Statistical Methods for Measuring Agreement Involving Longitudinal Data

BY

HAIRONG SHI
BS, Nanjing University, Nanjing, China, 1994
MS, University of Illinois at Chicago, Chicago, Illinois, 2004

THESIS

Submitted in Partial Fulfilment of the requirements
for the Degree of Doctor of Philosophy in Public Health Sciences
in the Graduate College of the
University of Illinois at Chicago, 2018

Chicago, Illinois

Defense Committee:
Dulal K. Bhaumik, Chair and Advisor
Sanjib Basu, Biostatistics
Sally A. Freels, Biostatistics
Bikas K. Sinha, Indian Statistical Institute, Kolkata, India
Domenic J. Reda, CSPCC, Hines VA Hospital

To my parents and my son

ACKNOWLEDGMENTS

It is my pleasure to have an opportunity to express my gratitude to all the people who have been accompanying and supporting me during my thesis work.

First I would like to express my deepest gratitude to my thesis advisor, Dr. Dulal Bhaumik, and my director at work, Dr. Domenic J. Reda, for providing stimulating environment that has helped me grow as a biostatistician. I have benefited greatly from both professors' enthusiasm, inspiration, and patience and I am especially grateful for their willingness to provide help when needed.

I would like to thank other members of my thesis committee, Dr. Sanjib Basu, Dr. Sally A. Freels, and Dr. Bikas K. Sinha, for their guidance and help during my thesis work. This thesis would not have become what it is today without their valuable insights and suggestions. Especially I want to thank Dr. Sinha for sharing his papers, methods and ideas for analyzing this type of data. I would like to thank Dr. Clegg for unconditionally sharing the data from the trial that he led. I want to acknowledge the faculty and staff members of the Biostatistics at UIC School of Public Health for their great support that has made my graduate studies an enjoyable experience, especially Dr. Don Hedeker for unconditionally sharing his published papers on longitudinal data, mixed models, pattern mixed models and unpublished manuscripts on three level mixed models. And I would like to. I also want to thank all the fellow students in the division for their friendship and help.

Last but not the least, I would like to express my deepest gratitude to my family and friends. This dissertation would not have been possible without their warm love, continued patience, and endless support. HS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

OA	Osteoarthritis
GAIT	Glucosamine/Chondroitin Arthritis Intervention Trial
CSP	VA Cooperative Studies Program
CCC	Concordance Correlation Coefficient
ICC	Intraclass Correlation Coefficient
MSD	Mean Squared Deviation
CP	Coverage Probability
TDI	Total Deviation Index
GEE	Generalized Estimating Equation
JSW	Joint Space Width
MRM	Mixed-effects Regression Model
BMEM	Bayesian Mixed-Effects Model
EB	Empirical Bayes
EM	Expectation-Maximization
MCMC	Markov Chain Monte Carlo
MML	Maximum Marginal Likelihood
REML	Restricted Maximum Likelihood Estimation
RMSE	Root Mean Square Error
CP	Coverage Probabilities

SUMMARY

Recent research concerning measurements of agreement between different methods or different raters have received wide attention. The concordance correlation coefficient (CCC) has been used to assess agreement between two raters or two measuring methods while the measurements are taken on the same continuous scale. However, the circumstances of repeated measurements may arise, e.g. longitudinal studies in clinical trials or bioassay data with sub-samples. The random variables are not independent nor identically distributed in that kind of situation. To appropriately account for the covariance between measurements, we have fitted three-level linear mixed-effect models with random intercepts at two levels. The model parameters are estimated using an expectation-maximization [E-M] like approach by iterating between the Empirical Bayes [EB] estimates of the random effects and maximum marginal likelihood estimates of the fixed and covariance parameters. For comparing agreement between two raters, we utilize two-level and three-level models to estimate CCC and observe that three-level models fit better for the dataset we collected in GAIT study. In order to handle missing data, we did the analysis with missing values by using mixed-effects model, model imputation, multiple imputation and pattern mixed model. We have achieved at consistent results among all the methods handling missingness in the dataset. The proposed model also gives us the opportunities to evaluate agreement after adjusting for the other covariates. We also use an approach to get the generalized confidence interval of CCC for further statistical inference. Our approach represents a first attempt in evaluating CCC for data with multiple level variations.

1. INTRODUCTION

Osteoarthritis (OA) is the most common form of all arthritides, afflicting at least 16 million persons in the United States. Based on 2010-2012 data from the National Health Interview Survey (NHIS), an estimated 52.5 million (22.7%) of adults have self-reported doctor-diagnosed arthritis. Based on 2003 NHIS data, a projected 67 million (25%) adults aged 18 years or older will have doctor-diagnosed arthritis by the year 2030. Large, weight bearing large joints are often involved, especially the knee, leading to limited activities of daily routine. Effective therapies for OA are limited but include general health measures such as exercise, weight loss, medications including non-steroidal anti-inflammatory drugs (NSAIDs), analgesics, and surgery, most often total joint replacement. Current medical therapies have only been proven to provide symptomatic relief in OA; however, the potential to prevent structural deterioration in OA has been postulated for some medical treatments.

Currently, plain radiography is the 'structure' outcome measure recommended by consensus committees for OA of the knee. Efforts to study the potentially disease-modifying therapies have been hampered by the challenges with accurately measuring disease progression, because of large variability in the rate of expected radiographic joint space narrowing, and the very small differences in Joint Space Width [JSW] that are being measured (2008, [1]; 2004, [2]). Unfortunately, these limitations of measurements remain a barrier to studies of structural modification in OA, as accurate and reproducible measurements are not uniformly obtained.

A number of different protocols to standardize study radiographs have been proposed, some including fluoroscopic guidance with the aim of improving precision (2003, [3]; 2009, [4]). In addition, manual and computer based techniques to measure the JSW

have been used. Some computerized methods have shown better accuracy compared to manual readings (1996, [5]; 1994, [6]).

Glucosamine/Chondroitin Arthritis Intervention Trial, (GAIT) N01-AR-9-2236, is an NIH funded, placebo-controlled, parallel, double-blind, five-arm randomized clinical trial, which was designed to determine whether glucosamine, chondroitin sulfate and/or the combination of glucosamine and chondroitin sulfate are more effective than placebo and whether the combination is more effective than glucosamine or chondroitin sulfate alone in the treatment of knee pain associated with osteoarthritis (OA) of the knee. Daniel O. Clegg, M.D, Professor of Medicine and Chief of the Division of Rheumatology, University of Utah School of Medicine, directed the coordinating center which oversaw the research, patient recruitment, and data collection efforts of thirteen study centers across the country. The Biostatistical Center led by Domenic Reda, Ph.D., was located at the Edward Hines Jr. VA Hospital in Hines, IL. The Biostatistical Center assisted in the planning and development of the study protocol; developed a randomization procedure and distributed information about randomization to all clinical sites; developed data forms and an Operations Manual; established procedures for data entry, data monitoring, and data quality control; monitored patient recruitment and provided monthly enrollment reports; maintained data files; monitored quality control, protocol compliance, and prepared all statistical reports. To further explore issues encountered in measuring Joint Space Width (JSW) from radiographs of knees affected by OA, the GAIT ancillary structural radiographic study is designed to compare results among two manual readers and a computer measurement system.

Subjects enrolled in the GAIT ancillary structural study met the original GAIT inclusion criteria, summarized as aged 40 years or older with clinical evidence of painful OA of the knee for at least six months, and radiographic evidence of OA as determined

by having a Kellgren & Lawrence grade 2- or 3-rated radiograph of the index knee. Six hundred and sixty-two of the 1583 original GAIT study participants were also enrolled in the structural study (2008, [7]; 2006, [8]). For each participant signed informed consent as approved by the IRB and HIPAA was procured. The parent trial was registered (Clinical Trials.gov NCT00032890).

In the GAIT structural study, each entering patient had a semi-flexed weight-bearing radiograph of the knees before starting investigational therapy in the primary study and following 12 and 24 months of therapy. For each knee, the JSW was defined as the narrowest dimension in the medial compartment of the knee. This location was not necessarily at the middle of the weight-bearing surface of the medial tibia (the point used for measuring the rim to floor distance) and/or at any specified distance from the medial condyle. The direction or line upon which the JSW measurement was taken was perpendicular to the plane of the joint surfaces and not necessarily in parallel with the axis of the lower extremity. The site used to measure JSW could not include an osteophyte or involve the tibial spine. All manual measurements were performed using the Mitutoyo Digimatic Calipers (Mitutoyo Products) and recorded to the hundredth millimeter.

Figure 1 is a photograph showing the position for the MTP view of the knee. The first metatarsophalangeal joint is positioned directly below the front of the film cassette. The knees are flexed until the knees touch front and middle of the film cassette. The X-ray beam is horizontal and parallel to the floor. The patient stands on the sheet of paper. The outline of the feet are marked on the sheet so as to facilitate repositioning at subsequent visits.

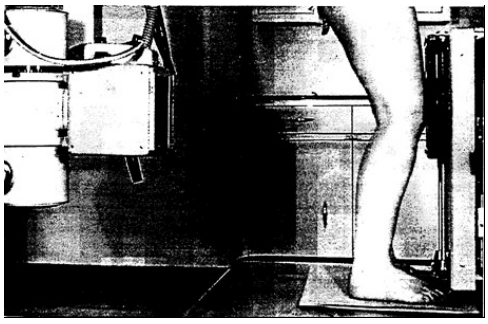

Figure 1: Photograph showing the position for the metatarsophalangeal (MTP) view of the knee

Figure 2 shows both knees appearing in the center of the film. The long axes of the tibiae are parallel to the margins of the film and exposure parameters allow adequate depiction of anatomy. Medial and lateral joint spaces of both knees are correctly delineated. In this case bicompartmental osteoarthritis of both knees predominant in the medial compartments is seen.

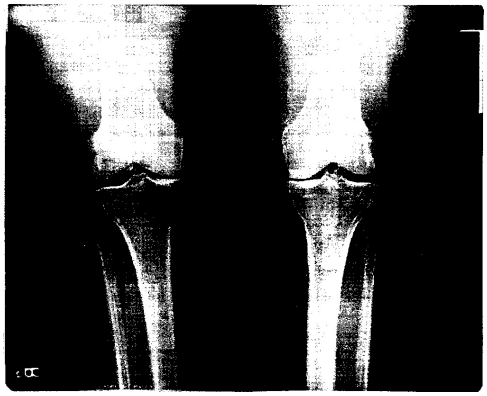

Figure 2: Acceptable: Good Quality Radiograph of the Knees

Figure 3 is a diagram illustrating the normal variability in the posterior inclination of the medial tibial plateau (range 0°-10°) among knees positioned for radiography in the semiflexed MTP view. The tibiofemoral angle is determined by the radiographic position. The angle between the 2 limb bones is similar between examinations and is only changed if the patient unusually alters his/her pelvic tilt. The site of JSW measurement remains the same between knees and coincides with the load transmission region across the joint (arrows), and is parallel to the film and perpendicular to the femoral and tibial margins.

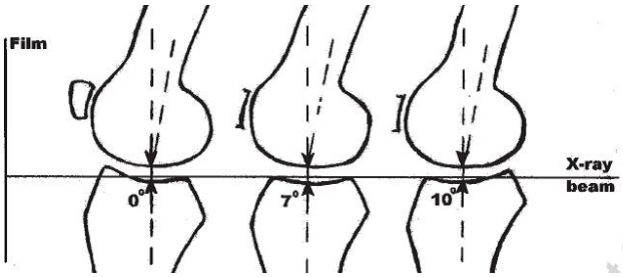

Figure 3: Diagram illustrating the normal variability in the posterior inclination of the medial tibial plateau (range 0°-10°) among knees positioned for radiography in the semiflexed MTP view.

Each radiograph was evaluated for meeting inclusion criteria for the structural study. The images were blinded and coded at the Hines CSP Coordinating Center as to patient name, participating clinic, treatment group and date the X-ray was taken. The X-rays were read by physician investigators without knowledge of the patient name, participating clinic, treatment, and date of X-ray. These blinded images were read by physician investigators in matched pre-post pairs, but read in a randomly assigned order. Each X-ray was interpreted independently by two investigators: Dr. Williams, MD, a rheumatologist with extensive experience in clinical investigation and previous experience in

radiographic interpretation of clinical trials, and Dr. Julia Crim, MD, a musculoskeletal radiologist. Each reader reviewed plain hard copy radiographs and measured the JSW with calipers. An additional non-technical rater used the computer program Mdisplay by Buckland-Wright (1994, [9]) to measure JSW on digitized images of each film. Mdisplay is a semi-automated program requiring a user to mark the endpoints of the medial tibial and femoral condyles and then an edge finding algorithm determines the joint space borders.

To determine whether manual measurement of JSW on plain radiographs is equally as reproducible as computer generated measurements of digitalized radiographs, it is necessary to perform statistical evaluation of the agreement. Ideally, new measurements would be compared with true values; however, in the case of measuring JSW - as is often the case - true values of the measurement in question are unknown. Typically, new measurements are compared with measurements from the gold standard methods of the field in question. In statistics, the term 'Agreement' refers to both accuracy when a true value is known and precision when the true value is unknown.

In Chapter 2, the general statistical literature review of agreement will be addressed. In Chapter 3, we review the methodology of evaluating agreement. In Chapter 4, we provide the three-level mixed-effect models we used to estimate the extent of agreement. In Chapter 5, we provide the results of simulations. In Chapter 6, we apply the method to the example of JSW. In Chapter 7, we apply the different methods for handling missing data. In Chapter 8 we adjust other covariates to evaluate agreement between raters. In Chapter 9 we estimate the generalized confidence interval for CCC by generalized pivot statistics. Finally, we draw conclusions and provide some discussions in Chapter 10.

2. STATISTICAL LITERATURE REVIEW ON AGREEMENT

Cohen (1960, [10]) defined kappa as a coefficient of inter-rater agreement to evaluate the degree of agreement for nominal scales. Nominal scales were at the time popular in the clinical-social-personality areas of psychology. Historically, two or more raters independently categorized items, and simple agreement between raters was sufficient for categorization without being held to predetermined, outside rating criteria. There was no sense of the correctness of categorization, and the raters equally and independently applied their judgment. Additionally, there were no restrictions placed on the distribution of ratings over categories for either rater. There existed a need to determine the reliability of rater judgments, and Cohen's kappa attempted to fill this void. In order to calculate kappa, data structure first paired ordinal observations with a bivariate multinomial distribution. Assuming that the units were independent, the categories of the nominal scale were independent, and the raters operated independently; kappa was interpreted as the proportion of joint categorizations in which there is agreement after excluding the contribution due to chance agreement.

Cohen (1968, [11]) later developed a weighted kappa which counted disagreements differently and proved especially useful when categories were ordinal. Three matrices were involved : the matrix of observed scores, the matrix of expected scores based on chance agreement, and the weight matrix. Weight matrix cells located on the diagonal (upper-left to bottom-right) represent disagreement and thus contain zeros. Off-diagonal cells contain weights indicating the seriousness of that disagreement. Commonly, cells off the diagonal were weighted according to their distance from the diagonal; adjacent cells were weighted 1, cells two spaces from the diagonal were weighted 2, and so on. Cohen's

kappa is limited to the measurement of agreement between two raters and is applied to categorical data only.

Fleiss (1971, [12]) developed Fleiss' kappa for a similar measure of agreement when there are more than two raters. The Fleiss kappa, however, is a multi-rater generalization of Scott's π statistic, not Cohen's kappa, and was designed to take into account the possibility of guessing. However, if the assumptions of rater independence and other factors are not met, Fleiss' kappa may underestimate the true level of agreement (Zapf, 2016, [72]). Furthermore, values of Fleiss' kappa do not have a direct interpretation, and thus it has become common for researchers to accept low kappa values in their inter-rater reliability studies.

Low levels of inter-rater reliability are not acceptable in health care or in clinical research, especially when results of studies may change clinical practice in a way that leads to poor patient outcomes. The Pearson correlation coefficient measures a linear relationship only, and fails to detect any departure from the 45° line or any other relationships. The paired t-test and the least square analysis may mislead when data are scattered or data have different trends other than agreement. The traditional methods of measuring agreement on continuous variables such as the Pearson correlation coefficient, the paired t-test, or the least squares analysis of slope (= 1) and intercept (= 0), have obvious limitations.

Several recent statistical approaches to measure continuous variable agreement are discussed below:

(i) Descriptive tools, such as pairwise plots with a 45-degree line, and Bland-Altman plots (Bland and Altman, 1986, [13]);

(ii) Unscaled summary indices based on absolute differences of measurements, such as mean squared deviation including repeatability coefficient and reproducibility coefficient, limits of agreement (Bland and Altman, 1999, [33]), coverage probability, and total deviation index (Lin et al., 2002, [18]);

(iii) Scaled summary indices, such as the intraclass correlation coefficient (ICC), concordance correlation coefficient (CCC), coefficient of individual agreement, and dependability coefficient. (Lin (1989), (1992), [14, 15]).

The pairwise plot with a 45^0 line passing through the origin is the simplest display comparing the results of two methods of measurement. Ideally, all the observations would be on or around the 45^0 line. In practice, however, data points will be clustered near the line and it will be difficult to assess between-method differences. The Bland-Altman plot [13] displays the differences of methods against their mean. If the differences in measurements are normally distributed, 95% of differences would lie between the mean difference ± 1.96 x SD. In practice, two times standard deviation is used.

When using the simple graphical methods changed to describing the 95% confidence interval of the difference of measurement, the mean and standard deviation of the differences were assumed to be the same throughout the range of measurement. However, the mean difference may also be approximately proportional to the magnitude of the measurement. In other words, the variances may increase when mean difference between raters increases.

Bland and Altman (1986, [13)] extended the basic approach to resolve the issue of inconsistent variance with a simple logarithmic transformation approach and a regression approach. They used one-way analysis of variance, with subject as the factor to estimate the within-subject standard deviation from the square root of the residual mean square

and assumed the mean difference between replicates to be zero. A nonparametric approach by using the percentage of change was used when the distribution of data was skewed. This method displays the difference and magnitude of agreement, but still does not show the scale.

Lin (1989, [14]; 1992, [15]) proposed a new index, Concordance Correlation Coefficient (CCC), to evaluate reproducibility. CCC evaluates the agreement between two readings from the same sample by measuring the variation from the 45^0 line through the origin, or the concordance line. Departure from the standard is measured by how far the observations deviate from the concordance line on a scale of 1 (perfect agreement) to -1 (perfect reversed agreement or, perfect disagreement), and including 0 (no agreement). It consists of a measure of precision (not correctable) multiplied by a measure of accuracy (correctable), and was referred to as ρ_c. This index measured a scale shift (ratio of 2 standard deviations) and a location shift.

We will use CCC to evaluate agreement between different raters in this dissertation. CCC depends largely on analytical range and the intra-sample variation. In certain cases where there are practical difficulties using CCC, the mean of squared difference (MSD) is the recommended approach. MSD is a good statistical index when individual departure of paired data is of particular interest. Lin et al. (2002, [18]) introduced the total deviation index (TDI $(1-p)$) and coverage probability (CP). This proposed method aims to capture a proportion of data (difference of two observers) called coverage probability within a boundary, or total deviation index, from target value. For example, we will evaluate that at least 80% of observations are located within 10% relative deviation of their target values. In this case, we set the TDI at 10% and test whether or not CP exceeds 80%. The TDI $(1-p)$ describes a boundary such that a majority $100(1-p)$ percent of the differences (or percentages for the log transformed data) of paired observations are within

the boundary. The limit is similar to that of the prediction interval.

Since 1989, CCC has been widely used as a measure of reproducibility in practice. Lin's method was applicable for studies evaluating two raters or methods without replication. Chinchilli et al. (2001, [29]; 2009, [50]) extended Lin's approach to repeated measures designs by using a weighted concordance correlation coefficient. However, those methods cannot adjust the effects of covariates, especially when one needs to model agreement of multiple readings. Barnhart et al. (2001, [28]; 2010, [57]) modeled CCC via three sets of estimating equations by a generalized estimating equation (GEE) approach. The proposed approach is flexible for several reasons: first, it can incorporate more than two correlated readings and test for the equality of pairwise concordant correlation coefficient estimates; second, it can incorporate covariates towards prediction in marginal distributions; third, it can be used to identify covariates towards prediction of concordance correlation coefficients; and finally, it requires minimal distribution assumptions. However, the proposed GEE method for estimating the CCC has the tendency to underestimate the true concordance correlation coefficient when sample size is small. This may be due to the fact that the empirically corrected standard error is smaller than the actual standard deviation. This is also the limitation of GEE approach.

Instead of moment estimation methods, Carrasco (2003, [17]) proposed a mixed effects model to estimate the CCC. He also demonstrated that ICC and CCC are the same measure of agreement estimated by assuming the raters had a fixed effect, and so the contribution of the variability of the raters' means to the ICC will be a sum of squares rather than a variance. The CCC can be extended to more than two raters by using the variance components approach. It can be adjusted for confounding covariates by incorporating them into the mixed model. The CCC of variance components estimation results in a more accurate point estimate than does the moment method.

In summary, the pairwise plot and the Bland-Altman plot (1986, [13]) convey an intuitive sense of agreement visually, but do not quantify the degree of agreement. Lin's CCC measures both scale and location shifts to quantify the agreement between two raters. He also assumes the samples to be independent. Chinchilli (2001, [29]) extended Lin's approach for repeated measures. Barnhart (2001, [28]) modeled CCC by a generalized estimating equation (GEE) approach, but this method may result in underestimation of the CCC. Carrasco (2003, [17]) proposed a mixed effects model to estimate ICC. By assuming a fixed rater effect, he demonstrated that ICC and CCC are the same. In our study, three raters evaluated subjects' joint space width at baseline and one/two years post-treatment. In order to find the intra-rater CCC, we randomly selected 29 subjects at each time point for repeated readings. There are three levels of correlation in this study: 1) multiple X-rays at the same time point; 2) individual person level correlation over time; and 3) group level correlations. We propose a three-level mixed model to estimate the CCC.

3. METHODOLOGY

3.1 Cohen's kappa on categorical data

Cohen's kappa measures the agreement between two raters classifying subjects into distinct, mutually exclusive categories. Suppose raters X and Y are evaluating one variable with k categories from N independent subjects. Let p_{11} represent the percentage of the subjects both raters assigned to category 1 given all subjects,..., p_{tt} represent the percentage of the subjects both raters assigned to category t given all subjects, and p_{ij} represent the percentage of the subjects assigned to category i by one rater and assigned to category j by the other given all subjects. Further, let $p_{.j}$ denote the marginal percentage of column j and $p_{i.}$ denote the marginal percentage of row i.

Table I: Agreement between two raters on one categorical variable with T categories

		Rater Y				
Rater X	Category	1	2	...	k	Total
	1	p_{11}	p_{12}	...	p_{1k}	$p_{1.}$
	2	p_{21}	p_{22}	...	p_{2k}	$p_{2.}$
	...					
	T	p_{k1}	p_{k2}	...	p_{kk}	$p_{k.}$
	total	$p_{.1}$			$p_{.k}$	1.0

Let

$$p_o = p_{11} + p_{22} + ... + p_{kk}, \tag{3.1}$$

and

$$p_e = p_{.1}p_{1.} + p_{.2}p_{2.} + ... + p_{.k}p_{k.}, \tag{3.2}$$

Cohen's Kappa is

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{3.3}$$

where p_o is the sum of percentage at diagonals, or observed agreement between raters; p_e is the sum of the products of marginal distributions, or the estimated hypothetical probability of chance agreement; κ represents Cohen's kappa, adjusting for the possibility of agreement occurring by chance.

If κ is less than 0 (i.e., if the observed agreement is less than what would have been expected by chance), there is no further practical interest. If $\kappa = 1$, then the raters are in complete agreement. For a κ less than 0.70, generally the inter-rater agreement is considered to be poor; on the other hand, when κ is greater than 0.70, inter-rater agreement is considered to be satisfactory. Below, we examine two examples showing when Kappa is adequate (example 1) and when kappa is inadequate (example 2) for evaluating agreement between 2 raters involving k = 2 categories of classification only.

Table II: Example 1: Cohen's Kappa evaluation showing perfect agreement

	Rater Y				
Rater X	Category	1	2	Total	
	1	30	0	30	$n_{.1}$
	2	0	70	70	$n_{.2}$
	total	30	70	100	$n_{..}$
		$n_{1.}$	$n_{2.}$		

In this case, Cohen's kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{1 - 0.58}{1 - 0.58} = 1.$$

Table III: Example 2: Cohen's Kappa evaluation showing inadequate agreement

	Rater Y				
Rater X	Category	1	2	Total	
	1	1	6	7	$n_{.1}$
	2	13	80	93	$n_{.2}$
	total	14	86	100	$n_{..}$
		$n_{1.}$	$n_{2.}$		

Here, Cohen's kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0.81 - 0.8096}{1 - 0.8096} = 0.002.$$

We can see that kappa works perfectly well when all the n subjects are rated by the two raters with 100 percent 'matching' between the two categories. That simply means that the sum of the diagonal frequencies is n — irrespective of the decomposition. However, Cohen's kappa performs poorly about inter-rater agreement in case of marginal heterogeneity as in the second example — even though there is predominantly high percentage of match [81 out of 100] between the two raters.

We reiterate that in case of 100 percent matching along the main diagonal, Cohen's kappa always takes the value 1 - as it should. Is it also equally true that in case of 100 percent 'mismatch' i.e., 100 percent matching along the 'anti-diagonals', kappa assumes

the value '-1' ? It is tempting to assert that kappa is -1, that means it is a case of completely reverse agreement or perfect disagreement. Sinha et al (2006) [26] pointed out an interesting feature of kappa in this scenario. Example 3 shows a situation wherein there is not even a single agreement case on the diagonals.

Table IV: Example 3: Cohen's Kappa evaluation under 100 percent mismatch

	Rater Y				
Rater X	Category	1	2	Total	
	1	0	60	60	$n_{.1}$
	2	40	0	40	$n_{.2}$
	total	40	60	100	$n_{..}$
		$n_{1.}$	$n_{2.}$		

Here, Cohen's kappa is:

$$\kappa = \frac{p_o - p_e}{1 - p_e} = \frac{0 - 0.48}{1 - 0.48} = -0.923.$$

Sinha et al (2006, [26]) observed that kappa captures 100 percent reverse agreement i.e., assumes the value of '-1' if and only if $p_e = 0.5$, that means $p_{12} = p_{21} = 0.5$ for the two categories.

This reflects on a 'defect' or imperfection in the definition of kappa. Sinha et al (2006, [26]) went ahead to modify the original kappa formula as:

$$\kappa_{M1} = \frac{p_o - p_e}{\pi_{1.}\pi_{2.} + \pi_{.1}\pi_{.2}}.$$

Here, $\pi_{i.} = n_{i.}/n$, $\pi_{.j} = n_{.j}/n$. The modification is based on the analysis of situation

leading to total disagreement between two raters. Computation now leads to $\kappa = \frac{0-0.48}{0.48} = -1$.

