \beta_1, \beta_2)'$ a 3-dimensional vector containing the fixed effects. We also choose the vector variable M_var as random effects, i.e., M_var is random across individual subjects; thus the design matrix $\boldsymbol{Z}_i$ has two columns. Therefore, $\boldsymbol{b}_i = (b_{1i}, b_{2i})'$ a 2-dimensional vector containing the random effects. Thus, for the above mentioned example of two methods and each with two replicates, we have

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1}^e \\ y_{i1}^n \\ y_{i2}^e \\ y_{i2}^n \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} + \begin{bmatrix} \epsilon_{i1}^e \\ \epsilon_{i1}^n \\ \epsilon_{i2}^e \\ \epsilon_{i2}^n \end{bmatrix},$$

where $\epsilon_{i1}^e, \epsilon_{i1}^n, \epsilon_{i2}^e$, and ϵ_{i2}^n are the random errors corresponding to the responses $y_{i1}^e, y_{i1}^n, y_{i2}^e$, and y_{i2}^n for the ith subject. In general, the quantity $\boldsymbol{\epsilon}_i$ is a n_i-dimensional

vector of residual components, and the $n_i \times 3$ and $n_i \times 2$ dimensional design matrices $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are:

$$\boldsymbol{X}_i = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ . & . & . \\ . & . & . \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{Z}_i = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ . & . \\ . & . \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

We observe that $\boldsymbol{X}_i$ has a shortage of rank, and to get an estimable solution, *SAS* imposes $\beta_2 = 0$. The solution for $\boldsymbol{\beta}$ gives the means of the two methods μ_e and μ_n. The between-subject variance–covariance matrix $\boldsymbol{D}$ of the established method and a new method is a general (2×2)-dimensional matrix, and $\boldsymbol{R}_i$ is a $(n_i \times n_i)$-dimensional covariance matrix which depends on i only through its dimension n_i. The marginal density function is $\boldsymbol{y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i)$. Suppose the matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix}$ represents the within-subject variance–covariance matrix of the established method (e) and a new method (n) at any replicate, where σ_e^2 and σ_n^2 are the within-subject variances of the established method and a new method respectively, and σ_{en} is the within-subject covariance between the two methods. Also, suppose $\boldsymbol{V}$ represents the $p \times p$-dimensional correlation matrix of the replicated measurements on a given method, where $p = \max_i(p_i)$. It is assumed that the within-subject variance–covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications, and the correlation matrix $\boldsymbol{V}$ is the same for both methods (Timm, 2002, p. 401; Timm and Mieczkowski, 1997, p. 279). That is, the variances σ_e^2 and σ_n^2 are the within-subject variances of the two methods at any particular replication and are the same for all replications. We assume $\boldsymbol{R}_i = \dim_{n_i}(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, where $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ respectively are positive definite matrices as described above, and $\otimes$ represents the Kronecker product structure. The notation $\dim_{n_i}(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, represents a $(n_i \times n_i)$-dimensional submatrix obtained from the $(2p \times 2p)$-dimensional matrix $(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, by appropriately keeping the columns and rows corresponding to the n_i-dimensional response vector $\boldsymbol{y}_i$. Since the equicorrelated or compound symmetry (CS) correlation structure assumes equal correlation among all replicated measurements, we assume that the correlation matrix $\boldsymbol{V}$ of the replicated measurements has equicorrelated correlation structure. For the above design matrix $\boldsymbol{Z}_i$ and between-subject $\boldsymbol{D}$ and within-subject $\boldsymbol{R}_i$ sources of variation, the observed $(n_i \times n_i)$-dimensional overall variance–covariance matrix $\boldsymbol{\Omega}_i$ for the ith individual is given by

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{\Omega}_i &= \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i, \\ &= \boldsymbol{Z}_i \begin{bmatrix} d_e^2 & d_{en} \\ d_{en} & d_n^2 \end{bmatrix} \boldsymbol{Z}_i' + \dim_{n_i}\left(\boldsymbol{V} \otimes \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix}\right). \end{aligned}$$

Thus, the covariance matrix has the same structure for each subject, except that of the dimension. The 2×2 block diagonals $\mathrm{Block}\,\boldsymbol{\Omega}_i$ in the overall variance–covariance matrix $\boldsymbol{\Omega}_i$ represent the overall variance–covariance matrix between the two methods. Similarly, the 2×2 block diagonals in the overall correlation matrix

$\boldsymbol{\Omega}_i$-Correlation represent the overall correlation matrix between the two methods. Thus, the off-diagonal element in this 2×2 overall correlation matrix gives the overall correlation between the two methods. From the above equation we see that

$$\text{Block } \boldsymbol{\Omega}_i = \begin{bmatrix} \omega_e^2 & \omega_{en} \\ \omega_{en} & \omega_n^2 \end{bmatrix} = \begin{bmatrix} d_e^2 & d_{en} \\ d_{en} & d_n^2 \end{bmatrix} + \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix}.$$

Therefore, the overall variability is the sum of between-subject variability and within-subject variability. Thus, we see that if there is a disagreement in overall variabilities, then it may be due to the disagreement in either between-subject variabilities or within-subject variabilities, or both. If the investigator is not interested in knowing the exact cause of disagreement between the two methods, then he or she can go for the analysis of overall variabilities.

