Statistica Sinica **23** (2013), 1743-1759
doi:http://dx.doi.org/10.5705/ss.2012.036s

SEQUENTIAL TESTING OF MEASUREMENT ERRORS IN INTER-RATER RELIABILITY STUDIES

Mei Jin[1,3], Aiyi Liu[2], Zhen Chen[2] and Zhaohai Li[1]

[1]*George Washington University,*
[2]*National Institute of Child Health and Human Development and* [3]*Capital One*

Abstract: Inter-rater reliability is usually assessed by means of the intraclass correlation coefficient. Using two-way analysis of variance to model raters and subjects as random effects, we derive group sequential testing procedures for the design and analysis of reliability studies in which multiple raters evaluate multiple subjects. Compared with the conventional fixed sample procedures, the group sequential test has smaller average sample number. The performance of the proposed technique is examined using simulation studies and critical values are tabulated for a range of two-stage design parameters. The methods are exemplified using data from the Physician Reliability Study for diagnosis of endometriosis.

Key words and phrases: Interim analysis, inter-rater reliability, intraclass correlation coefficient, measurement errors, sample size and power, two-way ANOVA.

1. Introduction

Inter-rater reliability studies are conducted to investigate the reproducibility and level of agreement on assessments among different raters. The analysis of reliability is a common feature in research and practice since most measurements involve measurement errors, particularly those made by humans. Measurement error can seriously affect statistical analysis and interpretation; it therefore becomes important to assess the amount of such error by calculating a reliability index (Shrout and Fleiss (1979)).

Inter-rater reliability refers to the reproducibility of the raters when repeated at random on the same subject or specimen under the same condition by the same rater or different ones, either simultaneously or at several time points. A common measure of the reliability of measurements is the so-called intraclass correlation coefficient, with larger values indicating higher level of consistency.

There are various versions for intraclass correlation coefficient derived from different statistical models. Shrout and Fleiss (1979) proposed a set of guidelines for choosing the appropriate model in reliability studies. In actual inter-rater reliability studies, multiple raters usually evaluate multiple subjects and the two-way analysis of variance (ANOVA) model is considered to be the appropriate

analytical model. Moreover, when raters can be considered as random sample from the target population of raters, the two-way random effects model can be applied (Fleiss (1999); Zou and McDermott (1999); Cappelleri and Ting (2003); Tian and Cappelleri (2003)).

For fixed sample design, there are two main approaches available to determine the sample size. One approach is to find the size such that the power of statistical test is met (Donner and Eliasziw (1987); Walter, Eliasziw, and Donner (1998)) while the other is to assure the precision of estimation (Shoukri, Asyali, and Walter (2003); Bonett (2002); Saito et al. (2006)).

Sequential methods were introduced in response to demands for more efficient testing of anti-aircraft gunnery during World War II, culminating in Wald's development of the sequential probability ratio test which had an immediate impact on weapons testing (Wald (1947); Siegmund (1985); Lai (2004)). Gradually, in studies involving human subjects, specialized statistical methods were called for to balance the ethical and financial advantages of stopping a study early against the risk of an incorrect conclusion (Jennison and Turnbull (2000)). Armitage (1954, 1958) and Bross (1952, 1958) introduced the use of sequential methods in the medical field. However, the early development for group sequential methods came from Pocock (1977), O'Brien and Fleming (1979), and Lan and DeMets (1983).

Motivated by the idea of sequential testing that is widely used in clinical trials, it is natural to adopt and extend these sequential testing methods in the design and analysis of reliability studies to reduce the sample size and study cost. In reliability studies evaluating the measurement error by applying the one-way ANOVA model, the multistage group sequential designs were proposed. Under one-way ANOVA, the sums of squares in the estimation of the intraclass correlation coefficient possess independent increments, thus simplifying the calculation of stopping boundaries (Liu, Schisterman, and Wu (2006)).

In this paper, we develop multistage testing procedures using two-way ANOVA for hypotheses concerning the intraclass correlation coefficient in a inter-rater reliability study. In Section 2, we state the hypotheses of interest, introduce the structure and assumption of the two-way ANOVA models, and propose simulation designs for the one-stage problem for sample size and power calculation. In Section 3, we develop methods to determine critical values, sample size, and power using Lan and DeMets's (1983) error spending approach. Realizing that the between-rater sum of squares violates the independent increments assumption, we develop simulation techniques to effectively calculate the critical values. The performance of the proposed methods is examined in Section 4 using simulation studies, and critical values are tabulated for a range of two-stage design parameters. In Section 5, we exemplify the methods using data from the

Physician Reliability Study for diagnosis of endometriosis. Finally, summary and discussion are given in Section 6.

2. Hypotheses and Fixed Sample Design

The intraclass correlation coefficient ρ is commonly used to assess inter-rater reliability. The designs developed in this paper are based on testing a null hypothesis $H_0 : \rho \leq \rho_0$ that the true intraclass correlation coefficient is less than some uninteresting level ρ_0. We require that under the null hypothesis the probability of concluding that the rating is sufficiently promising be less than α, the level of significance of the test. We also require that under a specified alternative hypothesis $H_1 : \rho \geq \rho_1$, the probability of rejecting the rating should be less than β. If the null hypothesis is not rejected, then we may conclude that the raters are not well trained or that the variable to be measured possesses unacceptable measurement errors.

2.1. Two-way random effects ANOVA model

Consider a random sample of n subjects for which a continuous variable y is measured independently by d raters randomly selected from a population of raters. Three sources of variation usually occur from such a design: subjects, raters, and random errors. Higher inter-rater reliability is achieved if the variations from random errors and raters are relatively smaller than the variation from subjects.

