Contemporary Clinical Trials 36 (2013) 244–253

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Covariate-adjusted confidence interval for the intraclass correlation coefficient

Mohamed M. Shoukri [a,c,*], Allan Donner [b], Abdelmoneim El-Dali [d]

[a] National Biotechnology Center, KFSHRC, Saudi Arabia
[b] Schulich School of Medicine and Dentistry, Canada
[c] Al-Faisal University College of Medicine, Saudi Arabia
[d] Department of Biostatistics, KFSHRC, Saudi Arabia

ARTICLE INFO

Article history:
Received 2 April 2013
Received in revised form 4 July 2013
Accepted 7 July 2013
Available online 16 July 2013

Keywords:
Multi-level models
Intra-class correlation
Generalized Estimating Equations
Percentile bootstrap confidence intervals
Monte-Carlo simulations

ABSTRACT

A crucial step in designing a new study is to estimate the required sample size. For a design involving cluster sampling, the appropriate sample size depends on the so-called design effect, which is a function of the average cluster size and the intracluster correlation coefficient (ICC). It is well-known that under the framework of hierarchical and generalized linear models, a reduction in residual error may be achieved by including risk factors as covariates. In this paper we show that the covariate design, indicating whether the covariates are measured at the cluster level or at the within-cluster subject level affects the estimation of the ICC, and hence the design effect. Therefore, the distinction between these two types of covariates should be made at the design stage. In this paper we use the nested-bootstrap method to assess the accuracy of the estimated ICC for continuous and binary response variables under different covariate structures. The codes of two SAS macros are made available by the authors for interested readers to facilitate the construction of confidence intervals for the ICC. Moreover, using Monte Carlo simulations we evaluate the relative efficiency of the estimators and evaluate the accuracy of the coverage probabilities of a 95% confidence interval on the population ICC. The methodology is illustrated using a published data set of blood pressure measurements taken on family members.

1. Introduction

Estimation of the intraclass correlation coefficient (ICC) is relevant to many applications in survey sampling, genetic epidemiology, reliability studies and other fields. In genetic epidemiology it is used as a measure of familial aggregation, i.e. as a measure of similarity of responses among siblings who belong to the same family [1,2], while in interobserver agreement studies, it is used as a measure of reliability [3]. In cluster randomized trials and observational studies that involve aggregates of individuals as sampling units; the ICC measures the degree of similarity among individuals belonging to the same cluster and must be taken into account in both the estimation of sample size and the statistical analysis.

The ICC may be defined as the ratio of the between cluster variance divided by the total variance (the sum of between cluster variance and within cluster variance). When the trait of interest is measured on quantitative scale (e.g. blood pressures, body mass index) the ICC may be estimated using standard expressions for variance components under the assumption of multivariate normality. The most common model used for this purpose is one-way random effects analysis of variance (ANOVA) [4,5]. When the trait is measured on a binary scale, the ANOVA model may be used as well to find a point estimator for the ICC.

* Corresponding author at: National Biotechnology Center, KFSHRC, Saudi Arabia. Tel.: +966 509491454.
E-mail address: shoukri@kfshrc.edu.sa (M.M. Shoukri).

1551-7144
http://dx.doi.org/10.1016/j.cct.2013.07.003

It may be desirable for the purpose of increasing precision to extend the one-way random effects model to include one or more covariates. In this case the selected covariate structure would be expected to affect the estimated ICC and its standard error. In particular we discuss three scenarios: when the covariate is measured at the cluster level, measured at the individual level and when measured at both levels of hierarchy.

In summary, the main objective of this paper is to derive a covariate adjusted variance components estimator for the ICC with corresponding standard error under the three proposed covariate structures and under the assumption of multivariate normality. For the case in which the response variable is measured on a binary scale, we use the Generalized Estimating Equations to find a working correlation estimate, accounting for the measured covariates. We construct confidence limits for the ICC using the non-parametric "Accelerated Bias-corrected percentile" bootstrap known as BCa interval [6,7]. The asymptotic relative efficiency of the ICC estimators corrected for the effect of measured covariates will be assessed relative to the estimator obtained when covariate effects are not accounted for, using Monte Carlo simulations. Moreover, we use simulations to evaluate the coverage probabilities of the 95% confidence intervals on the population parameter. We illustrate the methodology presented in this paper on published arterial blood pressure data collected from nuclear families.

The paper is structured as follows: In Section 2 we introduce the normal linear mixed model and the ANOVA estimator of the ICC. Covariate adjusted ICC estimators and their large sample standard errors are obtained using the delta method, with comparisons made with the standard errors obtained using bootstrap. In Section 3 we discuss the case of a binary outcome, and introduce a BCa confidence interval for the ICC. Section 4 presents an example using a published data set of arterial blood pressures taken on nuclear families.

In Section 5, we design a Monte Carlo study to evaluate the asymptotic efficiency of the proposed estimators and evaluate the adequacy of the constructed confidence intervals on the population values of the ICC. Two SAS macros for nested-bootstrap cluster re-sampling that may be used to facilitate the construction of confidence intervals for ICC in the continuous and binary case are available from the first author.

2. Effect of the covariate structure

In what follows we investigate four statistical models. The first, we call the baseline or the unconditional mean model [8]. The second includes one covariate measured at the cluster level; while the third includes a covariate measured at the individual level, and the fourth includes both types of covariates.

2.1. No covariates (baseline models)

The most commonly used model for estimating the ICC is the one-way random effects model given by:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{1}$$

where μ is the grand mean of all measurements in the population, τ_i reflects the effect of cluster i, and ϵ_{ij} is the error term ($i = 1, 2, \ldots k$; $j = 1, 2, \ldots n_i$). It is assumed that the cluster effects $\{\tau_i\}$ are normally and identically distributed with mean 0 and variance σ_τ^2, the errors $\{\epsilon_{ij}\}$ are normally and identically distributed with mean 0 and variance σ_ϵ^2, and the $\{\tau_i\}$ and $\{\epsilon_{ij}\}$ are independent. For this model the ICC, which may be interpreted as the correlation ρ between any two members of a cluster, may be defined as

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}. \tag{2}$$

It is seen by definition that the ICC is defined as non-negative in this model, a plausible assumption for the application of interest here. We also note that the variance components σ_τ^2 and σ_ϵ^2 can be estimated from the one-way ANOVA mean squares [9–11] given in expectation by

$$\mathrm{E(MSB)} = \sigma^2 + n_0\sigma_\tau^2, \tag{3}$$

where, $n_0 = \frac{1}{k-1}\left[N - \sum_{i=1}^{k} n_i^2/N\right]$, and $N = \sum_{i=1}^{k} n_i$.

$$\mathrm{E(MSW)} = \sigma_\epsilon^2.$$

The ANOVA estimator of the population intraclass correlation is thus given by:

$$\widehat{\rho}_0 = \frac{\mathrm{MSB} - \mathrm{MSW}}{\mathrm{MSB} + (n_0 - 1)\mathrm{MSW}} \tag{4}$$

where MSB and MSW are, obtained from the usual ANOVA table, with corresponding sums of squares

$$\mathrm{SSB} = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$$

$$\mathrm{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2.$$

Using the delta method, and to the first order of approximation, the variance of $\widehat{\rho}_0$ [5] is given by:

$$var(\widehat{\rho}_0) = \frac{2(1-\rho)^2(1 + (n_0 - 1)\rho)^2}{n_0^2(k-1)\left(1 - \frac{k}{N}\right)}. \tag{5}$$

Note that when $n_i = n$, $i = 1, 2, \ldots k$, Eq. (5) reduces to

$$var(\widehat{\rho}_0) = \frac{2(1-\rho)^2(1 + (n-1)\rho)^2}{n(n-1)(k-1)}. \tag{6}$$

This equation differs from the variance expression given in [12,13] by a factor $(1 - \frac{1}{k})$, which for large number of clusters is 1. Note also that when $n_i = 1$ (as in twin studies), $var(\widehat{\rho}_0) = k^{-1}(1-\rho^2)^2$.