3. *PROC MIXED* OF *SAS*

We use *PROC MIXED* of *SAS* to get the maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$, $\boldsymbol{D}$, $\boldsymbol{R}_i$, and $\boldsymbol{\Omega}_i$. *METHOD = ML* specifies *PROC MIXED* to calculate the maximum likelihood estimates of the parameters. *REML* is the default method of *SAS* that offers non-biased *REML* estimates of the covariance parameters. The *COVTEST* option requests hypothesis tests for the random effects. *CLASS* statement specifies the categorical variables. *DDFM = KR* specifies the Kenward and Roger (1997) correction for computing the denominator degrees of freedom for the fixed effects. The Kenward–Roger correction is suggested whenever one has replicated or repeated measures data, and also for missing data. The *SOLUTION (S)* option in the *MODEL* statement provides the estimate of the difference between the two mean readings (bias) of the two methods. *RANDOM* and *REPEATED* statements specify the structure of the covariance matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i$. See the sample program in Appendix A that demonstrates the use of *RANDOM* and *REPEATED* statements. The advantage of *PROC MIXED* is that it can handle the Kronecker product covariance structure ($\boldsymbol{V} \otimes \boldsymbol{\Sigma}$). *PROC MIXED* calculates the $(n_i \times n_i)$-dimensional submatrix $\boldsymbol{R}_i$ of the ith subject from the $(2p \times 2p)$-dimensional matrix ($\boldsymbol{V} \otimes \boldsymbol{\Sigma}$), and eventually calculates the $(n_i \times n_i)$-dimensional submatrix $\boldsymbol{\Omega}_i$. When the number of replications on each subject by respective methods is unequal, *PROC MIXED* considers the case as a missing value situation. At present *PROC MIXED* can only have option $\boldsymbol{\Sigma}$ as unstructured, and $\boldsymbol{V}$ as unstructured, autoregressive of order one (AR(1)) or CS structure. Options *V* and *VCORR* in the *RANDOM* statement give the estimate of the overall variance–covariance matrix $\boldsymbol{\Omega}_1$ and the corresponding $\boldsymbol{\Omega}_1$-Correlation matrix, i.e., for the first subject. The option *G* in the *RANDOM* statement gives the estimate of the between-subject variance–covariance matrix $\boldsymbol{D}$. Option *R* in the *REPEATED* statement gives the estimate of the variance–covariance matrix $\boldsymbol{R}_1$ for the first subject. One can get the $\boldsymbol{\Omega}_i$ variance–covariance matrix and the corresponding $\boldsymbol{\Omega}_i$-Correlation matrix for all subjects by specifying $V = 1$ to N, and $VCORR = 1$ to N in the *RANDOM* statement. For detailed information, see *SAS/STAT* User's Guide (Version 9, 2004). When the correlation matrix $\boldsymbol{V}$ on the replicated measurements assumes CS as structured and $\boldsymbol{\Sigma}$ as unstructured, we can either use the option *TYPE = UN @ CS* along with *SUBJECT = PATIENT*, or use the option *TYPE = UN* along with *SUBJECT = REPLICATE(PATIENT)* in the

REPEATED statement. We will use the second option in this article. This option does not give the whole $n_i \times n_i$-dimensional matrix $\boldsymbol{R}_i$ as does the first option, but only the 2×2 within-subject variance–covariance matrix $\boldsymbol{\Sigma}$. In this article we need only this information to calculate the within-subject variances (repeatability coefficients) of the two methods.

4. DISAGREEMENT BETWEEN THE TWO METHODS AND THE RELATED HYPOTHESES TESTINGS

As mentioned in the introduction, if there is a disagreement between the two methods, it is important to know whether it is due to the bias, due to the difference in between-subject variabilities, or due to the difference in within-subject variabilities of the two methods. If it is due to the bias between the two methods, it is easy to correct. The output of *PROC MIXED* always gives the bias, its t-value, and its p-value. Nonetheless, it is not straightforward to check the agreement or disagreement in between-subject variabilities and in within-subject variabilities of the two methods. We will accomplish these by the indirect use of *PROC MIXED* in two steps (described below) by using likelihood ratio tests. If the disagreement is due to the difference in between-subject variabilities then there is a true disagreement between the two methods, and the methods should not be used interchangeably. If the disagreement is due to the difference in within-subject variabilities, i.e., the disagreement is due to the difference in the repeatability coefficients of the two methods, then the methods should also not be used interchangeably.

Therefore, each of the above three factors of disagreement has separate interpretation. Disaggregated criteria examine these various possible sources of disagreement separately (Lee et al., 1989), whereas aggregated criteria are based on an index that combines different sources of possible disagreement between the two methods. Disaggregate criterion for assessing individual bioequivalence is discussed in Chen (1997). Chen mentioned that aggregated criteria have the advantage of balancing different sources of disagreement while disaggregated criteria may have the advantage of identifying the actual source of disagreement if the agreement is not satisfactory. From the practical point of view, we are interested in the factor-wise assessment between the two methods rather than a combined assessment, so that we know the exact cause of disagreement. Since we have more than one factor or multiple factors, whatever method of inference is used an appropriate adjustment for multiple-testing (Westfall et al., 1999) is needed. A common way to account for multiple-testing is to consider the family-wise error (FWE) rate. Roughly speaking, the FWE is the probability of making a false claim when the entire family of inferences is considered. Bonferroni correction provides an effective conservative control of FWE level, α. In this article we use Bonferroni adjustment procedure to calculate the adjusted p-values for multiple-testing of the three factors. An adjusted p-value is a p-value defined in a collection of hypothesis $\{H_j,\ j = 1, \ldots, k\}$, and we can simply compare its corresponding adjusted p-value with the desired FWE level, α. The Bonferroni procedure rejects any hypothesis, H_j, in the collection $\{H_j,\ j = 1, \ldots, k\}$, whose corresponding p-value, p_j, is less than or equal to α/k. This is equivalent to rejecting any H_j for which kp_j is less than or equal to α. Thus, kp_j is

the Bonferroni adjusted p-value for H_j. Bonferroni adjusted p-value using the "$\tilde{p}$" symbol to denote "adjusted p-value" for hypothesis H_j is defined as:

$$\tilde{p}_j = \begin{cases} kp_j : kp_j \leq 1, \\ 1 : kp_j > 1 \end{cases} .$$

4.1. Testing of Hypothesis of Difference Between the Means of the Two Methods

We are interested in testing the following hypothesis:

H_μ : the two methods do not have the same mean,

vs. K_μ : the two methods have the same mean,

which is equivalent to testing the following hypothesis:

$$H_\mu : |\mu_e - \mu_n| > \delta_\mu, \quad \text{vs.} \quad K_\mu : |\mu_e - \mu_n| \leq \delta_\mu,$$

where δ_μ is a pre-specified acceptable difference between the two population means μ_e and μ_n which is not set by a statistician, but by investigators or regulators. Output of *PROC MIXED* (Solution for Fixed Effects) gives the bias and the corresponding t-value and p-value.