Denote by y_{ij} the measurement made on the ith subject by the jth rater, $i = 1, \ldots, n$ and $j = 1, \ldots, d$. Then the two-way ANOVA model is

$$y_{ij} = \mu + s_i + r_j + \epsilon_{ij}, \tag{2.1}$$

where μ is the grand mean of all measurements. The subject effect s_i, the rater effect r_j, and the random error ϵ_{ij} are assumed to be independent and normally distributed with mean 0 and variances σ_s^2, σ_r^2 and σ_ϵ^2, measuring the magnitude of the variation from the three resources, respectively. Note that these assumptions lead to $\mathrm{Var}(y_{ij}) = \sigma_s^2 + \sigma_r^2 + \sigma_\epsilon^2$, $\mathrm{Cov}(y_{ij}, y_{i'j}) = \sigma_r^2$ if $i \neq i'$, $\mathrm{Cov}(y_{ij}, y_{ij'}) = \sigma_s^2$ if $j \neq j'$, and $\mathrm{Cov}(y_{ij}, y_{i'j'}) = 0$ if $i \neq i'$ and $j \neq j'$. The vector including measurements from all subjects $\mathbf{y}_{1\times nd}^T = (y_{11}, \cdots, y_{1d}, \cdots, y_{n1}, \cdots, y_{nd})$ is then distributed as $N_{nd}(\mathbf{u}, \Sigma)$,

$$\mathbf{u} = \mu \mathbf{1}_{nd},\ \Sigma = \sigma_\epsilon^2 \mathbf{I}_n \otimes \mathbf{I}_d + \sigma_s^2 \mathbf{I}_n \otimes \mathbf{J}_d + \sigma_r^2 \mathbf{J}_n \otimes \mathbf{I}_d, \tag{2.2}$$

where, throughout, the superscription T stands for the transpose of a matrix or vector, $\mathbf{1}_m$ is the vector of order m with all elements 1, $\mathbf{I}_n$ is the identity matrix or unit matrix of size n, and $\otimes$ is the Kronecker product of two matrices

(Rao (1973)), and $\mathbf{J}_{n\times m}$ is the $n \times m$ matrix with all elements 1, with a simpler notation $\mathbf{J}_n(=\mathbf{J}_{n\times n})$ being used instead if $m = n$.

2.2. Point estimation of intraclass correlation coefficient

The intraclass correlation coefficient in a two-way ANOVA model is given by

$$\rho = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_r^2 + \sigma_\epsilon^2}. \tag{2.3}$$

It provides a measure for the overall consistency among measurements from the same subject. The index reduces to the intraclass correlation coefficient in a one-way ANOVA model, if there is no variation among raters. Furthermore, it can be shown that ρ is the correlation coefficient between two measurements y_{ij_1} and y_{ij_2} on the same subject i by two different raters j_1 and j_2. Thus, larger values of ρ indicate higher coherence among measurements on the same subject by different raters. Perfect consistency among these measurements occurs with $\rho = 1$ when subjects are the only source of variation, in which case $\sigma_r = \sigma_\epsilon = 0$. A low reliability may indicate that the raters are not well trained or that the variable to be measured possesses unacceptable measurement errors.

Consider the sums of squares

$$\begin{aligned} SS_{subject} &= \sum_{i=1}^{n}\sum_{j=1}^{d}(\bar{Y_{i.}} - \bar{Y_{..}})^2, \\ SS_{rater} &= \sum_{i=1}^{n}\sum_{j=1}^{d}(\bar{Y_{.j}} - \bar{Y_{..}})^2, \\ SS_{error} &= \sum_{i=1}^{n}\sum_{j=1}^{d}(Y_{ij} - \bar{Y_{i.}} - \bar{Y_{.j}} + \bar{Y_{..}})^2, \end{aligned}$$

where

$$\bar{Y_{..}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{d} y_{ij}}{nd}, \quad \bar{Y_{i.}} = \frac{\sum_{j=1}^{d} y_{ij}}{d}, \quad \text{and} \quad \bar{Y_{.j}} = \frac{\sum_{i=1}^{n} y_{ij}}{n}.$$

Here (Rao (1973)) $SS_{subject}$, SS_{rater}, and SS_{error} are independent and can be written as

$$SS_{subject} \widetilde{=} (d\sigma_s^2 + \sigma_\epsilon^2)V_1, SS_{rater} \widetilde{=} (n\sigma_r^2 + \sigma_\epsilon^2)V_2, SS_{error} \widetilde{=} \sigma_\epsilon^2 V_3, \tag{2.4}$$

where V_1, V_2, and V_3 are independent chi-square variables with degrees of freedom $n-1$, $d-1$, and $(n-1)(d-1)$, respectively, and $\widetilde{=}$ means that its two sides have the same distribution.

Take $MS_{subject}$ as $SS_{subject}$ divided by the degrees of freedom of its corresponding χ^2-distribution, and similarly define MS_{rater} and MS_{error}. Then

$$E(MS_{subject}) = d\sigma_s^2 + \sigma_\epsilon^2,\ E(MS_{rater}) = n\sigma_r^2 + \sigma_\epsilon^2,\ E(MS_{error}) = \sigma_\epsilon^2,$$

yielding unbiased estimators of σ_s^2, σ_r^2 and σ_ϵ^2 as

$$\hat{\sigma_s^2} = \frac{(MS_{subject} - MS_{error})}{d}, \hat{\sigma_r^2} = \frac{(MS_{rater} - MS_{error})}{n}, \hat{\sigma_\epsilon^2} = MS_{error}. \quad (2.5)$$

An estimator of the intraclass correlation coefficient ρ can be obtained by substituting the variance estimators, so

$$\hat{\rho} = \frac{\hat{\sigma_s^2}}{\hat{\sigma_s^2} + \hat{\sigma_r^2} + \hat{\sigma_\epsilon^2}} = \frac{n(d-1)S - n}{n(d-1)S + d(n-1)R + (nd - n - d)}, \quad (2.6)$$

where

$$R = \frac{SS_{rater}}{SS_{error}} \simeq \frac{(n\omega + 1)V_2}{V_3} \sim \frac{n\omega + 1}{n-1} F_{d-1,(n-1)(d-1)}, \quad (2.7)$$

$$S = \frac{SS_{subject}}{SS_{error}} \simeq \frac{d\rho(\omega+1) + (1-\rho)}{1-\rho} \frac{V_1}{V_3}$$

$$\sim \frac{d\rho(\omega+1) + (1-\rho)}{(d-1)(1-\rho)} F_{n-1,(n-1)(d-1)}, \quad (2.8)$$

where $\omega = \sigma_r^2/\sigma_\epsilon^2$ and $F_{a,b}$ is the central F-distribution with degrees of freedom a and b. We refer to $\hat{\rho}$ as the sample intraclass correlation coefficient.

