Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature:

_____ _____

Natalie R. Daya Date

Development of Agreement Measures for Studying Consistency of Menstrual Cycle Length and its Relationship to Fertility

By

Natalie R. Daya
Master of Public Health
Department of Biostatistics

Ying Guo, Ph.D.

Development of Agreement Measures for Studying Consistency of Menstrual Cycle Length and its Relationship to Fertility

By

Natalie R. Daya
Bachelor of Science
Loyola University Maryland
2010

Thesis Advisor: Ying Guo, Ph.D.

An Abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in the Department of Biostatistics
2012

Abstract

Development of Agreement Measures for Studying Consistency of Menstrual Cycle Length and its Relationship to Fertility

By Natalie R. Daya

The menstrual cycle acts as an overt indicator of reproductive health. Irregular menstruation with inconsistency in cycle length may affect fertility and risk of chronic diseases such as breast cancer. In this study we are interested in quantifying the consistency of menstrual cycle length in terms of an agreement measure and evaluating the association between the consistency of cycle length and a woman's reproductive health and characteristics such as diet and lifestyle. There are several challenges in conducting agreement analysis for our study: there is an unequal number of cycle length observations across subjects; there are censored observations in recorded cycle lengths; standard agreement method only provides group-level summary statistics and does not allow modeling agreement in terms of subject-specific covariates. In this thesis, we aim to develop appropriate statistical methods to address these issues. To accommodate censoring, we propose to impute cycle lengths for censored observations using the mean residual life estimated based on a parametric model. We first consider a group-level agreement method based on the intra-class correlation coefficient. We then propose a novel within-subject agreement method for replicated measurements by generating paired data based on the replications within each subject and then estimating an agreement measure based on the generated pairs. We consider subject-specific agreement index based on two different measures: Concordance Correlation Coefficient (CCC) and Total Deviation Index (TDI) and compare their performance. We then model the subject-specific agreement measures in terms of subject's covariates effects. Our results show that the new within-subject agreement method provides more biologically meaningful results than the group-level method. The TDI which is a more straightforward agreement measure provides better results than the CCC which is a scaled agreement measure. Consistency of a woman's menstrual cycle length is significantly associated with pregnancy status, stress level and caffeine consumption.

Development of Agreement Measures for Studying Consistency of Menstrual Cycle Length and its Relationship to Fertility

By

Natalie R. Daya
Bachelor of Science
Loyola University Maryland
2010

Thesis Advisor: Ying Guo, Ph.D.

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in the Department of Biostatistics
2012

Acknowledgements

I would like to take this opportunity to thank everyone who has supported me through the completion of this degree. First, I would like to thank Ying Guo, Ph.D. for her guidance and direction. She has been a great mentor throughout this process and I have learned a great deal from her. I would like to thank Amita Manatunga, Ph.D. for her helpful comments and suggestions. I would also like to thank Michele Marcus, Ph.D. from the Department of Epidemiology for the motivating example of the Mount Sinai Study of Women Office Workers (MSSWOW). Lastly, I would like to thank my family for their continued love and support.

Table of Contents

List of Tables

List of Figures

1. Introduction

The objectives of this study are twofold. The first objective is to investigate the epidemiology of women's reproductive health via their menstrual cycle lengths. We seek to determine how the consistency of menstrual cycle length is associated with a subject's reproductive health, which is reflected by the ability to achieve pregnancy. We also examine how a subject's diet and lifestyle affect the consistency of menstrual cycle length. The second objective is to develop an appropriate agreement method for studying the consistency of menstrual cycle length. We aim to develop agreement methods to determine an appropriate measure to quantify and model agreement among repeatedly measured menstrual cycle lengths within a woman. In the following, we discuss these two objectives in greater detail.

The menstrual cycle has an important bearing on the fertility and health of a woman. In particular, irregular menstrual cycles have been used as a useful and noninvasive indicator of reproductive health. Menstrual dysfunction may decrease fertility and increase future risk of various chronic diseases such as breast cancer, cardiovascular disease and diabetes. As an indicator of women's reproductive health, menstrual cycles have the added advantage that they are easy to observe, cost-effective and can be monitored by women themselves. Altered patterns of menstruation may facilitate earlier detection and possibly treatment of potential reproductive dysfunction and diseases (Guo, Manatunga et al. 2006). In this thesis, we are interested in studying the consistency of menstrual cycle lengths and its association with reproductive health, since irregular menstruation with large variability in the cycle length often indicates anovulatory cycles.

Menstrual cycle length and regularity are likely to be influenced by a variety of endogenous and exogenous factors such as reproductive history and behavioral characteristics. Diet has been shown to influence circulating levels of estrogen and its metabolism in premenopausal women (ages 5-18), and cigarette smoking has often been associated with abnormality in menstruation (ages 19-23) (Kato, Toniolo et al. 1999). We are interested in studying the association between the consistency of cycle lengths with subject's characteristics such as diet and lifestyle. The covariates we are interested in include the number of cigarettes smoked, the amount of beer, wine, and liquor consumed, the amount of coffee and cola consumed, the intensity of exercise performed and stress level.

In this study we quantify and evaluate the consistency of menstrual cycle length in terms of an agreement measure. Agreement methods are often applied in clinical and biological studies to evaluate the similarity of measurements produced by different raters or methods on the same subjects. They are therefore a natural choice of approach for quantifying consistency or similarity of replicated cycles within the same woman. There are, however, several major challenges for conducting agreement analysis in our study. First, most agreement studies measure agreement between paired measurements or a fixed number of measurements that come from the same participant. In our study, however, each subject has repeatedly measured samples and the number of replications varies across subjects, which is different from the standard settings of existing agreement studies. In this thesis, we aim to develop appropriate agreement methods to evaluate the consistency of menstrual cycle lengths within a woman. Another challenge of our study is that there are censored observations in the recorded cycle length. A censored

observation most commonly occurred due to the following reasons: the woman was in the middle of her cycle when the study ended, she got pregnant during the study, or she missed recording her cycle entirely. Since the majority of existing agreement methods are developed on completed observations, we need to develop a statistical approach to accommodate censored observations in our agreement measures. The third challenge in our study is that we are interested in modeling subject's covariate effects on the consistency of the cycle length to evaluate how the cycle length regularity is influenced by a variety of endogenous and exogenous factors such as reproductive history and behavioral characteristics. Existing agreement measures typically provide a summary statistic of agreement for the whole group of subjects and hence do not allow modeling of the agreement in terms of subject-specific covariates. We aim to develop a new modeling framework for agreement studies so that we can examine the effects of subject-specific covariates on the strength of agreement among repeated measurements within a subject.

To achieve the aforementioned objectives, we have conducted the following research. We first develop a parametric approach to accommodate censored observations by providing an imputation for the cycle length of a censored observation using the mean residual life estimated based on the normal distribution. To develop appropriate agreement methods for our replicated samples, we considered two approaches: 1) We first define and estimate a group-level agreement measure, the intra-class correlation (ICC), based on replicated samples by fitting a random-effect ANOVA model. We then compare the ICC between different pregnancy and age groups; 2) We propose a novel within-subject agreement method for subjects with replicated measurements.

Specifically, we generate paired data from replicated samples for each subject, and estimate two existing agreement measures, the Concordance Correlation Coefficient (CCC) and Total Deviation Index (TDI) based on these pairs. This provided us with subject-specific agreement measures. Compared to the group-level agreement measure, the proposed within-subject agreement method has two major advantages: it is not susceptible to between-subject variability and it allows us to model the strength of agreement among replicated samples in terms of subjects' clinical and demographical information. We then model subject-specific CCC and TDI in terms of subject's covariates and conduct model selection steps to identify covariates with significant association with the consistency of cycle lengths. To accommodate missing covariate values in the modeling, we either use the mean covariate values for an identical cycle length for that subject if available or for all cycle lengths for that individual if an identical one is not available. We also conduct a sensitivity analysis to evaluate the effects of imputing values for missing covariates on our results in modeling.

Based on our results, we found that our proposed subject-specific agreement method provides biologically more sensible findings than the standard group-level agreement method. When comparing the subject-specific CCC and TDI, we find that the TDI to be a more reliable measure of agreement among replicated samples within a subject. Those who got pregnant during the study had lower values of TDI compared to those who did not get pregnant, indicating pregnant women had more consistent cycle lengths. Furthermore, among those who got pregnant, TDI increased with age indicating generally less consistent cycle lengths as a woman ages. When modeling agreement in

terms of subject-specific covariates, stress level and caffeine consumption were shown to have a significant effect on the consistency of menstrual cycle length.