4.2. Testing of Hypothesis of Difference in Between-Subject Variabilities of the Two Methods

From the between-subject variance–covariance matrix $\boldsymbol{D}$, d_e^2 and d_n^2 are the between-subject variances of the established method and a new method, respectively. We are interested in testing the following hypothesis:

H_d : the two methods do not have the same between-subject variabilities,

vs. K_d : the two methods have the same between-subject variabilities,

which is equivalent to testing the following hypothesis:

$$H_d : \frac{d_e}{d_n} < \delta_{d_1} \quad \text{or} \quad \frac{d_e}{d_n} > \delta_{d_2}, \quad \text{vs.} \quad K_d : \delta_{d_1} \leq \frac{d_e}{d_n} \leq \delta_{d_2},$$

where δ_{d_1} and δ_{d_2} $(0 < \delta_{d_1} < 1 < \delta_{d_2})$ are pre-specified acceptable thresholds for $\frac{d_e}{d_n}$. These limits are also not set by a statistician, but by investigators or regulators. We apply the likelihood ratio test for this hypothesis testing. To compute the test statistic $-2\ln\Lambda_d$, where

$$-2\ln\Lambda_d = \left[-2\ln\max_{K_d} L\right] - \left[-2\ln\max_{H_d} L\right],$$

the log likelihood function under both null hypothesis and alternating hypothesis must be maximized separately. We do this by setting the option *METHOD = ML* in *PROC MIXED* statement. The option *TYPE = UN* in the *RANDOM* statement, along with the option *TYPE = UN* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under H_d. Similarly, the option *TYPE = CS* in the *RANDOM* statement, along with the option *TYPE = UN* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under K_d. Since $\boldsymbol{D}$ is 2×2 dimensional, one can also use *TYPE = AR*(1) or *TOEP* in the *RANDOM* statement to calculate the "−2 Log Likelihood" for the covariance structure under K_d. *PROC MIXED* calculates this under the heading of "Fit Statistics". The above test statistic $-2\ln\Lambda_d$ under K_d follows a chi-square distribution with degrees of freedom (d.f.) v_d, where v_d is computed as

$$v_d = \text{LRT df (under } H_d) - \text{LRT df (under } K_d).$$

4.3. Testing of Hypothesis of Difference in Within-Subject Variabilities of the Two Methods

Following Bland and Altman (1999), we name $1.96\sqrt{2}\sigma_e$ as the *repeatability coefficient* of the established method, where σ_e^2 is the within-subject variance of the established method as defined earlier, and similarly, the *repeatability coefficient* of a new method. For 95% of subjects, two replicated measurements by the same method will be within this repeatability coefficient. We test the difference between the repeatability coefficients of the two methods by testing the following hypothesis:

H_σ : the two methods do not have the same within-subject variabilities,

vs. K_σ : the two methods have the same within-subject variabilities,

which is equivalent to testing the following hypothesis:

$$H_\sigma : \frac{\sigma_e}{\sigma_n} < \delta_{\sigma_1} \quad \text{or} \quad \frac{\sigma_e}{\sigma_n} > \delta_{\sigma_2}, \quad \text{vs.} \quad K_\sigma : \delta_{\sigma_1} \le \frac{\sigma_e}{\sigma_n} \le \delta_{\sigma_2},$$

where δ_{σ_1} and δ_{σ_2} $(0 < \delta_{\sigma_1} < 1 < \delta_{\sigma_2})$ are pre-specified acceptable thresholds for $\frac{\sigma_e}{\sigma_n}$ and are set by the investigators. As before, here also we apply the likelihood ratio test for this hypothesis testing, and maximize the log likelihood function under both null hypothesis and alternating hypothesis separately to compute the test statistic $-2\ln\Lambda_\sigma$, where

$$-2\ln\Lambda_\sigma = \left[-2\ln\max_{K_\sigma} L\right] - \left[-2\ln\max_{H_\sigma} L\right].$$

The option *TYPE = UN* in the *RANDOM* statement, along with *TYPE = UN* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under H_σ. *TYPE = UN* in the *RANDOM* statement, along with *TYPE = CS* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under K_σ. The test statistic $-2\ln\Lambda_\sigma$

under K_σ follows a chi-square distribution with d.f. v_σ = LRT df (under H_σ) − LRT df (under K_σ).

4.4. Testing of Hypothesis of Difference in Overall Variabilities of the Two Methods

From the overall variance–covariance matrix Block $\mathbf{\Omega}_i$, we see that ω_e^2 and ω_n^2 are the overall variances of the established method and a new method, respectively. We are interested in testing the following hypothesis:

H_ω : the two methods do not have the same overall variabilities,

vs. K_ω : the two methods have the same overall variabilities,

which is equivalent to testing the following hypothesis:

$$H_\omega : \frac{\omega_e}{\omega_n} < \delta_{\omega_1} \quad \text{or} \quad \frac{\omega_e}{\omega_n} > \delta_{\omega_2}, \quad \text{vs.} \quad K_\omega : \delta_{\omega_1} \le \frac{\omega_e}{\omega_n} \le \delta_{\omega_2},$$

where δ_{ω_1} and δ_{ω_2} $(0 < \delta_{\omega_1} < 1 < \delta_{\omega_2})$ are pre-specified acceptable thresholds for $\frac{\omega_e}{\omega_n}$. As before, here also we apply the likelihood ratio test to compute the test statistic $-2\ln\Lambda_\omega$, where

$$-2\ln\Lambda_\omega = \left[-2\ln\max_{K_\omega} L\right] - \left[-2\ln\max_{H_\omega} L\right].$$

The option *TYPE = UN* in the *RANDOM* statement, along with *TYPE = UN* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under H_ω. The option *TYPE = CS* in the *RANDOM* statement, along with *TYPE = CS* in the *REPEATED* statement, is used to calculate the "−2 Log Likelihood" for the covariance structure under K_ω. The test statistic $-2\ln\Lambda_\omega$ under K_ω follows a chi-square distribution with d.f. v_ω = LRT df (under H_ω) − LRT df (under K_ω).

5. SOME EXAMPLES

We demonstrate the proposed method by considering three real datasets. All the datasets are taken from different papers of Bland and Altman (1986, 1999, 2007). The first dataset is larger in size, whereas the second and the third datasets are of smaller sizes. The first and the second datasets have balanced replications, while the third dataset has unbalanced replications.