2.3. Calculation of operating characteristics

The two-way ANOVA model reduces to the popular one-way ANOVA model if $\sigma_r^2 = 0$, that is, if there is no variations among raters. In this case, observations from the same subject are treated as interchangeable repeated measurements, and observations from different subjects are independent. For the one-way ANOVA model, the sample intraclass correlation coefficient is a function of an F-statistic, and methods are directly available for computing the powers and sample sizes of the test based on the F-statistic; see, among others, Donner and Eliasziw (1987), Walter, Eliasziw, and Donner (1998), and Liu, Schisterman, and Wu (2006).

For the two-way ANOVA model observations from different subjects but the same rater are not independent, and methods for power and sample size calculations based on one-way ANOVA are not applicable because the sample intraclass correlation coefficient involves correlated F-statistics, S and R in (2.6), arising from the rater effects. Kraemer (1976) considered reliability in a two-way ANOVA model; however, the intraclass correlation coefficient there differed from the traditional one in that the rater variability was excluded.

For a pre-specified significance level α, the fixed sample size design rejects the null hypothesis $H_0 : \rho \leq \rho_0$ if the sample intraclass correlation coefficient exceeds some critical value $0 < c < 1$. The power function of the test

$$\Psi(\rho;\omega) = P_\rho\left(\hat{\rho} > c\right) \tag{2.9}$$

depends, in addition to ρ, on n, d, and the variance ratio ω. It can be shown (see Appendix) that, for given n, d, ω, Ψ is an increasing function in ρ, implying that we can control the type I error rate at the nominal level α by setting $\Psi = \alpha$ for $\rho = \rho_0$, and set the power to be $1 - \beta$ at ρ_1 for $\rho \geq \rho_1$.

To deal with the nuisance parameter ω, we can take the conventional approach of finding $0 < c < 1$ such that $\sup_\omega \Psi(\rho_0;\omega) = \alpha$. Similarly, given the power $1 - \beta$ at an alternative value $\rho = \rho_1$, the sample size n can be determined from $\inf_\omega \Psi(\rho_1;\omega) = 1 - \beta$.

Allowing the nuisance parameter ω to change freely in the entire range $[0, \infty]$ often results in a very conservative test for values of ω encountered in practice. Very often, data from other similar studies can provide an empirical estimate of ω, thus making it possible to have an upper bound ω_1 for ω. Taking these arguments into account, we propose to determine the critical value and sample size via the equations

$$\Psi(\rho_0;\omega_1) = \alpha, \quad \Psi(\rho_1;\omega_1) = 1 - \beta. \tag{2.10}$$

2.4. A simulation procedure

Exact calculation of the operating characteristics of the test statistic $\hat{\rho}$ appears to be complicated because it involves the joint distribution of two correlated F-statistics S and R in (2.6). As $\hat{\rho}$ is a function of three independent χ^2 random variables, we propose to calculate the operating characteristics based on independently generating large numbers of replicates of the χ^2 variables.

Specifically, for a given set (n, d, ω, ρ, c), the power $\Psi(\rho;\omega)$ is computed as the proportion of simulated values of $\hat{\rho}$ that are larger than c. To generate a random value $\hat{\rho}$ from a simulation, we generate a random observation of (V_1, V_2, V_3), which then yields random values of R and S according to (2.7) and (2.8). Substituting these values in (2.6) then yields the simulated value of $\hat{\rho}$.

Using this approach, we can find the critical value and sample size that satisfy the error requirements (2.10). Note that the sample size is taken as the smallest n that meets the error requirements.

Setting $n = 104$, $d = 4$, $\alpha = 0.05$, $\rho_0 = 0.5$ and $\omega_1 = 0.5$, we found $c = 0.6173$, using the approach. For $\omega \leq \omega_1$, Figure 1 plots the type I error $\Psi(\rho_0;\omega)$ and Figure 2 plots the power $\Psi(\rho_1;\omega)$ at $\rho_1 = 0.7$ and 0.8, both versus

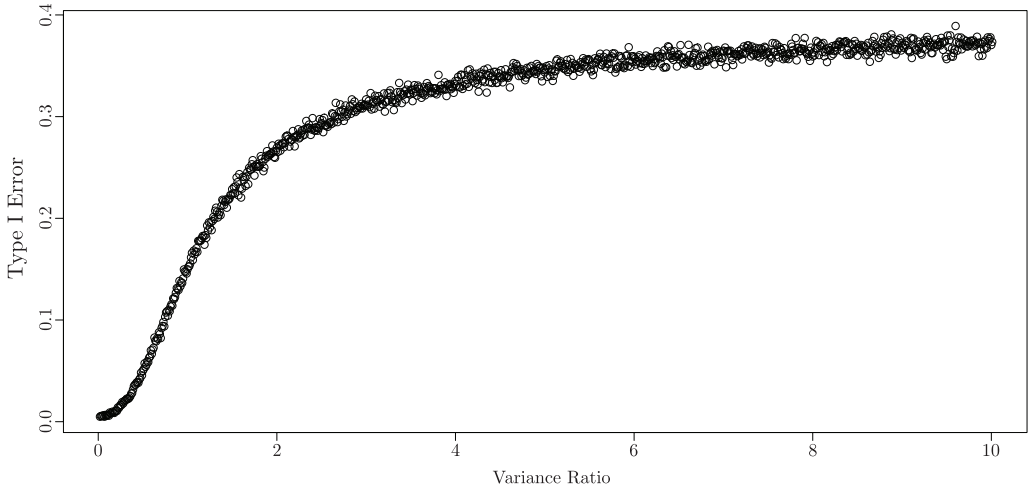

Figure 1. Type I Error Versus Variance Ratio-10,000 runs.