An approximate $(1 - \alpha)$ 100% confidence interval on ρ may then be constructed as:

$$\widehat{\rho}_0 \pm z_{1-\alpha/2}\sqrt{var(\widehat{\rho}_0)}. \tag{7}$$

Extensive simulations to evaluate the coverage probabilities of the above interval showed [14] that this approximation is adequate over a wide range of the parameter combinations (ρ, k, n). For the different estimators of ICC that will be

considered later in this paper and their variances, expression (7) provides an approximate $(1-\alpha)$ 100% confidence interval on the corresponding population parameter.

2.2. Effect of one measured covariate

The Stanish and Taylor [15] adjusted model (2) is given as follows:

$$y_{ij} = \mu + \tau_i + \beta\left(x_{ij} - \bar{x}\right) + \epsilon_{ij},$$

where x_{ij} represents a covariate measured without error and $\bar{x} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}$.

Again, from Searle et al. [11] we have

$$\text{E (SSW)} = (N-K-1)\,\sigma_\epsilon^2,\ E\ (\text{MSW}) = \sigma_\epsilon^2,$$
$$\text{and} \qquad E(\text{MSB}) = \frac{1}{k-1}\text{E(SSB)} = \sigma_\epsilon^2 + n_{01}\sigma_\tau^2$$

where $n_{01} = \frac{1}{k-1}\left[(k-1)n_0 - \frac{\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2}\right]$ and $\bar{x}_i = n_i^{-1}\sum_{j=1}^{n_i} x_{ij}$.

Therefore, the ANCOVA estimator of ρ is given by:

$$\widehat{\rho}_1 = \frac{\text{MSB}-\text{MSW}}{\text{MSB} + (n_{01}-1)\text{MSW}}. \tag{8}$$

The asymptotic variance of $\widehat{\rho}_1$, using the delta method is given by:

$$var(\widehat{\rho}_1) = \frac{2(1-\rho)^2(1+(n_{01}-1)\rho)^2}{n_{01}^2(k-1)\left(1-\frac{k}{N}\right)}.$$

Note that

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2 + \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2.$$

We have two remarks on the above set-up:

(i) If x_{ij} is measured at the cluster level, then $x_{ij} = \bar{x}_i$. Hence

$$n_{01} = n_0 - \frac{1}{k-1}\frac{\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}$$

and the expectation of the mean sum of squares between clusters can be written as:

$$\text{E(MSB)} = \sigma_\epsilon^2 + n_0\sigma_\tau^2 - \frac{\sigma_\tau^2}{k-1}\left[\frac{\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}\right]. \tag{9}$$

It is clear from Eq. (9) that when one covariate is measured at the cluster level, the expected mean square between clusters is reduced by the amount

$$\frac{\sigma_\tau^2}{k-1}\left[\frac{\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2}{\sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2}\right].$$

The covariate effect

$$\text{CEF} = \sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2 \Big/ \sum_{i=1}^{k} n_i(\bar{x}_i-\bar{x})^2$$

is the ratio of two quantities that measure the extent of the deviation of the cluster mean from the overall mean of the measured covariate. The numerator in this expression is seen to be weighted by the square of the cluster size, and the denominator by the cluster size. We also note that the degrees of freedom associated with the within cluster sum of squares is reduced by 1, due to the estimation of the regression coefficient β.

It is not generally clear how measuring a covariate on the cluster level will affect the estimated value of the ICC. However, in the case $n_i = n$, $i = 1, 2, \ldots k$, CEF $= n$ and $E(MSB) = \sigma_\epsilon^2 + n\sigma_\tau^2\left(\frac{k-2}{k-1}\right)$.

While it is clear that an estimated ICC may decrease in value, Stanish and Taylor [15] also identified situations when this estimate may increase in value. They based their argument on the quantity:

$$w = \frac{\text{effect of } x \text{ on within cluster variation}}{\text{effect of } x \text{ on between cluster variation}}.$$

In case $w < 1$ adjusting for x tends to decrease the estimated ICC, while when $w > 1$ adjusting for x tends to increase the estimated ICC.

2.3. Effect of two measured covariates

The one-way random effects model with two covariates may be written as:

$$y_{ij} = \mu + \tau_i + \beta_1\left(x_{ij}-\bar{x}\right) + \beta_2\left(z_{ij}-\bar{z}\right) + \epsilon_{ij}. \tag{10}$$

The expectations of the within and between mean squares are given respectively by:

$$\text{E(SSW)} = (N-k-2)\sigma_\epsilon^2$$
$$\text{E(MSB)} = \sigma_\epsilon^2 + n_{02}\sigma_\tau^2,$$

where, $n_{02} = n_0 - N'_{02}$, and

$$N'_{02} = \frac{1}{k-1}\left[\frac{1}{c}\left\{s_{zz}\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})^2 - \right.\right.$$
$$\left.\left. -2s_{xz}\sum_{i=1}^{k} n_i^2(\bar{x}_i-\bar{x})(\bar{z}_i-\bar{z}) + s_{xx}\sum_{i=1}^{k} n_i^2(\bar{z}_i-\bar{z})^2\right\}\right]$$
$$s_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)^2, s_{zz} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij}-\bar{z}\right)^2,$$
$$s_{xz} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)\left(z_{ij}-\bar{z}\right).$$

The derivation of these expressions is outlined in the Appendix.

The estimator of the ICC in this case will be given as:

$$\widehat{\rho}_2 = \frac{\text{MSB}-\text{MSW}}{\text{MSB} + (n_{02}-1)\text{MSW}}.$$

The large sample variance of $\widehat{\rho}_2$ is given by:

$$var(\widehat{\rho}_2) = \frac{2(1-\rho)^2(1+(n_{02}-1)\rho)^2}{n_{02}^2(k-1)\left(1-\frac{k}{N}\right)}.$$

In what follows we assume that x_{ij} is measured at the cluster level, that is $x_{ij} = \overline{x}_i$, implying $s_{xx} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})^2$ and $s_{xz} = \sum_{i=1}^{k} n_i(\overline{x}_i - \overline{x})(\overline{z}_i - \overline{z})$.