2. Motivation and Background

2.1 Motivation example

This study is based on the data obtained from the Mount Sinai Study of Women Office Workers (MSSWOW), a prospective cohort study to explore the effects of Video Display Terminal (VDT) use on rates of spontaneous abortion (Guo, Manatunga et al. 2006). The participants were originally recruited between 1991 and 1994 from fourteen different companies or governmental agencies in New York, New Jersey and Massachusetts. Invitations were sent to all women employees of reproductive age. 4,640 (63% of those contacted) women office workers completed a 23-page cross-sectional questionnaire during work hours concerning their work, lifestyle, health and demographic information.

Responses from the questionnaire determined eligibility for the prospective phase (Phase II) of MSSWOW. A woman was eligible if her responses indicated that she was currently trying to conceive a child, was planning to discontinue birth control in the next 12 months, or had sexual intercourse at least once in the past month without using any method of birth control. A woman was excluded if she had been attempting to conceive a child unsuccessfully for the past 12 months or longer, if she had a hysterectomy, or if her partner had a vasectomy. 895 of the women who completed the initial questionnaire met the eligibility criteria and 603 (67%) consented to take part in Phase II of the study.

An in-person interview was administered in Phase II of the study to closely examine the background of the eligible woman who consented to participate. Information on potential confounders of the relationship between VDT exposure and

spontaneous abortion was gathered. Medications taken in the last month, illicit drug use, beverage (especially caffeinated and alcoholic) consumption, smoking history, medical history, gynecological history, reproductive history, partner characteristics and demographics were collected for each participant.

After the interview process, if a woman was still considered eligible, she was asked to keep a daily diary of specific exposures on either a daily or monthly basis during her enrollment in the study. The exposures included the number of cigarettes smoked, the amount of beer, wine, and liquor consumed, the amount of coffee and cola consumed, the intensity of exercise performed and stress level. Periods of menstrual bleeding were also recorded. The information about menstrual bleeding was used to construct menstrual cycles for the duration of each woman's participation in the study, where menstrual cycle is defined as the interval from the first day of one bleeding episode up to and including the day before the next bleeding episode (WHO). A total of 524 women completed the study and kept daily diaries. They were asked to participate for one year or until clinical pregnancy.

2.2 Some existing agreement measures

In the following, we introduce several existing agreement measures for continuous outcomes in the literature.

Intra-class Correlation Coefficient (ICC)

The ICC, first proposed in 1925 by Fisher, has been a commonly used measure of agreement. ICC compares the variability of different ratings of the same subject with the total variation across all ratings and all subjects. Under various ANOVA models, ICC has different representations. In specific, the ICC is the ratio of between-sample variance

and the total variance (between- and within-sample variance) to measure precision under the model of equal marginal distributions.

Concordance Correlation Coefficient (CCC)

Lin (1989) introduced the CCC as an agreement index for continuous outcomes. The CCC is based on the scaled expected squared difference between two correlated variables. Let Y_1 and Y_2 denote a pair of continuous outcomes of the same individual by two different raters/methods. We assume that the joint distribution of Y_1 and Y_2 has finite second moments with means μ_1 and μ_2, respectively, and covariance matrix $\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$. Lin's (1989) CCC for Y_1 and Y_2 is defined as:

$$\rho_C = 1 - \frac{E[(Y_1 - Y_2)^2]}{E[(Y_1 - Y_2)^2 | Y_1 \; and \; Y_2 \; are \; independent]}$$

Lin (1989) showed that the CCC could be expressed as a function of the marginal mean, marginal variance and the covariance of the two correlated measurements.

CCC ranges from -1 to 1 with a value of 1 representing perfect agreement, a value of -1 representing perfect disagreement and a value of 0 representing no agreement beyond that expected by chance (Guo and Manatunga 2007). CCC evaluates the agreement between two readings from the same sample by measuring the variation from the 45˚ line through the origin (the concordance line). It contains measurements of accuracy and precision. Any departure from the 45˚ line would result in a CCC less than 1.

Total Deviation Index (TDI)

The Total Deviation Index (TDI) (Lin 2000) describes a boundary such that a majority of the difference between paired observations are within the boundary

(measurement unit and/or per cent) from their target values. TDI translates the mean squared differences (MSD) into an index that can be directly compared to a predetermined criterion. The Total Deviation Index (TDI) is an intuitively clear measurement of agreement which captures a large proportion of paired differences within a predetermined boundary, M (Lin, Hedayat et al. 2002). Let D be the difference of paired-measurement differences from two devices or observers. The TDI is a measure of the boundary M for a given coverage probability π (CP), such that,

$$\Pr(D^2 < M^2) = \pi$$

3. Descriptive statistics of the data set

Our 457 subjects of interest included 320 women who did not get pregnant and 137 women who successfully got pregnant during the study period of one year. The mean age of those who got pregnant was 30.66 (SD = 4.37), which was statistically significantly lower than the mean age of those who were not able to get pregnant (mean = 31.65, SD = 4.91) (p = .042). Our sample was broken down into age groups: 19-25, 26-30, 31-35 and 36-41. The age group classifications stratified by pregnancy status are shown in Table 1.

Table 1. Age group classifications of women by pregnancy status.

Age group	Pregnant	Not Pregnant	Total
19-25	19	40	59
26-30	45	94	139
31-35	50	101	151
36-41	23	85	108
Total	137	320	457

Descriptive statistics of cycle length

The range of the mean cycle length of those who did not get pregnant is significantly larger (12 to 132) than the mean cycle length range of those who got pregnant (24 to 84) (Appendix, Figure I). The range of the standard deviation of cycle length is also much wider in those who did not get pregnant (0 to 80) compared to those who did (0 to 16) (Appendix, Figure II). This pattern is also seen when comparing the distribution of the mean cycle length and the standard deviation of cycle length between age groups and pregnancy status (Appendix, Figures III-IV). Those in the youngest age group, 19 to 25 years of age, demonstrate the most striking contrast across pregnancy

status with those who got pregnant having a much narrower spread of mean cycle length and standard deviation of cycle length than those who did not get pregnant.

The mean cycle length among those who got pregnant was longest for those ages 19 to 25 and continually decreased with age while the standard deviation of cycle length among those who got pregnant was smallest for those ages 19 to 25 and continually increased with age. This pattern was not present in those who did not get pregnant. Furthermore, the average standard deviation of cycle length is consistent higher among those who did not get pregnant compared to those who did across all age groups (Table 2).

Table 2. Average mean and standard deviation of cycle length by pregnancy status and age group.

	Pregnant		Not pregnant	
Age group	**Mean cycle length**	**Standard deviation of cycle length**	**Mean cycle length**	**Standard deviation of cycle length**
19-25	32.94	2.44	31.81	4.92
26-30	29.88	2.89	31.84	5.05
31-35	29.36	3.31	32.75	6.33
36-41	29.01	3.43	28.65	3.66

4. Accommodation of censored observations

In this section, we present our method for accommodating censored observations in recorded cycle lengths. We let Y_{ij} represent the true cycle length for the i^{th} woman's j^{th} cycle, T_{ij} represent the observed cycle length, and $\delta_{ij} = 1$ if T_{ij} is censored and 0 if uncensored.

A censored observation indicates that the subject got pregnant during the study, was in the middle of her cycle when the study ended, missed recording her cycle entirely or failed to report the full length of the cycle for some unknown reason. We propose to

accommodate censoring using an imputation approach. If a subject had only one observation and the cycle length was censored or missing, the mean cycle length for that age group and pregnancy status was imputed. For an individual with multiple complete observations, we impute her censored observations using the mean residual life of the cycle length. We define Y as the true cycle length and T as the observed cycle length that may be censored. Assume $Y \sim N(\mu, \sigma^2)$, if the observed cycle length T is censored, the mean residual life of the true cycle length can be derived as follows;

$$E(Y \mid Y > T) = \int_T^{\infty} y \cdot P(Y = y | Y > T) dy$$

$$= \frac{1}{1-\Phi\left(\frac{T-\mu}{\sigma}\right)} \int_T^{\infty} y \cdot P(Y = y) dy, \text{ let } z = \frac{T-\mu}{\sigma}, \text{ then,}$$

$$= \frac{1}{1-\Phi(z)} [\mu \int_z^{\infty} \varphi(z) dz + \sigma \int_z^{\infty} z \cdot \varphi(z) dz]$$

$$= \mu - \frac{\sigma}{1-\Phi(z)} \cdot \int_z^{\infty} \varphi'(z) dz$$

$$= \mu + \sigma \cdot \frac{\varphi(z)}{1-\Phi(z)}$$

where $\varphi(\cdot)$ is the probability distribution function of N(0,1), $\Phi(\cdot)$ is the cumulative distribution function of N(0,1) and $\varphi'(\cdot)$ is the derivative of $\varphi(\cdot)$. The sample estimate of the mean residual life can then be obtained by plugging in the sample mean and sample standard deviation;

$$\bar{x} + s \cdot \frac{\varphi(z)}{1 - \Phi(z)}$$

where $\bar{x} = \frac{1}{K}\sum_{j=1}^{K} Y_{ij}$ and $s = \sqrt{\frac{1}{K}\sum_{j=1}^{K}(Y_{ij} - \bar{x})^2}$ are the mean and standard deviations, respectively, for the i^{th} individual and K is the total number of cycle lengths.