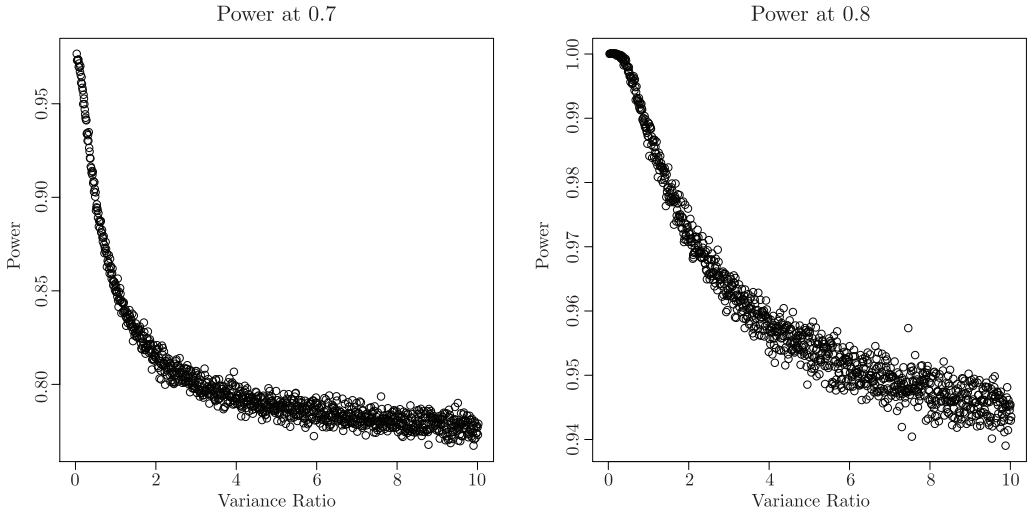

Figure 2.

ω. Although no analytic proof is available, it appears from the figures that, as ω increases, the type I error increases and the power decreases (for alternative values of ρ sufficiently away from the null value ρ_0). Thus for $\omega \leq \omega_1$, (2.10) implies

$$\Psi(\rho_0;\omega) \leq \alpha, \quad \Psi(\rho_1;\omega) \geq 1-\beta.$$

3. Multistage Group Sequential Testing and Simulation Technique

Sequential statistical methods were originally developed for economic bene-

fits. In any experiment or survey in which data accumulate steadily over a period of time, it is natural to monitor results as they occur with a view to taking such action as early termination or a modification of the study design. The many reasons for conducting interim analyses of accumulating data can be loosely categorized in three classes: ethical, administrative, and economic (Jennison and Turnbull (2000)).

For example (Rothstein (1990)), in job performance rating, such critically important decisions as pay, promotions, and firing are made on the basis that the ratings are reliable enough to produce useful and valid performance appraisals. An interim analysis is needed to ensure that the rating process is being executed as planned, and that the raters are well trained and satisfy eligibility criteria. An early examination of interim results can reveal the presence of problems that can be remedied before too much expense is incurred. Early stopping ensures that resources are not wasted. Sequential methods typically lead to savings in sample size, time, and cost when compared with standard fixed sample procedures.

3.1. Group sequential test

Suppose the subjects are taken into groups. Denote by n_k the number of subjects in the kth group, and let $N_k = n_1 + n_2 + \cdots + n_k$. Denote by $\hat{\rho}_k$ the estimate of the intraclass correlation coefficient ρ computed from data observed up to the kth group. Let K be the prespecified maximum number of interim analyses planned for the study. A K-stage testing procedure for H_0 is carried out as follows. At each stage $k = 1, \ldots, K-1$, with critical value c_k, stop if

$$\hat{\rho}_i \le c_i, i \le k-1, \hat{\rho}_k > c_k, \tag{3.1}$$

and reject H_0. Otherwise, d measurements are taken from the raters for each of the n_{k+1} subjects in the $(k+1)$th group, and $\hat{\rho}_{k+1}$ is computed based on all available data up to the $(k+1)$th stage. If the test has not stopped, stop at the Kth stage and reject H_0 if $\hat{\rho}_K > c_K$. Here $c_k, k = 1, \ldots, K$, are the critical values that satisfy the error requirements.

The power function of the group sequential test is

$$\sum_{k=1}^{K} P_\rho(\hat{\rho}_i \le c_i, i \le k-1, \hat{\rho}_k > c_k), \tag{3.2}$$

which depends on the variance ratio ω.

The design approach considered here is to specify the parameters ρ_0, ρ_1, α, and β, along with the nuisance parameter ω, and then determine the critical values and sample sizes that have power α at ρ_0 and $1-\beta$ at ρ_1. To determine

sample sizes and critical values for testing the null hypothesis H_0, we employ the Lan and DeMets error spending approach (Lan and DeMets (1983)).

We focus on two-stage designs that are more appealing in terms of complexity of computation and potential applications to a reliability study. However, the simulation approach to a two-stage test can be straightforwardly extended to more stages.

3.2. Two-stage design

In a two-stage study, the null hypothesis H_0 is rejected if (i) at the first stage $\hat{\rho}_1 > c_1$ or (ii) $\hat{\rho}_1 \le c_1$ at the first stage and $\hat{\rho}_2 > c_2$ at the second. Suppose that the type I error to spend at the first stage is α_1. We have

$$P_{\rho_0}(\hat{\rho}_1 > c_1) = \alpha_1, \tag{3.3}$$

$$P_{\rho_0}(\hat{\rho}_1 \le c_1, \hat{\rho}_2 > c_2) = \alpha - \alpha_1. \tag{3.4}$$

(The errors spent at each stage then add up to the total error of α.) The requirement that the power at ρ_1 be $1 - \beta$ yields

$$P_{\rho_1}(\hat{\rho}_1 > c_1) + P_{\rho_1}(\hat{\rho}_1 \le c_1, \hat{\rho}_2 > c_2) = 1 - \beta. \tag{3.5}$$

These three equations determine the critical values and the sample size assuming the allocation ratio, n_1/N_2, of subjects to the first group is specified. The critical value c_1 for the first stage and the first summand of (3.5) can be determined by the approach proposed in Section 2 for a fixed sample size test.

Computation of (3.4) and the second summand in (3.5) involves the joint distribution of $\hat{\rho}_1$ and $\hat{\rho}_2$ or, equivalently, the joint distribution of two correlated χ^2-statistics (apart from proper constants),

$(SS_{subject}(2), SS_{subject}(1)), (SS_{error}(2), SS_{error}(1)), (SS_{rater}(2), SS_{rater}(1))$.