It is not clear how the measured covariates affect the estimated ICC, because the cross product term may either be positive or negative. However, if the cluster size is relatively constant $n_i \simeq n$, and the two covariates are orthogonal, we may write:

$$\mathrm{E(MSB)} = \sigma_{\epsilon}^2 + n\left(1-\frac{2}{k-1}\right)\sigma_{\tau}^2. \tag{11}$$

3. The case of clustered binary data

Let $y_{ij} = 1(0)$ denote the presence (absence) of a condition in the *jth* observation from the *ith* cluster assume that:

$$\Pr\left[y_{ij}=1\right] = \mu_{ij}, \text{ and } \Pr\left[y_{ij}=0\right] = 1-\mu_{ij}. \tag{12}$$

Murray et al. [16] considered the logit-scale additive mixed effects model with random component b_i to account for the between clusters variations. This model is a special case of the family of generalized linear mixed models.

$$\log\left[\frac{\mu_{ij}}{1-\mu_{ij}}\right] = x_{ij}'\beta + b_i. \tag{13}$$

In this model, b_i is a random sample from a normal distribution with mean 0 and variance σ_b^2. Yelland et al. [17] considered a log-transformation on the success probability, hence converting the between cluster variance to the probability scale. They used Monte-Carlo simulation to compare the variance components estimates of the ICC to that obtained from the unadjusted ANOVA model. The main objective of their study was to evaluate the relative bias in the estimation of the ICC under different modeling strategies. Yelland et al. [17] indicated that their study had several limitations the most important of which was that the true value of the ICC was unknown. Therefore it was not possible to compare the estimated ICC values to the true value. Since our main concern in this paper is with the accuracy in estimation, to achieve this objective we use the well-developed GEE methodology [18], together with the bootstrap methodology to construct confidence interval on the ICC. The GEE method is semi-parametric, and estimates of the regression parameters, which is its main target, are derived without full specification of the multivariate joint distribution of $y_i' = \left(y_{i1}, y_{i2}, \ldots y_{in_i}\right)$. Instead, specification of the likelihood of the marginal distribution of y_{ij} is given together with a working correlation matrix for the vector of observations in each cluster. That is the GEE method avoids the need to specify a form for the multivariate distribution of the binary responses y_i by only assuming a functional form (Bernoulli distributions for each observation within a cluster). The covariance structure is then treated as a nuisance. Consistent estimates of the variance of the regression coefficients are obtained under the assumption of independence across clusters, even when the assumed correlation structure is incorrect.

First we relate the mean of the marginal response to a linear combination of the covariates, omitting the random component b_i from Eq. (13).

We choose the form of an $n_i \times n_i$ working correlation matrix Σ_i for each y_i ($i = 1, 2, \ldots k$). The (j; j') element of Σ_i is the known or estimated correlation between (y_{ij}, $y_{ij'}$). This working correlation may depend on a vector of unknown parameters α, which is assumed to be same for all clusters [18].

Although the correlation matrix Σ_i can differ from cluster to cluster, we assume here that the ICC in the working matrix is an average of intraclass correlations across clusters. Thus an exchangeable correlation $\Sigma_{ij'} = \rho$ is used for this purpose. In this case the working covariance matrix for y_i equals:

$$v(\alpha) = A_i^{1/2}\sum\nolimits_i(\alpha)A_i^{1/2},$$

where

$$\Sigma_i(\alpha) = \begin{bmatrix} 1 & \alpha & \ldots & \alpha \\ \alpha & 1 & \ldots & \alpha \\ \alpha & \alpha & \ldots & 1 \end{bmatrix},$$

$$A_i = \begin{bmatrix} \mathrm{var}\left(y_{ij}\right) & 0\ldots & 0 \\ 0 & \mathrm{var}\left(y_{ij}\right)\ldots & 0 \\ \vdots & & \\ 0 & & \mathrm{var}\left(y_{ij}\right) \end{bmatrix},$$

and $var(y_{ij}) = \mu_{ij}(1 - \mu_{ij})$.

The GEE estimator of β is the solution of

$$\sum_{i=1}^{k}\left(\frac{\partial\mu_i}{\partial\beta}\right)' \Sigma^{-1}(\widehat{\alpha})(y_i-\mu_i) = 0$$

where $\mu_i = E(y_i)$.

Liang and Zeger [18] proposed an estimator for α based on the residuals

$$\widehat{r}_{ij} = \frac{y_{ij}-\widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}\left(1-\widehat{\mu}_{ij}\right)}}.$$

Under the common correlation (exchangeable), that is $Corr(y_{ij}, y_{il}) = \alpha$ for all (i, j, l) on estimator for α is:

$$\widehat{\alpha} = \sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{n_i}\sum\nolimits_{l=j+1}^{n_i-1}\widehat{r}_{ij}\widehat{r}_{il}/\left\{\sum\nolimits_{i=1}^{k}\binom{n_i}{2}-p\right\} \tag{14}$$

where p is the number of covariates in the logistic regression model. Usually we take $\widehat{\alpha}$ in an approximation for the ICC [18].

4. Data analysis and bootstrap confidence intervals

4.1. Example: Miall and Oldham's blood pressures, family data

The data used for illustration here are obtained from a survey that aimed at assessing the levels of similarity in systolic and diastolic blood pressures among family members living within 25 miles of Rhonda Fach Valley in South Wales and published by Miall and Oldham [19]. Observations were made

on parents and their offspring, with each observation consisting of systolic and diastolic blood pressures measured to the nearest 5 mm Hg. However among 250 sampled families, only 204 contained information on brothers and sisters. Furthermore, because of the impossibly low systolic blood pressure (15 mm Hg) for one daughter, another family was omitted leaving 203 families for the analysis. Since these data were given on a continuous scale, we transformed the measurements into a binary scale. The dichotomization was such that for an individual whose blood level was above 130/85, the assigned binary score is $y_{ij} = 1$, else $y_{ij} = 0$. This dichotomization follows the definition of hypertension provided by the Institute of Medicine.

In a recent paper Field and Walsh [20] suggested several approaches to bootstrap clustered data. In a random sample of k clusters each of size n, they considered the observations as fixed with inferences made with respect to the random sampling mechanism. In this case their main concern was with the accommodation of different forms of cluster sampling. One of the simplest approaches is the so-called "cluster bootstrap". Roberts and Xitao [21] implemented a specific form of bootstrap cluster sampling which they named "nested bootstrap" using the PROC MIXED procedure in SAS [22]. Note that PROC MIXED is designed to fit hierarchical data with normally distributed responses, and is not appropriate for the analysis of binary response data. We therefore modified the bootstrap macro so that SAS PROC GLM is used to calculate the between and the within mean of squares of the appropriate analysis of covariance (ANCOVA) from which we obtain the bootstrap replications, the bootstrap estimate, and hence the bootstrap standard errors. The SAS bootstrap macros are available and may be requested from the first author.

In Tables 1 and 2, Model 1 is the unconditional mean model (no covariates), Model 2, is the regression model with cluster specific covariates (in this case we used the mother's systolic blood pressures), Model 3 is the regression model with within cluster varying covariate (in this case we used the sib's gender), while Model 4 contains both types of covariates. We note from both Tables that the covariate design structure clearly affects the estimated values of the ICC. While there is a clear pattern of decline in the values of ICC in Table 1, we found that Model 4 in the binary case has a higher ICC relative to Model 3. It is also important to note that, the standard errors under the four models are fairly stable. Moreover, in Table 1, the bootstrap standard errors are almost identical to their first order approximations.