Example 1 (Systolic Blood Pressure Data)**.** This dataset is taken from Bland and Altman (1999, Table 1). Simultaneous measurements of systolic blood pressure were made by each of the two experienced observers (denoted by J and R) using a sphygmomanometer, and by a semi-automatic blood pressure monitor (denoted by S). Three sets ($p = 3$) of readings were made in quick succession on 85 subjects. Blood pressures taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. We want to

examine whether either of the two observers can be replaced by the semi-automatic blood pressure monitor. To see this, we first analyze the dataset by taking the observer J and the machine S, and then analyze the observer R and the machine S.

Sample *SAS* program for the analysis of the observer J and the monitor S is given in Appendix A. Selected parts of the SAS output for covariance structure under H_d are given in Appendix B. From the output of *PROC MIXED* ("Solution for Fixed Effects") we see that the bias between the observer J and the automatic blood pressure machine S is -15.6196 mmHg with t-value $= -7.65$ and d.f. $= 85$. Therefore, the p-value $= 2.851E$-11. The MLE of the between-subject variance–covariance matrix of the observer J and the blood pressure machine S is given by

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 923.99 & 785.24 \\ 785.24 & 971.30 \end{bmatrix}.$$

To test the difference in between-subject variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_d = (4061.6) - (4061.5) = 0.1$, where 4061.6 and 4061.5 are the values of "-2 Log Likelihood" reported by *SAS* for the two models under K_d and H_d, respectively. The test statistic under K_d follows a χ^2 with d.f. $v_d = 5 - 4 = 1$. The corresponding p-value is 0.7518. The MLE of the within-subject variance–covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 37.4078 & 16.0627 \\ 16.0627 & 83.1412 \end{bmatrix}.$$

We see that the estimates of the within-subject variances are exactly the same as obtained by Bland and Altman (1999). We also observe (see Appendix B) that all the elements in the estimated between-subject variance–covariance matrix $\widehat{\boldsymbol{D}}$ and the estimated within-subject variance–covariance matrix $\widehat{\boldsymbol{\Sigma}}$ are highly statistically significant, assuring that the variance–covariance matrices D and $\boldsymbol{\Sigma}$ are indeed needed in the model. Also, from the model fitting information, we find that the model with Kronecker product structure is a good fit for the data. The estimated repeatability coefficients (defined in Section 4.3) for the observer J and the machine S are 16.9532 mmHg and 25.2743 mmHg, respectively. As a result, repeatability of the machine S is 49% more than the repeatability of the observer J. To see the difference between the repeatability coefficients or the difference in within-subject variabilities of the observer J and the machine S, we calculate the test statistic $-2\ln\Lambda_\sigma = (4090.1) - (4061.5) = 28.6$, where 4090.1 and 4061.5 are the values of "-2 Log Likelihood" under K_σ and H_σ, respectively. The test statistic under K_σ follows χ^2_1. The corresponding p-value $= 8.8982E$-8. The 2×2 block diagonals Block $\widehat{\boldsymbol{\Omega}}_i$ in the estimated 6×6 overall variance–covariance matrix $\widehat{\boldsymbol{\Omega}}_i$ give the overall variance–covariance matrix of the observer J and the machine S, and they are given by

$$\text{Block}\,\widehat{\boldsymbol{\Omega}}_i = \begin{bmatrix} 961.39 & 801.31 \\ 801.31 & 1054.44 \end{bmatrix}.$$

The estimated 6×6 overall $\widehat{\boldsymbol{\Omega}}_i$-Correlation matrix is given by

$$\widehat{\boldsymbol{\Omega}}_i\text{-Correlation} = \begin{bmatrix} \begin{bmatrix} 1.0000 & 0.7959 \\ 0.7959 & 1.0000 \end{bmatrix} & 0.9611 & 0.7799 & 0.9611 & 0.7799 \\ 0.7799 & 0.9212 & 0.7799 & 0.9212 & 0.7799 & 0.9212 \\ 0.9611 & 0.7799 & \begin{bmatrix} 1.0000 & 0.7959 \\ 0.7959 & 1.0000 \end{bmatrix} & 0.9611 & 0.7799 \\ 0.7799 & 0.9212 & 0.7799 & 0.9212 & 0.7799 & 0.9212 \\ 0.9611 & 0.7799 & 0.9611 & 0.7799 & \begin{bmatrix} 1.0000 & 0.7959 \\ 0.7959 & 1.0000 \end{bmatrix} \\ 0.7799 & 0.9212 & 0.7799 & 0.9212 & & \end{bmatrix}.$$

From the 2×2 block diagonals of the above $\widehat{\boldsymbol{\Omega}}_i$-Correlation matrix we see that the overall correlation coefficient between the observer J and the machine S is 0.7959. Also, the overall correlation coefficient between the observer J and the machine S for two different replicates is 0.7799. The overall correlation coefficients between two different replicates of the observer J and the machine S are 0.9611 and 0.9212, respectively. We observe that the overall correlation coefficient between two different replicates of the observer J is marginally higher than the overall correlation coefficient between two different replicates of the machine S. Interestingly, we also observe that the overall variability of the observer J is less than the overall variability of the machine S. To test the difference in the overall variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_\omega = (4090.4) - (4061.5) = 28.9$, where 4090.4 and 4061.5 are the values of "-2 Log Likelihood" under K_ω and H_ω, respectively. The test statistic under K_ω follows χ^2_2. The corresponding p-value $= 5.3021E\text{-}7$. The significant p-value for the difference in the overall variabilities of the observer J and the machine S is due to the significant p-value for the difference in the within-subject variabilities.

Now, the Bonferroni adjusted p-values for bias, between-subject, and within-subject variabilities of the observer J and the machine S are <0.0001, 1.0, and <0.0001, respectively. So, we see that there is a strong disagreement in within-subject variabilities of the observer J and the machine S. Hence, we do not recommend to switch the observer J and the machine S if needed. In this example we see that the methods do not have as high an overall correlation coefficient as 0.82.

For the observer R and the automatic blood pressure monitor S, again from the output of *PROC MIXED* we see that the bias is -15.7059 mmHg with t-value $= -7.75$ and d.f. $= 85$. Therefore, the corresponding p-value $= 1.798E\text{-}11$. The MLE of the between-subject variance–covariance matrix $\boldsymbol{D}$ is given by

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 906.13 & 778.62 \\ 778.62 & 971.30 \end{bmatrix}.$$

To test the difference in between-subject variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_d = (4059.9) - (4059.6) = 0.3$, where 4059.9 and 4059.6 are the values of "-2 Log Likelihood" for the two models under K_d and H_d, respectively. The test statistic under K_d follows χ^2_1 and the corresponding p-value $= 0.5839$.