For the one-way ANOVA model, Liu, Schisterman, and Wu (2006) showed that the independence between subjects warrants that the sums of squares involved in estimation of the intraclass correlation coefficient can be decomposed into independent increments. This independent-increment property substantially simplifies the simulation for calculation of critical values and power of the test.

For the two-way ANOVA model, the between-subject sum of squares $SS_{subject}$ and the within sum of squares SS_{error} have independent increments: $SS_{subject}(1)$, $SS_{subject}(2) - SS_{subject}(1)$, $SS_{error}(1)$, and $SS_{error}(2) - SS_{error}(1)$ are mutually independent. The four increments have the same distributions as $(d\sigma_s^2 + \sigma_\epsilon^2)U_1$, $(d\sigma_s^2 + \sigma_\epsilon^2)U_2$, $\sigma_\epsilon^2 U_3$, and $\sigma_\epsilon^2 U_4$, respectively, where U_1, U_2, U_3 and U_4 are independent chi-square random variables with degrees of freedom $n_1 - 1$, $N_2 - n_1$, $(n_1 - 1)(d - 1)$, and $(N_2 - n_1)(d - 1)$, respectively. These sums of squares at the

second stage can be expressed as functions of independent chi-square variables, similar to the one-way ANOVA model in Liu, Schisterman, and Wu (2006).

A major complication arises from calculation of the between-rater sum of squares $SS_{rater}(2)$. Because observations from different subjects made by the same rater are not independent, the between-rater sums of squares do not have the independent increments property; see the Appendix. Therefore, unlike the other two sums of squares, $SS_{rater}(1)$ and $SS_{rater}(2)$ cannot be generated using independent chi-square variables. We propose a more efficient method for simulation based on two independent multivariate distributions .

3.3. Simulation technique for two-stage design

With specified parameters the probabilities in (3.3)−(3.5) can be estimated by simulating a large number of realizations of the sums of squares. The probability $P(\hat{\rho}_1 > c_1)$ can be derived using simulation techniques described in Section 2.4. To compute the probability $P(\hat{\rho}_1 \leq c_1, \hat{\rho}_2 > c_2)$, we first simulate a large number of observations of $U = (U_1, U_2, U_3, U_4)$, defined as before. For each simulation, the between-subject and within sum of squares are

$$SS_{subject}(1) = (d\sigma_s^2 + \sigma_\epsilon^2)U_1,\ SS_{subject}(2) = (d\sigma_s^2 + \sigma_\epsilon^2)U_1 + (d\sigma_s^2 + \sigma_\epsilon^2)U_2,$$
$$SS_{error}(1) = \sigma_\epsilon^2 U_3,\ SS_{error}(2) = \sigma_\epsilon^2 U_3 + \sigma_\epsilon^2 U_4.$$

Empirical values of $SS_{rater}(1)$ and $SS_{rate}(2)$ can be generated from two independent multivariate normal variables. Details are given in the Appendix.

With each simulated value of $SS_{subject}(1)$, $SS_{sugject}(2)$, $SS_{error}(1)$, $SS_{error}(2)$, $SS_{rater}(1)$, and $SS_{rater}(2)$, we compute $\hat{\rho}_1$ and $\hat{\rho}_2$. The probability is then estimated as the proportion of $\{(\hat{\rho}_1, \hat{\rho}_2) : \hat{\rho}_1 \leq c_1, \hat{\rho}_2 > c_2\}$.

Remark 1. Although the probabilities can be estimated by simulating N_2-variate multivariate normal distributions under model (2.1), our experiences show that the proposed simulation approach based on independent chi-square and multivariate normal variables is much more time-efficient.

4. Simulation Results

To set up the simulation, set $n_1 = N_2/2$. The error spending function of Kim and DeMets (1987) with $\alpha_1 = \alpha/2$ is applied to allocate the error. The values of design parameters considered were $\alpha = (0.025, 0.05)$, $\rho_0 = (0.5, 0.6)$, and $1 - \beta = (0.8, 0.9)$ at $\rho_1 = (0.7, 0.8, 0.9)$. Here, according to Landis and Koch (1977), the chosen null values indicate moderate consistency in measurements. We considered the numbers of raters $(4, 6, 8)$, and the values of ω to be $(0.1, 0.5, 1, 10)$. For each set of design parameters $(\alpha, \beta, \rho_0, \rho_1, \omega)$, we searched a

Table 1. Sample size tabulation for a number of two-stage designs. $d = 4, \omega = 0.5, \alpha_1 = \alpha/2, n_1 + n_2 = N_2$ (10,000 replicates)

α	β	n_1	n_2	c_1	c_2	ASN	fixed size
$\rho_0 = 0.5, \rho_1 = 0.7$							
0.05	0.1	57	57	0.6526	0.6191	70.56	104
0.025	0.1	86	86	0.6548	0.6258	104.80	164
0.05	0.2	27	27	0.6866	0.6432	38.68	50
0.025	0.2	39	38	0.6904	0.6513	54.88	72
$\rho_0 = 0.5, \rho_1 = 0.8$							
0.05	0.1	11	10	0.7502	0.6886	14.35	20
0.025	0.1	14	14	0.7606	0.6999	18.73	25
0.05	0.2	7	7	0.7892	0.7151	10.56	13
0.025	0.2	9	9	0.7928	0.7284	13.47	17
$\rho_0 = 0.6, \rho_1 = 0.8$							
0.05	0.1	28	28	0.7639	0.7257	36.11	53
0.025	0.1	38	38	0.7687	0.7328	48.96	72
0.05	0.2	16	15	0.7902	0.7469	22.99	30
0.025	0.2	23	22	0.7940	0.7528	32.92	41
$\rho_0 = 0.6, \rho_1 = 0.9$							
0.05	0.1	6	6	0.8526	0.7962	7.98	11
0.025	0.1	7	7	0.8668	0.8090	9.59	13
0.05	0.2	4	4	0.8868	0.8238	6.03	7
0.025	0.2	5	5	0.8909	0.8295	7.54	9

potential range of (c_1, c_2, N_2) to find a combination that meets the error requirements.