The bootstrap provides an attractive approach for obtaining simple estimates of the standard error and bias of the ICC from the estimated working correlation. In this section we denote the working correlation estimator of ICC by $\widehat{\rho}_w$. The SAS code

Table 1
Effect of measured covariates on the estimated ICC and its variance under different covariate structures based on the ANCOVA analyses.

Statistic	Model 1	Model 2	Model 3	Model 4
$\widehat{\rho}$	0.360	0.295	0.247	0.226
Bootstrap variance	0.001	0.002	0.002	0.003
Analytic[a] variance	0.002	0.003	0.002	0.002

[a] Note that analytic variances of the estimated ICC under Models 2, 3, and 4 are obtained by substituting, $n_0 = 3.08$, for Model 1, $n_{01} = 3.16$ for Model 2, $n_{02} = 3.06$ for Model 3 and $n_{02} = 3.14$ for Model 4 in Eq. (5).

Table 2
Effect of measured covariates on the estimated working correlation under different covariate structures using the GEE method.

Statistic	Model 1	Model 2	Model 3	Model 4
$\widehat{\rho}$	0.169	0.154	0.017	0.022
Bootstrap variance	0.002	0.002	0.002	0.003
BCa 95% confidence interval	(0.165, 0.173)	(0.148, 0.160)	(0.013, 0.021)	(0.017, 0.026)

needed to obtain this estimator from the working correlation using a logit-link is given by

$$\begin{aligned} &Proc\ Genmod;\\ &Class\ cluster\_ID;\\ &Model\ y = gender\ mother\_sbp/dist = bin\ link = logit;\\ &Repeated\ subject = cluster\_ID/Type = exch\ corrw;\\ &Run; \end{aligned} \tag{15}$$

Note that gender is a subject specific covariate, while mother's systolic blood pressure is a cluster specific covariate. We should also note that the estimated regression coefficients under the logit-link have Population-Averaged (PA) log-odds ratio (OR) interpretation which is the recommended measure of association between covariates and the binary response under cross-sectional study design [18]. For prospective and cohort studies, the relative risk (RR) is used as a measure of association between response and risk factors. The above SAS code can be used to produce RR estimate when the $link = log$ is used in the option of the *Model* statement in place of the $link = logit$. We also noted that the estimated value of the working correlation under the GEE is the same regardless of the link function.

The nested bootstrap SAS-macro for clustered data produces bootstrap replicates $\widehat{\rho}_w = (\widehat{\rho}_{w1}, \widehat{\rho}_{w2}, \ldots \widehat{\rho}_{wk})$.

The corresponding bootstrap standard error is approximated by the empirical standard deviation of $\widehat{\rho}_w$, i.e.:

$$\widetilde{\sigma}(\widehat{\rho}_w) = \sqrt{\frac{1}{b-1}\sum_{l=1}^{b}\left(\widetilde{\rho}_{wl} - \overline{\overline{\rho}}_b\right)^2} \to \sigma(\widehat{\rho}_w)$$

as $b \to \infty$, where $\overline{\overline{\rho}}_b = \frac{1}{b}\sum_{l-1}^{b}\widetilde{\rho}_{wl}$.

We can see from the replicated bootstraps (Table 3) that there is a significant skewness and kurtosis in the empirical distribution of $\widehat{\rho}$. Hence the normal approximation used to construct confidence intervals is no longer valid. However there are several methods to construct bootstrap confidence intervals including the Gaussian bootstrap, bootstrap-t and the percentile bootstrap [6]. Again, due to the significant skewness in the distribution of $\widehat{\rho}$, the first two methods of calculating the confidence intervals are not suitable. The percentile bootstrap did not perform well even though it does not assume normality. We therefore used the so-called bias-corrected, accelerated (or BCa) percentile interval. Shao & Tu [7] provided a good review of this method. It is recommended that to obtain sufficiently accurate 95% BCa confidence intervals, the number of bootstrap samples, should be at least 1000. In our example, we used 2000 bootstrap samples to construct a BCa confidence interval for the ICC. We assessed the normality of the distribution of the estimated ICC under the four models first graphically using the Q–Q plots of the bootstrap samples, and using an approach proposed by D' Agostino et al. [23]. These

Table 3
Skewness and kurtosis of the bootstrap replicates for the 4 models.

	Continuous outcome models				Dichotomous outcome models			
	1	2	3	4	1	2	3	4
Skewness	.148	−.227	−.127	−.07	.191	.177	.132	.11
(se)	(.046)	(.035)	(.03)	(.026)	(.055)	(.055)	(.055)	(.055)
Kurtosis	.233	−.040	−.394	−.42	.153	.195	.063	.02
(se)	(.092)	(.071)	(.06)	(.05)	(.109)	(.109)	(.109)	(.109)
P-value	0.0002	0.00000	000000	00000	.0008	.02	.047	.017

authors provided a simple SAS macro to calculate the values of a chi-square omnibus test statistic that utilizes both the skewness and kurtosis of the bootstrap replicates to assess the departure from normality. For the example data, Figs. 1–4 for the continuous response case and Figs. 5–8 for the binary case, we find that the empirical distributions of the bootstrap samples are skewed and leptokurtic, with the corresponding p-values indicating that the hypothesis of normality is not supported. The results are shown in Table 3.

5. Asymptotic relative efficiency and coverage probabilities

5.1. Continuous variables

We consider Model 1 (no covariates included) as the baseline model. We shall assess the performance of other models relative to the baseline model using the concept of asymptotic relative efficiency (ARE). This is just the limit as $k \to \infty$ of $\mathrm{Var}(\widehat{\rho}_1)/\mathrm{Var}\left(\widehat{\rho}_j\right) j = 2, 3, 4$.

Since we have closed forms for the variance expressions we provide ARE plots in Figs. 9–11 of $\mathrm{EFF} = \mathrm{Var}(\widehat{\rho}_1)/\mathrm{Var}(\widehat{\rho}_2)$ for $k = 10, 20, 50$.

We used the combinations $n = 2, 5, 10, 20, 50$; and $\rho = 0.1, 0.2, 0.5, 0.7, 0.9$. As can be seen, ARE levels are low for small values of ρ and small average cluster sizes. However, the ARE rapidly increases and plateaus at about 99%. In Table 4 we show the ARE values for the case of two covariates, one measured at the cluster level, and the other is varying within cluster. In this case we designated the covariate x_{ij} as a within-cluster constant covariate taking only two values either 0 or 1. The covariate z_{ij} is taken as within-cluster varying covariate taking the values $(-1, 1)$. Values of $\mathrm{Var}(\widehat{\rho}_1)/\mathrm{Var}(\widehat{\rho}_2)$ are given in Table 4. As can be seen the ARE follows the same pattern shown in Figs. 9–11.