The MLE of the within-subject variance–covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 37.9804 & 17.3333 \\ 17.1412 & 83.1412 \end{bmatrix}.$$

Here also, the estimates of the within-subject variances are exactly the same as obtained by Bland and Altman (1999). The estimated repeatability coefficients for the observer R and the machine S are 17.0825 mmHg and 25.2743 mmHg, respectively. Therefore, the repeatability of the machine S is 48% more than the repeatability of the observer R. As before, to see the difference between the repeatability coefficients, or the difference in within-subject variabilities of the two methods, we calculate the test statistic $-2\ln\Lambda = (4087.5) - (4059.6) = 27.9$, where 4087.5 and 4059.6 are the values of "−2 Log Likelihood" under K_σ and H_σ, respectively. This test statistic under K_σ follows a χ^2_1. The corresponding p-value = 1.2775E-7. The 2×2 block diagonals Block $\widehat{\boldsymbol{\Omega}}_i$ in the estimated $\widehat{\boldsymbol{\Omega}}_i$ variance–covariance matrix give the overall variance–covariance matrix of the observer R and the machine S, and are given by

$$\text{Block}\,\widehat{\boldsymbol{\Omega}}_i = \begin{bmatrix} 944.11 & 795.95 \\ 795.95 & 1054.44 \end{bmatrix}.$$

From the estimated $\widehat{\boldsymbol{\Omega}}_i$-Correlation matrix (not shown here) we see that the overall correlation coefficient between the observer R and the machine S is 0.7977. Also, the overall correlation coefficient between the observer R and the machine S for two different replicates is 0.7804. The overall correlation coefficients between two different replicates of the observer R and the machine S are 0.9598 and 0.9212, respectively. As before, we observe that the overall correlation coefficient between two different replicates of the observer R is marginally higher than the overall correlation coefficient between two different replicates of the machine S. Here also, we observe that the overall variability of the observer R is less than the overall variability of the machine S. To test the difference in overall variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_\omega = (4088.0) - (4059.6) = 28.4$, where 4088.0 and 4059.6 are the values of "−2 Log Likelihood" under K_ω and H_ω, respectively. The test statistic under K_ω follows a χ^2_2. The corresponding p-value = 6.808E-7. Here also, the significant p-value for the difference in the overall variabilities of the observer R and the machine S is due to the significant p-value for the difference in the within-subject variabilities.

Now, the Bonferroni adjusted p-values for bias, between-subject, and within-subject variabilities of the observer R and the machine S are <0.0001, 1.0, and <0.0001, respectively. So, we see that there is a strong disagreement in within-subject variabilities of the observer R and the machine S. Hence, we do not recommend to switch the observer R and the machine S if needed. Here also, we see that the methods do not have a high overall correlation coefficient.

Example 2 (Peak Expiratory Flow Rate Data)**.** This dataset in Table 1 (Bland and Altman, 1986), compares two methods of measuring peak expiratory flow rate (PEFR). The sample was collected from a wide range of PEFR, but was not from

Table 1 PEFR measured with Wright peak flow and mini Wright peak flow meter

	Wright peak flow meter		Mini Wright peak flow meter	
Subject	First PEFR (1/min)	Second PEFR (1/min)	First PEFR (1/min)	Second PEFR (1/min)
1	494	490	512	525
2	395	397	430	415
3	516	512	520	508
4	434	401	428	444
5	476	470	500	500
6	557	611	600	625
7	413	415	364	460
8	442	431	380	390
9	650	638	658	642
10	433	429	445	432
11	417	420	432	420
12	656	633	626	605
13	267	275	260	227
14	478	492	477	467
15	178	165	259	268
16	423	372	350	370
17	427	421	451	443

any defined population. Two measurements ($p = 2$) were made with a Wright peak flow meter (X) and two measurements with a mini Wright peak flow meter (Y), in random order. All measurements were taken using the same two instruments.

From the output of *PROC MIXED* we see that the bias between the two methods is $-6.0294\,\text{min}^{-1}$ with p-value $= 0.4509$. The MLE of the between-subject variance–covariance matrix of the Wright peak flow meter (X) and the mini Wright peak flow meter (Y) is given by

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 12871 & 11803 \\ 11803 & 11459 \end{bmatrix}.$$

To test the between-subject variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_d = (688.9) - (688.2) = 0.7$, where 688.9 and 688.2 are the values of "−2 Log Likelihood" under K_d and H_d, respectively. The test statistic under K_d follows a χ^2_1. The corresponding p-value $= 0.4028$. The MLE of the within-subject variance–covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 234.29 & 2.0000 \\ 2.0000 & 396.44 \end{bmatrix}.$$

The coefficient of repeatability for the Wright peak flow meter is $42.4275\,\text{min}^{-1}$, and the coefficient of repeatability for the mini Wright peak flow meter is $55.1899\,\text{min}^{-1}$. Therefore, repeatability of the mini Wright peak flow meter is 30% more than the repeatability of the Wright peak flow meter. To see the difference between

the repeatability coefficients, or the difference in within-subject variabilities, of the two peak flow meters, we calculate the test statistic $-2\ln\Lambda_\sigma = (689.4) - (688.2) = 1.2$, where 689.4 and 688.2 are the values of "-2 Log Likelihood" under K_σ and H_σ, respectively. The above test statistic under K_σ follows a χ^2_1. The corresponding p-value $= 0.2733$. The 2×2 block diagonals Block $\widehat{\mathbf{\Omega}}_i$ in the estimated 4×4 overall variance–covariance matrix $\widehat{\mathbf{\Omega}}_i$ are given by

$$\text{Block}\,\widehat{\mathbf{\Omega}}_i = \begin{bmatrix} 13105 & 11805 \\ 11805 & 11855 \end{bmatrix}.$$