However, we will not discuss this issue any further nor we will adopt this modified formula for kappa in our study.

To overcome the limitation of Cohen's kappa being applicable for two raters, Fleiss' kappa (1971, [12]) addresses agreement for any number of raters giving categorical assignments to a fixed number of subjects/items. Again, it does not work well in case of marginal heterogeneity or reverse agreement when the cumulative product of marginals is not equal to 0.5.

As before, let N be the total number of subjects, and n be the number of raters, and let k be the number of categories into which assignments are made independently by each of the n raters of each of the N subjects. The subjects are indexed by i = 1, ... N and the categories are indexed by j = 1, ... k. Let n_{ij} represent the number of raters who assigned the i-th subject to the j-th category.

Let p_j be the proportion of all assignments which are attributed to the j-th category:

$$p_j = \frac{1}{Nn}(\sum_{i=1}^{N} n_{ij}), \sum_{j=1}^{k} p_j = 1;$$

where

$$\sum_{j=1}^{k} n_{ij} = n, \sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij} = Nn.$$

Since number of all possible pairwise raters among n raters is $n(n-1)/2$, the extent of agreement among the n raters for the ith subject would be the proportion of pairwise raters matching for the jth category, given the total number of pairs, for all the categories

combined. Therefore, p_i is defined as:

$$p_i = \frac{1}{n(n-1)/2} \sum_{j=1}^{k} n_{ij}(n_{ij}-1)/2,$$

$\bar{p}$ is the mean of the $p_i's$

$$\bar{p} = \frac{1}{N} \sum_{i=1}^{N} p_i.$$

If the raters made their assignments completely at random, the mean proportion of agreement is expected to be:

$$\bar{p_e} = \sum_{j=1}^{k} p_j^2 = \sum_{j=1}^{k} (\frac{1}{Nn} \sum_{i=1}^{N} n_{ij})^2.$$

Fleiss' kappa is defined as

$$\kappa = \frac{\bar{p} - \bar{p_e}}{1 - \bar{p_e}}. \tag{3.4}$$

The factor $1 - \bar{p_e}$ gives the degree of agreement that is attainable above that attributable to chance, and $\bar{p} - \bar{p_e}$ gives the degree of agreement actually achieved above chance. The statistic κ takes values between 0 and 1, where a value of 1 means complete agreement. Fleiss' initial goal was to extend kappa to 3 raters and more. But Fleiss' generalized statistic does not reduce to kappa if the number of raters is 2. Instead, it reduces to another agreement coefficient called Π, proposed by Scott (1955, [32]). Nonetheless, Fleiss decided to refer to his coefficient as a generalized kappa.

3.2 Paired t-test and Pearson correlation coefficient

Cohen's kappa and Fleiss' kappa are commonly used to evaluate agreement on categorical variables; however, evaluating the agreement between two raters on a continuous variable/scale is another common question. There are several traditional statistics or tests used for that purpose, the Pearson correlation coefficient being one of them. Pearson's correlation coefficient is defined as

$$\rho_{x,y} = \frac{Cov(x,y)}{\sigma_x \sigma_y}, \tag{3.5}$$

that is, the covariance of the two variables divided by the product of their standard deviations.

Another traditional method is developed by considering a paired vector of observations X and Y, with all paired data around the 45^0 concordance line $X = Y$. If each reading in X is identical to the corresponding one in Y, then we say X and Y are in perfect agreement. Bland and Altman (1986, [13]; 1999, [33]) used the paired t-test to see how closely the samples agree in paired samples.

The test statistic is calculated as:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}, \tag{3.6}$$

where $\bar{d}$ is the mean difference, s^2 is the sample variance based on the differences, n is the sample size and t is the Student t with n-1 degrees of freedom.

However, Pearson's correlation coefficient reflects the linear relationship. Lin (1989) pointed out that the results by those methods may be misleading. Figure 4 shows three

cases where Pearson's correlation coefficients are very high due to strong linear relationships, but in fact there is strong disagreement between raters. The upper plot shows a situation where Pearson's correlation coefficient shows highly significant linear relationship, however, the rater Y's measurements are always higher than X rater, that means, fails to detect the disagreement based on location; the middle plot shows strong linear relationsip, however, a failure to detect disagreement on scale, and the lower plot shows strong linear relationship, however, a failure to detect disagreement on both location and scale.

The paired t-test evaluates means rather than individual pairs. Figure 5 shows several examples where paired t-test misleads the agreement between measurements. The paired t-test data in the top four plots in Figure 5 will fail to reject H_0 even though there is strong disagreement between the raters because the mean of the difference between ratings in these examples is close to 0. The paired t-test fails to explain the nature of the relationship between the responses given by the two raters. In the final example in Figure 5 the difference between the two raters is very consistent, although one rater had a reliably higher evaluation than the other. Therefore when the standard deviation of the difference is small enough, and when the number of pairs is large enough, the result of mean difference divided by the standard error will be large enough that the test statistic will reject H_0 despite showing strong agreement between two raters. Yet the paired t-test fails to highlight the nature of the agreement, as the rater on the Y axis consistently rates the items higher than the rater on the X axis, a disagreement on location despite overall strong agreement on scale.

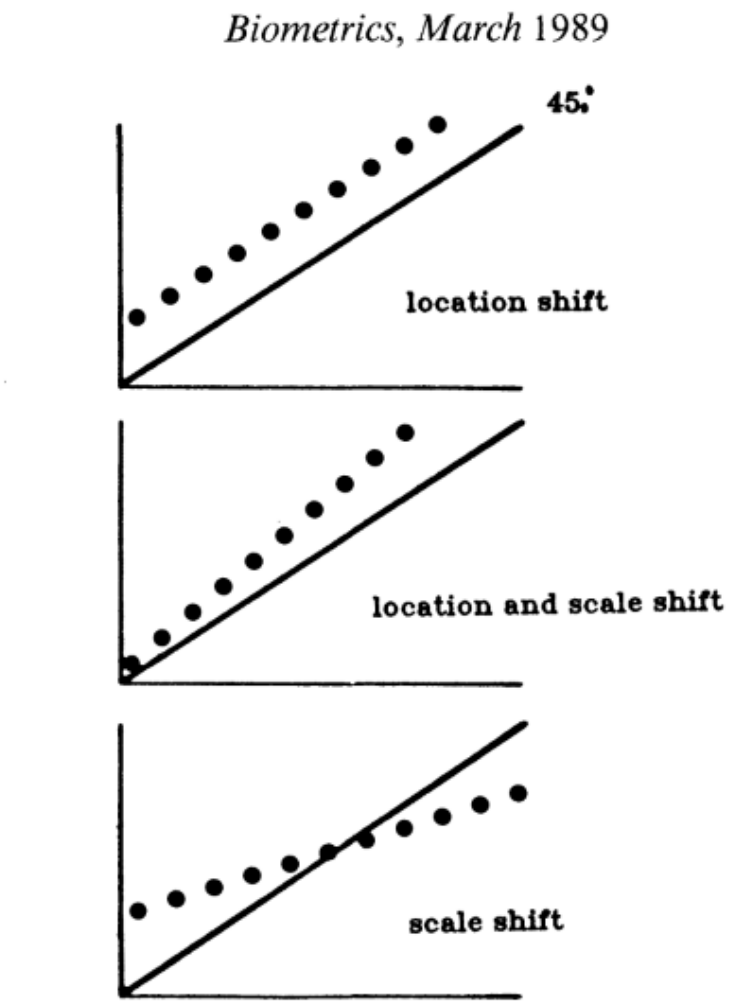

Figure 4: Three cases where Pearson's correlation coefficient fails to detect disagreement

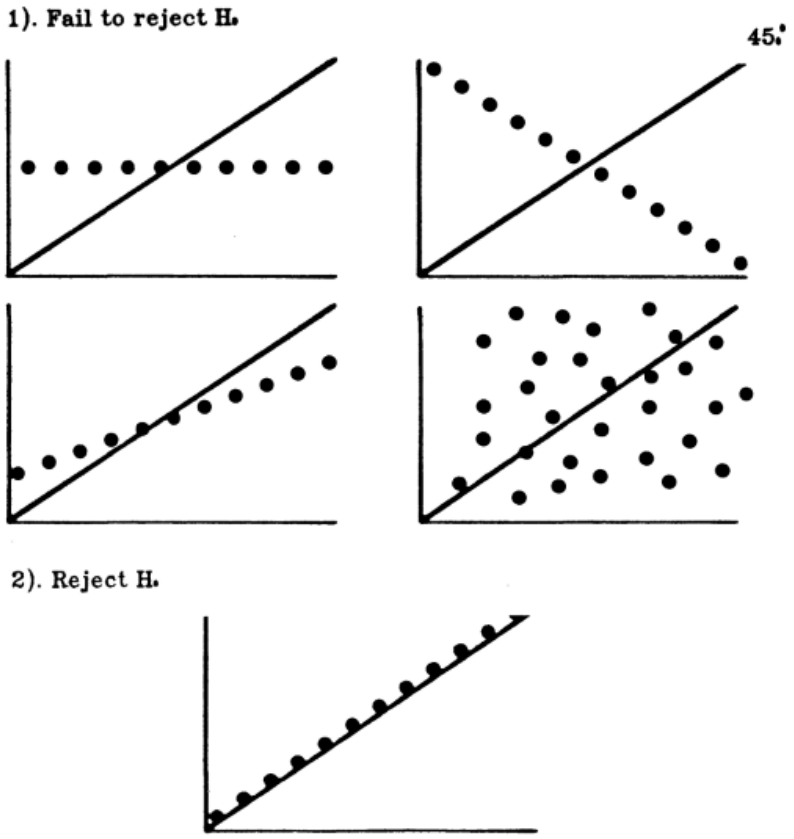

Figure 5: Several cases where paired t-test misleads to the conclusion of agreement measures

3.3 Lin's Mean Squared Deviation

In order to quantify the agreement on both location and scale, Lin (1989, [14]) introduced the Mean Squared Deviation [MSD] to evaluate an aggregated deviation from the identity line, a frequently- used measure for studying the differences between two vectors of observations. When there are two raters MSD is defined as $E(Y - X)^2$. It can

be expressed as:

$$MSD = (\mu_y - \mu_x)^2 + \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}(= \epsilon^2, say). \tag{3.7}$$

The notations are as usual. From equation (3.7), we can see that MSD is always ≥ 0, MSD increases when there is a large difference between means or variances, or both. The drawback of the MSD is that though it is intuitively meaningful, there is no easy way to set up an upper acceptable limit for MSD to evaluate the goodness of the agreement.

Naturally, when we have paired observations based on the ratings provided by the two raters independently on each of n subjects, the sample analogue of MSD is based on observed means, variances and the correlation coefficient. Note that σ_{yx} stands for population covariance between X and Y.

3.4 Lin's Concordance Correlation Coefficient

In a series of papers, Lin (1989, [14]; 1992, [15]; 2007, [16]) and Lin et al. (2002, [18]) studied various aspects of MSD involving two raters in a continuous scale. In order to avoid the drawbacks of MSD, Concordance Correlation Coefficient (CCC) was introduced to measure the agreement between two raters in a standardized scale. It is denoted by ρ_c and is defined as

$$\begin{aligned}
\rho_c =& 1 - \frac{E(Y-X)^2}{E(Y-X)^2 \mid \rho = 0} \\
=& 1 - \frac{\epsilon^2}{\sigma_y^2 + \sigma_x^2 + (\mu_y - \mu_x)^2} \\
=& \frac{2\sigma_{yx}}{\sigma_y^2 + \sigma_x^2 + (\mu_y - \mu_x)^2}.
\end{aligned} \tag{3.8}$$

Here, $E(Y-X)^2$ gives the mean square for within sample total deviation,and $E(Y-X)^2 \mid \rho = 0$ gives the mean square for total deviation under zero correlation set-up. So CCC is a standardized form of MSD lying between [-1, 1].

When two raters are grading n independent subjects, paired observations on Y and X are randomly collected, and it is assumed that

(i) Y_i and X_i have a bivariate distribution with mean $\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}$, and variance $\begin{bmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{bmatrix}$;

(ii) Y_i and Y_j are independent when $i \neq j$;

(iii) X_i and X_j are independent when $i \neq j$.

The sample counter-part of ρ_c is the so-called plug-in estimator based on sample means, sample variances and the sample covariance.

In order to achieve an approximation to the normal distribution, Lin (1989) transformed ρ_c using Fisher's z transformation to obtain

$$\hat{\lambda} = tanh^{-1}(\hat{\rho_c}) = \frac{1}{2} ln(\frac{1+\hat{\rho}_c}{1-\hat{\rho}_c}), \tag{3.9}$$

This quantity has an asymptotically normal distribution with mean $tanh^{-1}(\rho_c)$, and variance

$$\frac{1}{n-2}\{\frac{(1-\rho^2)\rho_c^2}{(1-\rho_c^2)\rho^2} + \frac{2\rho_c^3(1-\rho_c)\nu^2}{\rho(1-\rho_c^2)^2} - \frac{\rho_c^4\nu^4}{2\rho^2(1-\rho_c^2)^2}\},$$

where,

$$\nu = \frac{\mid \mu_y - \mu_x \mid}{\sqrt{\sigma_x \sigma_y}}.$$

From equation (3.8), since,

$$\frac{E(Y-X)^2}{E((Y-X)^2 \mid \rho = 0)} \geq 0; \tag{3.10}$$
$$1 - \frac{E(Y-X)^2}{E((Y-X)^2 \mid \rho = 0)} \leq 1;$$

and,

$$E[(y-\mu_y)+(x-\mu_x)]^2 \geq 0; \tag{3.11}$$
$$(\mu_x-\mu_y)^2 \geq 0;$$
$$\sigma_y^2+\sigma_x^2+2\sigma_{yx}+(\mu_y-\mu_x)^2 \geq 0;$$
$$2\sigma_{yx} \geq -(\sigma_y^2+\sigma_x^2+(\mu_y-\mu_x)^2);$$
$$\frac{2\sigma_{yx}}{\sigma_y^2+\sigma_x^2+(\mu_y-\mu_x)^2} \geq -1.$$

It follows that CCC is between -1 and 1. CCC will be 1 only when $\mu_y = \mu_x$ and $\sigma_{yx} = \sigma_y^2 = \sigma_x^2$, or the distribution of observations from two raters are identical. CCC will be 0 when $\sigma_{yx} = 0$, meaning the distribution of observations from two raters are independent. CCC will be -1 when $\mu_y = \mu_x$ and -$\sigma_{yx} = \sigma_y^2 = \sigma_x^2$, or the pair of y and x are on line of y = -x, -45^0. So we can see that CCC measures agreement on both location and scale between two raters, but it is limited to only two raters.

3.5 Mixed-Effects Regression Models

In our study, the JSW of participants was measured repeatedly at three time points. The covariance between repeated measurements on the same subject, covariance between categorizations assigned by the same rater, and covariance between the raters reading the same feature-related naturally come into the picture and all these have a direct role to play in the analysis of data. Traditional methods are limited due to restrictive assumptions concerning the variance-covariance structure of the repeated measures. The univariate "mixed-model" analysis of variance assumes that the variances and covariances of the dependent variable across time are equal. Mixed-effects regression models (MRMs) are used to estimate the sample mean and variance-covariance structure. Hedeker and Gibbons [19] point out that a basic characteristic of MRMs is the inclusion of random subject effects into the regression model in order to account for the influence of subjects on their repeated observations. When the same subjects are repeatedly measured over time, their responses are correlated over time, and their estimated trend line or curve can be expected to deviate systematically from the overall mean trend line. Additionally, they indicate the degree of subject variation that exists in the population of subjects.

If individuals have no influence on their repeated outcomes, then all of the random terms would equal 0. However, it is more likely that subjects will have positive or negative influences on their longitudinal data, and so the random terms will deviate from 0. In addition, it may be assumed that the errors of measurement are conditionally independent and this seems to be more reasonable than the ordinary independence assumption associated with the general linear model. Vide Laird and Ware (1982, [20]).

Firstly, the traditional 2-level model for longitudinal data is described as follows:

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{v}_i + \boldsymbol{\varepsilon_i},$$

where $i = 1, 2, ..., N$ subjects, $j = 1, 2, ..., n_i$ observations from subject i;

(i) $\boldsymbol{y}_i$ is the $n_i \times 1$ response vector for individual i;

(ii) $\boldsymbol{X}_i$ is the $n_i \times p$ design matrix for the fixed effects;

(iii) $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed effects parameters;

(iv) $\boldsymbol{Z}_i$ is the $n_i \times r$ design matrix for the random effects;

(v) $\boldsymbol{v}_i$ is an $r \times 1$ vector of unknown random effects following $N(0, \Sigma_\nu)$,

(vi) $\boldsymbol{\varepsilon_i}$ is a $n_i \times 1$ residual vector following $N(0, \sigma^2 I_{ni})$.

The within-subjects mixed effect level-1 model for typical random intercept and trend ($j = 1, ..., n_i$)

$$y_{ij} = b_{0i} + b_{1i}X_{1_{ij}} + \epsilon_{ij},$$

The between-subjects level-2 model ($i = 1, ..., N$)

$$b_{0i} = \beta_0 + \nu_{0i};$$

$$b_{1i} = \beta_1 + \nu_{1i}.$$

Here, β_0 is the group level intercept; β_1 is the group level-1 slope; ν_{0i} is the ith individual's deviation intercept from the average of the group, and ν_{1i} is the ith individual's

slope deviation from the group level slope. In practice, the number of levels of data must be considered when the multilevel data are collected for more than two levels. As long as the variance attributable to a higher-order level is perceived, the higher-order level may/should be included in the model.

A 3-level model for longitudinal data is described as follows:

$$\boldsymbol{y}_{ijk} = \boldsymbol{X}_{ijk}\boldsymbol{\beta} + \boldsymbol{Z}_{(3)ijk}\boldsymbol{v}_i + \boldsymbol{Z}_{(2)ijk}\boldsymbol{u}_{ij} + \boldsymbol{\varepsilon_{ijk}},$$

where $i = 1, 2, ..., N$ subjects, denote the level-3 units; $j = 1, 2, ..., n_i$ observations, denote level-2 units nested within the i-th level-3 unit; $k = 1, 2, ..., n_{ij}$, denote level-1 units nested within the pair (i, j). So there are N level-3 units, $\sum_{i=1}^{N} n_i$ level-2 units and $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij}$ level-1 units.

(i) $\boldsymbol{y}_{ijk}$ is the vector of $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij} \times 1$, where i denotes the level-3 unit, j denotes the level-2 unit nested within the i^{th} level-3 unit, and k denotes the level-1 unit nested within the pair (i, j);

(ii) $\boldsymbol{X}_{ijk}$ is the $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij} \times p$ design matrix for the fixed effects;

(iii) $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown fixed parameters;

(iv) $\boldsymbol{Z}_{(3)ijk}$ is the $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij} \times r$ design matrix for the level-3 random effects;

(v) $\boldsymbol{v}_i$ is an $r \times 1$ vector of unknown level-3 random effects following $N(0, \Sigma_\nu)$;

(vi) $\boldsymbol{Z}_{(2)ijk}$ is the $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij} \times q$ design matrix for the level-2 random effects;

(vii) $\boldsymbol{u}_{ij}$ is an $q \times 1$ vector of unknown level-2 random effects following $N(0, \Sigma_u)$;

(viii) $\boldsymbol{\varepsilon_{ijk}}$ is a vector of $\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij} \times 1$ standing for residual vector following $N(0, \sigma^2 I_{\sum_{i=1}^{N}\sum_{j=1}^{n_i} n_{ij}})$.

3.6 Estimation of Concordance Correlation Coefficient

When observations are not independent, Fleiss (1971, [12]) proposed the following model for continuous variables measured at m time periods from N subjects by k raters.

$$Y_{ijl} = \mu + \alpha_i + \beta_j + \epsilon_{ijl}, \tag{3.12}$$

where individual $i = 1, 2, ..., N$, observer $j = 1, 2, ..., k$, and measurement $l = 1, 2, ..., m$; μ is the overall mean, α_i is the individual effect (random), β_j is the fixed effect of the jth rater, and ϵ_{ijk} is the random error.

Assumption : $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$, and error terms are independent with any other component of the measurement model. Intraclass correlation coefficients (ICC) are measures of the relative similarity of quantities which share the same observational units of a sampling and/or measurement process. Carrasco (2003, [17]) assumed the rater's effect is fixed. Based on this model and assumption, the intraclass correlation coefficient (ICC) is:

$$ICC = \rho_{ICC} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2}. \tag{3.13}$$

After assuming that raters' effects are fixed, and that one measurement by each rater on each subject is taken (m =1), the following equalities are fulfilled.

Let $\mu_j = \mu + \beta_j$, and $\mu_i = \mu + \beta_i$.

$$\sigma_\alpha^2 = \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sigma_{ij},$$

$$\sigma_\beta^2 = \frac{1}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\mu_i - \mu_j)^2,$$

$$\begin{aligned}\sigma_\epsilon^2 =& \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \frac{1}{2}(\sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}) \\ =& \frac{1}{k} \sum_{i=1}^{k} \sigma_i^2 - \frac{2}{k(k-1)} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sigma_{ij}.\end{aligned}$$

where σ_i^2 and μ are the variance and mean of the measurements made by rater i; σ_{ij} is the covariance between raters i and j. Putting all of those together,

$$\rho_{ICC} = \frac{2 \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sigma_{ij}}{(k-1) \sum_{i=1}^{k} \sigma_i^2 + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} (\mu_i - \mu_j)^2}. \tag{3.14}$$

That is exactly the same expression as the CCC for k raters, implying that CCC can be estimated by variance components through a mixed effects model. The beauty of this method is that it measures the CCC for more than two raters. However, there is a strong assumption that the raters have fixed effects and that samples are independent.

3.7 Hypothesis testing of agreement

Sinha and Dutta (2013, [27]) developed a likelihood ratio test for a hypothesis of the form $H_0 : |\mu_x - \mu_y| \geq \epsilon_0$, $\frac{\sigma_x}{\sigma_y}$ or $\frac{\sigma_y}{\sigma_x} \geq \eta_0$, $\rho \leq \rho_0$ where ϵ_0 is close to 0, and η_0 and ρ_0 are close to 1 by assuming x and y are paired measurements by two raters. They follow bivariate normal distribution, denoted by:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N_2 \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \right). \tag{3.15}$$

This testing problem is equivalent to testing the union of four composite hypotheses:

$$
\begin{aligned}
H_{01} &: \mu_x = \mu, \mu_y = \mu + \epsilon_0, \sigma_x = \sigma, \sigma_y = \sigma\eta_0, \rho = \rho_0; \\
H_{02} &: \mu_x = \mu, \mu_y = \mu + \epsilon_0, \sigma_x = \sigma, \sigma_y = \frac{\sigma}{\eta_0}, \rho = \rho_0; \\
H_{03} &: \mu_x = \mu, \mu_y = \mu - \epsilon_0, \sigma_x = \sigma, \sigma_y = \sigma\eta_0, \rho = \rho_0; \\
H_{04} &: \mu_x = \mu, \mu_y = \mu - \epsilon_0, \sigma_x = \sigma, \sigma_y = \frac{\sigma}{\eta_0}, \rho = \rho_0.
\end{aligned} \tag{3.16}
$$

The likelihood function is written as

$$
\begin{aligned}
L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho|data) = & \frac{1}{(2\pi\sigma_x\sigma_y\sqrt{1-\rho^2})^n} exp[-\frac{1}{2(1-\rho^2)} \\
& \sum_{i=1}^{n} (\frac{x_i - \mu_x}{\sigma_x})^2 - 2\rho(\frac{x_i - \mu_x}{\sigma_x})(\frac{y_i - \mu_y}{\sigma_y}) + (\frac{y_i - \mu_y}{\sigma_y})^2].
\end{aligned} \tag{3.17}
$$

Let

$$
\lambda_1 = \frac{\max\limits_{\Theta_{01}} L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho|data)}{\max\limits_{\Theta} L(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho|data)}, \tag{3.18}
$$

where $\Theta = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$ and $\Theta_{01} = (\mu, \sigma)$. MLE of μ and σ are

$$
\begin{aligned}
\hat{\mu} = & a\bar{x} + (1-a)\bar{y^*}, \\
\hat{\sigma^2} = & \frac{Q(\hat{\mu})}{2n(1-\rho_0^2)}.
\end{aligned} \tag{3.19}
$$

where

$$\begin{aligned}
\bar{x} =& \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\bar{y^*} =& \bar{y} - \epsilon_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \epsilon_0, \\
S_{xx} =& \sum_{i=1}^{n}(x_i - \bar{x})^2, \\
S_{yy} =& \sum_{i=1}^{n}(y_i - \bar{y})^2, \\
S_{xy} =& \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \\
k =& 1 + \eta_0^2 - 2\rho_0\eta_0, \\
a =& \frac{\eta_0(\eta_0 - \rho_0)}{k}, \\
Q(\mu) =& Q_1(\mu) + Q_2, \\
Q_1(\mu) =& n[(\bar{x} - \mu)^2 - 2\frac{\rho_0}{\eta_0}(\bar{x} - \mu)(\bar{y^*} - \mu) + \frac{1}{\eta_0^2}(\bar{y} - \mu)^2)], \\
Q_2 =& S_{xx} - 2\frac{\rho_0}{\eta_0}S_{xy} + \frac{1}{\eta_0^2}S_{yy}.
\end{aligned} \tag{3.20}$$

Sinha and Dutta (2013, [27]) established that $\sqrt{n}T \underset{\rightarrow}{a} N(0,1)$ where

$$\begin{aligned}
Q_2^* =& S_{xx}S_{yy} - 2S_{xy}, \\
V =& \frac{\sqrt{n}(\bar{x} - \bar{y^*})}{\sigma\sqrt{k}}, \\
V_1^2V_2^2 =& \frac{Q_2^*}{\sigma^4\eta_0^2(1 - \rho_0^2)}, \\
T =& \frac{V}{(V_1^2V_2^2)^{\frac{1}{4}}}.
\end{aligned} \tag{3.21}$$

4. THREE-LEVEL MIXED-EFFECTS MODEL AND ESTIMATION OF CONCORDANCE CORRELATION COEFFICIENT

In our study, we will use CCC to evaluate agreement between the raters. As described in the introduction, independent subjects were followed for two years. X-rays were taken at baseline, one year and two years apart. Two radiologists manually measured joint space width based on X-rays. One computerized rating system read the same measurements. Each rater took second, repeated measurements on all 29 subjects' X-rays for all three time points. This presents data with multiple types of variation: within-subjects repeated measures at different time points, repeated evaluations at same time point, and variation between subjects and within raters. In our study, we assume individual subjects are independent, but that readings within subjects are not independent. Readings from the same rater are assumed to have variation between different time points and also within the same time point in case of repeat measurements. There are two levels of covariance between raters, one associated with patient-level variation and another associated with repeated readings of the same x-ray. We assume the ratings by one rater is X, and ratings by another rater is W. Let

$$X = \mu_X + \mu + \theta, \tag{4.1}$$

$$W = \mu_W + \nu + \gamma. \tag{4.2}$$

where μ_X and μ_W are the mean of X and W; μ and ν are denoted as the within subject variability. $\mu \sim \mathrm{N}(0,\ \sigma_\mu^2)$, $\nu \sim \mathrm{N}(0,\ \sigma_\nu^2)$; θ and γ are denoted as the within raters variability. $\theta \sim \mathrm{N}(0, \sigma_\theta^2)$, $\gamma \sim \mathrm{N}(0, \sigma_\gamma^2)$, and θ and γ are independent.