5.2. Clustered binary data

For Model 1, the asymptotic variance of the ICC estimator for clustered binary data was given in closed form by Mak [24], and Zou & Donner [25]. We denote this by $\mathrm{Var}(\widehat{\rho}_b)$. When covariates are included, there is no closed form expression for the asymptotic variance of the ICC estimator. To evaluate the performance under the models proposed in Section 4, we shall again use the concept of ARE. The ARE was used by Sutradhar and Das [26] to evaluate the effect of the misspecification in the correlation structure on the efficiency loss in the estimated regression coefficient using the GEE. Chaganty and Joe [27] demonstrated that consistency and ARE of the regression coefficient estimators are guaranteed under exchangeable correlation structure.

In this section we use Monte-Carlo simulation of clustered binary data for a fully specified probability model, assuming

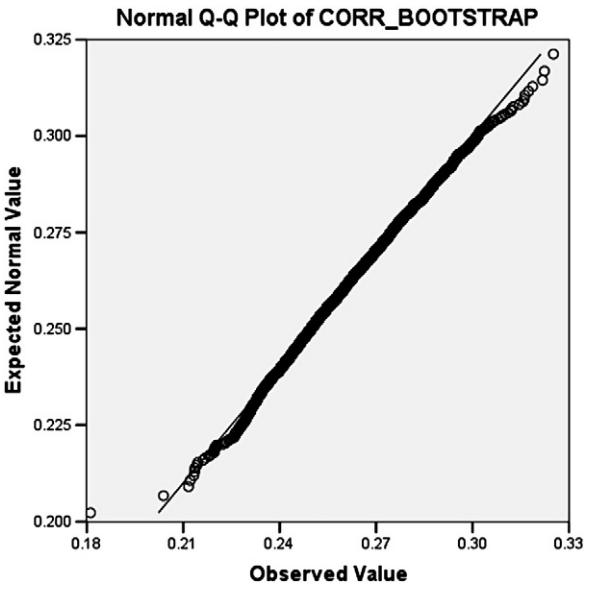

Fig. 1. Q–Q plot for 2000 bootstrap samples of ICC based on Model 1 (no covariates).

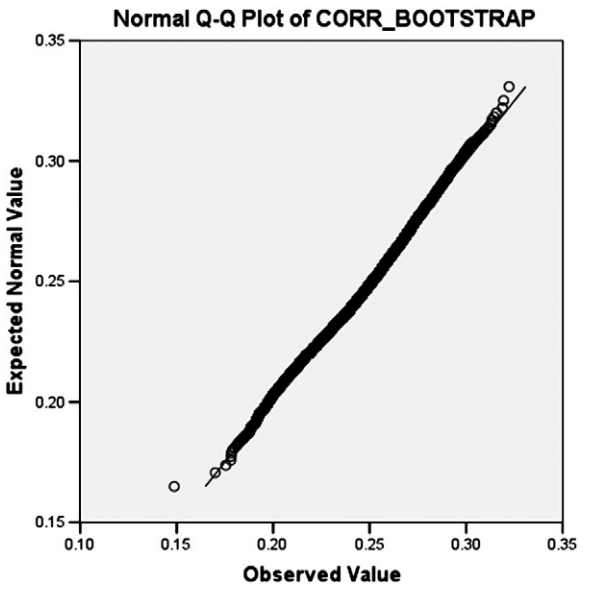

Fig. 2. Q–Q plot for 2000 bootstrap samples of ICC based on Model 2 (cluster level covariate is mother's systolic blood pressure levels).

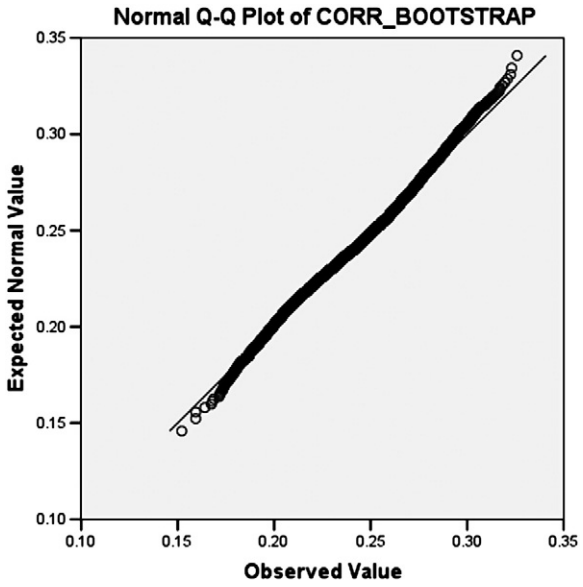

Fig. 3. Q–Q plot of 2000 bootstrap samples. Response measured on continuous scale, one covariate measured at the within cluster level (subject's diastolic blood pressure levels).

exchangeable correlation (common intraclass correlation). Thereafter evaluate the ARE of the estimated working correlation obtained by the GEE under the correct specification, when two measured covariates are included in the study design. We shall use the same covariate structure for x_{ij} and z_{ij} as indicated. We set $\rho = 0.3, 0.7$; $n = 5, 10$; and $k = 10, 20$.

The simulation steps are:

1. We define logit $(\pi_{ij}) = \beta_1 x_{ij} + \beta_2 z_{ij}$.
2. Set $\beta_1 = \beta_2 = 1$ and $\pi = \sum_{i=1}^{k} \sum_{j=1}^{n} \pi_{ij}/nk$, as in [26,27].

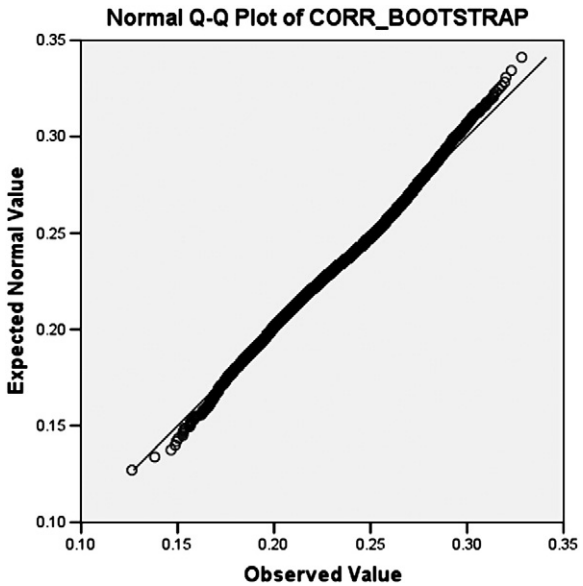

Fig. 4. Q–Q plot of 2000 bootstrap samples of ICC for the case of continuous response and two covariates (one measured at the cluster level and measured at the sib-within cluster level).

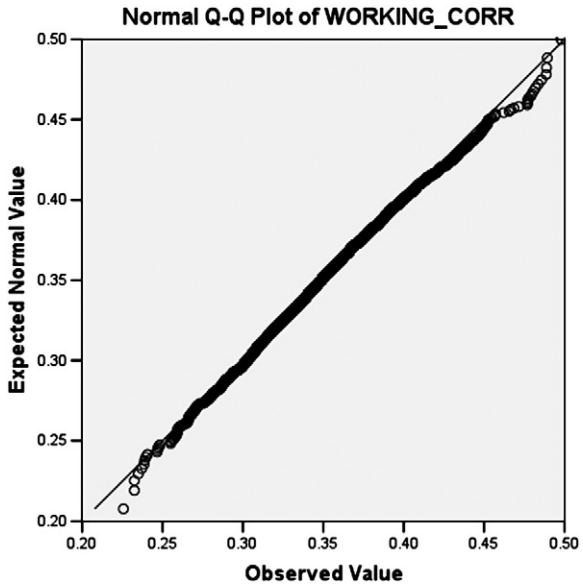

Fig. 5. Q–Q plot of 2000 bootstrap sample for the working correlation for binary response (GEE without covariates).