5. Group-level agreement method

We first conduct a group level agreement study, which is the typical type of analysis employed in existing agreement studies. We use the intra-class correlation coefficient, ICC, proposed by Fisher (1925) to assess intra-method agreement. To define the ICC, we consider the following random-effect ANOVA model for the repeated measured cycle lengths:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

The parameter μ represents the overall mean cycle length, the parameters α_i are the random individual effect and ε_{ij} are components of measurement error. It is assumed that $\alpha_i \sim N(0, \sigma_b^2)$ and $\varepsilon_{ij} \sim N(0, \sigma_w^2)$ where σ_b^2 is the between-subject variance, σ_w^2 is the within-subject variance and that all α_i and ε_{ij} are independent of each other. Under this model, the ICC can be expressed as:

$$\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}$$

ICC lies in the range of [0,1]. A value towards one represents the within-women variability is very small relative to the between-women variance and hence indicates higher agreement of repeated measurement within a woman. We obtained estimates for the within and between subject variance using the MIXED procedure in SAS. To study the association between consistency of cycle length and reproductive health, we evaluated the ICC for the pregnant group and the non-pregnant group. To account for the potential confounding effect of age on cycle length, we also stratified by the following age groups: 19-25, 26-30, 31-35 and 36-41.

Results

Our study population was comprised of 457 women, 137 of them got pregnant and 320 of them did not. The ICC of the pregnant and not pregnant groups were 0.469 and 0.510, respectively (Table 3). This signifies that the within-subject variance among those who got pregnant is smaller than the within-subject variance among those who did not get pregnant.

When stratified by age, the ICC of those who got pregnant and were between the ages of 19-25, 26-30, 31-35 and 36-31 were 0.898, 0.365, 0.381 and 0.298, respectively (Table 3). The ICC of those who got pregnant and were older than 25 was much lower than the ICC of those who got pregnant and were between the ages of 19-25.

The ICC of those who did not get pregnant and were between the ages of 19 to 25, 26-30, 31-35 and 36-31 were 0.349, 0.425, 0.859 and 0.441, respectively (Table 3). Those who did not get pregnant revealed no trend in ICC with age.

Table 3. ICC by pregnancy status and age group.

	Pregnant			**Not pregnant**		
Age group	$\boldsymbol{\sigma_b^2}$	$\boldsymbol{\sigma_w^2}$	**ICC**	$\boldsymbol{\sigma_b^2}$	$\boldsymbol{\sigma_w^2}$	**ICC**
19-25	118.81	13.44	0.898	22.33	41.70	0.349
26-30	9.14	15.94	0.365	32.10	43.44	0.425
31-35	13.46	21.84	0.381	92.82	15.30	0.859
36-41	8.21	19.37	0.298	18.71	23.75	0.441
Total	16.99	19.27	0.469	48.73	46.85	0.510

Discussion

From Table 3, the within-women variation of the cycle length is 19.27 for the pregnant group as compared to 46.85 for the non-pregnant group, indicating the variability of the within-women cycle length is much lower for pregnant women.

However, since the between-women variation of the cycle length is also much lower in the pregnant group, the estimated ICC is actually lower for the pregnant group, suggesting lower agreement among within-woman cycle lengths in the pregnant group. This contradicting result is due to the well-known susceptibility of ICC to the between-subject variability.

The higher ICC value of the pregnant group in all age groups (except those ages 31-35) compared to the non-pregnant group supports the idea that those who were able to get pregnant had more consistent menstrual cycle lengths over the course of the study. The dependency of the ICC on the between- subject variability, however, is a major drawback of this group-level agreement approach. If the between-subject variability varies greatly, as seen in Tables 3, the ICC values from the two groups (pregnant vs. not pregnant overall and stratified by age) can lead to misleading conclusions regarding the agreement among within-subject repeated measures.

6. Subject-specific agreement method

To address the aforementioned issue of group agreement method and also to allow modeling subjects' covariate effects on the strength of agreement, we propose a new subject-specific agreement method for measuring agreement among replicated measurements based on existing agreement measures.

6.1 Generation of paired-observations based on replicated measurements

Existing agreement measures are usually defined to measure the agreement between paired observations. To define a subject-specific agreement index, we first generated paired observations within an individual. The rationale was to restructure the replicated measurements into the paired-data framework and subsequently apply existing agreement measurements to the subject's paired-observations.

For each individual, all $\binom{K}{2}$ possible pairs of cycle lengths are represented by ($\tilde{Y}_{ik}^{(1)}$, $\tilde{Y}_{ik}^{(2)}$) such that $\tilde{Y}_{ik}^{(1)} \leq \tilde{Y}_{ik}^{(2)}$ where k = 1,…, $\binom{K}{2}$. We then apply two existing agreement measures on the generated paired-observations to evaluate the agreement of the cycle lengths within each subject. In the following, we introduce the two agreement measures and their estimation in our data.

6.2 Concordance correlation coefficient (CCC)

We adapted CCC to measure the consistency of repeated measures of menstrual cycle lengths recorded throughout the duration of the study. Subject specific level concordance correlation coefficients (CCC) were calculated to examine variance within a subject. To estimate the coefficient for an individual we adopted Lin's proposed estimator by substituting the sample counterparts of the marginal moments and covariance, i.e.,

$$\hat{\rho}_c = \frac{2S_{12}}{S_1^2 + S_2^2 + (\bar{P}_1 - \bar{P}_2)^2}$$

where $\bar{P}_j = \frac{1}{\binom{K}{2}} \sum_{k=1}^{\binom{K}{2}} \tilde{Y}_{ik}^{(j)}$, $S_j^2 = \frac{1}{\binom{K}{2}} \sum_{k=1}^{\binom{K}{2}} (\tilde{Y}_{ik}^{(j)} - \bar{P}_j)^2$ and

$S_{12} = \frac{1}{\binom{K}{2}} \sum_{k=1}^{\binom{K}{2}} (\tilde{Y}_{ik}^{(1)} - \bar{P}_1)(\tilde{Y}_{ik}^{(2)} - \bar{P}_2)$, j = 1, 2.

$\hat{\rho}_c$ is a consistent estimator of CCC and has an asymptotic normal distribution. The normal approximation of $\hat{\rho}_c$ can be improved by using Fisher's Z-transformation (Lin 1989),

$$\hat{Z} = tanh^{-1}(\hat{\rho}_C) = \frac{1}{2} ln \frac{1 + \hat{\rho}_C}{1 - \hat{\rho}_C}$$

The CCC has an appealing interpretation as the product of an accuracy coefficient X_a and precision coefficient ρ. The precision coefficient measures the strength of association and the accuracy coefficient measures the agreement between two marginal distributions. ρ is the Pearson correlation coefficient and $X_a = 2/(\omega + \frac{1}{\omega} + \nu^2)$ with $\nu^2 = (\mu_1 - \mu_2)^2/\sigma_1\sigma_2$ representing the location shift and $\omega = \sigma_1/\sigma_2$ representing the scale shift.

Results

The mean Fisher's Z-transformation of the subject-specific CCC for those ages 19-25, 26-30, 31-35 and 36-41 who got pregnant were 0.424, 0.347, 0.358 and 0.359, respectively. From youngest to oldest age group among those who did not get pregnant the mean transformed CCC values were: 0.388, 0.403, 0.398 and 0.423. There was no observable trend in transformed CCC between subjects of varying ages within or between pregnancy statuses. Surprisingly, among those who did not get pregnant, the eldest age group (36-41 years old) had the highest mean transformed CCC (mean = .423, SD =

.094) which was equal to the highest mean transformed CCC among those who got pregnant (age group 19-25: mean = .424, SD = .078) (Table 4). The data for all age groups (except for those who got pregnant, ages 19-25) is slightly skewed to the left which indicates there are outliers on the low end. In other words, within each age subgroup there are women with extremely low transformed CCC signifying poor consistency in cycle length for these subjects.