Thus even if the distributions of μ and ν are identical and $\mu_X = \mu_Y$, not necessarily $\theta \equiv \gamma$; and hence ρ_c will not measure perfect correlation. In fact, the goal is to find the agreement between the raters over k ratings cross multiple time points of n independent subjects. For non-identical ratings from two raters over subjects, the CCC is as follows:

$$\rho_c = \frac{2cov(X, W)}{var(W) + var(X) + (\mu_W - \mu_X)^2}. \tag{4.3}$$

Note that when X $\equiv$ W, that means, $cov(X, W) = var(X) = var(W)$, and $\mu_X = \mu_W$, $\rho_c = 1, or -1$, for same / opposite direction of covariance of X and W.

Let $i = 1, 2, \ldots, N$; i.e., there are N subjects;

$j(i) = 0, 1, 2, \ldots, T_i$; i.e., the i^{th} subject is measured at T_i time points ;

$k(ij) = 1, 2, \ldots, K_{ij}$; i.e., the i^{th} subject at the j^{th} time point is measured k_{ij} times ;

X_{ijk} represents readings from one rater; W_{ijk} represents readings from another rater.

The data structure is shown in figure 6.

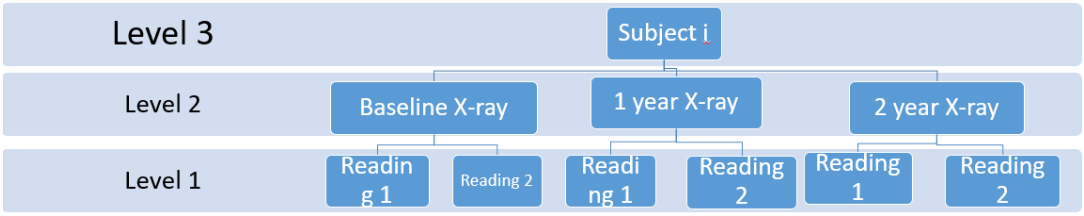

Figure 6: Data levels by one rater

We proposed a three-level mixed model in terms of replicates (level-1) nested within X-rays (level-2) which are nested within subjects (level-3):

$$\boldsymbol{x}_{ijk} = \mathrm{P}_{X1i}\boldsymbol{\beta}_X + \mathrm{P}_{X2i}\boldsymbol{v}_{Xij} + \mathrm{P}_{X3i}\boldsymbol{u}_{Xi} + E_{X_{ijk}}, \tag{4.4}$$

$$\boldsymbol{w}_{ijk} = \mathrm{P}_{W1i}\boldsymbol{\beta}_W + \mathrm{P}_{W2i}\boldsymbol{v}_{Wij} + \mathrm{P}_{W3i}\boldsymbol{u}_{Wi} + E_{W_{ijk}}. \tag{4.5}$$

Here, $\mathrm{P}_{.1i}$ is $n_i \times p$ design matrix for the fixed effects, p is the number of fixed variables in the model, $\mathrm{P}_{.3i}$ and $\mathrm{P}_{.2i}$ are the design matrices for the level-3, level-2 effects. $\boldsymbol{\beta}$ is an $(p \times 1)$ vector of regression coefficients, including intercept and slope, and $E_{.ijk}, \boldsymbol{v}_{.ij}, \boldsymbol{u}_{.i}$, denote level-1, level-2, and level-3 random effects, respectively. $\boldsymbol{u}_{.i}$ is random subject effect and explains the between-subject variability, denoted by G. $\boldsymbol{v}_{.ij}$ is the within-subject variability, denoted by Q, and measures ith subject effect different time point variability. $\boldsymbol{v}_{.ij} \sim N(0, Q); \boldsymbol{u}_{.i} \sim N(0, G); E_{ijk} \sim N(0, \Sigma)$.

The assumptions are:

(i) U_{Xi}, V_{Xij} and E_{Xi} are independent;

(ii) U_{Xi} and $U_{Xi'}$ are independent, $i \neq i'$;

(iii) V_{Xij} and $V_{Xi'j}$ are independent, $i \neq i'$;

(iv) $Cov(V_{Xij}, V_{Xij'}) \neq 0$, $j \neq j'$.

(v) U_{Wi}, V_{Wij} and E_{Wi} are independent;

(vi) U_{Wi} and $U_{Wi'}$ are independent, $i \neq i'$;

(vii) V_{Wij} and $V_{Wi'j}$ are independent, $i \neq i'$;

(viii) $Cov(V_{Wij}, V_{Wij'}) \neq 0$; $j \neq j'$.

(ix) $Cov(U_{Xi}, U_{Wi}) \neq 0$, for the two raters evaluating the same subject;

(x) $Cov(V_{Xij}, V_{Wij}) \neq 0$, for the two raters evaluating the same subject at the same time point;

(xi) Elements in $E_{X_{ijk}}$ and $E_{W_{ijk}}$ are independent, $E_{X_{ijk}}$ and $E_{W_{ijk}}$ are also independent.

Note that if $D(V_{Xij})= Cov(V_{Xij}, V_{Xij'})=0$, i.e., in the absence of within-rater variations, we arrive at the model discussed by Chinchille et al. (1996, [71]; 2001, [69]; 2007, [70]), Choudhary, P. K. et al. (2005, [53]), Carstensen, B. et al. (2008, [51]).

Let $Y_{ijk} = \begin{pmatrix} X_{ijk} & W_{ijk} \end{pmatrix}'$, combine those two equations together. The new model is:

$$Y_{ijk} = P_{1i}\beta + P_{3i}U_i + P_{2i}V_{ij} + E_{ijk}, \tag{4.6}$$

$$P_{1i} = \begin{pmatrix} \mathbf{P_{X1i}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_{W1i}} \end{pmatrix}, \beta = \begin{pmatrix} \beta'_X & \beta'_W \end{pmatrix}',$$

$$\mathrm{P}_{2i} = \begin{pmatrix} \mathbf{P_{X2i}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_{W2i}} \end{pmatrix}, U_i = \begin{pmatrix} \mathbf{U_{Xi}}' & \mathbf{U_{Wi}}' \end{pmatrix}',$$

$$\mathrm{P}_{3i} = \begin{pmatrix} \mathbf{P_{X3i}} & \mathbf{0} \\ \mathbf{0} & \mathbf{P_{W3i}} \end{pmatrix}, V_{ij} = \begin{pmatrix} \mathbf{V_{Xij}}' & \mathbf{V_{Wij}}' \end{pmatrix}'.$$

Typically, given the heterogeneity of the data, random intercept and trend are used in three-level models and random intercepts are used in two-level models. That is to say, subject i has subject-level difference from the average baseline and slope of the group. Nested within subject i, given one time point, there are k observations. The measurement refers to difference from the average of all the measurements at that time point. Figure 7 shows one example of that situation.

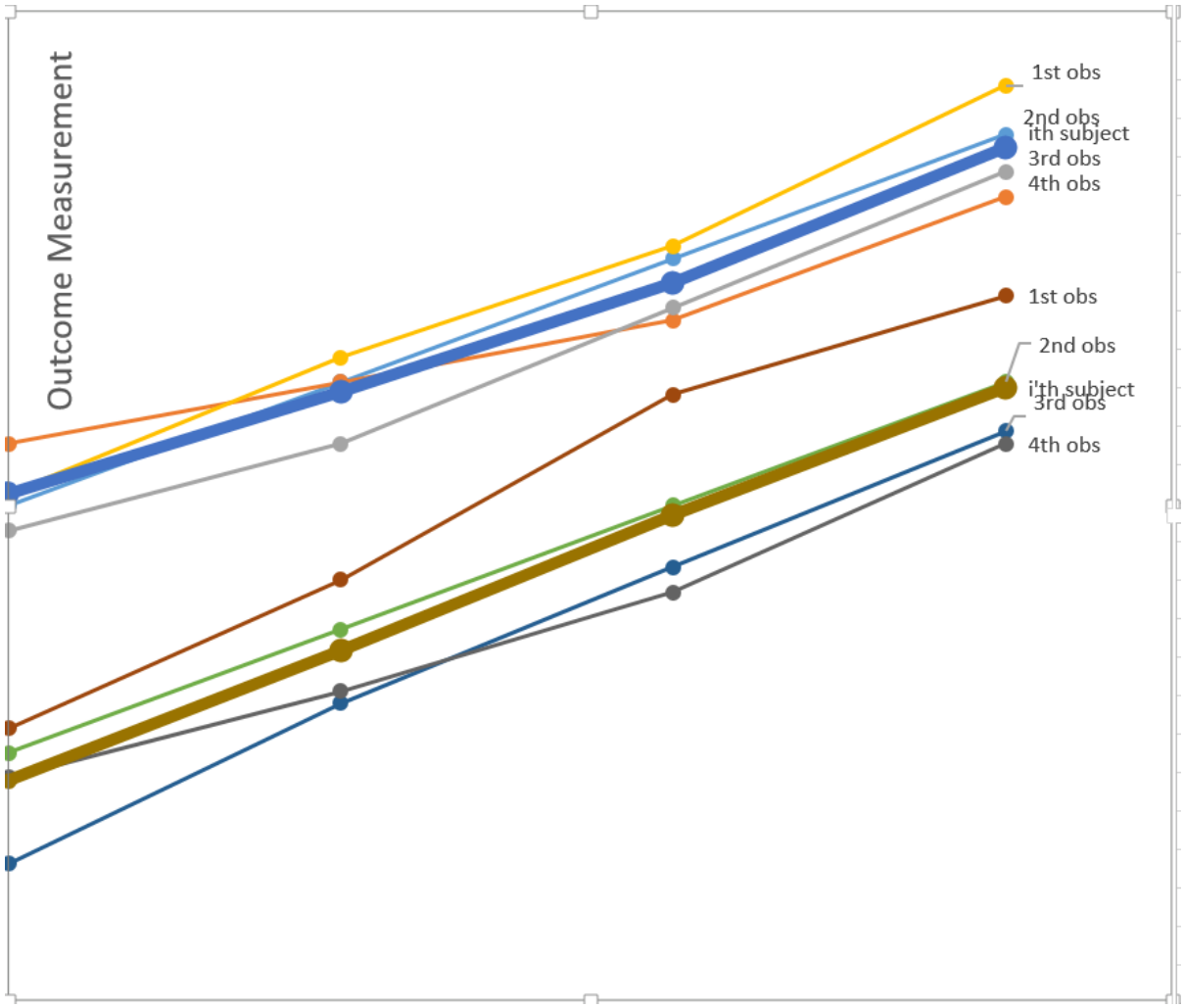

Figure 7: Two subjects multiple readings cross multiple time points by one rater

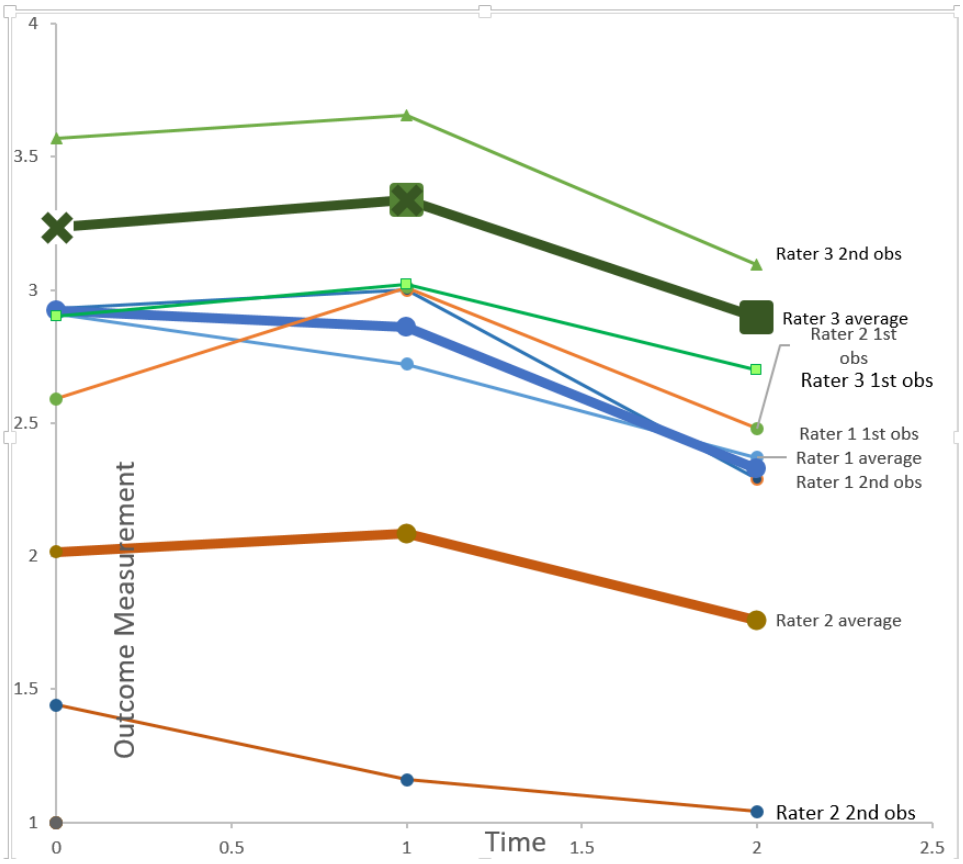

Figure 8: Subject i two readings by three raters at three time point - real data example

Figure 8 shows one example from real data. Subject was measured at three time points. Three raters read the measurement two times at each time point. The blue series represent rater one; the brown series represent rater two and the green series represent rater three. The highlighted lines are the average of each rater. We can see that first rater has smaller variation and the second rater has larger variation. The computer rater i.e., rater three has similar variation to rater one. The estimates of baseline, slope cross time by rater are varied and different.

Let x_{ijk} denote k-th measurement at time j for subject i rated by one rater. The model

expands to:

$$x_{ijk} = \beta_0 + \beta_1 t + \theta_{0i} + \theta_{1i} t + \nu_{0k(i)} + \epsilon_{ijk}, \tag{4.7}$$

where

i = subject 1, 2, ..., N;

j = time 0, 1, 2,..., T;

k = replication 1, 2,..., K.

In this model,

β_0 is the baseline average of x_{ijk}.

β_1 is the slope of x_{ijk} across time.

θ_{0i} is random intercept for subject i.

θ_{1i} is random slope for subject i.

$\nu_{0k(i)}$ is random intercept of kth observation nested within the i^{th} subject. $\nu_{0k(i)}$ and $\nu_{0j'(i)}$ are assumed to be uncorrelated

ϵ_{ijk} is random error and assumed to be autocorrelated.

Similarly, the model for the other rater is:

$$w_{ijk} = \beta'_0 + \beta'_1 t + \theta'_{0i} + \theta'_{1i} t + \nu'_{0k(i)} + \epsilon'_{ijk}, \tag{4.8}$$

with all the variables similarly defined as in x_{ijk}.

Define $y_{ijkl} = \begin{pmatrix} x_{ijk}|(l=1) & w_{ijk}|(l=2) \end{pmatrix}'$. Combining these two equations together, the new model is:

$$\begin{aligned} y_{ijkl} = {} & \beta_0 \delta_{(l=1)} + \beta_1 \delta_{(l=1)} t + \theta_{0i} \delta_{(l=1)} + \theta_{1i} t \delta_{(l=1)} + \nu_{0k(i)} \delta_{(l=1)} \\ & + \beta'_0 \delta_{(l=2)} + \beta'_1 \delta_{(l=2)} t + \theta'_{0i} \delta_{(l=2)} + \theta'_{1i} t \delta_{(l=2)} + \nu'_{0k(i)} \delta_{(l=2)} + \epsilon_{ijkl}. \end{aligned} \tag{4.9}$$

The outcome measurement vector for first subject:

$$\boldsymbol{x}_1 = \begin{pmatrix} \tilde{x_{11}} \\ \vdots \\ \tilde{x_{1T}} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} x_{111} \\ \vdots \\ x_{11K} \end{pmatrix} \\ \vdots \\ \begin{pmatrix} x_{1T1} \\ \vdots \\ x_{1TK} \end{pmatrix} \end{pmatrix}.$$

Mean vector for 1st subject:

$$\boldsymbol{u}_{X1} = E(X_1) = \beta_0 \mathbf{1}_{TK\times 1} + \beta_1 \begin{pmatrix} \mathbf{1}_K \\ 2\mathbf{1}_K \\ \vdots \\ T\mathbf{1}_K \end{pmatrix}_{TK\times 1}.$$

Here, $\mathbf{1}_K$ is defined as:

$$\mathbf{1}_K = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}'_{1\times K}.$$

Variance covariance structure for $\boldsymbol{x}_{1t}$:

$$\begin{aligned} \boldsymbol{\Sigma}_{tt} = \boldsymbol{\Sigma}_{x_{1t}} &= \begin{pmatrix} \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 & \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 & \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 \\ \vdots & \vdots & \vdots \\ \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 & \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 & \sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2 \end{pmatrix}_{K\times K} + \sigma^2_{\nu_0} J_{K\times K} + \sigma^2_e I_{K\times K} \\ &= (\sigma^2_{\theta_0} + \sigma^2_{\theta_1} t^2) J_{K\times K} + \sigma^2_{\nu_0} J_{K\times K} + \sigma^2_e I_{K\times K}. \end{aligned}$$

Covariance structure between two different time points of the same subject:

$$\boldsymbol{\Sigma}_{tt'} = Cov(x_{1t}, x_{1t'}) = (\sigma^2_{\theta_0} + \sigma^2_{\theta_1} tt')J_{K\times K} + \sigma^2_{\nu_0} J_{K\times K} + \sigma^2_e \Omega^{tt'}_{K\times K},$$

where $\Omega_{TK\times TK}$ defines an auto-correlated variance covariance structure.

Covariance structure for $\boldsymbol{x}_1$:

$$Cov(\boldsymbol{x}_1) = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} & \dots & \boldsymbol{\Sigma}_{1T} \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{\Sigma}_{T1} & \boldsymbol{\Sigma}_{T2} & \dots & \boldsymbol{\Sigma}_{TT} \end{pmatrix}_{KT\times KT}.$$

Covariance structure for $\boldsymbol{w}_1$ is similar to $\boldsymbol{x}_1$;

Covariance structure for $\boldsymbol{x}_i$ and $\boldsymbol{w}_{i'}$ is 0, where $i \neq i'$;

Covariance structure for $\boldsymbol{x}_i, \boldsymbol{w}_i$:

$$Cov(\boldsymbol{x}_1, \boldsymbol{w}_1) = Cov \begin{pmatrix} \theta_{0i} + \theta_{1i} t + \nu_{0k(i)}, \\ \theta'_{0i} + \theta'_{1i} t + \nu'_{0k(i)} \end{pmatrix}.$$

Let $\Sigma_{0xw} = Cov(\nu_{0k(i)}, \nu'_{0k(i)}) + Cov(\theta_{0i}, \theta'_{0i})$, for which between-raters variability is zero for two different subjects.

$$Cov(\boldsymbol{x}_1, \boldsymbol{w}_1) = I_T \otimes (\Sigma_{0xw} J_{k\times k}),$$
$$Cov(X, W) = I_N \otimes [I_T \otimes (\Sigma_{0XW} J_{k\times k})].$$

So,

$$CCC = \rho_c = \frac{2cov(X,W)}{var(W)+var(X)+(\mu_W-\mu_X)^2}, \text{when x and w are scaled.}$$
$$= \frac{2 \times \frac{1}{2} tr(\Sigma_{XX}+\Sigma_{WW}-\Sigma_{X-W})}{tr(\Sigma_{XX})+tr(\Sigma_{WW})+(\mu_x-\mu_W)'(\mu_X-\mu_W)}, \text{when x and w are not scaled.} \tag{4.10}$$

How do we claim $CCC \leq 1$? If and only if $\mu_X = \mu_W$, $tr(\mathbf{\Sigma}_{XX}) = tr(\mathbf{\Sigma}_{WW})$, $tr(\mathbf{\Sigma}_{X-W}) = 0$,

$$CCC \leq \frac{tr(\Sigma_{XX}+\Sigma_{WW}-\Sigma_{X-W})}{tr(\mathbf{\Sigma}_{XX})+tr(\mathbf{\Sigma}_{WW})+(\mu_X-\mu_W)'(\mu_X-\mu_W)} \leq \frac{2tr(\mathbf{\Sigma}_{XX})-0}{2tr(\mathbf{\Sigma}_{XX})} = 1.$$

If and only $\mu_X = \mu_W$, $tr(\mathbf{\Sigma}_{XX}) = tr(\mathbf{\Sigma}_{WW})$, $tr(\mathbf{\Sigma}_{X-W}) \leq 4tr(\mathbf{\Sigma}_{XX})$,

,

$$CCC \geq \frac{tr(\mathbf{\Sigma}_{XX}+\mathbf{\Sigma}_{WW}-4\mathbf{\Sigma}_{XX})}{tr(\mathbf{\Sigma}_{XX})+tr(\mathbf{\Sigma}_{WW})+(\mu_X-\mu_W)'(\mu_X-\mu_W)} \geq \frac{-2tr(\mathbf{\Sigma}_{XX})}{2tr(\mathbf{\Sigma}_{XX})} = -1.$$

4.1 Parameter Estimation

There are several ways for estimating the parameters in mixed-effects models. Newton-Raphson and EM Algorithms are the most popular (Lindstrom, 1998, [21]). The expectation-maximization (EM) algorithm is used to estimate the parameters in the proposed model (4.9). In the E-step, with the current values of the other parameters, we compute the "expected *a posteriori*" or empirical Bayes (EB) estimates of the random effects as well as the conditional variances of the random effects, given the data. In the M-step, given the current values of the random effects, we obtain the maximum marginal likelihood (MML) or restricted maximum [marginal] likelihood (REML) estimates of the regression coefficients, error variances, and the variances of the random effects. The algorithm iter-

ates between the EB and MML or REML estimates until convergence is achieved. That is to say:

(i) Give initial value(s) for parameter(s),

(ii) 'E' (Expectation)-step, calculate expected value based on parameters and given random variables, conditional variance of the random variables by expected values,

(iii) 'M' (Maximization)-step, estimate parameters and the variance terms by MML or REML based on random variables.

(iv) Repeat

- using estimated parameter values as true values to get expected values, and
- using expected values as observed values, iterating until convergence.

We have discussed before that the measurement $\boldsymbol{y}_i$ and random effect $\boldsymbol{u}_i, \boldsymbol{v}_{ij}$ have the following joint distribution :

$$\begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \\ \boldsymbol{v}_{ij} \end{bmatrix} \sim MVN \left(\begin{bmatrix} P_{1i}\boldsymbol{\beta} \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} P_{3i}QP_{3i}^T + P_{2i}GP_{2i}^T + \boldsymbol{\Sigma}_i & P_{2i}G & P_{3i}Q \\ GP_{2i}^T & G & 0 \\ QP_{3i}^T & 0 & Q \end{bmatrix} \right),$$

where Q, G are the variance of 1-level and 2-level random effects and $\boldsymbol{\Sigma}_i$ is a 12×12 block diagonal error covariance matrix.

In fact, since this is a linear mixed model for a continuous variable, as Gibbons and Hedeker pointed out, the 3-level linear model can be written as a 2-level mixed-effect

model for continuous variables. Let

$$\boldsymbol{y}_i^* = \begin{pmatrix} \boldsymbol{y}_{i1}, & \boldsymbol{y}_{i2}, & ..., & \boldsymbol{y}_{im_i} \end{pmatrix}',$$

$$\mathbf{P_{1i}} = \begin{pmatrix} \mathbf{p_{i1}}, & \mathbf{p_{i2}}, & ..., & \mathbf{p_{im_i}} \end{pmatrix}',$$

$$\mathbf{Z_i} = \begin{pmatrix} \mathbf{p_{(3)i1}} & \mathbf{p_{(3)i1}} & 0 & ... & 0 \\ \mathbf{p_{(3)i2}} & 0 & \mathbf{p_{(2)i1}} & ... & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ \mathbf{p_{(3)in_i}} & 0 & 0 & ... & \mathbf{p_{(2)in_i}} \end{pmatrix},$$

$$\boldsymbol{v}_i = \begin{pmatrix} \nu_i, & \nu_{i1}, & ..., & \nu_{in_i} \end{pmatrix}',$$

then this set of regression equations can be written as

$$\boldsymbol{y}_i^* = \mathbf{P_{1i}}\boldsymbol{\beta} + \mathbf{Z_i}\boldsymbol{v}_i + \epsilon_\mathbf{i}. \tag{4.11}$$

Then the observation $\boldsymbol{y}_i^*$ and random effects $\boldsymbol{v}_i$ have joint normal distribution

$$\begin{bmatrix} \boldsymbol{y}_i^* \\ \boldsymbol{v}_i \end{bmatrix} \sim MVN \left(\begin{bmatrix} P_{1i}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T + \Sigma_i & \boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \\ \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T & \Sigma_{\boldsymbol{v}} \end{bmatrix} \right), \tag{4.12}$$

where $\Sigma_{\boldsymbol{v}}$ is a $n_i \times n_i$ covariance matrix of random effects; in our case, n_i will be 2, and Σ_i is a 12×12 block diagonal matrix of the error covariance matrices of all links.

4.1.1 Empirical Bayes Estimation

The Empirical Bayes (EB) approach takes into account two stochastic processes, one for the data $f(\boldsymbol{y}^*|\boldsymbol{v};\eta)$, and the other one for the random effects $g(\boldsymbol{v})$. Here, f is the likelihood function (probability density), that describes probability of the data associated with the random effects $\boldsymbol{v}$ and parameters η (β, σ). Instead of considering $\boldsymbol{v}$ as fixed values, we assume a prior distribution $\pi(\boldsymbol{v}|\eta)$ is placed on $\boldsymbol{v}$. After the data is obtained, the distribution of $\boldsymbol{v}$ can be updated by combining the prior distribution with the observed data. The resulting distribution of $\boldsymbol{v}$ is called posterior distribution $\pi(\boldsymbol{v}|\boldsymbol{y}^*, \eta)$. The posterior distribution is the basis of all Bayesian inference. The estimate of $\boldsymbol{v}$ is the mean of the posterior distribution, that is,

$$E(\boldsymbol{v}|y^*, \eta) = \int \boldsymbol{v} \times \pi(\boldsymbol{v}|\boldsymbol{y}^*, \eta) d\boldsymbol{v}.$$

If η is known, the posterior distribution can be derived via the following Bayes' rule,

$$p(\boldsymbol{v}|y^*, \eta) = \frac{f(\boldsymbol{y}^*|\boldsymbol{v}) \times \pi(\boldsymbol{v}|\eta)}{\int f(\boldsymbol{y}^*|\boldsymbol{v}) \times \pi(\boldsymbol{v}|\eta) d\boldsymbol{v}} = \frac{f(\boldsymbol{y}^*|\boldsymbol{v}) \times \pi(\boldsymbol{v}|\eta)}{m(\boldsymbol{y}^*|\eta)}, \tag{4.13}$$

where $m(\boldsymbol{y}^*|\eta)$ is the marginal distribution of $\boldsymbol{y}^*$ given η.