3. Generate pseudo random variables μ_{ij} from the beta-distribution with parameters: $a = \frac{\pi(1-\rho)}{\rho}$ and $b = \frac{(1-\rho)(1-\pi)}{\rho}$.
4. Generate Bernoulli $(1,\mu_{ij})$ for set $i = 1, 2, k$ and $j = 1, 2, \ldots n$. We therefore have a sequence of Beta-Bernoulli trials.
5. Use the GEE to fit 2000 simulated data sets generated under the above set-up for each combination (k,n,ρ), specifying exchangeable correlation.
6. Compute the mean and variance of the estimated working correlation $\widehat{\alpha}$, say denoted by $var(\widehat{\alpha})$.

The ARE is measured by the ratio $\text{EFF} = \text{Var}(\widehat{\rho}_b)/\text{Var}(\widehat{\alpha})$. For a limited number of parameter combinations we summarize the results in Table 5. As can be seen, there is a similar trend to

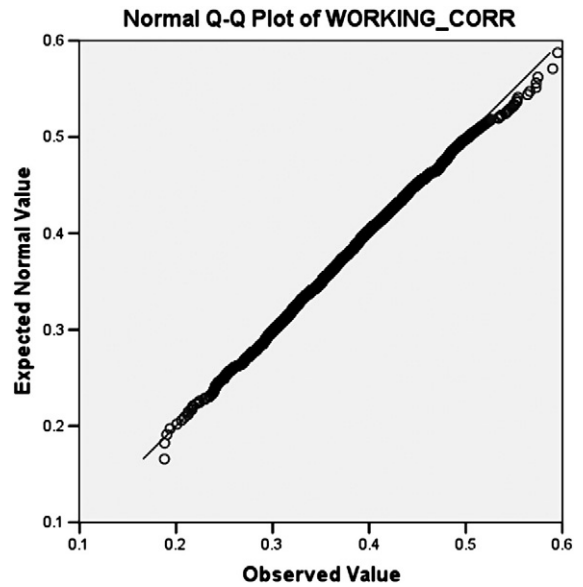

Fig. 6. Q–Q plot for 2000 bootstrap samples for the working correlation based on Model 2 (GEE with one cluster level covariate).

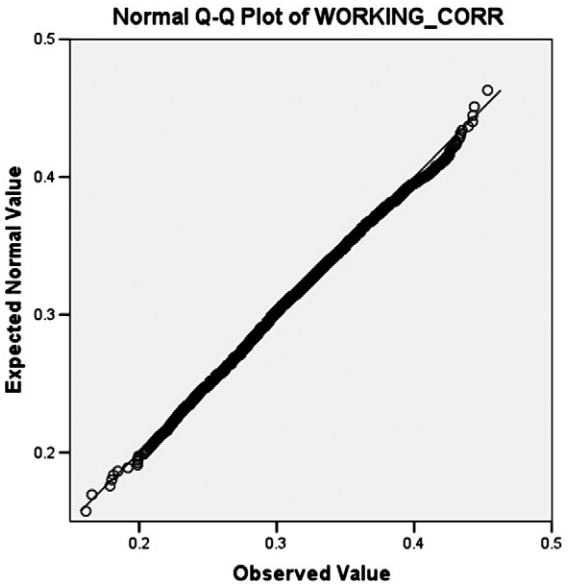

Fig. 7. Q–Q plot of 2000 bootstrap samples for the working correlation (GEE with one covariate measured at the subject within cluster level).

Table 4. Larger values of k, n, and ρ gave ARE values that are approximately 100 indicating that there is almost no efficiency loss in this case.

Thus inclusion of covariates, although produces different values of the estimated ICC under different models, the precisions of the estimates are still within acceptable level when compared to the model that does not include measured covariates.

We note that Crowder [28] demonstrated that the parameters involved in working correlation matrix are subject to "uncertainty of definition which can lead to a breakdown

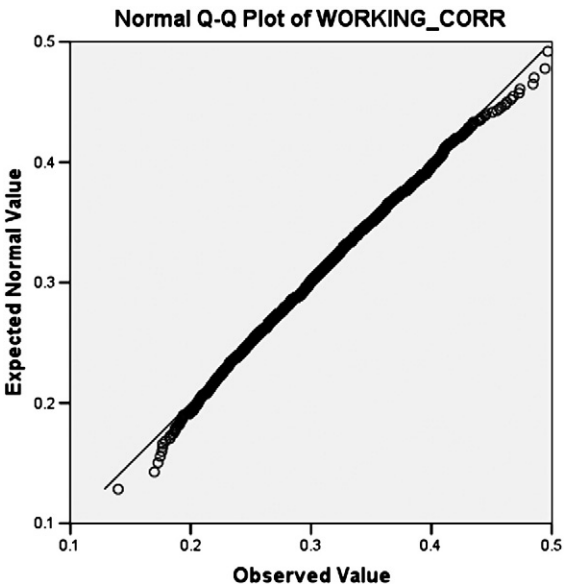

Fig. 8. Q–Q Plot of 2000 bootstrap samples for the working correlation (GEE with two covariates).

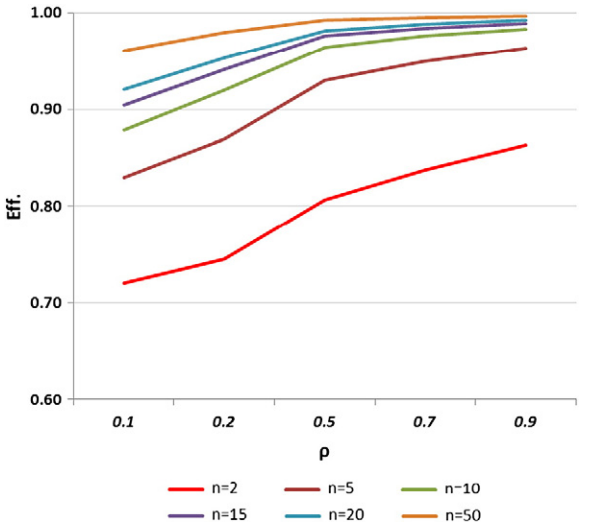

Fig. 9. Plots of ARE for $k = 10$.

of the asymptotic properties of the estimators". Therefore, in the above simulation, we simulated data under common (exchangeable) correlation to avoid the effect of misspecification on the estimation of the working correlation.