Table 4. Fisher's Z-transformation of the CCC stratified by pregnancy status and age group.

	Pregnant					Not pregnant				
Age group	**Mean**	**SD**	**Q1**	**Median**	**Q3**	**Mean**	**SD**	**Q1**	**Median**	**Q3**
19-25	0.424	0.078	0.349	0.413	0.455	0.388	0.080	0.324	0.390	0.442
26-30	0.347	0.147	0.240	0.363	0.459	0.403	0.096	0.335	0.424	0.485
31-35	0.358	0.140	0.277	0.405	0.472	0.398	0.105	0.335	0.410	0.477
36-41	0.359	0.113	0.269	0.361	0.464	0.423	0.094	0.366	0.427	0.502

Discussion

Transformed CCC was greatest for those aged 19-25 who got pregnant, suggesting the highest-level of agreement among these individuals. This is expected and provides support to the idea that consistency of a woman's menstrual cycle length is indicative of her reproductive health. Moreover, fertility declines quite rapidly with age, which is why we would see the highest-level of agreement in the lowest age group.

Other than finding that the youngest age group among those who got pregnant had the highest mean transformed CCC, there is a lack of trends in transformed CCC within and between pregnancy statuses and age groups. According to extensive biological research on the consistency of the length of a woman's menstrual cycle, we would expect transformed CCC to decrease with age and be significantly higher among those who got pregnant compared to those who did not. Our findings provide empirical evidence for the

existing criticism that the CCC is susceptible to the between-observation variability (where an observation is a pair of cycle lengths for a given individual) and hence may not be an accurate measure of the agreement between the paired measurements within an observation. Due to the sensitivity of the CCC to between-observation variability, we concluded that it was not a good measure of agreement in this study.

6.3 Total Deviation Index (TDI)

The Total Deviation Index (TDI) is an intuitively clear measurement of agreement which captures a large proportion of paired differences within a predetermined boundary, M (Lin, Hedayat et al. 2002). In our study, the paired differences D were calculated by taking the difference between all possible combinations of cycle lengths ($\tilde{Y}_{ik}^{(1)}$, $\tilde{Y}_{ik}^{(2)}$) for a given subject. We set CP at 80% but other probabilities can be considered as well. The M value was then defined as:

$$\Pr(D^2 < M^2) = \pi$$

Here, the agreement of observations with their target values is measured by the mean squared deviation (MSD) (Lin, Hedayat et al. 2002). We adapted this statistic to measure the agreement of $\tilde{Y}_i^{(1)}$ and $\tilde{Y}_i^{(2)}$within a subject:

$$\text{MSD} = \in^2 = \text{E}(D^2) = \text{E}(\tilde{Y}_i^{(2)} - \tilde{Y}_i^{(1)})^2 \text{, where D} = \tilde{Y}_i^{(2)} - \tilde{Y}_i^{(1)}$$

We assume that the joint distribution of $\tilde{Y}_i^{(2)}$ and $\tilde{Y}_i^{(1)}$ has finite second moments with means $\mu^{(2)}$ and $\mu^{(1)}$, variances σ_2^2 and σ_1^2 and covariance σ_{12}. Therefore, MSD can be expressed as:

$$\in^2 = (\mu^{(2)} - \mu^{(1)})^2 + \sigma_2^2 + \sigma_1^2 - 2\sigma_{12} \text{ ,}$$

The MSD for each individual can be computed using the sample counterparts:

$$e^2 = (\overline{P}_2 - \overline{P}_1)^2 + S_2^2 + S_1^2 - 2S_{12}$$

$\in$ can then be estimated using $\tilde{e}$ (Lin 1989),

$$\tilde{e} = \sqrt{\frac{1}{\binom{K}{2}-1}\sum_{k=1}^{\binom{K}{2}}(\tilde{Y}_{ik}^{(2)} - \tilde{Y}_{ik}^{(1)})^2}$$

where $\binom{K}{2}$ is the number of paired observations ($\tilde{Y}_{ik}^{(1)}$, $\tilde{Y}_{ik}^{(2)}$) for the i[th] individual.

Assuming that the distribution of D is normal with mean $\mu_d = \mu^{(2)} - \mu^{(1)}$ and $\sigma_d^2 = \sigma_2^2 + \sigma_2^2 - 2\sigma_{12}$, the proportion of the population with |D| less than M, M>0, becomes $\pi = \Pr(D^2 < M^2) = X^2[M^2, 1, \frac{\mu_d^2}{\sigma_d^2}]$ where $X^2(\cdot)$ is the cumulative noncentral chi-squared distribution with 1 degree of freedom and noncentrality parameter $\frac{\mu_d^2}{\sigma_d^2}$. This noncentrality parameter is the relative bias squared. The TDI for measuring the boundary M is defined as:

$$TDI_\pi = \sqrt{(X^{2(-1)}[\pi, 1, \frac{\mu_d^2}{\sigma_d^2}])}$$

where $X^{2(-1)}(\cdot)$ is the inverse function of $X^2(\cdot)$. Inference based on estimate of this TDI is intractable. According to Chebyshev's inequality, this probability has a lower bound of:

$$\Pr(D^2 < M^2) > 1 - (\in^2/M^2)$$

The lower bound of the M^2 value is therefore proportional to $\in^2$, the MSD. Lin (2000) suggested using the TDI to approximate the value of M that yields $P(D^2 < M^2) = \pi$, therefore:

$$TDI_\pi = M_\pi = \Phi^{-1}(1 - \frac{1-\pi}{2})\,|\in| \qquad (2)$$

where $\Phi^{-1}(\cdot)$ is the inverse cumulative normal distribution and $|\cdot|$ is the absolute value. TDI can be estimated by replacing $\in$ with $\tilde{e}$. The approximation is good under the following conditions which the data met, $\pi = .80$ and $\frac{\mu_d^2}{\sigma_d^2} \leq 8$.

TDI as defined in (2) is proportional to the square root of the MSD. TDI was calculated to describe a boundary such that a majority, 80%, of the differences in paired menstrual cycle lengths was within the boundary. In other words, it is a probability interval (Lin, Hedayat et al. 2002). We then examined and compared the total deviation index within and between pregnancy groups stratified by age.

Results

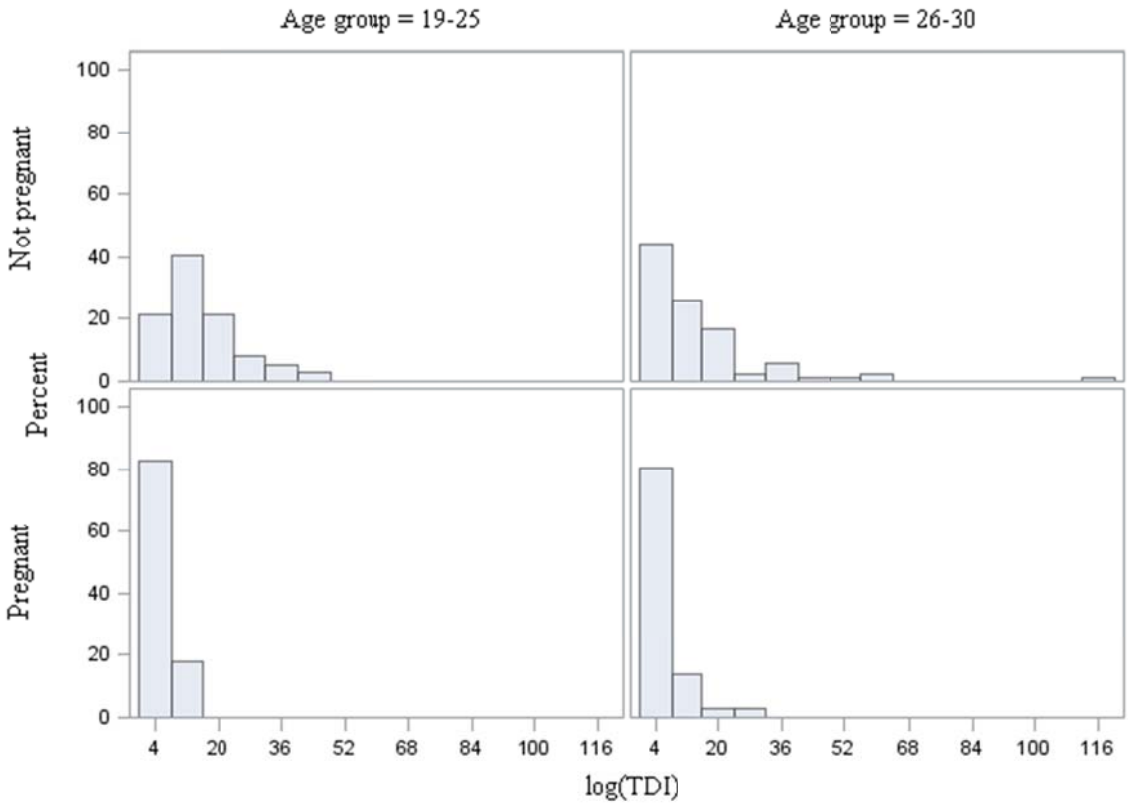
The average of the subject-specific TDI for those who got pregnant ages 19-25, 26-30, 31-35 and 36-41 were 4.87 (SD = 3.96), 5.85 (SD = 4.85), 6.50 (SD = 6.14) and 6.85 (SD = 7.68), respectively (Table 5). The TDI was smallest for those aged 19-25 who got pregnant. Among those who got pregnant, average TDI consistently increased as age group increased. The average of the subject-specific TDI for those who did not get pregnant ages 19-25, 26-30, 31-35 and 36-41 were 15.37 (SD = 10.17), 14.43 (SD = 16.31), 16.53 (SD = 17.69) and 12.35 (SD = 9.17), respectively. Among those who did not get pregnant, the eldest age group (36-41 years old) had the smallest average TDI (mean = 12.35, SD = 9.17) (Figure 1).