In almost all cases, however, η is unknown. Empirical Bayes and fully Bayesian approaches differ in how to proceed from here. Empirical Bayes would estimate $\boldsymbol{\varepsilon}$ from marginal distribution $m(\boldsymbol{y}^*|\eta)$ to get $\widehat{\eta}$ via techniques such as MLE and then plug in $\widehat{\eta}$ to get $p(\boldsymbol{v}|\boldsymbol{y}^*, \widehat{\eta})$. Empirical Bayes considers $p(\boldsymbol{v}|\boldsymbol{y}^*, \widehat{\eta})$ as the posterior distribution of $\boldsymbol{v}$.

A statistician who decides to take the fully Bayesian approach would place a distribution on η, or $\eta \sim h(\eta|\lambda)$ where λ is the hyperparameter. The posterior distribution of

vvec can be derived in the following way,

$$p(\boldsymbol{v}|\boldsymbol{y}^*,\lambda) = \frac{\int f(\boldsymbol{y}^*|\boldsymbol{v}) \times \pi(\boldsymbol{v}|\eta)h(\eta|\lambda)d\eta}{\int\ \int f(\boldsymbol{y}^*|\boldsymbol{v}) \times \pi(\boldsymbol{v}|\eta)h(\eta|\lambda)d\boldsymbol{v}d\eta} = \int f(\boldsymbol{v}|\boldsymbol{y}^*,\eta) \times h(\eta|y^*,\lambda)d\eta. \quad (4.14)$$

Now we apply EB to estimate $\boldsymbol{v}_i$ in model 4.9. As we described before, observed measurement $\boldsymbol{y}_i^*$ and random effect $\boldsymbol{v}_i$ have the following joint distribution,

$$\begin{bmatrix} \boldsymbol{y}_i^* \\ \boldsymbol{v}_i \end{bmatrix} \sim MVN\left(\begin{bmatrix} P_{1i}\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{Z}_i\Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T + \Sigma_i & \boldsymbol{Z}_i\Sigma_{\boldsymbol{v}} \\ \Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T & \Sigma_{\boldsymbol{v}} \end{bmatrix}\right). \quad (4.15)$$

The conditional distribution of $\boldsymbol{v}_i$ given $\boldsymbol{y}_i^*$ can be expressed as

$$\boldsymbol{v}_i|\boldsymbol{y}_i^* \sim N(\Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T(\boldsymbol{Z}_i\Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i)^{-1}(\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}), \Sigma_{\boldsymbol{v}} - \Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T(\boldsymbol{Z}_i\Sigma_{\boldsymbol{v}}\boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i)^{-1}\boldsymbol{Z}_i\Sigma_{\boldsymbol{v}}). \quad (4.16)$$

The EB posterior distribution of $\boldsymbol{v}_i$ can be obtained after replacing $\Sigma_{\boldsymbol{v}}$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$ with their corresponding estimates. There are several ways to estimate $\Sigma_{\boldsymbol{v}}$, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$. The most used are maximum likelihood estimation (MLE) and Restricted maximum likelihood estimation (REML) which will be discussed in the next section.

The posterior mean of $\boldsymbol{v}_i$ and posterior covariance of $\boldsymbol{v}_i$ provide EB estimates of $\boldsymbol{v}_i$ and its covariance, denoted as $\tilde{\boldsymbol{v}}_i$ and $\tilde{\Sigma}_{\boldsymbol{v}|\boldsymbol{y}_i^*}$, respectively.

$$
\begin{aligned}
\tilde{\boldsymbol{v}}_i &= \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T (\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T + \Sigma_i)^{-1} (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}) \\
&= \Sigma_{\boldsymbol{v}} [(\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T (\boldsymbol{Z}_i^T)^{-1} + \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1} (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}) \\
&= \Sigma_{\boldsymbol{v}} [\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} + \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1} \boldsymbol{Z}_i (\boldsymbol{Z}_i)^{-1} (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}) \\
&= \Sigma_{\boldsymbol{v}} [(\boldsymbol{Z}_i)^{-1} \boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} + (\boldsymbol{Z}_i)^{-1} \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1} (\boldsymbol{Z}_i)^{-1} (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}) \\
&= \Sigma_{\boldsymbol{v}} \Big[\Sigma_{\boldsymbol{v}} + (\boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i)^{-1}\Big]^{-1} (\boldsymbol{Z}_i^T \boldsymbol{Z}_i)^{-1} \boldsymbol{Z}_i^T (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}) \\
&= \boldsymbol{R} (\boldsymbol{Z}_i^T \boldsymbol{Z}_i)^{-1} \boldsymbol{Z}_i^T (\boldsymbol{y}_i^* - P_{1i}\boldsymbol{\beta}), \\
\tilde{\Sigma}_{\boldsymbol{v}|\boldsymbol{y}_i^*} &= \Sigma_{\boldsymbol{v}} - \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T (\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T + \Sigma_i)^{-1} \boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \qquad (4.17) \\
&= [I - \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T (\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_i^T + \boldsymbol{\Sigma}_i)^{-1} \boldsymbol{Z}_i] \Sigma_{\boldsymbol{v}} \\
&= \{I - \Sigma_{\boldsymbol{v}} [\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} \boldsymbol{Z}_j^T (\boldsymbol{Z}_i^T)^{-1} + \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1} \boldsymbol{Z}_i\} \Sigma_{\boldsymbol{v}} \\
&= \{I - \Sigma_{\boldsymbol{v}} [\boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} + \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1} \boldsymbol{Z}_i\} \Sigma_{\boldsymbol{v}} \\
&= \{I - \Sigma_{\boldsymbol{v}} [\boldsymbol{Z}_i^{-1} \boldsymbol{Z}_i \Sigma_{\boldsymbol{v}} + \boldsymbol{Z}_i^{-1} \boldsymbol{\Sigma}_i (\boldsymbol{Z}_i^T)^{-1}]^{-1}\} \Sigma_{\boldsymbol{v}} \\
&= \{I - \Sigma_{\boldsymbol{v}} [\Sigma_{\boldsymbol{v}} + (\boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i)^{-1}]^{-1}\} \Sigma_{\boldsymbol{v}} \\
&= (\boldsymbol{I} - \boldsymbol{R}) \Sigma_{\boldsymbol{v}},
\end{aligned}
$$

where $\boldsymbol{R} = \Sigma_{\boldsymbol{v}} \Big[\Sigma_{\boldsymbol{v}} + (\boldsymbol{Z}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{Z}_i)^{-1}\Big]^{-1}$.

4.2 Maximum (Marginal) Likelihood Estimation

As noted in the previous section, in order to obtain EB estimates of random effect $\boldsymbol{v}$ and its covariance, $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$ in model 4.11 need to be estimated from the marginal distribution of $Y*$. There are several techniques to obtain estimates of $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\beta}$, including maximum likelihood estimation and restricted maximum likelihood estimation which

will be discussed in this section and the next. According to model 4.9,

$$\boldsymbol{y}^*|\boldsymbol{v} \sim N(P_1\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{v}, \boldsymbol{\Sigma}), \tag{4.18}$$

and

$$\boldsymbol{v} \sim N(0, \Sigma_{\boldsymbol{v}}). \tag{4.19}$$

So the joint distribution of $\boldsymbol{y}^*$ and $\boldsymbol{v}$ is:

$$\begin{aligned} f(\boldsymbol{y}^*, \boldsymbol{v}) =& f(\boldsymbol{y}^*|\boldsymbol{v})f(\boldsymbol{v}) \\ \propto& exp[-\frac{1}{2}(\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v})^T\Sigma^{-1}(\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v})] \\ & exp[-\frac{1}{2}\boldsymbol{v}^T\Sigma_{\boldsymbol{v}}^{-1}\boldsymbol{v}]. \end{aligned} \tag{4.20}$$

And the log likelihood is :

$$\begin{aligned} l = logf(\boldsymbol{y}^*, \boldsymbol{v}) =& -\frac{1}{2}(\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v})^T\Sigma^{-1}(\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v}) \\ & -\frac{1}{2}\boldsymbol{v}^T\Sigma_{\boldsymbol{v}}^{-1}\boldsymbol{v} + c. \end{aligned} \tag{4.21}$$

In order to find the MLE of $\boldsymbol{\beta}_i$,, we need to minimize this function:

$$\begin{aligned} Q(\beta, \boldsymbol{v}) =& (\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v})^T\Sigma^{-1}(\boldsymbol{y} - P_1\boldsymbol{\beta} - \boldsymbol{z}\boldsymbol{v}) \\ & - \boldsymbol{v}^T\Sigma_{\boldsymbol{v}}^{-1}\boldsymbol{v} \\ =& \boldsymbol{y}^T\Sigma^{-1}\boldsymbol{y} - 2\boldsymbol{\beta}^T P_1^T\Sigma^{-1}\boldsymbol{y} + 2\boldsymbol{\beta}^T P_1^T\Sigma^{-1}\boldsymbol{z}\boldsymbol{v} - 2\boldsymbol{v}^T\boldsymbol{z}^T\Sigma^{-1}\boldsymbol{y} \\ & + \boldsymbol{\beta}^T P_1^T\Sigma^{-1}P_1\boldsymbol{\beta} + \boldsymbol{v}^T\boldsymbol{z}^T\Sigma^{-1}\boldsymbol{z}\boldsymbol{v} - \boldsymbol{v}^T\Sigma_{\boldsymbol{v}}^{-1}\boldsymbol{v}. \end{aligned} \tag{4.22}$$

That means, to get the MLE of $\boldsymbol{\beta}$, we differentiate the previous function with respect to $\boldsymbol{\beta}$.

$$\frac{\partial Q(\beta, \boldsymbol{v})}{\partial \boldsymbol{\beta}} = -2P_1^T\Sigma^{-1}\boldsymbol{y} + 2P_1^T\Sigma^{-1}\boldsymbol{z}\boldsymbol{v} + 2P_1^T\Sigma^{-1}P_1\boldsymbol{\beta} = 0. \tag{4.23}$$

Hence,

$$\widehat{\boldsymbol{\beta}} = (P_1^T \Sigma^{-1} P_1)^{-1} P_1^T \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{z}\boldsymbol{v}). \tag{4.24}$$

The covariance of $\boldsymbol{\beta}$ is derived in the following steps.

$$Cov(\widehat{\boldsymbol{\beta}}) = [(P_1^T (\boldsymbol{Z}\Sigma_{\boldsymbol{v}}\boldsymbol{Z}^T + \boldsymbol{\Sigma})^{-1} \mathrm{P}_1]^{-1}. \tag{4.25}$$

For the covariance of $\boldsymbol{\Sigma}$, we can go back to model 4.9 again, and we rewrite the model as the following.

$$Y^* = P_1\beta + Z\boldsymbol{v} + \epsilon = P_1\beta + \epsilon^*. \tag{4.26}$$

$$\epsilon^* \sim N(0, \boldsymbol{V}). \tag{4.27}$$

Here,

$$\boldsymbol{V} = \boldsymbol{Z}\Sigma_{\boldsymbol{v}}\boldsymbol{Z}^T + \boldsymbol{\Sigma}. \tag{4.28}$$

So the likelihood for $(\beta, \boldsymbol{V})$ will be:

$$l(\beta, \boldsymbol{v}) = -\frac{1}{2} Ln|\boldsymbol{V}| + (\boldsymbol{y} - P_1\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - P_1\boldsymbol{\beta}) + c. \tag{4.29}$$

To maximize the log likelihood, let σ_{ii}^2 be the diagonal elements of $\boldsymbol{\Sigma}$, σ_{ij} be the off-diagonal elements of $\boldsymbol{\Sigma}$, $\sigma_{\boldsymbol{v}(ii)}^2$ be the diagonal elements of $\boldsymbol{G}$, and $\sigma_{\boldsymbol{v}(ij)}$ be the off-diagonal elements of $\boldsymbol{G}$.

$$\frac{\partial|\boldsymbol{V}|}{\partial\sigma_{ij}} = (2 - \delta_{ij})|\boldsymbol{\Sigma}_{ij}|, \tag{4.30}$$

$$\frac{\partial|\boldsymbol{V}|}{\partial\sigma_{\boldsymbol{v}(ij)}} = ZZ^T(2 - \delta_{ij})|\boldsymbol{G}_{ij}|, \tag{4.31}$$

where δ_{ij} is the Kronecker delta, $\delta_{ij} = 0$ for $i \neq j$, and $\delta_{ij} = 1$ for $i = j$. A general

principle for maximizing the likelihood with respect to each element of $\boldsymbol{\Sigma}$ and $\boldsymbol{G}$ is to equate the derivative of the log likelihood with respect to that particular element of the matrix to zero and solve the resulting equations for all the elements.

In the case of this study, we assume $\boldsymbol{\Sigma} = \sigma_{ii}^2 I$ and $\boldsymbol{G} = \sigma_{\boldsymbol{v}(ii)}^2 I$.

$$\widehat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}\Bigg[(y_i - P_{1i}\widehat{\beta})^T(y_i - P_{1i}\widehat{\beta}) + tr(\boldsymbol{Z}_i\Sigma_{\boldsymbol{v}|y_i}\boldsymbol{Z}_i^T)\Bigg], \tag{4.32}$$

$$\widehat{\sigma}_{\boldsymbol{v}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg(\tilde{\boldsymbol{v}}_i\tilde{\boldsymbol{v}}_i^T + \tilde{\Sigma}_{\boldsymbol{v}|\boldsymbol{y}_i}\Bigg). \tag{4.33}$$

Interestingly enough, as $\tilde{\boldsymbol{v}}_i \to 0$, and $\Sigma_{\boldsymbol{v}|y_i} \to 0$, $\widehat{\sigma}^2$ goes to the MLE, ignoring random subject effects. In other words, $\widehat{\sigma}^2$ takes into account both estimated residuals and the EB estimate of the uncertainty of the random effect.

4.3 Restricted Maximum Likelihood Estimation (REML)

One property of ML estimation is that it does not take into account the degrees of freedom that are involved in estimating fixed effects in variance components estimation. When the sample size is large enough, the resulting bias can be ignored. However, when the sample size is not sufficiently large, we will consider Restricted Maximum Likelihood Estimation (REML). Rather than using the observed data (Y vector) directly, REML is based on linear combinations of elements of y after fitting the fixed effects equivalent to residuals.[Searle and Casella, Variance Components (2006, [34])].

That means, we choose k so that $k^T P_1 \beta = 0 \ \ \forall \beta$. Hence, $kP_1 = 0$. Therefore, the form of k must be $k^T = c(I - P_1 P_1^-)$ or $k^T = c^T(I - P_1 P_1^+)$ for any c. So K^T would be chosen to be full row rank of $N - r_{P_1}$.

Given $Y \sim N(P_1\beta, \boldsymbol{V})$, we have $K^TY \sim N(0, K^T\boldsymbol{V}K)$.

The REML equation can therefore be derived from the ML equation as in 3.42 by replacements as:

(i) Replace Y by K^TY.

(ii) Replace P_1 by $K^TP_1 = 0$.

(iii) Replace Z by K^TZ.

(iv) Replace V by $K^TVK = 0$.

With l_R being the likelihood function of K^TY,

$$l_R = -\frac{1}{2}(N-r)Ln(2\pi) - \frac{1}{2}Ln|K^T\boldsymbol{V}K| - \frac{1}{2}\boldsymbol{y}^TK(K^TVK)^{-1}K^T\boldsymbol{y}. \tag{4.34}$$

Let

$$P = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}P_1(P_1^T\boldsymbol{V}^{-1}P_1)^-P_1^T\boldsymbol{V}^{-1} = K(K^TVK)^{-1}K^T, \tag{4.35}$$

then,

$$\begin{aligned}
\frac{\partial P}{\partial \sigma_i^2} &= -K(K^TVK)^{-1}frac\partial K^TVK\partial\sigma_i^2(K^TVK)^{-1}K^T \\
&= -K(K^TVK)^{-1}K^T\frac{\partial V}{\partial \sigma_i^2}K(K^TVK)^{-1}K^T \\
&= -P\frac{\partial V}{\partial \sigma_i^2}P \\
&= -PZZ^TP.
\end{aligned} \tag{4.36}$$

Hence,

$$\begin{aligned}\frac{\partial l_R}{\partial \sigma_i^2} &= -\frac{1}{2}tr\Big[(K^TVK)^{-1}K^TZZ^TK] + \frac{1}{2}Y^TPZZ^TPY \\ &= -\frac{1}{2}tr(PZZ^T) + \frac{1}{2}Y^TPZZ^TPY.\end{aligned} \tag{4.37}$$

For the information matrix, the second derivative of l_R:

$$\frac{\partial l_R}{\partial \sigma_i^2 \partial \sigma_j^2} = -\frac{1}{2}tr(PZ_jZ_j^TPZ_iZ_i^T) - Y^TPZ_jZ_j^TPZ_iZ_iPY. \tag{4.38}$$

As we can see, ML estimators for the variance components are biased because they do not take into account the loss in degrees of freedom from the estimation of fixed effects. The REML method corrects this, but REML estimation has the same procedure for estimating the fixed effects as ML.

SAS version 9.4 PROC MIXED procedure is used for this project.

5. SIMULATION STUDY

A simulation study was carried out to evaluate the accuracy and reliability of the parameter estimation algorithm. For this purpose, 1000 data sets—each data set containing 30 subjects having X-rays taken at three time-points, and each x-ray subsequently reviewed twice by two different raters—were generated according to the following model. The proposed values for the model parameters are based on several situations: CCC close to 1, close to 0 and the estimated value close to the real data. Twelve observations for each subject were generated according to the following model,

$$\begin{aligned} y_{ijkl} = {} & \beta_0\delta_{(l=1)} + \beta_1\delta_{(l=1)}t + \theta_{0i}\delta_{(l=1)} + \nu_{0k(i)}\delta_{(l=1)} \\ & + \beta_0'\delta_{(l=2)} + \beta_1'\delta_{(l=2)}t + \theta_{0i}'\delta_{(l=2)} + \nu_{0k(i)}'\delta_{(l=2)} + \epsilon_{ijkl}. \end{aligned} \tag{5.1}$$

Assumptions were used to generate the true values for fixed and random effects parameters:

(i) The parameters $\beta_0, \beta_0', \beta_1, \beta_1'$ were fitted in fixed values.

(ii) θ_{0i}, $\nu_{0j(i)}$ and ϵ_{ijkl} are independent.

(iii) $\begin{bmatrix} \theta_{0i} \\ \theta_{0i}' \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\theta^2 & \sigma_{\theta\theta'} \\ \sigma_{\theta\theta'} & \sigma_{\theta'}^2 \end{bmatrix} \right)$.

(iv) $\begin{bmatrix} \nu_{0j(i)} \\ \nu_{0j(i)}' \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\nu^2 & \sigma_{\nu\nu'} \\ \sigma_{\nu\nu'} & \sigma_{\nu'}^2 \end{bmatrix} \right)$, but $\nu_{0j'(i)}$ and $\nu_{0j(i)}'$ are uncorrelated when $j' \neq j$.

(v) $\epsilon \sim N(0, \sigma^2)$, and independent.

For each simulated data set, fixed effect and random effect parameters were estimated according to the estimation method described in last section. The performance of this method is evaluated using biases and root mean squared errors (RMSE). The evaluation quantities are defined as follows:

5.1 Bias

Bias is the expected difference between the estimated value and the true value, i.e. $Bias(\widehat{\theta,\theta}) = E(\widehat{\theta} - \theta)$. We took the mean of the estimated parameters as estimators and calculated the difference between the estimators and the true values.

5.2 Root mean squared error

Mean squared error(MSE), $Var(\widehat{\theta}) + Bias(\theta, \widehat{\theta})^2$, is debatably the most reliable quantity to evaluate the performance of estimators with as it holds both accuracy(bias) and precision (variance). Root Mean Squared error (RMSE), maintaining the same unit as the quantity being estimated, is derived by finding the square root of MSE. For an unbiased estimator(bias=0), the RMSE is the standard error (se) of the estimate.

Intra-rater CCC is used for measuring the agreement among multiple replications measured by a single rater. Inter-rater CCC is used for measuring the agreement among different raters. In this simulated data set, we assume there are two raters. So intra- and inter-CCC are calculated following formula (3.24). In the following tables, the abbreviations are as follows: EST=estimate, STD=Standard Deviation, RMSE=root mean squared error, $Intra_{l=1}$=Intra-CCC for rater 1, $Intra_{l=2}$=Intra-CCC for rater 2,

Inter=Inter-CCC between two raters.

As shown in Tables V-VI, the estimated fixed and random effects β_0, β_1, $var(\beta_{0i})$, $var(\beta_{1i})$, $\sigma^2_{0i}, \sigma^2_{1i}$, and σ^2_γ are very close to their respective true values when simulate the extreme situation. The biases in absolute values are small (less than 0.2); and RMSEs are less than 0.4.

Table VII shows the result simulated from the true data distribution, the estimated fixed and random effects are also very close to their respective true values. The biases in absolute values and RMSEs are better than the extreme situation. Taken together, this simulation study demonstrates the accuracy and precision in estimating the model parameters.

Table V: Simulation to evaluate performance of estimators when CCC is close to 1

Category	Parameter	True Value	EST(STD)	Bias	RMSE
Fixed	β_0	4.00	3.9828 (0.2481)	-0.0172	0.2487
	β_0'	4.00	3.9830 (0.2482)	-0.0170	0.2488
	β_1	-0.20	-0.1755 (0.1295)	0.0245	0.1318
	β_1'	-0.20	-0.1756 (0.1295)	0.0244	0.1318
Random variance	σ^2_ν	1.00	0.9844 (0.3544)	-0.0156	0.3547
	σ^2_θ	1.00	1.0058 (0.1883)	0.0058	0.1883
	$\sigma^2_{\nu'}$	1.00	0.9843 (0.3549)	-0.0157	0.3552
	$\sigma^2_{\theta'}$	1.00	1.0063 (0.1883)	0.0063	0.1884
	σ^2	0.05	0.0025 (0.0003)	-0.0475	0.0475
	$\sigma_{\nu\nu'}$	1.00	0.9843 (0.3546)	-0.0157	0.3549
	$\sigma_{\theta\theta'}$	1.00	1.0061 (0.1882)	0.0061	0.1883
CCC	$Intra_{l=1}$	0.98	0.9987 (0.0003)	0.0231	0.0231
	$Intra_{l=2}$	0.98	0.9987 (0.0003)	0.0187	0.0187
	$Inter$	0.98	0.9987 (0.0003)	0.0231	0.0231

Table VI: Simulation to evaluate performance of estimators when CCC is close to 0

Category	Parameter	True Value	EST(STD)	Bias	RMSE
Fixed	β_0	4.00	3.9822 (0.2485)	-0.0178	0.2491
	$\beta_0^{'}$	8.00	8.0010 (0.2524)	0.0010	0.2524
	β_1	-0.20	-0.1751 (0.1302)	0.0249	0.1325
	$\beta_1^{'}$	-0.20	-0.2003 (0.1309)	-0.0003	0.1309
Random variance	σ_{ν}^2	1.00	0.9836 (0.3546)	-0.0164	0.3550
	σ_{θ}^2	1.00	1.0052 (0.1908)	0.0052	0.1908
	$\sigma_{\nu^{'}}^2$	1.00	0.9955 (0.3478)	-0.0045	0.3478
	$\sigma_{\theta^{'}}^2$	1.00	0.9908 (0.1890)	-0.0092	0.1892
	σ^2	0.20	0.0399 (0.0043)	-0.1601	0.1601
	$\sigma_{\nu\nu^{'}}$	0.00	0.0102 (0.2533)	0.0062	0.2533
	$\sigma_{\theta\theta^{'}}$	0.01	0.0124 (0.1297)	0.0074	0.1299
CCC	$Intra_{l=1}$	0.91	0.9795 (0.0049)	0.0704	0.0706
	$Intra_{l=2}$	0.91	0.9798 (0.0048)	0.0707	0.0709
	$Inter$	0.00	0.0021 (0.0263)	0.0015	0.0263

Table VII: Simulation to evaluate performance of estimators when CCC corresponds almost to real data

Category	Parameter	True Value	EST(STD)	Bias	RMSE
Fixed	β_0	3.65	3.6385 (0.2238)	-0.0076	0.2240
	$\beta_0^{'}$	4.11	4.0998 (0.1932)	-0.0053	0.1933
	β_1	-0.10	-0.0861 (0.0552)	0.0090	0.0560
	$\beta_1^{'}$	-0.16	-0.1502 (0.0545)	0.0077	0.0551
Random variance	σ_{ν}^2	1.36	1.3540 (0.3653)	-0.0083	0.3654
	σ_{θ}^2	0.11	0.1115 (0.0354)	0.0001	0.0354
	$\sigma_{\nu^{'}}^2$	0.97	0.9669 (0.2663)	-0.0050	0.2663
	$\sigma_{\theta^{'}}^2$	0.08	0.1042 (0.0354)	0.0211	0.0412
	σ^2	0.16	0.1613 (0.0173)	-0.0007	0.0173
	$\sigma_{\nu\nu^{'}}$	1.12	1.1131 (0.3044)	-0.0065	0.3045
	$\sigma_{\theta\theta^{'}}$	0.10	0.1085 (0.0283)	0.0060	0.0289
CCC	$Intra_{l=1}$	0.90	0.8951 (0.0284)	-0.0059	0.0290
	$Intra_{l=2}$	0.93	0.8640 (0.0358)	-0.0664	0.0754
	$Inter$	0.81	0.8014 (0.0434)	-0.0107	0.0447

6. DATA ANALYSIS

In this section, we will apply the methods discussed in the previous chapters to evaluate agreement between different raters.

Back to the GAIT trial, subjects enrolled in the GAIT ancillary structural study met the original GAIT inclusion criteria, summarized as being at least 40 years of age with clinical evidence of painful OA of the knee for at least immediate past six months and radiographic evidence of OA as determined by having a Kellgren Lawrence grade 2- or 3-rated radiograph of the index knee. As mentioned earlier, 662 of the 1583 original GAIT study participants were also in the structural study. Patients were asked to continue following-up even if they stopped taking their assigned treatment. The study treatments were glucosamine hydrochloride (HCl) 500 mg three times daily, sodium chondroitin sulfate 400 mg three times daily, both glucosamine and chondroitin sulfate as above, celecoxib 200 mg once daily or placebo daily.