From the 4×4 estimated $\widehat{\mathbf{\Omega}}_i$-Correlation matrix (not shown here) we see that the overall correlation coefficient between the two Wright peak flow meters is 0.9471. Also, the overall correlation coefficient between the two Wright peak flow meters for two different replicates is 0.9469. The overall correlation coefficients between two different replicates of the Wright peak flow meter and the mini Wright peak flow meter are 0.9821 and 0.9666, respectively. Here we observe that the overall correlation coefficient between two different replicates of the Wright peak flow meter is slightly higher than the overall correlation coefficient between two different replicates of the mini Wright peak flow meter. In this example we observe that the overall variability of the Wright peak flow meter is more than the overall variability of the mini Wright peak flow meter. The test statistic $-2\ln\Lambda_\omega = (690.0) - (688.2) = 28.4$, where 690.0 and 688.2 are the values of "-2 Log Likelihood" under K_ω and H_ω, respectively. The test statistic under K_ω follows a χ^2_2. The corresponding p-value $= 0.40657$. Here the non-significant p-value for the difference in the overall variabilities of the two methods of measuring PEFR is due to the non-significant p-value for both the between-subject and the within-subject variabilities.

Now, the Bonferroni adjusted p-values for bias, between-subject, and within-subject variabilities of the two methods are 1, 1, and 0.8199, respectively. So, we conclude that there is a satisfactory agreement between the two methods of measuring PEFR, and we do recommend to switch the two methods if needed. In this example we see that the methods do have a high overall correlation coefficient.

Example 3 (Cardiac Data)**.** The data in Table 2 have been considered in Bland and Altman (1999, Table 4; 2007, Table 1). This dataset has measurements of left ventricular cardiac ejection fraction (%) by two methods, impedance cardiography (IC) and radionuclide ventriculography (RV), on 12 patients or subjects. The number of repeated observations differs by patient. Such data may occur if patients are measured at regular intervals during surgery. In this example, the repeated observations may not be replicates in the true sense, however, we still assume they are the true replicates for the purpose of this paper. It is reasonable to assume that the observations in Table 2 are reported in the order in which they are made, as the observations have not been arranged in any order. It is known that in any clinic, observations are usually reported in the order in which they are made. Since the dataset has an unbalanced number of observations (p_i) per patient, the overall variance–covariance matrix $\mathbf{\Omega}_i$ for each patient will have different dimensions. For instance, patient 1 will have a 10×10 dimensional variance–covariance matrix $\mathbf{\Omega}_1$,

Table 2 Cardiac ejection fraction (%) by two methods RV and IC for 12 subjects (Data provided by Dr. L. S. Bowling)

Subject	RV	IC	Subject	RV	IC	Subject	RV	IC
1	7.83	6.57	5	3.13	3.03	9	4.48	3.17
1	7.42	5.62	5	2.98	2.86	9	4.92	3.12
1	7.89	6.90	5	2.85	2.77	9	3.97	2.96
1	7.12	6.57	5	3.17	2.46	10	4.22	4.35
1	7.88	6.35	5	3.09	2.32	10	4.65	4.62
2	6.16	4.06	5	3.12	2.43	10	4.74	3.16
2	7.26	4.29	6	5.92	5.90	10	4.44	3.53
2	6.71	4.26	6	6.42	5.81	10	4.50	3.53
2	6.54	4.09	6	5.92	5.70	11	6.78	7.20
3	4.75	4.71	6	6.27	5.76	11	6.07	6.09
3	5.24	5.50	7	7.13	5.09	11	6.52	7.00
3	4.86	5.08	7	6.62	4.63	11	6.42	7.10
3	4.78	5.02	7	6.58	4.61	11	6.41	7.40
3	6.05	6.01	7	6.93	5.09	11	5.76	6.80
3	5.42	5.67	8	4.54	4.72	12	5.06	4.50
4	4.21	4.14	8	4.81	4.61	12	4.72	4.20
4	3.61	4.20	8	5.11	4.36	12	4.90	3.80
4	3.72	4.61	8	5.29	4.20	12	4.80	3.80
4	3.87	4.68	8	5.39	4.36	12	4.90	4.20
4	3.92	5.04	8	5.57	4.20	12	5.10	4.50

as there are 5 repetitions for patient 1. For patient 2, it will be 8×8 dimensional, as patient 2 has only 4 repetitions. We see that the bias between the two methods RV and IC is 0.7040 with p-value $= 0.0204$. The MLE of the between-subject variance–covariance matrix of the two methods RV and IC is given by

$$\widehat{\boldsymbol{D}} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix}.$$

To test the between-subject variabilities of the two methods we calculate the test statistic $-2\ln\Lambda_d = (173.1) - (173.1) = 0.0$, where 173.1 and 173.1 are the values of "-2 Log Likelihood" under K_d and H_d, respectively. The test statistic under K_d follows a χ^2_1. The corresponding p-value $= 1$. The estimate of the within-subject variance–covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix}.$$

The estimates of the variances are exactly the same as obtained by Bland and Altman (1999). The repeatability coefficient for the RV method is 0.9080, and the repeatability coefficient for the IC method is 1.0293. Thus, the repeatability of the IC method is little more than (13%) that of the RV method. The test statistic $-2\ln\Lambda_\sigma = (173.9) - (173.1) = 0.8$, where 173.9 and 173.1 are the values of "-2 Log Likelihood" under K_σ and H_σ, respectively. This test statistic under K_σ follows a χ^2_1. The corresponding p-value $= 0.3711$. The estimated overall variance–covariance

matrix of the RV and IC methods is given by

$$\text{Block}\,\widehat{\mathbf{\Omega}}_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}.$$

From the estimated $\widehat{\mathbf{\Omega}}_i$-Correlation matrix (not shown here), we see that the overall correlation coefficient between the two methods RV and IC is 0.7100. Also, the overall correlation coefficient between the two methods RV and IC for two different replicates is 0.6876. The overall correlation coefficients between two different replicates of RV and IC are 0.9384 and 0.9131, respectively. Here we observe that the overall correlation coefficient between two different replicates of the RV method is slightly higher than the overall correlation coefficient between two different replicates of the IC method. In this example, we observe that the overall variability of the RV method is little more than the overall variability of the IC method. The test statistic $-2\ln\Lambda_\omega = (174.0) - (173.1) = 0.9$, where 174.0 and 173.1 are the values of "-2 Log Likelihood" under K_ω and H_ω, respectively. The test statistic under K_ω follows a χ^2_2. The corresponding p-value $= 0.6376$. Here also, the non-significant p-value for the difference in the overall variabilities of the two methods, impedance cardiography (IC) and radionuclide ventriculography (RV), is due to the non-significant p-values for both the between-subject and the within-subject variabilities of the two methods.