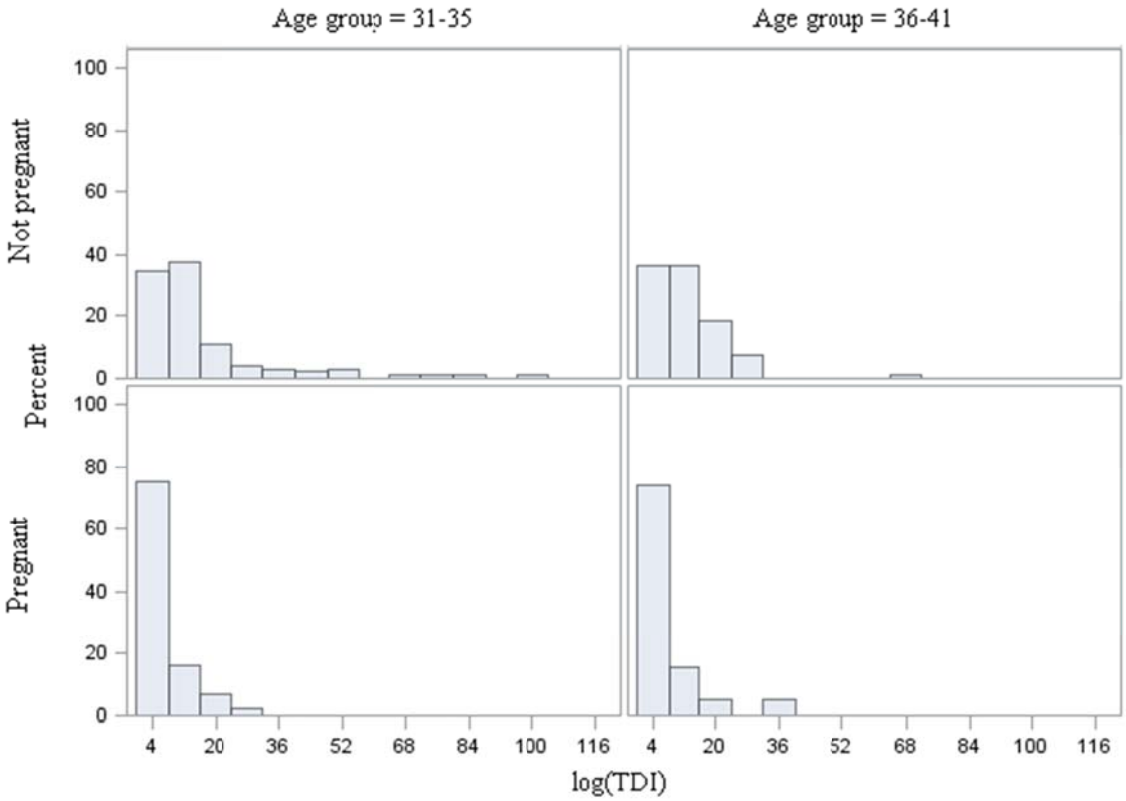
Table 5. Logarithm of TDI stratified by pregnancy status and age group.

	Pregnant					**Not pregnant**				
Age group	**Mean**	**SD**	**Q1**	**Median**	**Q3**	**Mean**	**SD**	**Q1**	**Median**	**Q3**
19-25	4.87	3.96	2.72	3.41	6.51	15.37	10.17	8.80	12.82	21.02
26-30	5.85	4.85	2.91	4.41	7.19	14.43	16.31	4.91	9.80	17.21
31-35	6.50	6.14	3.06	4.52	7.87	16.53	17.69	6.28	11.37	18.08
36-41	6.85	7.68	2.22	3.74	8.30	12.35	9.17	6.39	10.86	16.47

For all age groups, the TDI is consistently lower for the pregnant group as compared to non-pregnant group, indicating the within-women cycle lengths is more consistent for pregnant women after adjusting for age. The distribution of TDI is right skewed (Figure 1). In other words, within each age subgroup there are women with extremely high TDI values, signifying poor consistency in cycle length for these subjects. The right skewness is more obvious for the non-pregnant group as compared to the pregnant group.

Figure 1. Distribution of the logarithm of $TDI_{.8}$ by pregnancy status and age group.





Discussion

The TDI has a straightforward interpretation since it results in the same measurement scale as that of the variable considered for agreement purposes. The TDI sets a boundary such that 80% of the absolute differences in paired cycle lengths fall within this boundary. Therefore, among those participants aged 19-25 who got pregnant, 80% of the cycle length observations were within a 4.87 change in days from their paired cycle length observation. This is clinically sound, as we would expect those in the youngest age group who were able to get pregnant to have the most consistent menstrual cycle length. This is an indication of good reproductive health. Furthermore, the results in Table 5 are in agreement with the distributions of the summary statistics of the data presented in Chapter 3 and also with the results in the Appendix showing that those who

got pregnant have a narrower spread of mean cycle length and standard deviation of cycle length.

An appealing characteristic of the TDI is that it does not depend on the data range and thus avoids the inconvenience of not taking into account potential covariates that explain between-subject variation. The TDI will, however, depend on covariates explaining within-subject variation (Escaramis, Ascaso et al. 2010).

7. Modeling

One major advantage of our proposed new subject-specific agreement method is that it provides a subject-specific agreement index so that we can model the effects of subject's covariates on the strength of agreement. In the following, we model the effects of a woman's behavioral and demographic covariates on the subject-specific agreement index of her cycle lengths.

7.1 Covariates

The following covariates were measured and included in our modeling: pregnancy status (0/1, no/yes), age, stress level, exercise level, number of cigarettes, caffeine consumed and alcohol consumed. Mean age was calculated for each subject. Stress level, exercise level, number of cigarettes, caffeine consumed and alcohol consumed were recorded on a daily basis. Therefore, the mean of these covariates were calculated for each subject. Stress level was measured as 1 = low, 2 = medium low, 3 = medium high, 4 = high. For our analyses, stress level was categorized into two levels: 1 = low (including low and medium low) and 0 = high (including medium high and high). Exercise level was measured as 0 = did not exercise enough to sweat, 1 = exercised enough to sweat. Number of cigarettes is a continuous variable which signified number of cigarettes smoked each day of a recorded cycle. Caffeine consumed is a continuous variable which combined number of cups of coffee and cups of cola drunk each day of a recorded cycle. Alcohol consumed is a continuous variable which combined number of beers and glasses of wine drunk each day of a recorded cycle.

There are some missing values in the recorded covariates. The percentage of missing values for each covariate of interest was calculated. These percentages were

insignificant: stress = 1.75%, exercise = 1.09%, caffeine = 1.09%, alcohol = 1.09% and cigarette = 1.09%.

We adopt the following strategy to deal with missing covariates. If covariates were missing for an individual's cycle, the mean covariate values for an identical cycle length for that subject was used. If more than one identical cycle length exists for a given subject, the average of these covariates were used. If an identical cycle length did not exist within that individual then the average of covariates for all cycles was used. If subject had only one observation and the covariate set was missing, we consulted the original data set to determine if other cycles excluded from cleaned data set had values for covariates. If so, their averages were used. If not, covariates were left missing.