The metatarsophalangeal films were obtained as previously described, using the method of Buckland-Wright on participants in the GAIT radiographic substudy at baseline, at one year and two years or until study exit. The films were blinded and digitized for computer joint space width measurement, but were also read as hard copy films by two expert readers using digital calipers. During the course of study, a sample of all films was re-read by the raters and a computer rater. These were used to establish intra-and inter-rater reliability. Blinded radiographs were manually read by two raters. All measurements were recorded by the Mitutoyo Digimatic Calipers (Mitutoyo Products) and recorded as millimeters and to the hundredth millimeter. As has been elaborated before, for each knee, the joint space width (JSW) was the narrowest dimension in the medial compartment of the knee. This location was not necessarily at the middle of the weight-

bearing surface of the medial tibia (the point used for measuring the TIRD and rim to floor distance) and/or at any specified distance from the medial condyle. The direction or line upon which the JSW measurement was taken was perpendicular to the plane of the joint surfaces and not necessarily in parallel with the axis of the lower extremity (the line used to measure TIRD and rim to floor). The site used to measure JSW did not include an osteophyte nor involve the tibial spine.

6.1 Baseline analysis

The baseline JSWs of 281 participants were used to evaluate the agreement between raters. The mean and standard deviation of the difference were presented in the readings to provide information regarding the average magnitude of the difference. Inter-rater agreement between readers is assessed statistically by an unconditional intra-class correlation (ICC). The differences in readings between readers were also examined using Bland-Altman plots to graphically investigate systematic differences in disagreement[Bland, 1986, 1993]. Agreement in a Bland-Altman plot is indicated by differences that fall within the 95% confidence limits of mean difference. The results in this section are based on one reading by each rater. Table VIII shows that raters 1 and 2 have higher ICC and lower mean difference on JSW at baseline.

Examination of the Bland-Altman plots for all evaluated films shows that rater 1 was the most reliable as most readings fell within the 2SD boundary, whereas rater 2 was less so and the computer poorest of all. The plots (Figures 9 - 11)were done based on one reading by each rater at baseline.

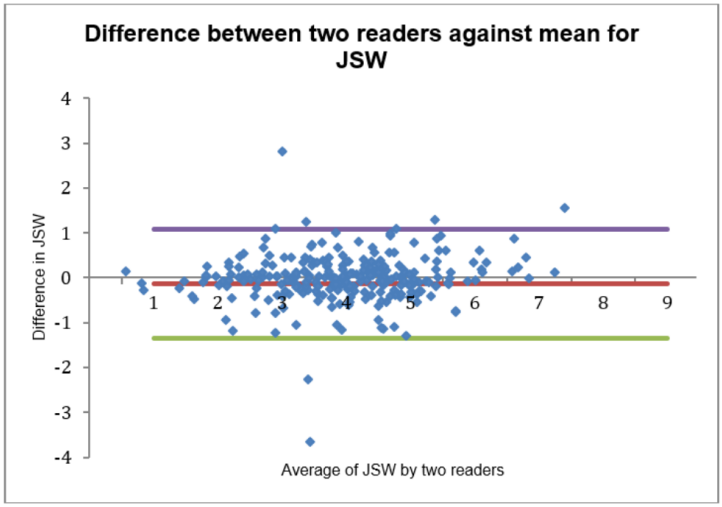

Figure 9: Rater 1 vs Rater 2 Bland-Altman plot n=281 radiographs.

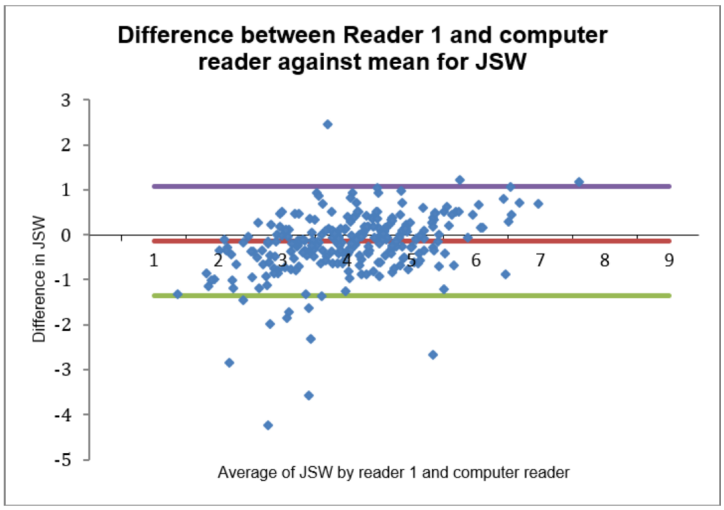

Figure 10: Rater 1 vs Computer Bland-Altman plot n=281 radiographs.

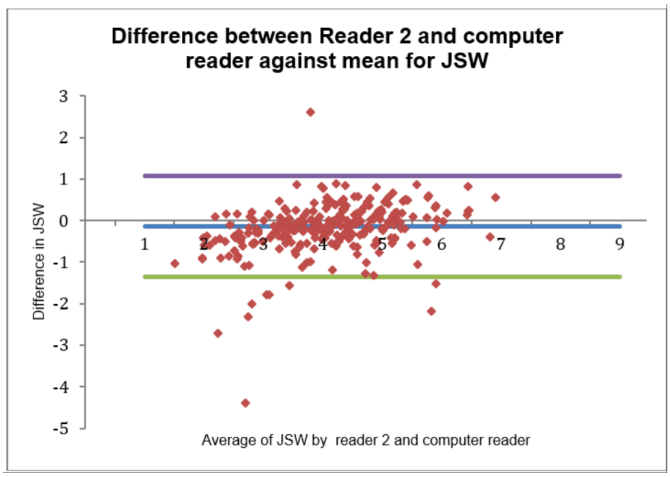

Figure 11: Rater 2 vs Computer Bland-Altman plot n=281 radiographs.

Table VIII: Descriptive statistics for human and computer: Joint Space Width n=281 radiographs

	ICC	Mean difference	95%CI
Rater 1 vs Rater 2	0.90	-0.02	-0.08, 0.04
Rater 1 vs Computer	0.80	-0.20	-0.28, -0.12
Rater 2 vs Computer	0.83	-0.18	-0.26, -0.11

6.2 Agreement on the three time points

Twenty-nine participants had completed X-rays at three time point and the quality of the films were accepted.

In figures 12, 13, and 14:

1. $JSW0\_1$: Reader 1 first reading for baseline X-ray
2. $JSW0\_2$: Reader 2 first reading for baseline X-ray

3. $JSW0$: Reader 3 (computer reader) first reading for baseline X-ray
4. $JSW0\_1\_2$: Reader 1 second reading for baseline X-ray
5. $JSW0\_2\_2$: Reader 2 second reading for baseline X-ray
6. $JSW0\_3\_2$: Reader 3 (computer reader) second reading for baseline X-ray
7. $JSW1\_1$: Reader 1 first reading for one year X-ray
8. $JSW1\_2$: Reader 2 first reading for one year X-ray
9. $JSW1$: Reader 3 (computer reader) first reading for one year X-ray
10. $JSW1\_1\_2$: Reader 1 second reading for one year X-ray
11. $JSW1\_2\_2$: Reader 2 second reading for one year X-ray
12. $JSW1\_3\_2$: Reader 3 (computer reader) second reading for one year X-ray
13. $JSW2\_1$: Reader 1 first reading for one year X-ray
14. $JSW2\_2$: Reader 2 first reading for one year X-ray
15. $JSW2$: Reader 3 (computer reader) first reading for one year X-ray
16. $JSW2\_1\_2$: Reader 1 second reading for two year X-ray
17. $JSW2\_2\_2$: Reader 2 second reading for two year X-ray
18. $JSW2\_3\_2$: Reader 3 (computer reader) second reading for two year X-ray

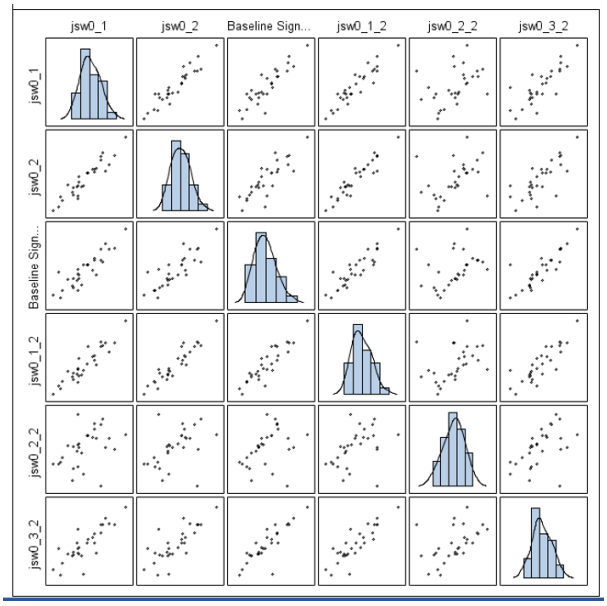

Figure 12: Correlation between observations at baseline

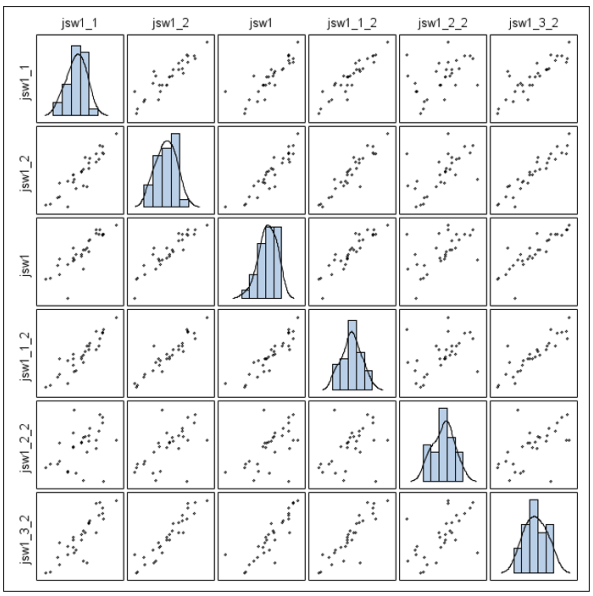

Figure 13: Correlation between observations at one year follow up

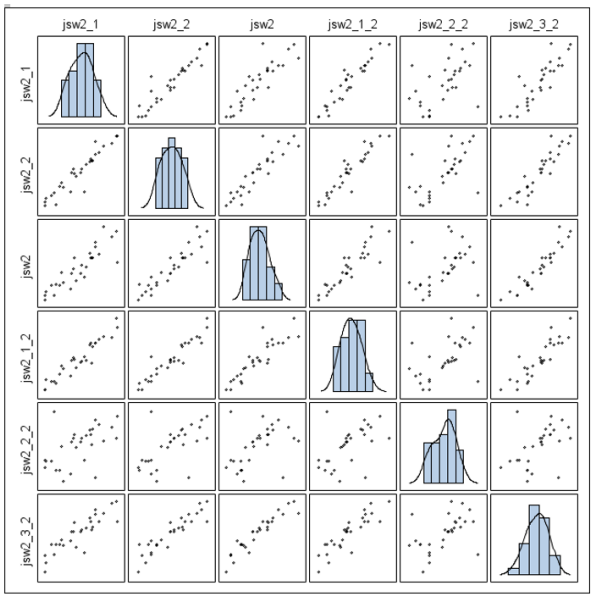

Figure 14: Correlation between observations at two year follow up

From the three plots, we can see that the correlation of the second reading of the second rater at three time points compared to the first reading are relatively low. The correlation of the second reading of this rater with the two readings of the other two raters are lower than the first reading of this rater with the two readings of the other two raters. It may indicate the degree of agreement for this rater with the other raters is relatively lower.

When we look closely at the agreement between two readings from the same rater, scatter plots in Figures 15-17 indicate the intra-CCC for each rater. Two readings by rater 1 at the three time points are very similar to each other; they narrowly cross $y = x$ lines. The variation by rater two is larger than that for rater 1, the range of spots at figure 16 being wide. The first reading of the third rater at the second time point is relatively larger than the second reading, most spots being above the line $y = x$; however, the reverse happened at the third time point. Table IX shows the Intra-CCC for each

rater. The value of Intra-CCC by the mixed model suggests that rater one is the most reliable rater of the three. Table X presents results of CCC between rater 1 and 2, rater 1 and 3, and rater 2 and 3. We use inter-CCC to show the quantitative agreement level between the raters. CCC between the second rater with the other raters are relatively smaller. Also, we estimate CCC by three-level (5.1) and two-level mixed effects models (similar to 5.1 but without random terms at the third level). The estimates of fixed variables by three-level model are almost the same as those by the two-level mixed effects model, but the third level variance-covariance removes some variation from the model error term variance and increases the covariance between raters slightly. Therefore, we get a higher CCC estimation by three-level models than the two-level models. In fact, based on the data structure as we showed before, the three-level model captures different levels of variance and covariance, and the results are closer to the real data. Therefore, CCC based on the three-level model is better than the one by a two-level model.

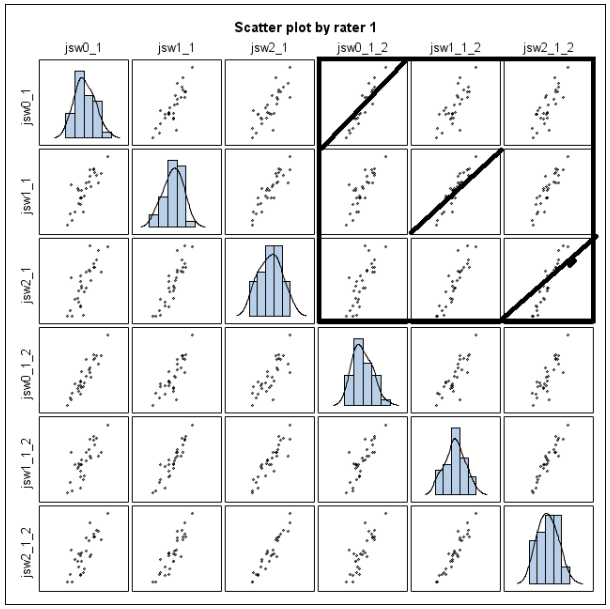

Figure 15: Scatter plot for rater 1 between two readings

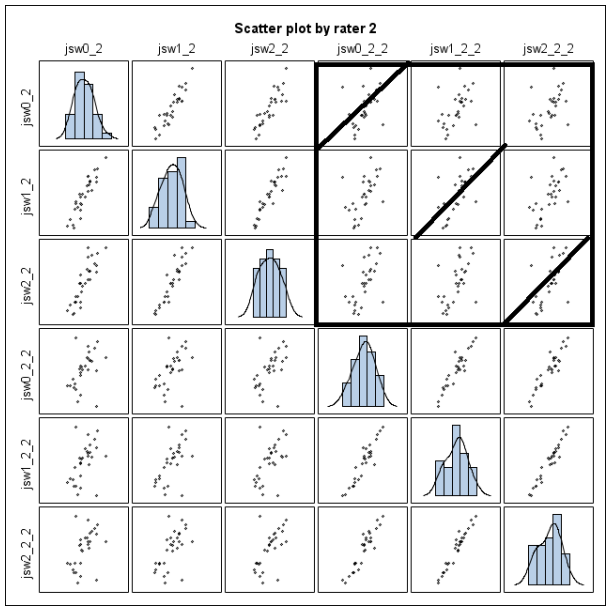

Figure 16: Scatter plot for rater 2 between two readings

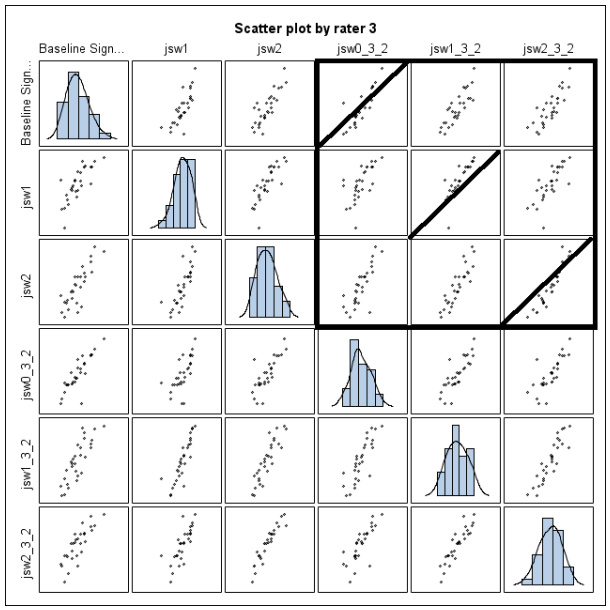

Figure 17: Scatter plot for rater 3 between two readings

7. MISSING DATA

Missing data is very common in longitudinal studies, even in very well designed and controlled clinical trials. It can be caused by participants dropping out or if they are assessed at a given study time point, only providing responses to part of the study variables. The reasons for dropout are varied such as randomly missed, adverse event, disappointed with the treatment effects, and so on.

7.1 Missing data mechanisms and main methods

To decide how to handle missing data, it is helpful to know how they are missing. Rubin (1976, [36]) described three missing data mechanisms.

(i) Missingness Completely At Random (MCAR). MCAR means if missingness does not depend on the data values, missing or observed. If data are missing completely at random, then the analysis based on complete data does not introduce bias.

(ii) Missing At Random (MAR). MAR is defined as when missingness only depends on the observed data values, and is independent from missing values. Both MCAR and MAR allow missingness to depend on observed correlated variables.

(iii) Missing Not At Random (MNAR). MNAR means that missingness depends on the missing data values. Missingness with the MNAR mechaism is also called nonignorable missingness.

Table IX: Intra-ICC estimation for three raters

Intra-ICC	Rater 1	Rater 2	Rater 3
	0.9190	0.8778	0.8436

Table X: Inter-CCC estimation by three-level model and two-level model

Inter-CCC estimation between rater one and two			
Category	Parameter	Three-level model Est(Std)	Two-level model Est(Std)
Fixed	β_0	3.6461 (0.2333)	3.6461 (0.2304)
	$\beta_0^{'}$	3.3876 (0.2342)	3.3876 (0.2339)
	β_1	-0.0952 (0.0723)	-0.0952 (0.0623)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0632 (0.0623)
Random variance	σ_θ^2	1.3252 (0.3818)	1.3514 (0.3813)
	$\sigma_{\theta\theta'}$	1.0882 (0.3511)	1.1596 (0.3509)
	$\sigma_{\theta'}^2$	1.3964 (0.3940)	1.3991 (0.3940)
	σ_ν^2	0.0706 (0.0616)	
	$\sigma_{\nu\nu'}$	0.2142 (0.0374)	
	$\sigma_{\nu'}^2$	0.0000 (.)	
	σ^2	0.4659 (0.0434)	0.4498 (0.0375)
CCC	*Inter*	0.6900	0.6266
Inter-CCC estimation between rater one and three			
Fixed	β_0	3.6461 (0.2291)	3.6461 (0.2264)
	$\beta_0^{'}$	4.1051 (0.1955)	4.1051 (0.1935)
	β_1	-0.0952 (0.0576)	-0.0952 (0.0454)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1579 (0.0454)
Random variance	σ_θ^2	1.3623 (0.3814)	1.3866 (0.3812)
	$\sigma_{\theta\theta'}$	1.1196 (0.3161)	1.1538 (0.3160)
	$\sigma_{\theta'}^2$	0.9719 (0.2746)	0.9868 (0.2744)
	σ_ν^2	0.1114 (0.0371)	
	$\sigma_{\nu\nu'}$	0.1025 (0.0272)	
	$\sigma_{\nu'}^2$	0.0831 (0.0319)	
	σ^2	0.1620 (0.0174)	0.2390 (0.0199)
CCC	*Inter*	0.8121	0.7670
Inter-CCC estimation between rater two and three			
Fixed	β_0	3.3876 (0.2345)	3.3876 (0.2340)
	$\beta_0^{'}$	4.1051 (0.1989)	4.1051 (0.1983)
	β_1	-0.0632 (0.0645)	-0.0632 (0.0628)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1579 (0.0628)
Random variance	σ_θ^2	1.3935 (0.3940)	1.3979 (0.3940)
	$\sigma_{\theta\theta'}$	0.9504 (0.3002)	1.0034 (0.3000)
	$\sigma_{\theta'}^2$	0.9461 (0.2745)	0.9504 (0.2744)
	σ_ν^2	0.0000 (.)	
	$\sigma_{\nu\nu'}$	0.1589 (0.0282)	
	$\sigma_{\nu'}^2$	0.0000 (.)	
	σ^2	0.4832 (0.0439)	0.4572 (0.0381)
CCC	*Inter*	0.6006	0.5497

Little and Rubin (2002, [35]) summarized four main methods to handle missing data.

The first method, called complete case analysis, is the simplest approach; it is also known as list-wise deletion. This method excludes missing cases and only analyzes the complete cases. In other wards, the inference drawn is based on the subjects with the observed data only. This method usually leads to biased estimates and loses power, especially when there is a large amount of missing data. If the observed data is a random sample from the full data with a small portion of missingness, unbiased parameter estimates might be achieved. However the analysis still has potential loss of precision.

The second method is by weighting. This is extended from complete case analysis. The approach adapts sampling method in survey data and involves an estimate of the probability of completeness. The weight is inversely proportional to the probability of selection multiplied by the probability of completeness. That is to say, the estimate of population mean is expressed as:

$$\bar{y} = \sum \frac{y_i}{\pi_i \hat{p_i}} / \sum (\pi_i \hat{p_i})^{-1}.$$

where y_i is observed data, π_i is the known probability of selection into the sample, and $\hat{p_i}$ is the probability of completeness. This method can reduce estimates' bias from the complete case analysis. However, the computation of variance using this method is not straight-forward; in addition, that computation is intensive.

The third method is imputation. Imputation means filling in the missing data with a value generated by different imputation approaches. After imputation, the new data will be analyzed as if they were complete. There are many ways of imputation. Imputations can be single imputation and multiple imputation. One of the common

imputation approaches is mean imputation. That means, the average of the completed data is filled in as the missing data. Another common imputation approach is Hot-deck imputation. It imputes the missing values by matching the observed part of the unit with an unit that is fully observed. Many variants of the Hot-deck imputation approach exist. These approaches differ mainly in how they find a match, including matching through stratification, neighborhood matching, weighting based on distance, and so on. That means the previous observed data is filled in for the missing data. A relatively advanced imputation is regression imputation by using a regression model to predict the missing data based on the other observed variables. The missing data will be filled by the predicted value after model parameter(s) estimation. The advantages for imputation are:

(i) Can handle a mixture of discrete and continuous variables;

(ii) Imputation is relatively easy to conduct;

(iii) The method can be relatively easily implemented.

The disadvantage are:

(i) Under improper imputation, accurate variance is not easy to find;

(ii) Conditionally specified models may be incompatible and the algorithm may not be convergent.

The single imputation methods do not take into account uncertainty created by missing data. To solve this problem, Rubin (1987, [37]) developed Multiple Imputation (MI). The procedure of MI is:

(i) Imputation based on the joint normal model.

(ii) Multiple imputed data sets can be analyzed by any statistical models.

(iii) Combination of model fit results from the imputed full data by Rubin's rule.

Monte Carlo (MC) simulation is often used in MI. Tanner and Wong (1987, [38]) and Gelfand and Smith (1990, [39]) showed MC is :

1. The circular formulas can be used to iteratively solve the computation problem through Monte Carlo simulation. Gibbs sampler is a Monte Carlo simulation approach to the computation problem.

2. In Gibbs sampler, an initial value for model parameter, say θ or for the missing data are given. For example, suppose that an initial θ is given, denoted by $\theta^{(0)}$.

3. The next step is to impute the missing values based on $y_{i(1)}^{mis} \sim f(y_{i(1)}^{mis}|y_i^{obs}, \theta^0)$.

4. Once $y_{i(1)}^{mis}, i = 1, \ldots, n$ are imputed, θ is updated by $\theta^1 \propto \prod_{i=1}^{n} f(y_{i(1)}^{mis}|y_i^{obs}, \theta)p(\theta)$.

5. The simulation is continued until convergence: when the distribution for the simulated random variables is (virtually) unchanged from one round to the next.

Rubin's combination rule:

1. With m imputed data sets, $y_1^{obs}, y_1^{mis(k)}, \ldots, y_n^{obs}, y_n^{mis(k)}$, fit a model $f(y, \theta)$ to obtain the maximum likelihood estimator $\hat{\theta}_k$ and the associated variance estimates $\hat{V}_k$, for $k = 1, \ldots, M$.

2. Combine the estimates of θ to obtain $\hat{\theta} = \frac{1}{M} \sum_1^M \hat{\theta}_k$.

3. The variance of $\hat{\theta}$ is estimated by $V = \frac{1}{M}\sum_1^M \hat{V}_k + (1+\frac{1}{M})\frac{1}{M-1}\sum_1^M(\hat{\theta_k} - \hat{\theta})^2$, where $(1+\frac{1}{M})$ is an adjustment for finite number of imputations.

The last method Rubin described is a model-based procedure, otherwise known as a likelihood based procedure. This approach models the observed data, draws inferences based on likelihood, and estimates the parameters of interest by maximizing likelihood. The estimates of variance by this procedure will consider missingness in the data. To take into account the effect of missing data, one could generate a likelihood including a model of missing data mechanism besides the model for observed data. Therefore, this method has an advantage of flexibility. If models are correctly specified, the inferences drawn based on the model are more efficient compared to the methods mentioned above. Rubin (1976, [36]) defined a full model including both the distributions of data and missing-data mechanism. One issue is related to the closed-form solution of the maximum likelihood problem under monotone missing data. Variance of the maximum likelihood estimator needs to be obtained in making inference about model parameters. When the missing data forms arbitrary missing data patterns, the close-form solution may not exist. When the parameters for the different conditional models are related, separate maximization cannot be done, even if we have monotone missing data patterns. Expectation-Maximization algorithm and related methods will be used to solve this problem.

Properties of methods are strongly influenced by assumptions made about missing mechanism. Mixed effect models(MRM), Generalized Estimating Equations(GEE), Covariance Pattern Model (CPM) analyze data with missing as incomplete. However, different methods have different mechanism assumptions. Generalized Estimating Equations assumes special case of MCAR. Likelihood based methods (MRM, CPM) assume MAR. When the missing data are nonignorable, Little (1995, [41]) discussed the selec-

tion models and pattern-mixture models for handling it. The selection model combines a model for the ideal complete data with a model of missingness processes. Hedeker and Gibbons (1997, [40]) developed pattern mixture models by using missing data pattern information in the longitudinal modeling. The pattern-mixture models stratify the missing data into different patterns and construct a corresponding complete-data model within strata. In the pattern-mixture model framework, subjects in the same strata share the same pattern of missing data. The complete data model is estimated for each pattern and the pattern-specific estimates are averaged into an overall result.