We note also that Yelland et al. [17] investigated the effect of covariate adjustment on the relative bias, which turned out to be severe under a variety of conditions. Thus far, we have demonstrated through the example that the covariate adjustment affected the estimated values of the ICC. It is therefore desirable to investigate the effect of the adjustment on the coverage probabilities of a 95% confidence interval. The results of the limited simulations are summarized in Table 6. For $100(1-\alpha)\%$ confidence interval with α typically 0.05, the coverage error will be $P_r(\rho_l < \rho < \rho_u) = (1-\alpha) + \varepsilon$, for some unknown constant ε, where $\varepsilon \to 0$ as the sample size gets larger. Allowing for the Monte Carlo error, we may declare the confidence intervals with coverage probabilities within $0.95 \pm 1.96\sqrt{\frac{0.95(.05)}{2000}} = (0.940, .959)$ limits as satisfactory. In

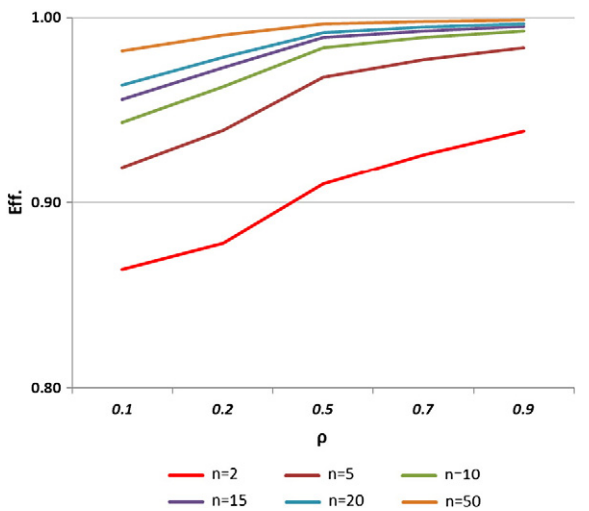

Fig. 10. Plots of ARE for $k = 20$.

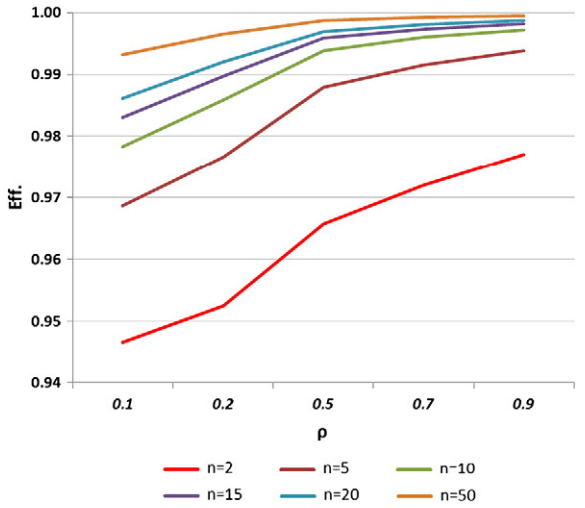

Fig. 11. Plots of ARE for $k = 50$.

Table 6, the limited number of simulated coverage indicates that while for the model with no covariates, the estimated probabilities are all within the allowable limits, few of such probabilities are outside the set limits for the model with two covariates. Improvement of the estimated coverage may be achieved by increasing the number of clusters.

6. Discussion

Previous work [29,30] has demonstrated that the efficiency of statistical estimation in the generalized linear mixed model, depends on both cluster size and the magnitude of the ICC. The synergetic effect of both cluster size and the ICC is known in the survey sample literature as the design effect (DEF) and is given by $\text{DEF} = 1 + (n - 1)\rho$, where n is the average cluster size. In studying the effect of covariate structure on statistical inferences arising from generalized linear models for clustered data we found it useful to distinguish between two types of covariates. The first type, a cluster constant or cluster-level covariate, does not vary between units within cluster, i.e. $x_{ij} = \bar{x}_i$ for $j = 1, 2, \ldots n_i$. An example would be mother or father blood pressure level in family studies. The other type, a within cluster covariate which varies across the subjects within a cluster is gender. We found that these sources of variations have varying effects on the estimated ICC. We also found that the effect of measured covariates is the same whether they are measured on the continuous or the categorical scale. Second; in fitting GEE to obtain estimates of the working correlation for clustered binary response data, using either the *logit-link* or *the log-link* has no effect on the estimated ICC. Our final recommendation is that in the design stage of studies where the sampling units are clusters of individuals, investigators should decide in advance on the number of measured covariates, together with the covariate structure. The available SAS macros can then be used to assess the uncertainty about the estimated ICC whether the response is measured on the continuous scale, or categorical binary.

Table 4
ARE of the ICC estimator: Two measured covariates and continuous response.

k	n	ρ	% EFF
10	2	0.0	96
10	2	0.3	97
10	2	0.5	98
10	4	0.0	97
10	4	0.3	99
10	4	0.5	99
20	2	0.0	98
20	2	0.3	99
20	2	0.5	99

Table 5
ARE of the GEE estimator in the case of two measured covariates.

k	n	ρ	% EFF
10	5	0.3	82
10	10	0.3	97
20	5	0.7	92
20	10	0.7	97

Table 6
Simulated coverage probabilities of 95% confidence intervals on the population ICC for binary response data.

		Model 1 (no covariates)		Model 4 (2-covariates)	
k	n	$\rho = 0.3$	0.7	$\rho = 0.3$	0.7
10	5	0.956	0.954	0.932[a]	0.948
10	10	0.951	0.953	0.940	0.985
20	5	0.955	0.951	0.948	0.962[a]
20	10	0.942	0.944	0.934[a]	0.953

[a] Means that the coverage is outside the desired limits.

Acknowledgment

The authors acknowledge the constructive comments made by the EIC, and three anonymous reviewers. All authors contributed to this work and agreed to the contents of the manuscripts.

Appendix. Derivation of the expectation of mean square between clusters in the case of two measured covariates

To obtain the estimators of the variance components σ_ϵ^2 and σ_τ^2 model (11) can be written as:

$$y = x_1\beta + x_2\tau + \epsilon \equiv xb + \epsilon$$

where $y = (y_1, y_2, \ldots y_k)', y_i = \left(y_{i1}, y_{i2}, \ldots y_{in_i}\right)'$

$$x_1 = \begin{bmatrix} e_1 & x_1 & z_1 \\ e_2 & x_2 & z_2 \\ \vdots & \vdots & \vdots \\ e_k & x_k & z_k \end{bmatrix}, x_2 = \begin{bmatrix} e_1 & 0 & \ldots & 0 \\ 0 & e_2 & \ldots & 0 \\ \vdots & & & \\ 0 & 0 & \ldots & e_k \end{bmatrix}$$

$$\beta = \begin{bmatrix} \mu_1 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \tau = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix}, \quad b = \begin{bmatrix} \beta \\ \tau \end{bmatrix}$$

$$x_i = \left[(x_{i1} - \bar{x}), (x_{2i} - \bar{x}), \ldots \left(x_{in_i} - \bar{x}\right)\right]'$$

$$z_i = \left[(z_{i1} - \bar{z}), (z_{i2} - \bar{z}), \ldots z_{in_i} - \bar{z}\right)]'$$

and finally $x = (x_1\ x_2)'$.