T-tests were employed to compare caffeine consumption, exercise intensity, alcohol consumption and number of cigarette smoked between those who got pregnant and those who did not. Chi-square tests were conducted to examine the distribution of stress level (low vs. high) among those who got pregnant and those who did not get pregnant. The only statistically significant difference between those who got pregnant and those who did not was in cigarette use (Table 6). These tests were conducted at the .05 significance level.

Table 6. Covariates of pregnant and not pregnant women.

Covariate	**Pregnant (n= 137)**		**Not pregnant (n= 320)**		
Stress	**n(%)**	**p-value**	**n(%)**	**p-value**	
Low (<2)	76 (55.47)	0.20	161 (50.31)	0.91	
High (≥2)	61 (44.53)		159 (49.69)		
	Mean (SD)	**Median (Range)**	**Mean (SD)**	**Median (Range)**	**p-value**
Caffeine	1.86 (1.62)	1.44 (9.01)	1.75 (1.39)	1.66 (10.86)	0.484
Exercise	0.26 (0.25)	0.20 (1.16)	0.23 (0.22)	0.16 (1.00)	0.152
Alcohol	0.45 (0.52)	0.34 (3.60)	0.52 (0.80)	0.25 (7.68)	0.302
Cigarette	0.46 (1.26)	0.00 (10.08)	0.82 (1.44)	0.00 (6.75)	0.012

Those who were pregnant smoked an average of 0.46 cigarettes (SD = 1.26) and those who were not pregnant smoked an average of 0.82 cigarettes (SD = 1.44) (p = 0.012) (Figure 2). A positive trend was found between caffeine and stress (r = 0.06, p = 0.187) (Figure 3).

Figure 2. Distribution of smoking by pregnancy status.

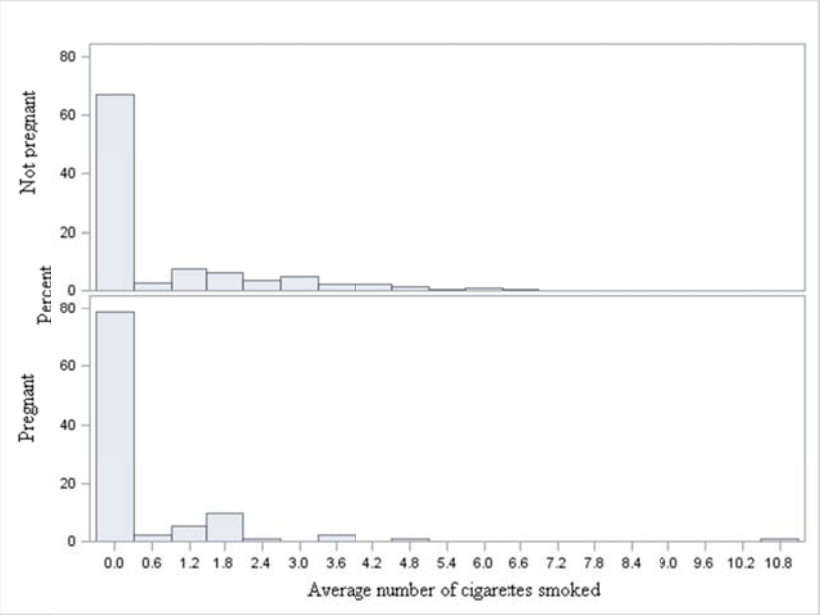
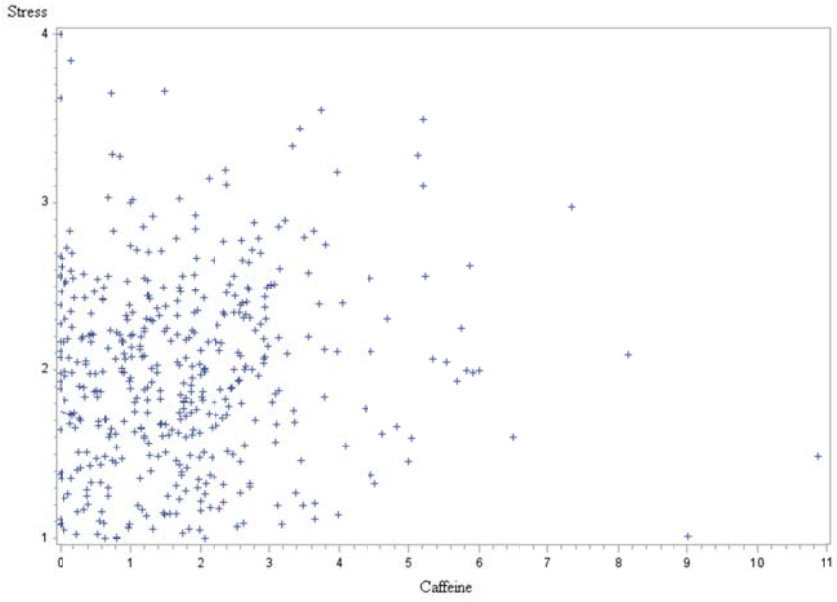


Figure 3. Scatterplot of caffeine and smoke.



7.2 Modeling covariates

In order to determine whether the consistency of cycle lengths was confounded by between subject variability, the CCC and TDI were modeled in terms of the covariates of interest: pregnancy status (X_1), age (X_2), stress level (X_3), exercise level (X_4), number of cigarettes (X_5), caffeine consumed (X_6) and alcohol consumed (X_7). Due to the positive skewness of the TDI estimates (Figure 1), the natural log transformation of the estimate was used (log(TDI)). The Fisher's Z-transformation of the CCC was used. Regression analyses were conducted to model the log(TDI) and transformed CCC using PROC REG in SAS. A regression analysis using stepwise selection was conducted with TDI as the dependent variable and the covariates as the independent variables. This allowed us to determine if changes in TDI was significantly associated with changes in any one of the covariates, while other covariates were held fixed. We fit a similar model using transformed CCC as the dependent variable.

Interaction terms were then tested one at a time using stepwise selection. Lower order covariates of the tested interaction term as well as the significant lower order covariates found in the previous step, were forced into these models. Lastly, the full model including significant covariates, interaction terms and their lower order covariates was fitted using PROC REG.

For the stepwise procedures the significance for entry of a covariate into the model was set at 0.15. The significance level for staying in the model was also set at 0.15.

Results

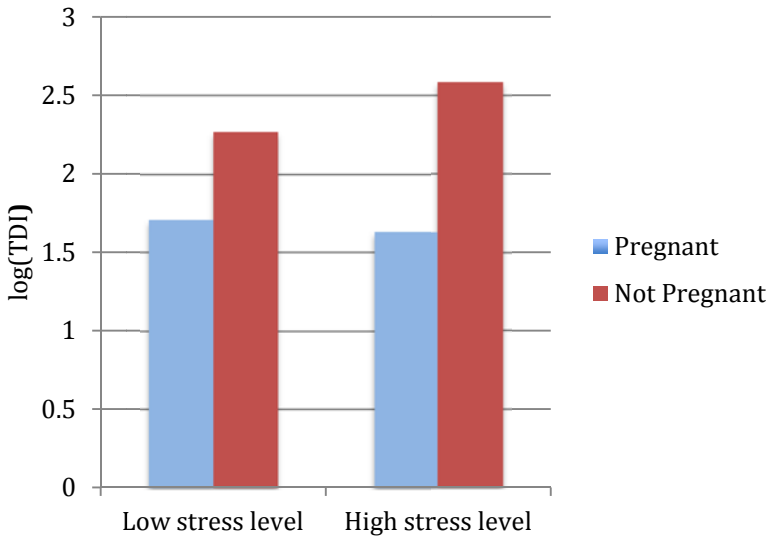
Pregnancy status, stress level, caffeine consumption and the interaction of pregnancy status and stress were all found to be significant predictors of the TDI (Table

7). The TDI on the log-scale is 0.32 higher among those who report a high level of stress (average stress level of 2 or higher), compared to those who report a low level of stress (p = 0.001). For each one-cup increase in caffeine intake, there was a 0.06 decrease in TDI on the log-scale (p = 0.0421). Among those who report a low stress level, TDI on the log-scale is 0.56 higher in the non-pregnant group as compared to the pregnant group. In high stress level, TDI on the log-scale is 0.96 higher in the non-pregnant group than in the pregnant group (Figure 4). Therefore, the TDI is higher for non-pregnant group as compared to pregnant group for both stress levels. The difference in TDI between the two pregnancy groups, however, is more significant for women with the high stress level (p = 0.028).