7.2 GAIT Study and JSW Missing Data Mechanisms

In GAIT study, we have different types of missing reasons, such as patient withdrawn from the study due to adverse events, due to not willing to do follow up X-rays, lost follow up, the quality of the X-rays, the different rater's judgment of the quality of the X-rays.

Figure 18 is the missing pattern of the three rater's first readings. Computer reading at baseline (JSW0) was used to judge whether the patients qualified for this study or not. None of the observations of this variable is missing. However, besides variable JSW0, there are unique combinations of 16 missing patterns, as showing in Figure 18. Out of 328 subjects, the 109 who have completed three time points X-rays and got JSW measured by three rater. All the others have at least one missing data by at least one rater. Among 109 completed subjects, 30 subjects were randomly selected. Three raters were asked to perform the second reading for all the X-rays of the 30 selected subjects. However, they completed the second readings for 29 subjects. In other words, only 29 subjects have completed data with two readings at three time points by three raters.

Missing Data Patterns											
Group	jsw0	jsw0_1	jsw0_2	jsw1	jsw1_1	jsw1_2	jsw2	jsw2_1	jsw2_2	Freq	Percent
1	X	X	X	X	X	X	X	X	X	109	33.23
2	X	X	X	X	X	X	X	X	.	1	0.30
3	X	X	X	X	X	X	.	X	X	12	3.66
4	X	X	X	X	X	X	.	X	.	5	1.52
5	X	X	X	X	X	X	.	.	X	1	0.30
6	X	X	X	X	X	X	.	.	.	34	10.37
7	X	X	X	X	.	X	.	.	.	1	0.30
8	X	X	X	.	X	X	X	X	X	5	1.52
9	X	X	X	.	X	.	X	X	X	3	0.91
10	X	X	X	.	.	.	X	X	X	10	3.05
11	X	X	.	X	X	.	X	X	X	1	0.30
12	X	X	.	X	X	.	.	X	.	1	0.30
13	X	X	.	X	X	.	.	.	.	3	0.91
14	X	X	.	.	X	.	X	X	.	1	0.30
15	X	.	.	X	.	.	X	.	.	88	26.83
16	X	.	.	X	.	.	.	.	.	37	11.28
17	X	.	.	.	.	.	X	.	.	16	4.88

Figure 18: Missing pattern

7.3 Mixed Effect model used for handling missing data

Among 328 subjects who have at least one x-ray rated, only 29 subjects have completed data. We randomly select 10 subjects from 299 subjects who have missing data and combine their data with the 29 completers. We do this 10 times and get 10 samples. Each sample data has 29 completers and 10 subjects with missing data. Assuming missing pattern is MAR, mixed effect model (same as model described in 5.1) is used for

handling missing data by EM algorithm conditional on observed values of the dependent variable. After we get the estimates from the 10 samples, the average of the estimates and inter-CCC between raters are calculated.

Similarly, we randomly select 15 subjects from the subjects with missing, and combine with the 29 completers, do the same process 10 times and get another 10 samples. The average inter-CCCs are estimated by these 10 samples.

Table XI lists the results from complete data (n=29), average estimates from combining missing data and complete data(n=39, 44). When the sample size gets larger, the means of the JSW are similar at each time point although the variance of the JSW by each rater and the covariance between raters are larger, the speed of covairance increasing is slightly larger than each rater's variance increasing. The inter-CCC is pretty stable between rater 1 and 3 when adding more subjects with missing values. That may be due to those two raters being more consistent and the missingness being at random. The inter-CCCs are getting better between 1 and 2, 2 and 3. That may be due to rater 2 having a larger variation on the second readings. We add more data which provides more information and helps to confirm the agreement of the second rater with the other raters.

In summary, when adding more subjects, the average of mean differences are similar, the variance of each rater and the covariance between raters are getting slightly larger, the average inter-CCCs from 10 samples with 10 subjects with missing or 15 with missing by handling missing using mixed models are slightly larger compared with the results from complete data.

Table XI: CCC estimation including missing data by three-level mixed effects models

Inter-CCC estimation between rater one and two				
Category	Parameter	Complete Est(Std)	Adding 10 with missing Est(Std)	Adding 15 with missing Est(Std)
Fixed	β_0	3.6461 (0.2333)	3.7440 (0.0493)	3.7322 (0.0490)
	$\beta_0^{'}$	3.3876 (0.2342)	3.5065 (0.0488)	3.5091 (0.0477)
	β_1	-0.0952 (0.0723)	-0.0959 (0.0045)	-0.0975 (0.0043)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0655 (0.0035)	-0.0651 (0.0033)
Random variance	σ_θ^2	1.3252 (0.3818)	1.4589 (0.3831)	1.5376 (0.3922)
	$\sigma_{\theta\theta'}$	1.0882 (0.3511)	1.2338 (0.3523)	1.2996 (0.3586)
	$\sigma_{\theta'}^2$	1.3964 (0.3940)	1.4963 (0.3857)	1.5420 (0.3878)
	σ_ν^2	0.0706 (0.0616)	0.0638 (0.0555)	0.0642 (0.0544)
	$\sigma_{\nu\nu'}$	0.2142 (0.0374)	0.2007 (0.0338)	0.1986 (0.0329)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)
	σ^2	0.4659 (0.0434)	0.4434 (0.0397)	0.4385 (0.0388)
Total variance	σ_1^2	1.8617	1.9660	2.0404
	σ_{12}	1.8623	1.9397	1.9805
	σ_2^2	1.3024	1.4345	1.4982
CCC	*Inter*	0.6900	0.7237	0.7345
Inter-CCC estimation between rater one and three				
Fixed	β_0	3.6461 (0.2291)	3.7519 (0.2047)	3.7203 (0.1976)
	$\beta_0^{'}$	4.1051 (0.1955)	4.1631 (0.1697)	4.1273 (0.1622)
	β_1	-0.0952 (0.0576)	-0.0918 (0.0542)	-0.1005 (0.0537)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1515 (0.0509)	-0.1581 (0.0492)
Random variance	σ_θ^2	1.3623 (0.3814)	1.4405 (0.3511)	1.4958 (0.3458)
	$\sigma_{\theta\theta'}$	1.1196 (0.3161)	1.1453 (0.2807)	1.1833 (0.2740)
	$\sigma_{\theta'}^2$	0.9719 (0.2746)	0.9629 (0.2372)	0.9911 (0.2302)
	σ_ν^2	0.1114 (0.0371)	0.1137 (0.0363)	0.1240 (0.0386)
	$\sigma_{\nu\nu'}$	0.1025 (0.0272)	0.1111 (0.0276)	0.1201 (0.0289)
	$\sigma_{\nu'}^2$	0.0831 (0.0319)	0.1005 (0.0341)	0.1042 (0.0340)
	σ^2	0.1620 (0.0174)	0.1613 (0.0171)	0.1621 (0.0172)
Total variance	σ_1^2	1.6357	1.7155	1.7820
	σ_{13}	1.2170	1.2246	1.2575
	σ_3^2	1.2221	1.2564	1.3034
CCC	*Inter*	0.8121	0.8179	0.8210
Inter-CCC estimation between rater two and three				
Fixed	β_0	3.3876 (0.2345)	3.5154 (0.0445)	3.4975 (0.0416)
	$\beta_0^{'}$	4.1051 (0.1989)	4.1654 (0.0298)	4.1290 (0.0272)
	β_1	-0.0632 (0.0645)	-0.0641 (0.0036)	-0.0649 (0.0034)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1543 (0.0036)	-0.1576 (0.0032)
Random variance	σ_θ^2	1.3935 (0.3940)	1.4450 (0.3628)	1.4822 (0.3579)
	$\sigma_{\theta\theta'}$	0.9504 (0.3002)	0.9769 (0.2670)	1.0132 (0.2615)
	$\sigma_{\theta'}^2$	0.9461 (0.2745)	0.9289 (0.2362)	0.9593 (0.2300)
	σ_ν^2	0.0000 (.)	0.0000 (.)	0.0000 (.)
	$\sigma_{\nu\nu'}$	0.1589 (0.0282)	0.1530 (0.0271)	0.1487 (0.0259)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0053 (0.0695)	0.0025 (0.0596)
	σ^2	0.4832 (0.0439)	0.4618 (0.0399)	0.4539 (0.0383)
Total variance	σ_2^2	1.8767	1.9068	1.9361
	σ_{23}	1.4293	1.3959	1.4157
	σ_3^2	1.1093	1.1299	1.1619
CCC	*Inter*	0.6006	0.6221	0.6338

7.4 Use of imputation for handling missing data

As we discussed in the previous section, there are many methods for imputation. Filling the data using the mean is not appropriate in our case since we want to evaluate the agreements between raters; simply using the mean will force the evaluation to be the similar, that will introduce bias on evaluating agreement. We also know that JSW will decrease by time for everybody, and the decreasing trend is varied by subjects. Thus, the last observation carried forward is not appropriate either. If we recall the scatter plots in Figures 6.7-6.9, we find a high linear correlation between the second readings and the first readings for each rater at each timepoint. Hence, we will get the model estimation based on complete data for the second readings through the first readings after adjusting for the baseline characteristics. The second reading missing data will be filled by the predicted value given subjects baseline characteristics, such as age and WOMAC pain stratum.

The model used for imputation is the following:

$$y_2 = \beta_0 + \beta_1 y_1 + \beta_2 x_1 + \beta_3 x_2,$$

Here, y_2 means the second reading, and y_1 means the first reading by the same rater and at the same time point. x_1 means age and x_2 means WOMAC pain stratum.

Figures 19-21 are the plots of model fit diagnostics for each rater at baseline. The model fits very well for Rater 1 (R-square = 0.87) and 3(R-square = 0.74), but not for rater 2 (R-square = 0.27). The predicted second readings are closer to the observed value by first and the third raters. This is due to the second reading of the second rater being inconsistent with the first reading, unlike the other raters.

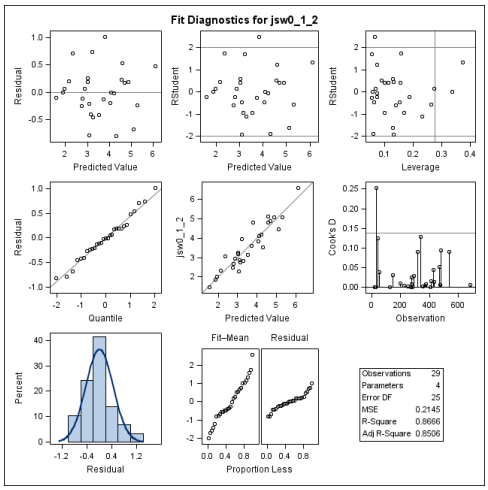

Figure 19: Model used for imputation rater 1

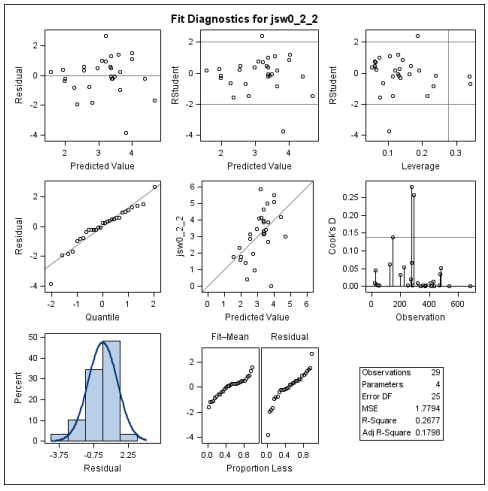

Figure 20: Model used for imputation rater 2

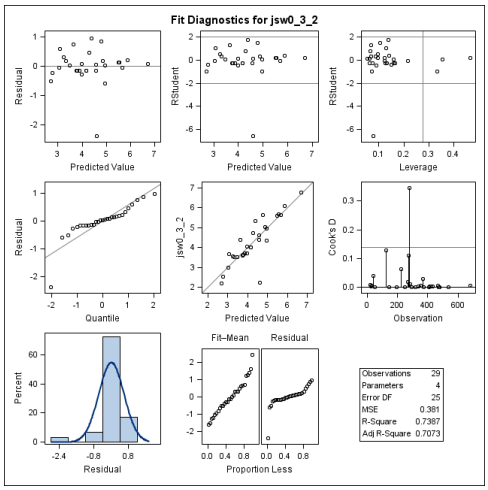

Figure 21: Model used for imputation rater 3

The information in Figures 19-21 confirms the conclusion we got from the previous scatter plots. The models fitting for each rater at the other time points are similar to the baseline.

After the missing data of second readings are imputed, the data is applied to the mixed model (4.10) to evaluate the inter-CCC between raters. Repeating this for all sample data described in the previous section, we get the average of the estimation. TableXII lists the results from complete data (n=29) and average estimates from datasets with additional 10 samples with imputed data (n=39, 44). When the sample size gets larger, the estimation of mean of the JSW are similar, the first-level and second-level variance are getting larger and that leaves the third-level variance, left-over variance decreasing. The estimated inter-CCCs are very close of both completed data and imputed data.

Table XII: CCC estimation including missing data by model imputation

Inter-CCC estimation between rater one and two				
Category	Parameter	Complete Est(Std)	Adding 10 with missing Est(Std)	Adding 15 with missing Est(Std)
Fixed	β_0	3.6461 (0.2333)	3.7185 (0.0478)	3.7044 (0.0474)
	$\beta_0^{'}$	3.3876 (0.2342)	3.4456 (0.0444)	3.4346 (0.0424)
	β_1	-0.0952 (0.0723)	-0.0947 (0.0044)	-0.1006 (0.0041)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0629 (0.0033)	-0.0595 (0.0030)
Random variance	σ_θ^2	1.3252 (0.3818)	1.4190 (0.3702)	1.4958 (0.3783)
	$\sigma_{\theta\theta'}$	1.0882 (0.3511)	1.1463 (0.3291)	1.1949 (0.3308)
	$\sigma_{\theta'}^2$	1.3964 (0.3940)	1.3628 (0.3494)	1.3749 (0.3434)
	σ_ν^2	0.0706 (0.0616)	0.0769 (0.0531)	0.0804 (0.0523)
	$\sigma_{\nu\nu'}$	0.2142 (0.0374)	0.2031 (0.0320)	0.1943 (0.0311)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)
	σ^2	0.4659 (0.0434)	0.4368 (0.0375)	0.4278 (0.0358)
Total variance	σ_1^2	1.8617	1.9327	2.0040
	σ_{12}	1.8623	1.7996	1.8028
	σ_2^2	1.3024	1.3494	1.3892
CCC	$Inter$	0.6900	0.7104	0.7180
Inter-CCC estimation between rater one and three				
Fixed	β_0	3.6461 (0.2291)	3.7511 (0.2023)	3.7311 (0.1949)
	$\beta_0^{'}$	4.1051 (0.1955)	4.2026 (0.1693)	4.1843 (0.1608)
	β_1	-0.0952 (0.0576)	-0.0941 (0.0542)	-0.1061 (0.0542)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1583 (0.0516)	-0.1669 (0.0501)
Random variance	σ_θ^2	1.3623 (0.3814)	1.4082 (0.3425)	1.4499 (0.3346)
	$\sigma_{\theta\theta'}$	1.1196 (0.3161)	1.1302 (0.2767)	1.1531 (0.2669)
	$\sigma_{\theta'}^2$	0.9719 (0.2746)	0.9554 (0.2356)	0.9677 (0.2252)
	σ_ν^2	0.1114 (0.0371)	0.1330 (0.0367)	0.1560 (0.0401)
	$\sigma_{\nu\nu'}$	0.1025 (0.0272)	0.1182 (0.0284)	0.1325 (0.0308)
	$\sigma_{\nu'}^2$	0.0831 (0.0319)	0.1289 (0.0339)	0.1424 (0.0336)
	σ^2	0.1620 (0.0174)	0.1425 (0.0137)	0.1371 (0.0126)
Total variance	σ_1^2	1.6357	1.6837	1.7430
	σ_{13}	1.2170	1.2268	1.2472
	σ_3^2	1.2221	1.2484	1.2856
CCC	$Inter$	0.8121	0.8138	0.8143
Inter-CCC estimation between rater two and three				
Fixed	β_0	3.3876 (0.2345)	3.4730 (0.0405)	3.4555 (0.0366)
	$\beta_0^{'}$	4.1051 (0.1989)	4.2021 (0.0295)	4.1844 (0.0265)
	β_1	-0.0632 (0.0645)	-0.0631 (0.0032)	-0.0628 (0.0028)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1570 (0.0035)	-0.1665 (0.0031)
Random variance	σ_θ^2	1.3935 (0.3940)	1.3323 (0.3312)	1.3295 (0.3162)
	$\sigma_{\theta\theta'}$	0.9504 (0.3002)	0.9412 (0.2552)	0.9576 (0.2441)
	$\sigma_{\theta'}^2$	0.9461 (0.2745)	0.9348 (0.2355)	0.9513 (0.2251)
	σ_ν^2	0.0000 (.)	0.0000 (.)	0.0000 (.)
	$\sigma_{\nu\nu'}$	0.1589 (0.0282)	0.1531 (0.0285)	0.1538 (0.0268)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0336 (0.0549)	0.0392 (0.0477)
	σ^2	0.4832 (0.0439)	0.4422 (0.0366)	0.4291 (0.0343)
Total variance	σ_2^2	1.8767	1.7744	1.7586
	σ_{23}	1.4293	1.4105	1.4196
	σ_3^2	1.1093	1.0943	1.1114
CCC	$Inter$	0.6006	0.6075	0.6196

7.5 Use of multiple imputation for handling missing data

The Markov chain Monte Carlo (MCMC) method is used for missing data multiple imputation. MCMC is a sequence of random variables in which the distribution of each element depends only on the value of the previous element. When a Markov chain is long enough for the distribution of the elements to stabilize to a stationary distribution, we get the distribution of interest. Repeatedly simulating steps of the chain, simulates drawing from the distribution of interest. In our situation, we use this method to impute all the missing values. Each sample with missing values is imputed 5 times, resulting in 5 datasets for each sample data. Each dataset is analyzed by the same mixed model to get model fit. Rubin's combination rule is used to get the combination results for each sample. Then, the average of 10 samples is calculated in tableXIII. The results suggest that when data have 25% missingness (10 subjects with missing), the average of model estimation is very close to the complete data; however, when missingness gets worse (15 subjects with missing, 34%), the MCMC methods may introduce more uncertainty, as a result, agreement between the raters becomes smaller.

7.6 Use of pattern mixture model for handling missing data

When missing data are nonignorable (MNAR), standard statistical models can yield badly biased results. The problem is that the observed data provide no information to either confirm or refute ignorability. In other words, we cannot test the missingness is MAR vs. MNAR. Pattern mixture model is used as the sensitivity analysis.

Table XIII: Inter-CCC estimation including missing data by multiple imputation

Inter-CCC estimation between rater one and two				
Category	Parameter	Complete Est(Std)	Adding 10 with missing Est(Std)	Adding 15 with missing Est(Std)
Fixed	β_0	3.6461 (0.2333)	3.7650 (0.2152)	3.7421 (0.2054)
	$\beta_0^{'}$	3.3876 (0.2342)	3.4534 (0.2348)	3.4562 (0.2283)
	β_1	-0.0952 (0.0723)	-0.0896 (0.0747)	-0.1082 (0.0766)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0550 (0.0685)	-0.0584 (0.0715)
Random variance	σ_θ^2	1.3252 (0.3818)	1.4264 (0.3676)	1.4039 (0.3527)
	$\sigma_{\theta\theta'}$	1.0882 (0.3511)	1.1117 (0.3499)	1.0804 (0.3714)
	$\sigma_{\theta'}^2$	1.3964 (0.3940)	1.4655 (0.4164)	1.6046 (0.5172)
	σ_ν^2	0.0706 (0.0616)	0.0804 (0.0770)	0.1413 (0.1422)
	$\sigma_{\nu\nu'}$	0.2142 (0.0374)	0.2697 (0.0667)	0.3262 (0.1077)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)
	σ^2	0.4659 (0.0434)	0.6005 (0.1103)	0.7175 (0.1989)
Total variance	σ_1^2	1.8617	2.1073	2.2627
	σ_{12}	1.8623	2.0659	2.3221
	σ_2^2	1.3024	1.3814	1.4066
CCC	$Inter$	0.6900	0.6501	0.6062
Inter-CCC estimation between rater one and three				
Fixed	β_0	3.6461 (0.2291)	3.7650 (0.2107)	3.7421 (0.2000)
	$\beta_0^{'}$	4.1051 (0.1955)	4.1998 (0.1773)	4.1890 (0.1660)
	β_1	-0.0952 (0.0576)	-0.0896 (0.0606)	-0.1082 (0.0607)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1477 (0.0594)	-0.1625 (0.0599)
Random variance	σ_θ^2	1.3623 (0.3814)	1.4757 (0.3676)	1.4686 (0.3510)
	$\sigma_{\theta\theta'}$	1.1196 (0.3161)	1.1731 (0.2910)	1.1509 (0.2726)
	$\sigma_{\theta'}^2$	0.9719 (0.2746)	0.9840 (0.2454)	0.9570 (0.2292)
	σ_ν^2	0.1114 (0.0371)	0.1199 (0.0472)	0.1391 (0.0598)
	$\sigma_{\nu\nu'}$	0.1025 (0.0272)	0.1163 (0.0358)	0.1398 (0.0444)
	$\sigma_{\nu'}^2$	0.0831 (0.0319)	0.0962 (0.0476)	0.1196 (0.0639)
	σ^2	0.1620 (0.0174)	0.2001 (0.0297)	0.2178 (0.0358)
Total variance	σ_1^2	1.6357	1.7957	1.8255
	σ_{13}	1.2170	1.2803	1.2944
	σ_3^2	1.2221	1.2895	1.2907
CCC	$Inter$	0.8121	0.8014	0.7885
Inter-CCC estimation between rater two and three				
Fixed	β_0	3.3876 (0.2345)	3.4534 (0.2351)	3.4562 (0.2285)
	$\beta_0^{'}$	4.1051 (0.1989)	4.1998 (0.1816)	4.1890 (0.1715)
	β_1	-0.0632 (0.0645)	-0.0550 (0.0694)	-0.0584 (0.0723)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1477 (0.0715)	-0.1625 (0.0739)
Random variance	σ_θ^2	1.3935 (0.3940)	1.4623 (0.4161)	1.6010 (0.5172)
	$\sigma_{\theta\theta'}$	0.9504 (0.3002)	0.9432 (0.2868)	0.9145 (0.3015)
	$\sigma_{\theta'}^2$	0.9461 (0.2745)	0.9434 (0.2456)	0.9023 (0.2301)
	σ_ν^2	0.0000 (.)	0.0000 (.)	0.0000 (.)
	$\sigma_{\nu\nu'}$	0.1589 (0.0282)	0.2130 (0.0641)	0.2757 (0.2757)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)
	σ^2	0.4832 (0.0439)	0.6196 (0.1040)	0.7391 (0.1944)
Total variance	σ_2^2	1.8767	1.7744	2.3402
	σ_{23}	1.4293	1.4105	1.6415
	σ_3^2	0.6006	0.6059	0.5439
CCC	$Inter$	0.6006	0.6059	0.5439

One sample with 10 subjects with missing data and another sample with 15 are applied by pattern mixture model. TableXIV shows that for the first sample, 29 have complete data, 5 have all first readings but have all second readings missing, and the other 5 have one or more missing values at first readings and all second readings. Similarly, the second sample has 29 completed data, 3 have all first readings but missing all second readings, the other 12 have one or more missing at first readings and all second readings. In this case, we use two dummy variables D1 and D2. Indicator variable of D1 is 0 for complete data at all first readings and 1 for any types missing at first readings. Similarly D2 is 0 for complete data at second readings and 1 for any types of missing at second readings.

Table XIV: Missing data pattern for two sample data

Data of 29 completers and 10 with missing												
jsw0_1	jsw0_2	jsw0	jsw1_1	jsw1_2	jsw1	jsw2_1	jsw2_2	jsw2	2^{nd}	n	D1	D2
X	X	X	X	X	X	X	X	X	X	29	0	0
X	X	X	X	X	X	X	X	X	.	5	0	1
X	X	X	X	X	X	.	.	.	.	2	1	1
X	X	X	X	X	.	X	X	X	.	1	1	1
.	.	X	.	.	X	.	.	X	.	1	1	1
.	.	X	.	.	X	.	.	.	.	1	1	1
Data of 29 completers and 15 with missing												
X	X	X	X	X	X	X	X	X	X	29	0	0
X	X	X	X	X	X	X	X	X	.	3	0	1
X	X	X	X	X	X	X	.	X	.	1	1	1
X	X	X	X	X	X	.	.	.	.	3	1	1
.	.	X	.	.	X	.	.	X	.	7	1	1
.	.	X	.	.	X	.	.	.	.	1	1	1

Mixed-effects pattern mixture model is written:

$$\begin{aligned} y_{ijkl} &= \beta_0\delta_{(l=1)} + \beta_1\delta_{(l=1)}t + \theta_{0i}\delta_{(l=1)} + \nu_{0j(i)}\delta_{(l=1)} \\ &+ \beta'_0\delta_{(l=2)} + \beta'_1\delta_{(l=2)}t + \theta'_{0i}\delta_{(l=2)} + \nu'_{0j(i)}\delta_{(l=2)} \\ &+ \beta_{0m1}\delta_{(l=1)}D_1 + \beta_{1m1}\delta_{(l=1)}D_1t + \beta'_{0m1}\delta_{(l=2)}D_1 + \beta'_{1m1}\delta_{(l=2)}D_1t \\ &+ \beta_{0m2}\delta_{(l=1)}D_2 + \beta_{1m2}\delta_{(l=1)}D_2t + \beta'_{0m2}\delta_{(l=2)}D_2 + \beta'_{1m2}\delta_{(l=2)}D_2t + \epsilon_{ijkl}. \end{aligned} \tag{7.1}$$

In this model:

(i) β_0, β_1, β'_0 and β'_1 are for completers.

(ii) β_{0m1}, β_{1m1}, β'_{0m1} and β'_{1m1} indicate how the group with first reading missing differ from the completers.

(iii) β_{0m2}, β_{1m2}, β'_{0m2} and β'_{1m2} indicate how the group with second reading missing differ from the completers.