Following Searle et al. [11] we can show that:

$$E(SSW) = (N-k-2)\sigma^2.$$

However E(SSB) may be obtained in several steps. Omitting details, we note that

$$E(MSB) = \frac{1}{k-1}\left[\sigma_\tau^2 \,\mathrm{Tr}\left(x'_2\left(I - x_1\left(x'_1x_1\right)^{-1}x'_1\right)x_2\right)\right.$$
$$\left.+\sigma_e^2[\mathrm{rank}(x)-\mathrm{rank}(x_1)]\right]$$
$$x'_1x_1 = \begin{bmatrix} N & 0 & 0 \\ 0 & S_{xx} & S_{xz} \\ 0 & S_{xz} & S_{zz} \end{bmatrix}$$
$$x'_2x_2 = \begin{bmatrix} n_1 & 0 & \dots & 0 \\ 0 & n_2 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & n_k \end{bmatrix}$$
$$\left(x'_1x_1\right)^{-1} = \begin{bmatrix} N^{-1} & 0 & 0 \\ 0 & \dfrac{S_{zz}}{c} & -\dfrac{S_{xz}}{c} \\ 0 & -\dfrac{S_{xz}}{c} & \dfrac{S_{xx}}{c} \end{bmatrix}.$$

The corrected sum of squares $S_{xx} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x})^2$

$$S_{zz} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij}-\bar{z}\right)^2$$
$$s_{xz} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}\right)\left(z_{ij}-\bar{z}\right), \text{and}$$
$$c = s_{xx}s_{zz}-s_{xz}^2.$$

Moreover, $\mathrm{rank}(x) = \mathrm{rank}(x'x) = k+2$, $\mathrm{rank}(x_1) = 3$, $\mathrm{Trace}(x'_2x_2) = N$

$$\mathrm{Trace}\left(x'_2x_1\left(x'_1x_1\right)^{-1}x'_1x_2\right)$$
$$= \sum_{i=1}^{k}\frac{n_i^2}{N} + \frac{1}{c}\left[s_{zz}\sum_{i=1}^{k}n_i^2(\bar{x}_i-\bar{x})^2 -\right.$$
$$\left.-2s_{xz}\sum_{i=1}^{k}n_i^2(\bar{x}_i-\bar{x})(\bar{z}_i-\bar{z}) + s_{xx}\sum_{i=1}^{k}n_i^2(\bar{z}_i-\bar{z})^2\right].$$

Finally we obtain:

$$E(MSB) = \sigma_\epsilon^2 + n_{02}\sigma_\tau^2$$

where

$$n_{02} = n_0 - \frac{1}{k-1}\left[\frac{1}{c}\left\{s_{zz}\sum_{i=1}^{k}n_i^2(\bar{x}_i-\bar{x})^2 -\right.\right.$$
$$\left.\left.-2s_{xz}\sum_{i=1}^{k}n_i^2(\bar{x}_i-\bar{x})(\bar{z}_i-\bar{z}) + s_{xx}\sum_{i=1}^{k}n_i^2(\bar{z}_i-\bar{z})^2\right\}\right].$$

References

[1] Higgins M, Keller J. Familial occurrence of chronic respiratory disease and familial resemblance in ventilatory capacity. J Chronic Dis 1975;28:239–51.
[2] Shoukri MM, Ward RH. Use of linear models to estimate genetic parameters and measures of familial resemblance in man. J R Stat Soc Ser C 1989;3:467–79.
[3] Bartko JJ. The intraclass correlation reliability coefficients. Psychol Bull 1966;83:762–5.
[4] Shoukri MM, Ward RH. Estimation of intraclass correlation. Commun Stat Theory Methods 1985;13(10):1239–55.
[5] Donner A. A review of inference procedures for the intraclass correlation in the one-way random effects model. Int Stat Rev 1986;54(1):67–82.
[6] Efron B, Tibshirani RJ. An introduction to bootstrap. Chapman and Hall; 1993.
[7] Shao J, Tu D. The jackknife and bootstrap. NY: Springer; 1996.
[8] Raudenbush SW, Bryk SA. Hierarchical linear model: applications and data analysis methods. 2nd ed. London, UK: Sage Publications; 2002.
[9] Donner A, Koval JJ. The large sample variance of intraclass correlation. Biometrika 1981;67:719–22.
[10] Smith CAB. Estimating genetic correlation. Ann Hum Genet 1980;43:265–84.
[11] Searle RS, Casella G, McCulloch CE. Variance components. NY: Wiley; 1992.
[12] Shoukri MM, Asyali MH, Walter SW. Issues of cost and efficiency in the design of reliability studies. Biometrics 2003;59(4):1109–14.
[13] Shoukri MM, Asyali MH, Donner A. Sample size requirements for the design of reliability study: review and new results. Stat Methods Med Res 2004;13:251–71.
[14] Mian IUH, Shoukri MM. Statistical analysis of intraclass correlation from multiple samples with applications to arterial blood pressure data. Stat Med 1997;16(13):1497–514.
[15] Stanish WM, Taylor N. Estimation of the intraclass correlation coefficient for the analysis of covariance. Am Stat 1983;37(3):221–4.
[16] Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological development. Am J Public Health 2004;94:423–32.
[17] Yelland LN, Salter AB, Ryan P, Laurence CO. Adjusted intra-class correlation coefficients for binary data: methods and estimates from a cluster-randomized trial in primary care. Clin Trials 2011;8(1):48–58.
[18] Liang K-Y, Zeger SL. Longitudinal data using generalized linear models. Biometrika 1986;73:13–22.
[19] Miall WE, Oldham PO. A study of arterial blood pressure and its inheritance in a sample of the general population. Clin Sci 1955;14:459–87.
[20] Field CA, Welsh AH. Bootstrapping clustered data. J R Stat Soc B 2007;69(Part 3):369–90.
[21] Roberts JK, Xitao Fan. Bootstrapping within the multilevel/hierarchical linear modeling framework: a primer for use with SAS and SPLUS. Mult Linear Regression Viewpoints 2004;30(1):23–34.
[22] SAS/STAT version 9.3. SAS Institute: Rayleigh, North Carolina, USA; 2004.
[23] D' Agostino R, Belanger A, D' Agostino JR. A suggestion for powerful and informative tests of normality. Am Stat 1990;44:316–21.
[24] Mak TK. Analyzing intraclass correlation for dichotomous variables. J R Stat Soc C 1988;37(3):344–52.
[25] Zou A, Donner A. Confidence interval estimation of the intraclass correlation coefficient for binary outcome data. Biometrics 2004;60(3):807–11.
[26] Sutradhar BC, Das K. On the efficiency of regression estimators in generalized linear models for longitudinal data. Biometrika 1999;86(2):459–65.
[27] Chaganty NR, Joe H. Efficiency of generalized estimating equations for binary responses. J R Stat Soc B 2004;66(4):851–60.
[28] Crowder M. On the use of a working correlation matrix in using generalized linear models for repeated measures. Biometrika 1995;82(2):407–10.
[29] TanHave TR, Landis JR, Weaver SL. Association models for periodontal disease progression: a comparison of methods for clustered binary data. Stat Med 1995;14:413–29.
[30] Neuhaus JM, Lesperance ML. Estimation efficiency in a binary mixed-effects model setting. Biometrika 1996;83:441–6.