Table 7. Final model of the logarithm of TDI in terms of woman's covariate effects.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.59	0.08	31.30	<.000
Pregnant	1	-0.96	0.13	-7.24	<.000
Stress (low)	1	-0.32	0.09	-3.43	0.001
Caffeine	1	-0.06	0.03	-2.04	0.042
Pregnant*Stress (low)	1	0.40	0.18	2.20	0.028

Figure 4. The distribution of the logarithm of TDI by pregnancy status and stress level.



Low stress level < 2, high stress level ≥ 2

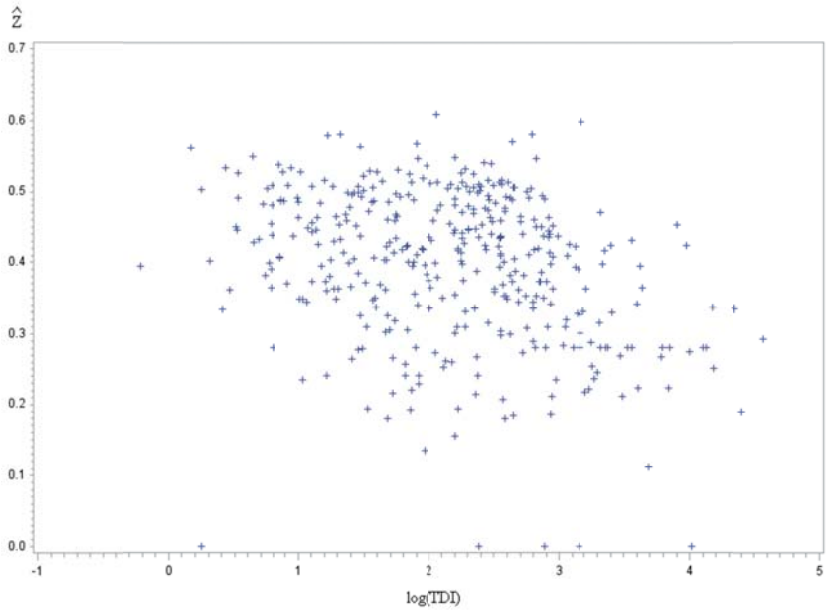
As with our findings in Chapter 5, the results of modeling transformed CCC do not correspond well with current scientific beliefs about the factors that influence the consistency of the menstrual cycle and the way in which they do so. For example, the results suggested that those who are pregnant have a smaller value of transformed CCC which indicates relatively less consistent menstrual cycle lengths. Therefore, the CCC does not adequately capture the agreement among within-women cycle lengths (Table 8).

Table 8. Final model of Fisher's Z-transformation of CCC in terms of women's covariate effects.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.38	0.04	9.21	<0.000
Pregnant	1	-0.04	0.01	-3.46	0.001
Age	1	-0.00	0.00	-0.58	0.560
Alcohol	1	-0.10	0.04	-2.51	0.012
Exercise	1	0.00	0.03	0.01	0.990
Cigarette	1	-0.01	0.01	-1.03	0.305
Age*alcohol	1	0.00	0.00	2.57	0.011
Cigarette*exercise	1	0.03	0.02	1.41	0.159

The transformed CCC, however, was statistically significantly negatively correlated with TDI which is what we would expect (r^2 = -0.26, p< .000). Lower agreement should be revealed by a relatively low CCC but high TDI (Figure 5).

Figure 5. Scatterplot of the logarithm of TDI and Fisher's Z-transformation of CCC.



7.3 Sensitivity Analysis

In order to determine if imputing values for missing covariates significantly affected our results, a regression was conducted using the data set prior to imputation. We concluded that imputations had no significant effect on our modeling results for log(TDI) or the transformed CCC with regards to the direction and magnitude of coefficients (Tables 9 and 10).

Table 9. Model of the logarithm of TDI in terms of women's covariate effects prior to imputation.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	2.51	0.08	656.85	1016.05	<.000
Pregnant	-0.74	0.09	42.79	66.19	<.000
Stress (low)	-0.20	0.08	3.98	6.15	0.014
Caffeine	-0.05	0.03	2.18	3.37	0.067

Table 10. Model of Fisher's Z-transformation of CCC in terms of women's covariate effects prior to imputation.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.36	0.01	13.08	1144.74	<.000
Pregnant	-0.04	0.01	0.13	11.31	0.001

8. Conclusions

8.1 Summary

In this agreement study, we measure the agreement among replicated samples where the number of replications varies across subjects. This is a non-standard setting in agreement studies. Other challenges in this project include accommodating censored observations and dealing with missing covariates values.

We present two types of agreement analysis for replicated sample. We first consider the group-level agreement analysis using the intra-class correlation coefficient, ICC. The estimate of the ICC is obtained through the between-subject and within-subject variability based on a repeated measures ANOVA model. We then propose a new subject-specific agreement method which provides a subject-specific agreement index and allows modeling subject's covariate effects on the strength of agreement. Our results show that the proposed new subject-specific agreement method provides more biologically sound results as compared to the commonly used group-level agreement method. Furthermore, compared to existing agreement method, the subject-specific agreement measure provides the important advantage of allowing us to assess the effect of subject's relevant characteristics on the agreement measure. This within-subject agreement method provides a new statistical method to evaluate and model the consistency of replicated measurements within the same subject.

We have considered and compared different agreement indices in our study. Scaled indices, such as the ICC and CCC, have been the most widely used agreement assays. These scaled indices, however, are shown to be sensitive to between-subject variability and may not accurately reflect the agreement between replicated

measurements within a subject. This shortcoming has been mentioned in previous papers. Our biologically illogical results provide further empirical evidence that demonstrate the drawbacks of these scaled indices.

Under the normal or log-normal distribution, the agreement measurements TDI and CCC essentially measure the same information but from different perspectives. The advantage of the TDI is that it does not depend on between-subject variability and therefore avoids the need to take into account potential covariates that explain between-subject variability. The TDI provided us with an unscaled procedure that describes a boundary such that a majority percent of the differences in paired measurements are within the boundary. Another advantage of the TDI is its straightforward interpretation. The TDI results in the same measurement scale as that of the variable considered for agreement purposes. The TDI also offers better power for inference than the CCC. Our results provide empirical support for the TDI as a more accurate measure of agreement among replicated samples within a subject.

Our scientific focus in this paper was to examine and compare menstrual cycle consistency among those who got pregnant and those who did not get pregnant and to determine the change in consistency with respect to subject-specific covariates. We found a woman's ability to get pregnant, her stress level and caffeine consumption were all significant factors influencing the consistency of her menstrual cycle length as measured by the TDI.

A limitation of the study is that the covariates we used in the modeling are cycle specific which is the average of the covariate within a cycle. A covariate value for a given day or cycle may be unusually high or low for a variety of reasons (e.g. death in the

family that causes an abnormally high level of stress for an individual), which would cause the average covariate value to be misrepresentative for that individual. Another limitation is that women who participated in diary keeping may not be representative of the larger population. They may be a subpopulation with more regular cycles.

8.2 Future research

While our study explored the possible covariates affecting the consistency of menstrual cycle length, which acts as an overt indicator of underlying reproductive health, further research needs to be conducted on the topic. We hope that the statistical method derived in this paper will enable researchers to make better use of menstrual cycle data as indicators of underlying biological function while considering demographical and clinical information that might influence a woman's menstrual cycle. The statistical methods we present in this thesis will be helpful to measure agreement among replicated samples where the number of replications can potentially vary across subjects and also to model subject's covariate effects on agreement.

References

Escaramis, G., C. Ascaso, et al. (2010). "The total deviation index estimated by tolerance intervals to evaluate the concordance of measurement devices." BMC medical research methodology **10**: 31.

Fisher, R. A. (1925). Statistical methods for research workers. Edinburgh, London,, Oliver and Boyd.

Guo, Y. and A. K. Manatunga (2007). "Nonparametric estimation of the concordance correlation coefficient under univariate censoring." Biometrics **63**(1): 164-172.

Guo, Y., A. K. Manatunga, et al. (2006). "Modeling menstrual cycle length using a mixture distribution." Biostatistics **7**(1): 100-114.

Kato, I., P. Toniolo, et al. (1999). "Epidemiologic correlates with menstrual cycle length in middle aged women." European journal of epidemiology **15**(9): 809-814.

Lin, L., A. Hedayat, et al. (2002). "Statistical Methods in Assessing Agreement: Models, Issues, and Tools." Journal of the American Statistical Association **97**: 257-270.