After model estimation, the average results are obtained by using sample proportions as estimates of missing-data pattern proportions. Vide Little (1995, [41]). That is to say,

$$\hat{\bar{\beta}} = \hat{\pi_c}\hat{\beta_c} + \hat{\pi_{m1}}\hat{\beta_{dm1}} + \hat{\pi_{m2}}\hat{\beta_{dm2}} = \hat{\beta_c} + \hat{\pi_{m1}}\hat{\beta_{m1}} + \hat{\pi_{m2}}\hat{\beta_{m2}} \tag{7.2}$$

Here, $\hat{\beta_c}$ corresponds to the coefficients of the group with completed data in the current model formulation; $\hat{\beta_{dm1}}$ and $\hat{\beta_{dm2}}$ correspond to the coefficients of the group with missing data at first readings or second readings in the current model; $\hat{\beta_{m1}}$ corresponds to the group with missing values at first readings coefficients differing from the completers group in the current model formulation; $\hat{\beta_{m2}}$ correspond to the group with missing values differing from the completers group at second readings coefficients in the current model

formulation; $\hat{\pi_{m1}}$, $\hat{\pi_{m2}}$ are the sample proportion of missing first readings and second readings.

Delta Method is used for estimating asymptotic variance of averaged estimates [Hedeker's handout on pattern mixture model].

$$\hat{V}(\hat{\bar{\beta}}) = \hat{V}(\hat{\beta_c}) + \hat{\pi_{m1}}^2 V(\hat{\beta_{m1}}) + \hat{\beta_{m1}}^2 V(\hat{\pi_{m1}}) + \hat{\pi_{m2}}^2 V(\hat{\beta_{m2}}) + \hat{\beta_{m2}}^2 V(\hat{\pi_{m2}}) \tag{7.3}$$

In this case, under marginal model for completion,

$$V(\hat{\pi_m}) = \hat{\pi_m}(1 - \hat{\pi_m})/N = n_m n_c / N^3. \tag{7.4}$$

After taking both average estimation and the estimated asymptotic variance of averaged estimates into consideration, the results are shown in tableXV. TableXV also includes the results from the mixed-effect model estimation on the data with 10, 15 subjects missing. We can see that the inter-CCCs from pattern mixture model estimation are slightly lower than the results from the mixed effect model. That may be due to the uncertainty reducing the agreement between raters which are counted to the consideration.

In summary, mixed model, model based imputation, multiple imputation, and pattern mixture imputation give us very similar results. That may suggest that the missing at random assumption is met in this case.

Table XV: Inter-CCC estimation including missing data by pattern mixture model

Inter-CCC estimation between rater one and two				
Parameter	Adding 10 with missing Est(Std)		Adding 15 with missing Est(Std)	
	Mixed model	Pattern mixed model	Mixed model	Pattern mixed model
β_0	3.7460 (0.2087)	3.4547 (0.3424)	3.7626 (0.2254)	3.6105 (0.5075)
$\beta_0^{'}$	3.5271 (0.2130)	3.2081 (0.2727)	3.5021 (0.2204)	3.4222 (0.4981)
β_1	-0.0680 (0.0635)	-0.1225 (0.1320)	-0.0902 (0.0655)	-0.1594 (0.1949)
$\beta_1^{'}$	-0.0407 (0.0570)	-0.1083 (0.1188)	-0.0602 (0.0587)	-0.1381 (0.2060)
σ_θ^2	1.3634 (0.3462)	1.2778 (0.3344)	1.5758 (0.4041)	1.5160 (0.4009)
$\sigma_{\theta\theta'}$	1.1934 (0.3282)	1.0724 (0.3090)	1.3161 (0.3657)	1.2656 (0.3654)
$\sigma_{\theta'}^2$	1.4752 (0.3683)	1.3357 (0.3437)	1.5448 (0.3912)	1.5235 (0.3971)
σ_ν^2	0.0533 (0.0511)	0.0527 (0.0520)	0.0559 (0.0533)	0.0574 (0.0547)
$\sigma_{\nu\nu'}$	0.1884 (0.0318)	0.1897 (0.0325)	0.1935 (0.0330)	0.1968 (0.0338)
$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)	0.0000 (.)
σ^2	0.4280 (0.0375)	0.4319 (0.0382)	0.4365 (0.0389)	0.4422 (0.0397)
σ_1^2	1.8447	1.7624	2.0682	2.0156
σ_{12}	1.9032	1.7676	1.9813	1.9657
σ_2^2	1.3818	1.2621	1.5096	1.4624
Inter-CCC	0.7302	0.7043	0.7359	0.7295
Inter-CCC estimation between rater one and three				
β_0	3.7727 (0.1967)	3.5742 (0.3424)	3.8222 (0.1980)	3.6835 (0.3579)
$\beta_0^{'}$	4.1898 (0.1635)	4.0891 (0.2727)	4.2164 (0.1604)	4.1333 (0.1405)
β_1	-0.0678 (0.0511)	-0.1338 (0.1320)	-0.0771 (0.0511)	-0.1884 (0.1074)
$\beta_1^{'}$	-0.1318 (0.0478)	-0.2363 (0.1188)	-0.1217 (0.0446)	-0.1127 (0.8266)
σ_θ^2	1.3367 (0.3225)	1.2786 (0.3177)	1.5114 (0.3465)	1.4269 (0.3354)
$\sigma_{\theta\theta'}$	1.0646 (0.2591)	1.0278 (0.2574)	1.1880 (0.2720)	1.1456 (0.2689)
$\sigma_{\theta'}^2$	0.8982 (0.2195)	0.8786 (0.2202)	0.9912 (0.2265)	0.9752 (0.2284)
σ_ν^2	0.0993 (0.0317)	0.0963 (0.0314)	0.0956 (0.0313)	0.0980 (0.0320)
$\sigma_{\nu\nu'}$	0.0971 (0.0236)	0.0922 (0.0229)	0.0866 (0.0216)	0.0882 (0.0220)
$\sigma_{\nu'}^2$	0.0793 (0.0287)	0.0742 (0.0278)	0.0675 (0.0259)	0.0675 (0.0259)
σ^2	0.1568 (0.0163)	0.1573 (0.0164)	0.1593 (0.0166)	0.1587 (0.0166)
σ_1^2	1.5928	1.5322	1.7663	1.6836
σ_{13}	1.1343	1.1101	1.2180	1.2014
σ_3^2	1.1617	1.1200	1.2746	1.2338
Inter-CCC	0.8148	0.7965	0.8206	0.7806
Inter-CCC estimation between rater two and three				
β_0	3.5559 (0.2073)	3.3079 (0.3424)	3.5524 (0.2024)	3.4748 (0.3735)
$\beta_0^{'}$	4.1952 (0.1678)	4.0883 (0.2727)	4.2197 (0.1637)	4.1320 (0.2051)
β_1	-0.0449 (0.0582)	-0.1211 (0.1320)	-0.0451 (0.0578)	-0.1394 (0.1452)
$\beta_1^{'}$	-0.1368 (0.0579)	-0.2338 (0.1188)	-0.1267 (0.0557)	-0.1088 (0.8626)
σ_θ^2	1.4280 (0.3515)	1.3090 (0.3329)	1.4588 (0.3496)	1.4200 (0.3483)
$\sigma_{\theta\theta'}$	0.9469 (0.2555)	0.8811 (0.2475)	1.0071 (0.2570)	0.9750 (0.2567)
$\sigma_{\theta'}^2$	0.8801 (0.2222)	0.8565 (0.2220)	0.9531 (0.2264)	0.9381 (0.2285)
σ_ν^2	0.0000 (.)	0.0000 (.)	0.0000 (.)	0.0000 (.)
$\sigma_{\nu\nu'}$	0.1433 (0.0248)	0.1432 (0.0253)	0.1380 (0.0247)	0.1395 (0.0251)
$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)	0.0000 (.)	0.0000 (.)
σ^2	0.4461 (0.0379)	0.4478 (0.0384)	0.4374 (0.0368)	0.4413 (0.0374)
σ_2^2	1.8741	1.7568	1.8962	1.8613
σ_{23}	1.3262	1.3043	1.3905	1.3794
σ_3^2	1.0902	1.0243	1.1451	1.1145
Inter-CCC	0.6230	0.5842	0.6309	0.6002

8. INFLUENCE OF COVARIATES ON CCC

Choudhary (2017, [31]) showed that covariates may affect either the mean or the variance of the estimated agreement. The covariates may include both subject-level, such as the gender of the subject or other baseline characteristics, and model-level covariates, such as the quality of measurement method itself, as method-level categorical covariates in the analysis.

8.1 Subject-level covariates adjusted for estimating CCC

In model 4.6 we proposed that we can include extra covariates in the model. In the GAIT study, baseline characteristics were collected. Those variables are subject-level. We try to apply gender, age, baseline pain level and several other variables in the model; only age and baseline pain level have significant effects on the JSW measurements. Therefore, the inter-CCC are evaluated after adjusting the two covariates.

TablesXVI-XVIII show the results of agreement on mean, variance and total CCC estimates between rater 1 and 2, 1 and 3, 2 and 3. We can see that the baseline of JSW estimates have changed after including two covariates; in fact, for the same age and same pain severity level patients, the mean difference between raters are the same as not adjusting the two covariates. However, the variance and covariance of JSW by different raters are decreasing after adjusting age and baseline pain effects. The level of covariance decreasing is even more than variance. That means the inter-CCC estimates are lower than the estimates not adjusting covariates.

Table XVI: Inter-CCC estimation between rater 1 and 2 after adjusting age and baseline pain

Category	Parameter	no covariates Est(Std)	adjust two co-variates Est(Std)
Fixed	β_0	3.6461 (0.2333)	7.1573 (1.2859)
	$\beta_0^{'}$	3.3876 (0.2342)	6.8987 (1.2860)
	β_1	-0.0952 (0.0723)	-0.0952 (0.0723)
	$\beta_1^{'}$	-0.0632 (0.0634)	-0.0632 (0.0634)
	β_{pain}		-0.1761 (0.4323)
	β_{age}		-0.0642 (0.0224)
Random variance	σ_θ^2	1.3252 (0.3818)	1.0903 (0.3313)
	$\sigma_{\theta\theta'}$	1.0882 (0.3511)	0.8502 (0.2987)
	$\sigma_{\theta'}^2$	1.3964 (0.3940)	1.1555 (0.3422)
	σ_ν^2	0.0706 (0.0616)	0.0706 (0.0616)
	$\sigma_{\nu\nu'}$	0.2142 (0.0374)	0.2142 (0.0374)
	$\sigma_{\nu'}^2$		
	σ^2	0.4659 (0.0434)	0.4659 (0.0434)
Agreement	$\bar{y_1} - \bar{y_2}$	0.2265	0.2266
	V_1	1.8617	1.6268
	V_2	1.8623	1.6214
	$Covariance$	1.3024	1.0644
	$Inter$	0.6900	0.6452

Table XVII: Inter-CCC estimation between rater 1 and 3 after adjusting age and baseline pain

Category	Parameter	no covariates Est(Std)	adjust two co-variates Est(Std)
Fixed	β_0	3.6461 (0.2291)	7.2741 (0.9257)
	$\beta_0^{'}$	4.1051 (0.1955)	7.7330 (0.9170)
	β_1	-0.0952 (0.0576)	-0.0952 (0.0576)
	$\beta_1^{'}$	-0.1579 (0.0532)	-0.1579 (0.0532)
	β_{pain}		-0.5655 (0.3076)
	β_{age}		-0.0712 (0.0159)
Random variance	σ_θ^2	1.3623 (0.3814)	1.0890 (0.3122)
	$\sigma_{\theta\theta'}$	1.1196 (0.3161)	0.8203 (0.2389)
	$\sigma_{\theta'}^2$	0.9719 (0.2746)	0.6466 (0.1924)
	σ_ν^2	0.1114 (0.0371)	0.1114 (0.0371)
	$\sigma_{\nu\nu'}$	0.1025 (0.0272)	0.1025 (0.0272)
	$\sigma_{\nu'}^2$	0.0831 (0.0319)	0.0831 (0.0319)
	σ^2	0.1620 (0.0174)	0.1620 (0.0174)
Agreement	$\bar{y_1} - \bar{y_2}$	-0.3963	-0.3962
	V_1	1.6357	1.3624
	V_2	1.2170	0.8917
	$Covariance$	1.2221	0.9228
	$Inter$	0.8121	0.7655

Table XVIII: Inter-CCC estimation between rater 2 and 3 after adjusting age and baseline pain

Category	Parameter	no covariates Est(Std)	adjust two co-variates Est(Std)
Fixed	β_0	3.3876 (0.2345)	6.8448 (1.0569)
	$\beta_0^{'}$	4.1051 (0.1989)	7.5623 (1.0479)
	β_1	-0.0632 (0.0645)	-0.0632 (0.0645)
	$\beta_1^{'}$	-0.1579 (0.0645)	-0.1579 (0.0645)
	β_{pain}		-0.5108 (0.3526)
	β_{age}		-0.0675 (0.0182)
Random variance	σ_θ^2	1.3935 (0.3940)	1.1840 (0.3538)
	$\sigma_{\theta\theta'}$	0.9504 (0.3002)	0.6880 (0.2384)
	$\sigma_{\theta'}^2$	0.9461 (0.2745)	0.6306 (0.1974)
	σ_ν^2	0.0000 (.)	0.0000 (.)
	$\sigma_{\nu\nu'}$	0.1589 (0.0282)	0.1589 (0.0282)
	$\sigma_{\nu'}^2$	0.0000 (.)	0.0000 (.)
	σ^2	0.4832 (0.0439)	0.4832 (0.0439)
Agreement	$\bar{y_1} - \bar{y_2}$	-0.6228	-0.6228
	V_1	1.8767	1.6672
	V_2	1.4293	1.1138
	$Covariance$	1.1093	0.8469
	$Inter$	0.6006	0.5345

9. STATISTICAL INFERENCE OF CONCORDANCE CORRELATION COEFFICIENT

To find the distribution of the CCC of whole data is very hard. Hypothesis testing of CCC is even more challenging (2013, [27]). Thus, we take a simple case of two raters with no covariates, as the standard value of CCC is unknown. The golden standard of CCC in our problem is not available in the literature. Instead of hypothesis testing, we construct the confidence interval of CCC(2017, [31]). This is because confidence intervals provide more information than p-values.

9.1 Generalized confidence interval estimated for CCC

To keep the case simple and avoid repeated measurements at different time point in this initial try, we use three raters' first readings at the two years X-ray. For each paired observations of two raters, they follow the bivariate normal distribution, denoted by:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix} \right),$$

let $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$.

In order to evaluate the agreement, it is enough to construct a confidence interval for the CCC, say θ, given by

$$CCC = \theta = \frac{2Cov(y_1, y_2)}{\sigma_{y_1}^2 + \sigma_{y_2}^2 + (\mu_{y_1} - \mu_{y_2})^2}.$$

Let $(\bar{Y_1}, \bar{Y_2})$ denote the sample mean vector and define

$$S = \sum_{i=1}^{n} \begin{pmatrix} Y_{1i} - \bar{Y_1} \\ Y_{2i} - \bar{Y_2} \end{pmatrix} \begin{pmatrix} Y_{1i} - \bar{Y_1} \\ Y_{2i} - \bar{Y_2} \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{12} & S_{22} \end{pmatrix}.$$

Then

$$\begin{pmatrix} \bar{Y_1} & \bar{Y_2} \end{pmatrix}' \sim N_2(\mu, (1/n)\Sigma),$$

and $S \sim W_2(\Sigma, n-1)$, the bivariate Wishart distribution with scale matrix Σ and degree of freedom n-1. We shall now construct a generalized pivot statistic for θ, whose percentile will be used to obtain a generalized confidence interval for θ. The generalized pivot statistic will be a function of the random variables $\begin{pmatrix} \bar{Y_1} & \bar{Y_2} \end{pmatrix}'$ and S, defined above, and the corresponding observed values, say $(\bar{y_1}, \bar{y_2})'$ and s, and is required to satisfy the conditions:

(i) Given $(\bar{y_1}, \bar{y_2})'$ and s, the distribution of the generalized pivot statistic is free of any unknown parameters;

(ii) The observed value of the generalized pivot statistic obtained by replacing $(\bar{Y_1}, \bar{Y_2})'$ and S with $(\bar{y_1}, \bar{y_2})'$ and s, respectively is simply θ, the parameter of interest;

We shall actually construct two generalized pivot statistics that satisfy the above properties and numerically investigate their performance. Since S $\sim W_2(\Sigma, n-1)$, we shall use the following properties of the Wishart distribution:

$$\begin{aligned} U_{22} &= \tfrac{S_{22}}{\sigma_{22}} \sim \chi^2_{n-1}, \\ U_{11.2} &= \tfrac{S_{11.2}}{\sigma_{11.2}} \sim \chi^2_{n-2}, \\ Z_1 &= (S_{12} - \tfrac{\sigma_{12}}{\sigma_{22}} S_{22}) / \sqrt{\sigma_{11.2} S_{22}} \sim N(0,1), \end{aligned}$$

where $S_{11.2} = S_{11} - S_{12}^2/S_{22}$, $\sigma_{11.2} = \sigma_{11} - \sigma_{12}^2/\sigma_{22}$ and χ_r^2 denote a central chi-square distribution with r degree of freedom. Here, the random variables $U_{22}, U_{11.2}$ and Z_1 are independently distributed. Denote the observed value of $S_{ij}(i,j=1,2)$ by s_{ij} and the observed value of $S_{11.2}$ by $s_{11.2}$. Define

$$\begin{aligned} R_{22} =& \frac{\sigma_{22}}{S_{22}} s_{22} = \frac{s_{22}}{U_{22}}, \\ R_{12} =& \frac{\sigma_{22}}{S_{22}} s_{12} - [\sqrt{s_{11.2} s_{22}} \frac{S_{12} - \frac{\sigma_{12}}{\sigma_{22}} S_{22}}{\sqrt{\sigma_{11.2} S_{22}}} \sqrt{\frac{\sigma_{11.2}}{S_{11.2}}} \frac{\sigma_{22}}{S_{22}}], \\ =& \frac{s_{12}}{U_{22}} - [\sqrt{s_{11.2} s_{22}} \frac{Z_1}{\sqrt{U_{11.2}} U_{22}}], \\ R_{11} =& \frac{\sigma_{11.2}}{S_{11.2}} s_{11.2} + \frac{R_{12}^2}{R_{22}} = \frac{s_{11.2}}{U_{11.2}} + \frac{R_{12}^2}{R_{22}}. \end{aligned} \tag{9.1}$$

The observed values of R_{11}, R_{12} and R_{22} are obtained by replacing the $S_{ij}'s$ with $s_{ij}'s$ and $S_{11.2}$ with $s_{11.2}$, and these observed values are easily seen to be σ_{11}, σ_{12} and σ_{22}, respectively. In other words, the matrix R given by $R = \begin{pmatrix} R_{11} & R_{12} \\ R_{12} & R_{22} \end{pmatrix}$ has an observed value. Note that R is positive definite. Now define

$$\begin{aligned} T_{11} =& (\bar{y_1} - \bar{y_2}) - \frac{(\bar{Y_1} - \bar{Y_2}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_{11} - 2\sigma_{12}^2 + \sigma_{22})/n}} \sqrt{(R_{11} - 2R_{12} + R_{22})/n} \\ =& (\bar{y_1} - \bar{y_2}) - \frac{Z_2}{\sqrt{n}} \sqrt{R_{11} - 2R_{12} + R_{22}}, \end{aligned} \tag{9.2}$$

where $Z_2 = \frac{(\bar{Y_1} - \bar{Y_2}) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_{11} - 2\sigma_{12}^2 + \sigma_{22})/n}} \sim N(0,1)$ and is independent of U_{22}, $U_{11.2}$ and Z_1. We note that the observed value of T_{11} is a generalized pivot statistic for $\mu_1 - \mu_2$. Similarly, R_{11}

is a generalized pivot statistic for σ_{11}; R_{22} is a generalized pivot statistic for σ_{22}; R_{12} is a generalized pivot statistic for σ_{12}. Thus, T, a generalized pivot statistic for θ is given by

$$\begin{aligned} T =& \frac{2[\frac{s_{12}}{U_{22}} - [\sqrt{s_{11.2}s_{22}}\frac{Z_1}{\sqrt{U_{11.2}U_{22}}}]]}{\frac{s_{11.2}}{U_{11.2}} + \frac{R_{12}^2}{R_{22}} + \frac{s_{22}}{U_{22}} + [(y_1 - y_2) - \frac{Z_2}{\sqrt{n}}\sqrt{R_{11} - 2R_{12} + R_{22}}]^2} \\ =& \frac{2R_{12}}{R_{11} + R_{22} + T_{11}^2} \end{aligned} \tag{9.3}$$

.

The percentile of T provide confidence limits for CCC. We note that T is a function of the independent random variables U_{22}, $U_{11.2}$, Z_1 and Z_2, and the observed quantities $(\bar{y_1} - \bar{y_2})'$ and the $s_{ij}'s$. The following algorithm is used to determine the percentiles of T:

(i) For a given sample of first X-ray readings by two raters $(y_{1i}, y_{2i})'$, $i = 1, \ldots, 29$, compute the sample mean $(\bar{y_1}, \bar{y_2})'$ and the sum of squares and the sum of products matrix

$$s = \sum_{i=1}^{n} \begin{pmatrix} y_{1i} - \bar{y_1} \\ y_{2i} - \bar{y_2} \end{pmatrix} \begin{pmatrix} y_{1i} - \bar{y_1} \\ y_{2i} - \bar{y_2} \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}.$$

(ii) For $j = 1, \ldots, 10000$, generate $U_{22} \sim \chi^2_{n-1}$, $U_{11.2} \sim \chi^2_{n-2}$, $Z_1 \sim N(0,1)$ and $Z_2 \sim N(0,1)$ and compute R_{11}, R_{12}, R_{22} and T_{11}.

(iii) Compute T.

(iv) End j loop. That means, repeat 10000 and get T with 10000 values.

(v) Order the T's to get the histogram, the $100(\alpha/2)$th and $100(1 - \alpha/2)$th percentile

of T. Provide a Monte Carlo estimate of the two sided $100(1-\alpha)$ confidence limits for CCC.

Figures 22-24 show the histogram of simulated CCCs between different raters. The mean of CCC between two manual raters are higher than between rater and computer.

TableXIX gives the simulated coverage probabilities (CP) of the three generalized confidence intervals for CCC between rater 1 and rater 2, rater 1 and computer, rater 2 and computer, based on the percentiles of T, for a 95 percent nominal level. The second column is the estimate CCC from samples between the different raters. The third column is the 95% generalized confidence interval by the method described. The last column is the coverage probabilities using 1000 simulations. From the results, we can see that all estimated CCC by sample are included in the 95% generalized confidence interval; CP is around 94% for the confidence interval after simulating 1000 times. The results also show that CCC between two raters' evaluations is better than each of them with computer at first readings of two year X-rays.

Table XIX: Inter-CCC for human and computer: Joint Space Width

Between raters	CCC estimate value	95%CI	CP
Rater 1 vs Rater 2	0.9388	0.8677, 0.9667	94.7%
Rater 1 vs Computer	0.8662	0.7308, 0.9235	93.9%
Rater 2 vs Computer	0.8712	0.7543, 0.9242	94.2%

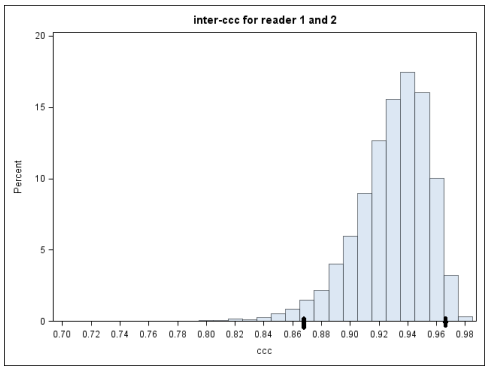

Figure 22: Histogram of 10000 simulated CCCs between raters 1 and 2

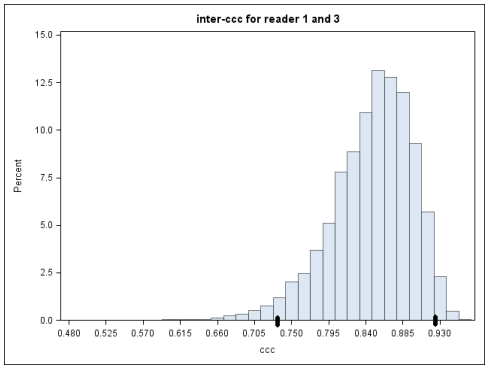

Figure 23: Histogram of 10000 simulated CCCs between rater 1 and computer

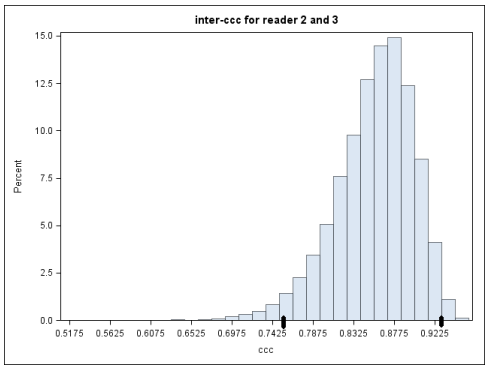

Figure 24: Histogram of 10000 simulated CCCs between rater 2 and computer

10. CONCLUSION

Assessment of agreement has been studied in many areas. Lin's concordance correlation coefficient has been commonly used in evaluating agreement for continuous variables. We follow Lin's idea of CCC, and apply it for longitudinal data with multiple replicates. A three level mixed-effect model is developed to evaluate the three level covariance between longitudinal data with replicate readings of the three raters, furthermore we estimate the variance covariance structure and sample means in order to estimate the CCC. Our approach is designed not only for complete data, but also for data with missing values. In addition, it can also be used for evaluating agreement after adjusting for covariates.

The expectation-maximization (EM) algorithm is used to estimate the parameters of the proposed model. The restricted maximum likelihood approach is used to take into account the degrees of freedom that are involved in estimating fixed effects in variance components estimation. One of the advantages of using EM is that the likelihood-based approach can handle unbalanced data, missing values, and covariates naturally. The reason to use REML rather than MML is that it's very common the sample size is limited in real data.

Simulation results show the excellent performance of our approach using three-level mixed-effect models for several situations: CCC close to 1, close to 0, and the estimated value close to the real data. In addition, the sample size we simulated is 30, not large; after running 10000 times, it demonstrates the accuracy and precision in estimating parameters.

The proposed model is applied to the complete data set sample we collected in the GAIT trial. We observe that the estimates of fixed variable by the three-level mixed

effects model are almost the same as by the two-level mixed effects model. However, the third level variance-covariance removes some variation from the model error term variance and increases the covariance between raters slightly. Thus, we get better CCC estimation by three-level models than the two-level models. In this situation, the three-level model captures different levels of variances and covariances, and the results are closer to the real data. Therefore, CCC based on three-level model is better.

We also analyze the data with missing values collected in the GAIT trial. Four methods for handling missing values are used including mixed model, model based imputation, multiple imputation, and pattern mixed model. Ten data samples with 10 and 15 subjects with missing data are combined with the complete data sets. The average inter-CCCs for first two methods are slightly larger than the estimates by complete data. It may be due to the covariance estimated by EM or the model imputation is larger than it is supposed to be. The latter two methods generate slightly smaller CCC estimation than the one by complete data. It may be caused by the uncertainty introduced. The assumption of the first three methods is missing at random and the last one assumes missing not at random. All methods gave us very similar results. That may suggest that the missing at random assumption is met in this case. In general, it will be good to use the pattern mixture model including missing pattern information to estimate CCC as the sensitivity analysis.