Now, the Bonferroni adjusted p-values for bias, between-subject, and within-subject variabilities of the two methods are 0.0612, 1.0, and 1.0, respectively. So, we conclude that there is a satisfactory agreement between the two methods IC and RV, and we do recommend to switch the two methods if needed. In this example we see that the methods do not have a high overall correlation coefficient.

We also apply our proposed method to coronary artery calcium data (Haber et al., 2005), as mentioned in the introduction. For this dataset the estimated 4×4 overall $\widehat{\mathbf{\Omega}}_i$-Correlation matrix is given by

$$\widehat{\mathbf{\Omega}}_i\text{-Correlation} = \begin{bmatrix} \begin{bmatrix} 1.0000 & 0.9957 \\ 0.9957 & 1.0000 \end{bmatrix} & \begin{matrix} 0.9919 & 0.9954 \\ 0.9954 & 0.9999 \end{matrix} \\ \begin{matrix} 0.9919 & 0.9954 \\ 0.9954 & 0.9999 \end{matrix} & \begin{bmatrix} 1.0000 & 0.9957 \\ 0.9957 & 1.0000 \end{bmatrix} \end{bmatrix}.$$

From the 2×2 block diagonals of the above estimated $\widehat{\mathbf{\Omega}}_i$-Correlation matrix we see that the overall correlation coefficient between the two radiologists is 0.9957. Also, the overall correlation coefficient between the two radiologists for two different replicates is 0.9954. The overall correlation coefficients between two different replicates of the first radiologist and the second radiologist are 0.9919 and 0.9999, respectively. We notice that all four overall correlation coefficients are almost perfect and they are almost the same. However, due to the existence of the random error (p-value $= 6.5095E\text{-}9$) our recommendation is not to switch the two radiologists. In other words, our method concludes that there is a perfect disagreement between the two radiologists even though all four overall correlation coefficients are almost perfect. For the cardiac dataset, on the other hand, we see

that there is a satisfactory agreement between the two methods IC and RV, but the overall correlation coefficient between the two methods is only 0.7100. These observations, along with Haber et al.'s (2005) observation as mentioned in the introduction, justify our decision that correlation coefficient type index is indeed misleading, and steers us not to use any kind of correlation coefficient for assessing agreement for data with replications. Correlation coefficient may depend on the internal structure of the dataset. Also, it is very sensitive to outliers (Roy, 2006). Further study is needed in this direction. Also, *PROC MIXED* only gives all overall correlation coefficients, but does not give their variances. Hence, at this time it is not possible to carry out any inference based on all overall correlation coefficients. One may use the bootstrap approach, where sampling with replacement is taken for each subject, to estimate the variances of all overall correlation coefficients.

6. CONCLUSIONS

In this article we propose a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects. The topic is of practical relevance in many practical fields, especially in medical and biomedical sciences. Our method has some advantages over other previous approaches: it is easily understandable by either a statistician or a non-statistician, and it is very easy to implement using *PROC MIXED* of *SAS*. The interpretation of the results is also easy. A few lines of computer program can be used by any person with a little bit of programming expertise. The powers of the likelihood ratio tests mentioned in this paper may depend on specific sample size and the specific number of replicated observations. One needs to do some simulation study for this. Since *PROC MIXED* can handle covariates, our model can easily be extended when covariate information is available. In this article we assume that the replicated measurements are true replicates. Sometimes true or genuine replicates cannot be obtained. In these cases, the correlation matrix on the replicates may assume other structures, such as AR(1). However, the mathematics involved with AR(1) structure are computationally very demanding, as it is not possible to find any closed-form solutions for the maximum likelihood estimates of the required parameters. We are currently working on this method and will report in a future correspondence.

APPENDIX

A SAS CODE

```
/* We here give the SAS code for fitting the linear mixed effects
model to the data of Bland and Altman (1999, Table 1) under the null
hypothesis that the two methods, the observer J and the blood
pressure monitor S, do not have the same within-subject
variabilities. y1, y2, and y3 denote the three replicates of the
observer J. y4, y5, and y6 denote the three replicates of the
observer R. Similarly y7, y8, and y9 denote the three replicates
of the blood pressure monitor S. We define the observer J and the
```

```
blood pressure monitor S together by a vector variable m_var;
m_var = J for the observer J and m_var = S for the blood
pressure monitor S.*/

options nocenter ls=80 ps=50 nodate nonumber;
data BloodPressure;
infile 'c: \ BloodPressure.dat';
input subj y1-y9;

data BloodPressure1; set BloodPressure;
y=y1; m_var='J'; rep=1; output;
y=y7; m_var='S'; rep=1; output;
y=y2; m_var='J'; rep=2; output;
y=y8; m_var='S'; rep=2; output;
y=y3; m_var='J'; rep=3; output;
y=y9; m_var='S'; rep=3; output;
drop y1-y9;
proc mixed data=BloodPressure1 method=ml covtest;
classes subj m_var rep;
model y=m_var /s ddfm=kr;
random m_var /type=un subject=subj v vcorr g;
repeated m_var/type= un subject=rep(subj) r;
run;

/* Here we give the SAS code for fitting the linear mixed effects
model to the same data under the alternative hypothesis that the
two methods have the same within-subject variabilities.*/

proc mixed data=BloodPressure1 method=ml covtest;
classes subj m_var rep;
model y=m_var /s ddfm=kr;
random m_var /type= un subject=subj v vcorr g;
repeated m_var/type= cs subject=rep(subj) r;
run;
```

B SAS OUTPUT FOR COVARIANCE STRUCTURE UNDER H_d

Estimated V Matrix for subj 1

Row	Col1	Col2	Col3	Col4	Col5	Col6
1	961.39	801.31	923.99	785.24	923.99	785.24
2	801.31	1054.44	785.24	971.30	785.24	971.30
3	923.99	785.24	961.39	801.31	923.99	785.24
4	785.24	971.30	801.31	1054.44	785.24	971.30
5	923.99	785.24	923.99	785.24	961.39	801.31
6	785.24	971.30	785.24	971.30	801.31	1054.44