For each (c_1, c_2, N_2), c_1 could be derived using the simulation techniques described in Section 2.4. A total of $m = 10,000$ random vectors $U = (U_1, U_2, U_3, U_4)$ were generated. Simultaneously, two samples each with $m = 10,000$ observations were drawn from independent multivariate normal variables as defined in the Appendix. The empirical probabilities in (3.3), (3.4) , and (3.5) were computed as the proportion that fell into the rejection regions. The combination that satisfied the error requirements was thus the desired solution to the design. To help narrow the search range, we first limited the values of c_2 and N_2 close to those required in a fixed sample design computed by methods mentioned in Section 2.4, and then gradually expanded the range until a solution is found.

Table 1 presents the simulation results of critical values and sample size required, for a selected range of designs. For each design, the final critical value c_2 is smaller than the interim critical value c_1, partly reflecting the increase of sample size. The Average Sample Number (ASN) under the alternative is smaller than the fixed sample size. Note that for a two-stage design the ASN at ρ is ASN $= n_1 P(\hat{\rho}_1 > c) + N_2 P(\hat{\rho}_1 \leq c_1)$.

All simulations were conducted using the R software. Codes are available from the authors upon request.

5. Example

We illustrate the proposed group sequential testing procedure on data from a study on endometriosis diagnosis. Endometriosis is a gynecological medical condition occurring in roughly 5%–10% of women. Despite its relatively high prevalence, substantive and methodological challenges exist, including diagnostic proficiency. The Physician Reliability Study (Schliep et al. (2012)) addressed this issue by examining agreement in diagnosing endometriosis and its association with both rater- and subject-specific information.

We used the review results of four experts who independently examined study subjects' intraoperative photos for signs of endometriosis or other gynecologic pathology. The specific outcome we used was the Revised American Society for Reproductive Medicine's classification that ranges from 1 to 150 with higher scores corresponding to severe symptoms of endometriosis.

The interim analysis was conducted based on the first 24 (about half of the 47 samples that have complete ratings) subjects' scores from the four experts. The interim error α_1 was set to $\alpha/2$ based on an error spending function of Kim and DeMets (1987). Thus, $n_1 = 24, N_2 = 47$, and $\alpha_1 = \alpha/2 = 0.025$. The null value is $\rho_0 = 0.4$, and the variance ratio ω is 0.013 based on the real data. With these design parameters, we applied the simulation approach of Section 3.3, and found $c_1 = 0.5953$ and $c_2 = 0.5359$. The power of the test at $\rho_1 = 0.55, 0.60, 0.65$ was 0.6088, 0.8150, and 0.9548, respectively. The average sample size (ASN) at these alternative values was 40, 36, and 30, respectively, reflecting the cost-effective nature of the design.

We then applied the test based on the scores on the first 24 subjects in the endometriosis study. The sample intraclass correlation coefficient computed at the first stage was $\hat{\rho}_1 = 0.6077$. We then could terminate the test at the first stage and reject the null hypothesis because the test statistic $\hat{\rho}_1$ exceeded the critical value c_1. Early evidence is that the four experts demonstrated quite satisfactory agreement in their scoring.

6. Discussion

In this paper, we develop multistage testing procedures using two-way ANOVA for hypothesis concerning the intraclass correlation coefficient in a inter-rater reliability study. The designs are proposed to determine critical values, sample size, and power using Lan and DeMets's error spending function. To our best knowledge there have been no exact tests available for the intraclass correlation coefficient based on two-way ANOVA framework due to the involvement of

two correlated F statistics. Realizing that the between-rater sum of squares violates the independent increments assumption, we develop simulation techniques to effectively calculate the operating characteristics of the sequential test. The performance of the proposed methods is examined using simulation studies and critical values are tabulated for a range of two-stage design parameters.

The operating characteristics of the test are affected substantially by the variance ratio ω. Our proposed approach requires an upper bound of ω be specified.

The methods discussed in this paper could be applied to the field of Genetics in the sense of the heritability analysis that estimates the relative contributions of differences in genetic and non-genetic factors to the total phenotypic variance in a population. Future research is needed to develop and compare various group sequential designs, and to propose methods for inference following the sequential testing such as the point and interval estimation of the intraclass correlation coefficient.

Acknowledgement

Research of Zhen Chen and Aiyi Liu was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH). The authors thank an associate editor and the referees for their valuable comments.

Appendix A: Increments of Sums of Squares

Recall that (Rao (1973)) if $\mathbf{y}$ has a multivariate normal distribution with a mean vector $\mathbf{u}$ and a positive definite variance-covariance matrix Δ, then an necessary and sufficient condition that $(\mathbf{y}-\mathbf{u})^T\mathbf{A}(\mathbf{y}-\mathbf{u})$ has a χ^2 distribution is that $\mathbf{A}\Delta\mathbf{A} = \mathbf{A}$ in which case the degrees of freedom is the rank of $\mathbf{A}\Delta$. Furthermore, $(\mathbf{y}-\mathbf{u})^T\mathbf{A}(\mathbf{y}-\mathbf{u})$ and $(\mathbf{y}-\mathbf{u})^T\mathbf{B}(\mathbf{y}-\mathbf{u})$ are independent if and only if $\mathbf{A}\Delta\mathbf{B} = \mathbf{0}$.