The three-level mixed-effect model can get the CCC estimation after adjusting the baseline characteristics. However, the estimation is smaller than the one without adjusting. This is because the covariance and variance are decreasing after adjusting the baseline characteristics variables, especially covariance which decreases more. Choudhary (2017, [31]) pointed out there are some types of variables called “measurement method” variables which may cause the mean difference; however, we have not included this type

of covariate in the models yet.

Statistical inference of CCC is performed by constructing the generalized confidence interval based on a simple case without covariates. This procedure simplifies the hypothesis testing and other inference procedures.

The model we proposed is based on a bivariate distribution. More work to include multiple raters will be done in the future. The other limitation is that construction of a generalized conference interval is based on one reading by each rater at last time point without covariates. The further statistical inference work on complicated datasets including multiple levels of covariance adjust by covariates will be taken into account later.

Taken together, our work attempts to estimate CCC by three-level mixed-effects model for multiple level covariance data. To illustrate our methods, we used a data set from the GAIT clinical trial.

CITED LITERATURE

1. Le Graverand, M.P., et al., Head-to-head comparison of the Lyon Schuss and fixed flexion radiographic techniques. Long-term reproducibility in normal knees and sensitivity to change in osteoarthritic knees. Ann Rheum Dis, 2008. 67(11): p. 1562-6.

2. Mazzuca, S.A., et al., Pitfalls in the accurate measurement of joint space narrowing in semiflexed, anteroposterior radiographic imaging of the knee. Arthritis Rheum, 2004. 50(8): p. 2508-15.

3. Dupuis, D.E., et al., Precision and accuracy of joint space width measurements of the medial compartment of the knee using standardized MTP semi-flexed radiographs. Osteoarthritis Cartilage, 2003. 11(10): p. 716-24.

4. Hunter, D.J., M.P. Le Graverand, and F. Eckstein, Radiologic markers of osteoarthritis progression. Curr Opin Rheumatol, 2009. 21(2): p. 110-7.

5. Ravaud, P., et al., Assessment of joint space width in patients with osteoarthritis of the knee: a comparison of 4 measuring instruments. J Rheumatol, 1996. 23(10): p. 1749-55.

6. Buckland-Wright, J.C., et al., Quantitative microfocal radiographic assessment of osteoarthritis of the knee from weight bearing tunnel and semiflexed standing views. J Rheumatol, 1994. 21(9): p. 1734-41.

7. Sawitzke, A.D., et al., The effect of glucosamine and/or chondroitin sulfate on the progression of knee osteoarthritis: a report from the glucosamine/chondroitin arthritis intervention trial. Arthritis Rheum, 2008. 58(10): p. 3183-91.

8. Clegg, D.O., et al., Glucosamine, chondroitin sulfate, and the two in combination for painful knee osteoarthritis. N Engl J Med, 2006. 354(8): p. 795-808.

9. Buckland-Wright, J.C., Quantitative radiography of osteoarthritis. Ann Rheum Dis, 1994. 53(4): p. 268-75.

10. Jacob Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, 1960, Vol. 20, p. 37-46.

11. Cohen, J., Weighed kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 1968, 70 (4): p. 213-220.

12. Fleiss, J.L., Measuring nominal scale agreement among many raters. Psychological Bulletin 1971, 76 (5): p. 378-382.

13. Bland, J. Martin and Altman, Douglas G., Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, Feb 8 1986: p. 307-310.

14. Lawrence I-Kuei Lin, A Concordance Correlation Coefficient to Evaluate Reproducibility, Biometrics 1989, 45, p. 255-268.

15. Lawrence I-Kuei Lin, Assay Validation Using the Concordance Correlation Coefficient, Biometrics 1992, 48, p. 599-604.

16. Lawrence I-Kuei Lin, Overview of Agreement Statistics for Medical Devices, Journal of Biopharmaceutical Statistics, 2007, 18:1, p. 126-144.

17. Josep L. Carrasco and L. Jover, Estimating the Generalized Concordance Correlation Coefficient through Variance Components, Biometrics 2003, 59, p. 849-858.

18. Lawrence Lin, A. S. Hedayat, Bikas Sinha, and Min Yang, Statistical Methods in Assessing Agreement: Models, Issues, and Tools, ASA Theory and Methods 2002, Vol.97, p.257-270.

19. Hedeker, R. D. Gibbons, and B. R. Flay. Random-effects regression models for clustered data: With an example from smoking prevention research. Journal of Consulting and Clinical Psychology 1994, 62: p. 757-765.

20. N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. Biometrics 1982, 38: p. 963-974.

21. M.J. Lindstrom and D.M. Bates. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. Journal of the American Statistical Association 1998, 83: p. 1014-1022.

22. Bikas Kumar Sinha, Erki Liski and Arto Luoma , Optimal designs in random coefficient linear regression models, Calcutta Statist. Assoc. Bull. 1996, 46, p. 211-229.

23. Bikas Kumar Sinha, Erkki Liski, Arto Luoma & N K Mandal, Optimal design for an inverse prediction problem under random coefficient, Jour. Indian Soc. Agricultural Statist. 49 1996/97, p. 277-288.

24. Bikas Kumar Sinha, Erkki Liski, Arto Luoma & N K Mandal, Optimal designs for prediction in random coefficient linear regression model, J. N. Srivastava Felicitation Volume. J. Combin. Inform. System Sci. 23, 1998.

25. L Lin, AS Hedayat, W Wu, Statistical Tools for Measuring Agreement, Springer Science and Business Media 2012.

26. Bikas K. Sinha, Pornpis Yimprayoon, Montip Tiensuwan, Cohen's Kappa Statistic: A Critical Appraisal and Some Modifications, Calcutta Statistical Association Bulletin, 2006, Volume: 58 issue: 3-4, p. 151-170.

27. G Dutta, BK Sinha, Some Further Aspects of Assessment of Agreement involving Bivariate Normal Responses, International Journal of Statistical Sciences, 2013f Vol. 13, 2013, p. 1-19.

28. Barnhart, H. X. and Williamson, J. M., Modelling concordance correlation via GEE to evaluate reproducibility. Biometrics 57, 2001, p. 931-940.

29. King TS, Chinchilli VM, A generalized concordance correlation coefficient for continuous and categorical data, Stat Med. 2001 Jul 30;20(14): p. 2131-47.

30. Ionut Betu and Thomas Mathew, Comparing the means and variances of a bivariate log-normal distribution, Statistics in Medicine, 2008; 27: p. 2684-2669

31. Choudhary, Pankaj K. and Nagaraja, Haikady N., Measuring Agreement Models, Methods, and Applications. Wiley, 2017.

32. Scott, W., Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19(3), p. 321-325.

33. Bland, J. Martin and Altman, Douglas G., Measuring agreement in method comparison studies, Stat Methods Med Res, 1999, Jun; 8(2): p. 135-60.

34. Searle, Shayle R., Casella, George, Variance Components, John Wiley Sons, Inc., 2006.

35. Little, Roderick J. A., and Rubin, Donald B., Statistical Analysis with Missing Data, Wiley, 2nd version, 2002.

36. Rubin, Donald B., Inference and missing data, Biometrika, Volume 63, Issue 3, 1976, p. 581-592.

37. Rubin DB. Multiple imputation for nonresponse in surveys. New York: John Wiley Sons, Inc.; 1987.

38. Tanner, Martin A., and Wong, Wing Hung, The Calculation of Posterior Distributions by Data Augmentation, Journal of the American Statistical Association, Vol. 82, No. 398, 1987, p. 528-540.

39. Gelfand, Alan E., and Smith, Adrian F. M., Sampling-Based Approaches to Calculating Marginal Densities, Journal of the American Statistical Association, Vol. 85, No. 410. 1990, p. 398-409.

40. Hedeker, D., and Gibbons, R. D., Application of random-effects pattern-mixture models for missing data in longitudinal studies, Psychological Methods, 2(1), 1997, p. 64-78.

41. Little, Roderick J. A., Modeling the Drop-Out Mechanism in Repeated-Measures Studies, Journal of the American Statistical Association, 90:431, 1995, p. 1112-1121.

42. Lin, Lawrence, S Hedayat, A, Wu, Wenting, A Unified Approach for Assessing Agreement for Continuous and Categorical Data, Journal of biopharmaceutical statistics 17, 2007, p. 629-52.

43. Barnhart, H. X., Haber, M. J. and Lin, L. I., An overview on assessing agreement with continuous measurement, Journal of Biopharmaceutical Statistics 17, 2007, p. 529-569.

44. Barnhart, H. X., Haber, M. J. and Song, J., Overall concordance correlation coefficient for evaluating agreement among multiple observers. Biometrics 58, 2002, p. 1020-1027.

45. Barnhart, H. X., Kosinski, A. S. and Haber, M. J., Assessing individual agreement. Journal of Biopharmaceutical Statistics 17, 2007b, p. 697-719.

46. Barnhart, H. X., Lokhnygina, Y., Kosinski, A. S. and Haber, M. J., Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement, Journal of Biopharmaceutical Statistics 17, 2007c, p. 721-738.

47. Barnhart, H. X., Song, J. and Haber, M. J., Assessing intra, inter and total agreement with replicated readings. Statistics in Medicine 24, 2005, p. 1371-1384.

48. Carrasco, J. L., Caceres, A., Escaramis, G. and Jover, L., Distinguishability and agreement with continuous data. Statistics in Medicine 33, 2014, p. 117-128.

49. Carrasco, J. L., Jover, L., King, T. S. and Chinchilli, V. M., Comparison of concordance correlation coefficient estimating approaches with skewed data. Journal of Biopharmaceutical Statistics 17, 2007, p. 673-684.

50. Carrasco, J. L., King, T. S. and Chinchilli, V. M., The concordance correlation coefficient for repeated measures estimated by variance components. Journal of Biopharmaceutical Statistics19, 2009, p. 90-105.

51. Carstensen, B., Simpson, J. and Gurrin, L. C., Statistical models for assessing agreement in method comparison studies with replicate measurements. The International Journal of Biostatistics 4, 2008, article 16.

52. Choudhary, P. K. and Nagaraja, H. N., Assessment of agreement using intersection-union principle. Biometrical Journal 47, 2005a, p. 674-681.

53. Choudhary, P. K. and Nagaraja, H. N., Selecting the instrument closest to a gold standard, Journal of Statistical Planning and Inference 129, 2005b, p. 229-237.

54. Choudhary, P. K. and Nagaraja, H. N., A two-stage procedure for selection and assessment of agreement of the best instrument with a gold standard. Sequential Analysis 24, 2005c, p. 237-257.

55. Choudhary, P. K. and Nagaraja, H. N., Tests for assessment of agreement using probability criteria. Journal of Statistical Planning and Inference 137, 2007, p. 279-290.

56. Choudhary, P. K. and Ng, H. K. T., A tolerance interval approach for assessment of agreement using regression models for mean and variance. Biometrics 62, 2006, p. 288-296.

57. Choudhary, P. K. and Yin, K., Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. Statistics in Biopharmaceutical Research 2, 2010, p. 122-132.

58. Choudhary, P. K., Sengupta, D. and Cassey, P., A general skew-t mixed model that allows different degrees of freedom for random effects and error distributions. Journal of Statistical Planning and Inference 147, 2014, p. 235-247.

59. Donner, A., Eliasziw, M. and Klar, N., Testing the homogeneity of kappa statistics. Biometrics 52, 1996, p. 176-183.

60. Donner, A., Shoukri, M. M., Klar, N. and Bartfay, E., Testing the equality of two dependent kappa statistics. Statistics in Medicine 19, 2000, p. 373-387.

61. Fay, M. P., Random marginal agreement coefficients: Rethinking the adjustment for chance when measuring agreement. Biostatistics 6, 2005, p. 171-180.

62. Feuerman, M. and Miller, A. R., Relationships between statistical measures of agreement: Sensitivity, specificity and kappa. Journal of Evaluation in Clinical Practice 14, 2008, p. 930-933.

63. Finney, D. J., A note on the history of regression. Journal of Applied Statistics 23, 1996, p. 555-557.

64. Fitzmaurice, G. M., Laird, N. M. and Ware, J. H., Applied Longitudinal Analysis, 2nd edn., 2011, John Wiley, Hoboken, NJ.

65. Gamer, M., Lemon, J., Fellows, I. and Singh, P., irr: Various Coefficients of Interrater Reliability and Agreement, 2012, R package version 0.84.

66. Guo, Y. and Manatunga, A. K., Nonparametric estimation of the concordance correlation coefficient under univariate censoring. Biometrics 83, 2007, p. 164-172.

67. Haber, M. J. and Barnhart, H. X., Coefficients of agreement for fixed observers, Statistical Methods in Medical Research 15, 2006, p. 255-271.

68. Hutson, A. D., A multi-rater nonparametric test of agreement and corresponding agreement plot, Computational Statistics and Data Analysis 54, 2010, p. 109-119.

69. King, T. S. and Chinchilli, V. M., Robust estimators of the concordance correlation coefficient, Journal of Biopharmaceutical Statistics 11, 2001b, p. 83-105.

70. King, T. S., Chinchilli, V. M. and Carrasco, J. L., A repeated measures concordance correlation coefficient, Statistics in Medicine 26, 2007a, p. 3095-3113.

71. Chinchilli, V. M., Martel, Juliann K., A Weighted Concordance Correlation Coefficient for Repeated Measurement Designs, Biometrics Vol. 52, No. 1, 1996, p. 341-353.

72. Zapf, Antonia, Castell, Stefanie, Morawietz, Lars, Measuring inter-rater reliability for nominal data - which coefficients and confidence intervals are appropriate? BMC Medical Research Methodology, 2016, p. 16-93.

APPENDIX

```
*****   08/29/2012       mixed model to get the corresponding variance                    ****;
*****   using 29 films three readers and two readings data - agree                       ****;
*****   get the distribution for each reader one by one                                  ****;
*************************************************************************************;
options nodate nonumber nocenter ;


libname in1 'K:\NIHGAIT\sasmast\saslib';
libname in2 'K:\NIHGAIT\sasmast\st\data';
libname in3 'K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\data';
data agree;
set in3.agree;
*if reader=1;
idfilm=id*100+time;
run;
proc means data=agree;
class idfilm;
var jsw;
run;
data subjectone;
set agree;
if id = 102154;
run;
proc sort data=subjectone;
```

```
by id reader time reading;
run;
proc means data=agree;
class idreader time;
var jsw;
output out=std_film;
run;
proc print data=std_film;
run;
data test;
set std_film;
if _stat_='STD' and jsw=0;
run;
* take two std with 0 at two time point away;
data agree_t;
set agree;
 if idreader=1042292 and time=0 then delete;
 if idreader=1071461 and time=1 then delete;
run;


*first try;
/*PROC MIXED METHOD=REML COVTEST DATA=agree;
  CLASS ID idreader idfilm;
  MODEL jsw =  time /SOLUTION;
  RANDOM INTERCEPT  /SUB=ID TYPE=UN G;
```

```
 RANDOM INTERCEPT  /SUB=idreader(id) TYPE=UN G;

 RANDOM INTERCEPT  /SUB=idreading(idreader) TYPE=UN G;

RUN; */*model estimation works, have to find the meaning of this model, three level or four level
model;

*second try;

/*PROC MIXED METHOD=REML COVTEST DATA=agree;

 CLASS ID idreader idreading;

 MODEL jsw =  time /SOLUTION;

 RANDOM INTERCEPT  /SUB=ID TYPE=UN G;

 RANDOM INTERCEPT  /SUB=idreader(id) TYPE=UN G;

 RANDOM INTERCEPT  /SUB=idreading(idreader(id)) TYPE=UN G;

RUN;*/* there's Syntax error;

*get the intra-ccc for first reader;

data agree1;

set agree;

if reader=1;

TIMEC=TIME;

run;

proc means data=agree1 mean std min max;

class timec reading;

var jsw;

run;

data agree1_plot;

set agree1;

if timec=1 then jsw=jsw+8;

if timec=2 then jsw=jsw+16;
```

```
run;
data agree1_1(keep=id idfilm jsw1_1);
set agree1_plot;
if reading=1;
jsw1_1=jsw;
run;
data agree1_2(keep=id idfilm jsw1_2);
set agree1_plot;
if reading=2;
jsw1_2=jsw;
run;
data agree1_p;
merge agree1_1 agree1_2;
by id idfilm;
run;

ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Reader one_all_09122018.rtf';
Title1 'Intra CCC for reader one';
PROC MIXED METHOD=REML COVTEST DATA=agree1;
  CLASS ID idfilm reading TIMEc;
  MODEL jsw =  READing READing* time/SOLUTION;
  RANDOM reading  /SUB=ID TYPE=UN G;

 ods output SolutionF=e1;*fixed effect estimates for estimating mean;
 ods output   CovParms=r1;
```

```
RUN;
proc transpose data=r1 out=r1transpose;
run;

data r1transpose (keep=v1 v12 v2);
set r1transpose;
v1=col1+col4;
v12=col2;
v2=col3+col4;
if _n_=1;
run;
proc transpose data=e1 out=e1transpose;
run;

data e1transpose(keep=dif0 dif1 dif2 dif);
set e1transpose;
dif0=col2;
dif1=dif0+(col4-col5);
dif2=dif0+(col4-col5)*2;
dif=(dif0+dif1+dif2)/3;
if _n_=2;
run;
data intra1;
merge r1transpose e1transpose;
inter=(2*v12)/(v1+v2+(dif*dif));
```

```
run;


ods rtf close;


PROC MIXED METHOD=REML COVTEST DATA=agree1;
  CLASS ID idfilm;
  MODEL jsw =  time /SOLUTION;
  RANDOM INTERCEPT TIME  /SUB=ID TYPE=UN G;
  RANDOM INTERCEPT TIME  /SUB=idfilm(id) TYPE=UN G;
RUN;*not positive deifined stop here;
/*PROC MIXED METHOD=REML COVTEST DATA=agree1;
  CLASS ID idfilm TIMEc;
  MODEL jsw =  time/SOLUTION;
  RANDOM INTERCEPT  /SUB=ID TYPE=UN G;
  RANDOM INTERCEPT  /SUB=idfilm(id) TYPE=UN G;
  REPEATED timec / SUB=ID TYPE=AR(1) R RCORR;
RUN;*/ *nonpositive definite estimated R matrix for id 102154;



* get the intra-ccc for the second reader;
data agree2;
set agree;
if reader=2;
timec=time;
run;
```

```
proc sort data=agree2;
by id time;
run;
proc means data=agree2;
var jsw;
by id time;
run;
data agree2_check;
set agree2;
if id in (102154, 102172, 102193, 102233, 104299, 104301, 104327)then delete;
run;

proc means data=agree2 mean std min max;
class timec reading;
var jsw;
run;
data agree2_plot;
set agree2;
if timec=1 then jsw=jsw+8;
if timec=2 then jsw=jsw+16;
run;
data agree2_1(keep=id idfilm jsw2_1);
set agree2_plot;
if reading=1;
jsw2_1=jsw;
```

```
run;
data agree2_2(keep=id idfilm jsw2_2);
set agree2_plot;
if reading=2;
jsw2_2=jsw;
run;
data agree2_p;
merge agree2_1 agree2_2;
by id idfilm;
run;



/*PROC MIXED METHOD=REML COVTEST DATA=agree2_check nobound;
  CLASS ID idfilm timec;
  MODEL jsw =  time /SOLUTION;
  RANDOM INTERCEPT  /SUB=ID TYPE=UN G;
 REPEATED timec / SUB=ID TYPE=AR(1) R RCORR;
 run;*/
ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Reader two_all09122018.rtf';
Title1 'Intra-CCC for Reader two';
PROC MIXED METHOD=REML COVTEST DATA=agree2;
  CLASS ID idfilm reading TIMEc;
  MODEL jsw =  READing READing* time/SOLUTION;
  RANDOM reading  /SUB=ID TYPE=UN G;
```

```
 ods output SolutionF=e2;*fixed effect estimates for estimating mean;
 ods output   CovParms=r2;
RUN;
proc transpose data=r2 out=r2transpose;
run;

data r2transpose (keep=v1 v12 v2);
set r2transpose;
v1=col1+col4;
v12=col2;
v2=col3+col4;
if _n_=1;
run;
proc transpose data=e2 out=e2transpose;
run;

data e2transpose(keep=dif0 dif1 dif2 dif);
set e2transpose;
dif0=col2;
dif1=dif0+(col4-col5);
dif2=dif0+(col4-col5)*2;
dif=(dif0+dif1+dif2)/3;
if _n_=2;
run;
data intra2;
```

```
merge r2transpose e2transpose;

intra=(2*v12)/(v1+v2+(dif*dif));

run;


ods rtf close;

* get the intra-ccc for the third reader;

data agree3;

set agree;

if reader=3;

run;

proc means data=agree3 mean std min max;

class timec reading;

var jsw;

run;

data agree3_plot;

set agree3;

if timec=1 then jsw=jsw+8;

if timec=2 then jsw=jsw+16;

run;

data agree3_1(keep=id idfilm jsw3_1);

set agree2_plot;

if reading=1;

jsw3_1=jsw;

run;

data agree3_2(keep=id idfilm jsw3_2);
```

```
set agree2_plot;
if reading=2;
jsw3_2=jsw;
run;
data agree3_p;
merge agree3_1 agree3_2;
by id idfilm;
run;
ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Reader three_all09122018.rtf';
Title1 'Intra-CCC for Reader three';
PROC MIXED METHOD=REML COVTEST DATA=agree3;
  CLASS ID idfilm reading TIMEc;
  MODEL jsw =  READing READing* time/SOLUTION;
  RANDOM reading  /SUB=ID TYPE=UN G;

 ods output SolutionF=e3;*fixed effect estimates for estimating mean;
 ods output   CovParms=r3;
RUN;
proc transpose data=r3 out=r3transpose;
run;

data r3transpose (keep=v1 v12 v2);
set r3transpose;
v1=col1+col4;
v12=col2;
```

```
v2=col3+col4;
if _n_=1;
run;
proc transpose data=e3 out=e3transpose;
run;

data e3transpose(keep=dif0 dif1 dif2 dif);
set e3transpose;
dif0=col2;
dif1=dif0+(col4-col5);
dif2=dif0+(col4-col5)*2;
dif=(dif0+dif1+dif2)/3;
if _n_=2;
run;
data intra3;
merge r1transpose e1transpose;
intra=(2*v12)/(v1+v2+(dif*dif));
run;

ods rtf close;
PROC MIXED METHOD=REML COVTEST DATA=agree3;
  CLASS ID idfilm;
  MODEL jsw =  time /SOLUTION;
  RANDOM INTERCEPT time /SUB=ID TYPE=UN G;
  RANDOM INTERCEPT  /SUB=idfilm(id) TYPE=UN G;
```

```
RUN;* log likelyhood test nosignificant difference with the previous random intercept model stop here;

DATA AGREETEST13;

SET AGREE;

IF READER=1 OR READER=3;

timec=time;

RUN;

ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Agreement between Reader
one and three_all.rtf';

Title1 'Inter-CCC for Reader one and three';


PROC MIXED METHOD=REML COVTEST DATA=AGREEtest13;

  CLASS ID READER idfilm;

  MODEL jsw = READER READER* time /SOLUTION noint;

  RANDOM  READER/SUB=ID TYPE=UN G;

 *RANDOM READER /SUB=idfilm(id) TYPE=UN G;

ods output SolutionF=e13;*fixed effect estimates for estimating mean;

  ods output   CovParms=r13;

RUN;* G matrix for random slope and three level model is not positive definite;

proc transpose data=r13 out=r13transpose;

run;



data r13transpose;

set r13transpose;

v1=col1+col4+col7;

v12=col2+col5;
```

```
v2=col3+col6+col7;
if _n_=1;
run;
proc transpose data=e13 out=e13;
run;

data e13transpose(keep=dif0 dif1 dif2);
set e13transpose;
dif0=col1-col2;
dif1=dif0+(col3-col4);
dif2=dif0+(col3-col4)*2;
if _n_=2;
run;
data inter;
merge r13transpose e13transpose;
inter=(2*v12/(v1+v2+(dif0*dif0))+2*v12/(v1+v2+(dif1*dif1))+2*v12/(v1+v2+(dif2*dif2)))/3;
run;


ods rtf close;
DATA AGREETEST12;
SET AGREE;
IF READER=1 OR READER=2;
timec=time;
if id in (102154, 102172, 102193, 102233, 104299, 104301, 104327)and reader=2 and reading=2 then
delete;
```

```
RUN;

ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Agreement between Reader one and two_all.rtf';

Title1 'Inter-CCC for Reader one and two';


PROC MIXED METHOD=REML COVTEST DATA=AGREEtest12;

  CLASS ID READER idfilm;

  MODEL jsw = READER READER* time /SOLUTION noint;

  RANDOM  READER/SUB=ID TYPE=UN G;

 *RANDOM READER /SUB=idfilm(id) TYPE=UN G;

RUN;


ods rtf close;

DATA AGREETEST23;

SET AGREE;

IF READER=2 OR READER=3;

timec=time;

if id in (102154, 102172, 102193, 102233, 104299, 104301, 104327)and reader=2 and reading=2 then delete;

RUN;

ods rtf file='K:\NIHGAIT\SASMAST\ST\SAS\agreement_longitudinal\doc\Agreement between Reader two and three_all.rtf';

Title1 'Inter-CCC for Reader two and three';

title2 'Without the obs with two reading at the same value';

PROC MIXED METHOD=REML COVTEST DATA=AGREEtest23;

  CLASS ID READER idfilm;

  MODEL jsw = READER READER* time /SOLUTION noint;
```

```
 RANDOM  READER/SUB=ID TYPE=UN G;

 *RANDOM READER /SUB=idfilm(id) TYPE=UN G;

RUN;


ods rtf close;
```

VITA

NAME:	Hairong Shi
EDUCATION:	BS, Chemistry, Nanjing University, Nanjing, China, 1994 MS, Biostatistics, University of Illinois at Chicago, Chicago, IL, 2004 PhD, Biostatistics, University of Illinois at Chicago, Chicago, IL, 2018
EXPERIENCE:	Biostatistician, CSPCC of Hines VA hospital, 2007-present Statistical programer, CSPCC of Hines VA hospital, 2004-2007
PROFESSIONAL MEMBERSHIP:	American Statistical Association
SELECTED PUBLICATIONS:	Allen D Sawitzke, Helen Shi, Martha F Finco, Dorothy D Dunlop, Crystal L Harris, Domenic J Reda, Daniel O Clegg, Clinical efficacy and safety of glucosamine, chondroitin sulphate, their combination, celecoxib or placebo taken to treat osteoarthritis of the knee: 2-year results from GAIT. Ann Rheum Dis (2010), doi: 10.1136