Covariance parameter estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
UN(1, 1)	subj	923.99	143.65	6.43	<.0001
UN(2, 1)	subj	785.24	135.51	5.79	<.0001
UN(2, 2)	subj	971.30	153.27	6.34	<.0001
UN(1, 1)	rep(subj)	37.4078	4.0575	9.22	<.0001
UN(2, 1)	rep(subj)	16.0627	4.4511	3.61	0.0003
UN(2, 2)	rep(subj)	83.1412	9.0179	9.22	<.0001

Fit Statistics

−2 Log Likelihood	4061.5
AIC (smaller is better)	4077.5
AICC (smaller is better)	4077.8
BIC (smaller is better)	4097.0

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
5	912.82	<.0001

Solution for Fixed Effects

Effect	m_var	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		143.03	3.4283	85	41.72	<.0001
m_var	J	−15.6196	2.0416	85	−7.65	<.0001
m_var	S	0	.	.	.	.

ACKNOWLEDGMENTS

The author would like to thank the editor and the two anonymous reviewers for their valuable recommendations that greatly improved the quality of the manuscript. The author would also like to acknowledge the generous support for the summer research grant from the College of Business at the University of Texas at San Antonio.

REFERENCES

Argall, J. A. W., Wright, N., Mackway-Jones, K., Jackson, R. (2003). A comparison of two commonly used methods of weight estimation. *Archives of Disease in Childhood* 88:789–790

Barnhart, H. X., Song, J., Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Stat. Med.* 24:1371–1384.

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

Bartko, J. J. (1994). Measures of agreement: a single procedure. *Stat. Med.* 13:737-745.

Bland, J. M., Altman, D. G. (1983). Measurement in medicine: the analysis of method comparison studies. *Statistician* 32:307–17.

Bland, J. M., Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* 8:307–310.

Bland, J. M., Altman, D. G. (1990). A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput. Biol. Med.* 20(5):337–340.

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

Bland, J. M., Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. *Journal of Biopharmaceutical Statistics* 17:571–582.

Boik, J. B. (1991). Scheffe's mixed model for multivariate repeated measures: a relative efficiency evaluation. *Commun. Statist – Theory Meth.* 20:1233–1255.

Bowling, L. S., Sageman, W. S., O'Connor, S. M., Cole, R., Amundson, D. E. (1993). Lack of agreement between measurement of ejection fraction by impedance cardiography versus radionuclide ventriculography. *Crit. Care Med.* 21:1523–1527.

Chaganty, N. R., Naik, D. N. (2002). Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference* 103:421–436.

Chen, M. L. (1997). Individual bioequivalence—a regulatory update. *Journal of Biopharmaceutical Statistics* 7:5–11.

Choudhary, P. K. (2007). Semiparametric regression for assessing agreement using tolerance bands. *Computational Statistics and Data Analysis* 51:6229–6241.

Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* 138:1102–1115.

Choudhary, P. K., Nagaraja, H. N. (2005). Assessment of agreement using intersection-union principle. *Biometrical Journal* 47(5):674–681.

Choudhary, P. K., Ng, H. K. T. (2006). Assessment of agreement under non-standard conditions using regression models for mean and variance. *Biometrics* 62:288–296.

Choudhary, P. K., Nagaraja, H. N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* 137:279–290.

Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Commun. Statist – Theory Meth.* 22:3105–3120.

Haber, M., Barnhart, H. X., Song, J., Gruden J. (2005). Observer variability: a new approach in evaluating interobserver agreement. *Journal of Data Science* 3:69–83.

Hamlett, A., Ryan, L., Serrano-Trespalacios, P., Wolfinger, R. (2003). Mixed models for assessing correlation in the presence of replication. *Journal of the Air & Waste Management Association* 53:442–450.

Hamlett, A., Ryan, L., Wolfinge, R. (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *SAS Users Group International, Proceedings of the Statistics and Data Analysis Section* 198–229:1–7.

Kenward, M. G., Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53:983–997.

Lam, M., Webb, K. A., O'Donnell, D. E. (1999). Correlation between two variables in repeated measures. *American Statistical Association, Proceedings of the Biometric Section* 213–218.

Laird, N. M., Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* 38:963–974.

Lee, J., Koh, D., Ong, C. N. (1989). Statistical evaluation of agreement between two methods for measuring a quantitative variable. *Comput. Biol. Med.* 19(1):61–70.

Lin, L. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45:255–268.

Lin, L. I., Hedayat, A. S., Sinha, B., Yang, M. (2002). Statistical methods in assessing agreement: models, issues and tools. *J. Amer. Statist. Assoc.* 97:257–270.

Naik, D. N., Rao, S. S. (2001). Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 28(1):91–105.

Quiroz, J. (2005). Assessment of equivalence using a concordance correlation coefficient in a repeated measurements design. *Journal of Biopharmaceutical Statistics* 15:913–928.

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biometrical Journal* 48:286–301.

Roy, A., Khattree, R. (2007). Classification rules for repeated measures data from biomedical research. In: Khattree, R., Naik, D. N., eds. *Computational Methods in Biomedical Research*. Boca Raton, FL: Chapman & Hall, pp. 323–370.

SAS Institute, Inc. (2004). SAS/STAT User's Guide Version 9. Cary, NC: SAS Institute, Inc.

Shults, J., Morrow, A. L. (2002). The use of quasi-least squares to adjust for two levels of correlation. *Biometrics* 58:521–530.

Shults, J., Whitt, M. C., Kumanyika, S. (2004). Analysis of data with multiple sources of correlation in the framework of generalized estimating equations. *Statistics in Medicine* 23:3209–3226.

St. Laurent, R. T. (1998). Evaluating agreement with a gold standard in method comparison studies. *Biometrics* 54:537–545.

Timm, N. H. (2002). *Applied Multivariate Analysis*. New York: Springer-Verlag.

Timm, N. H., Mieczkowski, T. A. (1997). *Univariate & Multivariate General Linear Models: Theory and Applications Using SAS Software*. Cary, NC: SAS Institute, Inc.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D., Hochberg, Y. (1999). *Multiple Comparisons and Multiple Tests Using SAS*. Cary, NC: SAS Institute, Inc.