1. Independent Increments of Two-stage $SS_{subject}$

Put all the measurements from the group of d raters on N_2 subjects together in a vector $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T$, where $\mathbf{y}_1 = (y_{11}, y_{12}, \cdots, y_{n_1\, d})^T$ represents the measurements from the first stage and $\mathbf{y}_2 = (y_{(n_1+1)\,1}, y_{(n_1+1)\,2}, \cdots, y_{N_2\, d})^T$ the measurements taken after the first stage. Then $\mathbf{y}$ has a multivariate normal distribution with mean vector $\mathbf{u} = \mu\mathbf{1}_{N_2}$ and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \Sigma_{11}\, \Sigma_{12} \\ \Sigma_{21}\, \Sigma_{22} \end{pmatrix},$$

where

$$\Sigma_{11} = \sigma_\epsilon^2 \mathbf{I}_{n_1} \otimes \mathbf{I}_d + \sigma_s^2 \mathbf{I}_{n_1} \otimes \mathbf{J}_d + \sigma_r^2 \mathbf{J}_{n_1} \otimes \mathbf{I}_d,$$
$$\Sigma_{12} = \sigma_r^2 \mathbf{J}_{n_1 \times n_2} \otimes \mathbf{I}_d, \quad \Sigma_{21} = \sigma_r^2 \mathbf{J}_{n_2 \times n_1} \otimes \mathbf{I}_d,$$
$$\Sigma_{22} = \sigma_\epsilon^2 \mathbf{I}_{n_2} \otimes \mathbf{I}_d + \sigma_s^2 \mathbf{I}_{n_2} \otimes \mathbf{J}_d + \sigma_r^2 \mathbf{J}_{n_2} \otimes \mathbf{I}_d.$$

Write

$$U_1 = \frac{SS_{subject}(1)}{d\sigma_s^2 + \sigma_\epsilon^2} = \mathbf{y}^T \mathbf{S}_1 \mathbf{y},$$
$$U_2 = \frac{SS_{subject}(2) - SS_{subject}(1)}{d\sigma_s^2 + \sigma_\epsilon^2} = \mathbf{y}^T \left[\mathbf{S}_2 - \mathbf{S}_1\right] \mathbf{y},$$

where

$$\mathbf{S}_1 = \begin{pmatrix} \mathbf{S}(n_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{S}_2 = \mathbf{S}(N_2)$$

with

$$\mathbf{S}(n) = \frac{1}{d(d\sigma_s^2 + \sigma_\epsilon^2)} \left[(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n\times n}) \otimes \mathbf{J}_{d\times d} \right].$$

It follows from straightforward matrix manipulations that $\mathbf{u}^T\ \mathbf{S}_1 = \mathbf{0}$, $\mathbf{u}^T\ \mathbf{S}_2 = \mathbf{0}$, and $\mathbf{S}_1\Sigma[\mathbf{S}_2\text{-}\mathbf{S}_1] = \mathbf{0}$. Therefore $U_1 = \mathbf{y}^T\mathbf{S}_1\mathbf{y} = (\mathbf{y}-\mathbf{u})^T\mathbf{S}_1(\mathbf{y}-\mathbf{u})$ and $U_2 = \mathbf{y}^T[\mathbf{S}_2 - \mathbf{S}_1]\mathbf{y} = (\mathbf{y}-\mathbf{u})^T[\mathbf{S}_2 - \mathbf{S}_1](\mathbf{y}-\mathbf{u})$ are independent.

2. Independent Increments of Two-stage SS_{error}

Write

$$U_3 = \frac{1}{\sigma_\epsilon^2} SS_{error}(1) = \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{pmatrix} \mathbf{E}_1 \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$
$$U_4 = \frac{1}{\sigma_\epsilon^2} \left[SS_{error}(2) - SS_{error}(1)\right] = \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{pmatrix} \left[\mathbf{E}_2 - \mathbf{E}_1\right] \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where

$$\mathbf{E}_1 = \begin{pmatrix} \mathbf{E}(n_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{E}_2 = \mathbf{E}(N_2)$$

with

$$\mathbf{E}(n) = \frac{1}{\sigma_\epsilon^2} \left[(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n\times n}) \otimes (\mathbf{I}_d - \frac{1}{d}\mathbf{J}_{d\times d}) \right].$$

Then $SS_{error}(2) - SS_{error}(1)$ and $SS_{error}(1)$ are independent since $\mathbf{u}^T\ \mathbf{E}_1 = \mathbf{0}$, $\mathbf{u}^T\mathbf{E}_2 = \mathbf{0}$ and $\mathbf{E}_1\Sigma[\mathbf{E}_2\text{-}\mathbf{E}_1] = \mathbf{0}$.

3. Dependent Increments of Two-stage SS_{rater}

Write

$$SS_{rater}(1) = \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{pmatrix} \mathbf{R}_1 \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

$$[SS_{rater}(2) - SS_{rater}(1)] = \begin{pmatrix} \mathbf{y}_1^T & \mathbf{y}_2^T \end{pmatrix} [\mathbf{R}_2 - \mathbf{R}_1] \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

where

$$\mathbf{R}_1 = \begin{pmatrix} \mathbf{R}(n_1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{R}_2 = \mathbf{R}(N_2)$$

with

$$\mathbf{R}(n) = \frac{1}{n}\left[\mathbf{J}_n \otimes (\mathbf{I}_d - \frac{1}{d}\mathbf{J}_d)\right].$$

We can show that $\mathbf{u}^T\ \mathbf{R}_1 = \mathbf{0}$ and $\mathbf{u}^T\ \mathbf{R}_2 = \mathbf{0}$, yielding

$$SS_{rater}(1) = \mathbf{y}^T\mathbf{R}_1\mathbf{y} = (\mathbf{y}-\mathbf{u})^T\mathbf{R}_1(\mathbf{y}-\mathbf{u}),$$
$$[SS_{rater}(2) - SS_{rater}(1)] = \mathbf{y}^T[\mathbf{R}_2 - \mathbf{R}_1]\mathbf{y} = (\mathbf{y}-\mathbf{u})^T[\mathbf{R}_2 - \mathbf{R}_1](\mathbf{y}-\mathbf{u}).$$

However, the increments are not independent since $\mathbf{R}_1\Sigma[\mathbf{R}_2\text{-}\mathbf{R}_1] \neq \mathbf{0}$.

4. Simulation Technique for SS_{rater}

Define $\mathbf{Z} = \mathbf{y}_2 - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1$. Then $\mathbf{Z}$ is distributed as multivariate normal distribution with mean vector $\mu\left(\mathbf{1}_{n_2} - \Sigma_{21}\Sigma_{11}^{-1}\mathbf{1}_{n_1}\right)$ and variance-covariance matrix $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

Random observations of $SS_{rater}(1)$ can be generated from $\mathbf{y}_1$. To simulate $SS_{rater}(2)$, we can generate $\mathbf{Z}$ independently of $\mathbf{y}_1$ and then calculate $\mathbf{y}_2 = \mathbf{Z} + \Sigma_{21}\Sigma_{11}^{-1}\mathbf{y}_1$. This then gives

$$SS_{rater}(2) = \frac{1}{N_2}\mathbf{y}^T[\mathbf{J}_{N_2} \otimes (\mathbf{I}_d - \frac{1}{d}\mathbf{J}_{d\times d})]\mathbf{y},$$

where $\mathbf{y}^T = (\mathbf{y}_1^T, \mathbf{y}_2^T)$.