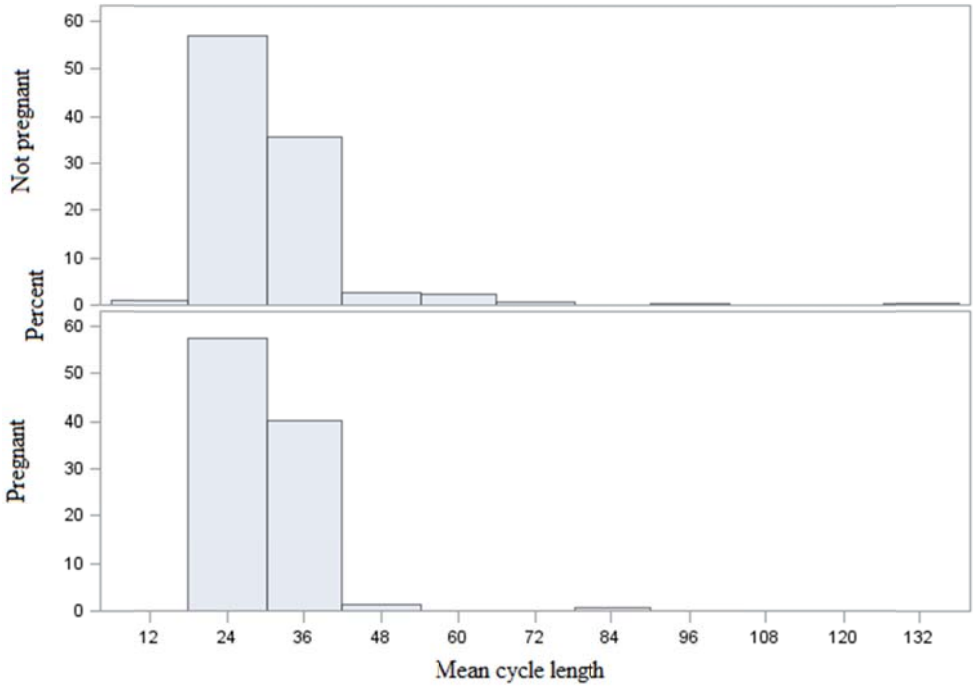
Lin, L. I. (1989). "A concordance correlation coefficient to evaluate reproducibility." Biometrics **45**(1): 255-268.

Lin, L. I. (2000). "Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence." Statistics in medicine **19**(2): 255-270.

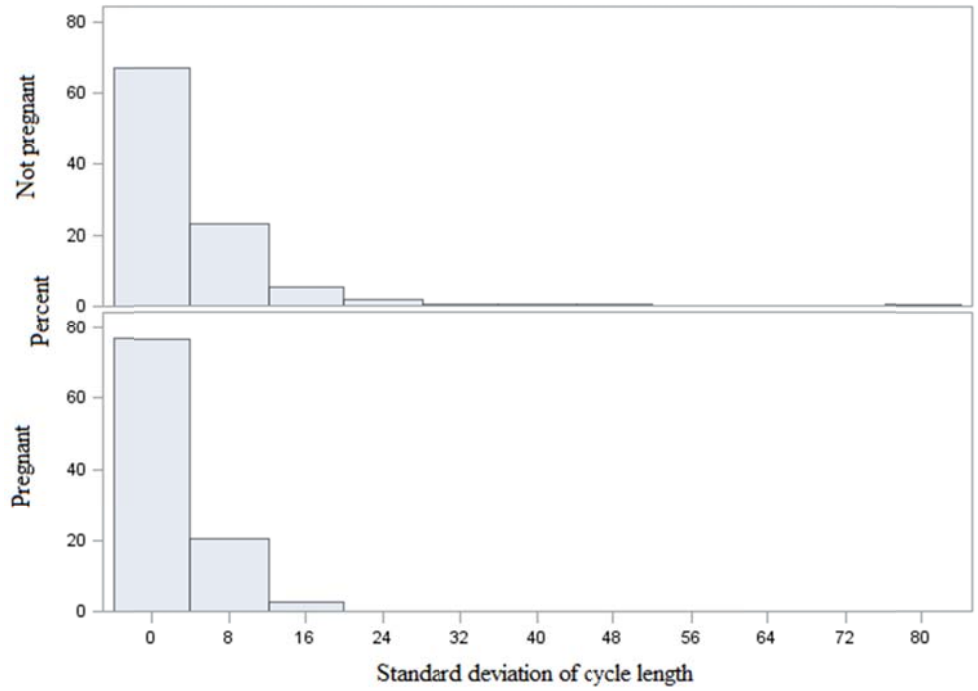
Verbeke, G. and G. Molenberghs (2000). Linear mixed models for longitudinal data. New York, Springer.

Appendix

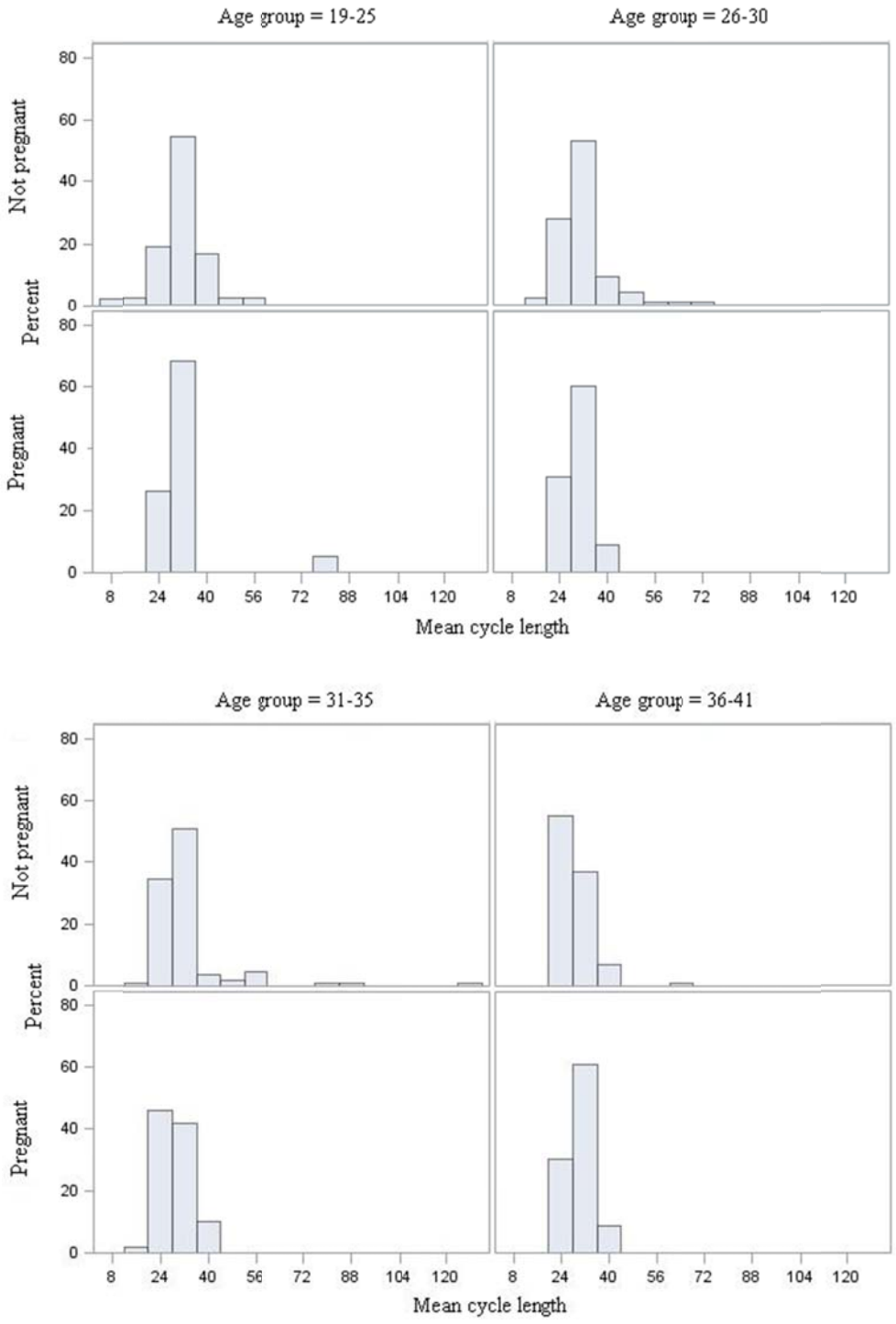
I. Distribution of the mean cycle length stratified by pregnancy status.



II. Distribution of the standard deviation of cycle length stratified by pregnancy status.



III. Distribution of the mean cycle length stratified by pregnancy status and age group.



IV. Distribution of the standard deviation of cycle length stratified by pregnancy status and age group.