Appendix B: Stochastic Ordering of $\hat{\rho}$

It follows from (6)−(8) that $\hat{\rho}$ has the same distribution as $W(\rho,\omega) = D_1/D_2$ where

$$D_1 = \frac{nd(d-1)\rho F_1}{1-\rho}\omega + \frac{n(d-1)(d\rho+1-\rho)F_1}{1-\rho} - n,$$
$$D_2 = \Big[\frac{nd(d-1)\rho F_1}{1-\rho} + nd(n-1)F_2\Big]\omega + \frac{n(d-1)(d\rho+1-\rho)F_1}{1-\rho}$$
$$+d(n-1)F_2 + (nd-n-d),$$

where $F_1 = V_1/V_3$ and $F_2 = V_2/V_3$.

Checking the derivative of W with respect to ρ, we can show that W is an increasing function of ρ. Hence, for $\rho_1 < \rho_2$, $W(\rho_1, \omega) > c$ implies $W(\rho_2, \omega) > c$. Therefore $P_{\rho_1}(\hat{\rho} > c) = P(W(\rho_1, \omega) > c) \leq P(W(\rho_2, \omega) > c) = P_{\rho_2}(\hat{\rho} > c)$.

References

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *Quarterly J. Medicine* **23**, 255-274.

Armitage, P. (1958). Sequential methods in clinical trials. *Amer. J. Public Health* **48**, 1395-1402.

Bonett, D. G. (2002). Sample size requirements for estimating intraclass correlations with desired precision. *Statist. Medicine* **21**, 1331-1335.

Bross, I. (1952). Sequential medical plans. *Biometrics* **8**, 188-205.

Bross, I. (1958). Sequential clinical trials. *J. Chronic Diseases* **8**, 349-365.

Cappelleri, J. C. and Ting, N. (2003). A modified large sample approach to approximate interval estimation for a particular intraclass correlation coefficient. *Statist. Medicine* **22**, 1861-1877.

Donner, A. and Eliasziw, M. (1987). Sample size requirements for reliability studies. *Statist. Medicine* **6**, 441-448.

Fleiss, J. L. (1999). *The Design and Analysis of Clinical Experiments*. Wiley, New York.

Jennison, C. and Turnbull, B. (2000). *Group Sequential Methods with Applications to Clinical Trials*. (2nd revised edn). Chapman and Hall/CRC, New York.

Kim, K., and DeMets, D. L. (1987). Design and analysis of group sequential tests based on the type I error spending rate function. *Biometrika* **74**, 149-154.

Kraemer, H. (1976). The small sample non-null properties of Kendall's coefficient of concordance for normal populations. *J. Amer. Statist. Assoc.* **71**, 608-613.

Lai, T. L. (2004). Interim and Terminal Analyses of Clinical Trials with Failure-Time Endpoints and Related Group Sequential Designs. In *Applications of Sequential Methodologies* (Edited by N. Mukhopadhyay, S. Datta and S. Chattopadhay), 193-218. Marcel Dekker, New York.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-663.

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.

Liu, A., Schisterman, E. F. and Wu, C. (2006). Multistage evaluation of measurement error in a reliability study. *Biometrics* **62**, 1190-1196.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrika* **35**, 549-556.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-200.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. Wiley, New York.

Rothstein, H. R. (1990). Interrater reliability of job performance ratings: growth to asymptote level with increasing opportunity to observe. *J. Appl. Psychology* **75**, 322-327.

Siegmund, D. (1985). *Sequential Analysis*. Springer-Verlag, New York.

Saito, Y., Sozu, T., Hamada, C. and Yoshimura, I. (2006). Effective number of subjects and number of raters for inter-rater reliability studies. *Statistics in Medicine* **25**, 1547-1560.

Schliep, K. C., Stanford, J. B., Chen, Z., Zhang, B., Dorais, J. K., Johnstone, E. B., Hammoud, A. O., Varner, M. W., Buck Louis, G. M., and Peterson, C. M. MD, on behalf of the Endometriosis: Natural History, Diagnosis and Outcomes (ENDO) Study Working Group. (2012). Interrater and intrarater reliability in the diagnosis and staging of endometriosis. *Obstetrics & Gynecology* **120**, 104-112.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations, uses in assessing rater reliability. *Psychological Bulletin* **86**, 420-428.

Shoukri, M. M., Asyali, M. H., and Walter, S. D. (2003). Issues of cost and efficiency in the design of reliability studies. *Biometrics* **59**, 1107-1112.

Tian, L. and Cappelleri, J. C. (2003). A new approach for interval estimation and hypothesis testing of a certain intraclass correlation coefficient: the generalized variable method. *Statistics in Medicine* **23**, 2125-2135.

Wald, A. (1947). *Sequential Analysis*. Wiley, New York.

Walter, S. D., Eliasziw, M. and Donner, A. (1998). Sample size and optimal designs for reliability studies. *Statist. Medicine* **7**, 101-110.

Zou, K. H., and McDermott, M. P. (1999). Higher-moment approaches to approximate interval estimation for a certain intraclass correlation coefficient. *Statist. Medicine* **18**, 2051-2061.

Department of Statistics, George Washington University, Washington, DC, U.S.A.
Credit Risk Management, Capital one, McLean, VA, U.S.A.

E-mail: mei.jin@capitalone.com

NICHD, National Institute of Health, Rockville, MD, U.S.A.

E-mail: liua@mail.nih.gov

NICHD, National Institute of Health, Rockville, MD, U.S.A.

E-mail: chenzhe@mail.nih.gov

Department of Statistics, George Washington University, Washington, DC, U.S.A.

E-mail: zli@gwu.edu

(Received February 2012; accepted August 2012)