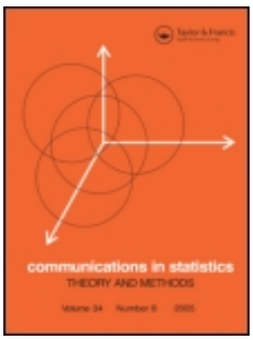

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Resampling method to estimate intra-cluster correlation for clustered binary data

Hrishikesh Chakraborty & Pranab K. Sen

To cite this article: Hrishikesh Chakraborty & Pranab K. Sen (2016) Resampling method to estimate intra-cluster correlation for clustered binary data, Communications in Statistics - Theory and Methods, 45:8, 2368-2377, DOI: 10.1080/03610926.2013.870202

To link to this article: https://doi.org/10.1080/03610926.2013.870202

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2016, VOL. 45, NO. 8, 2368–2377
http://dx.doi.org/10.1080/03610926.2013.870202

Taylor & Francis
Taylor & Francis Group

Resampling method to estimate intra-cluster correlation for clustered binary data

Hrishikesh Chakraborty[a] and Pranab K. Sen[b]

[a]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA; [b]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

ABSTRACT
Various methods have been proposed to estimate intra-cluster correlation coefficients (ICCs) for correlated binary data, and many are very sensitive to the type of design and underlying distributional assumptions. We proposed a new method to estimate ICC and its 95% confidence intervals based on resampling principles and U-statistics, where we resampled with replacement pairs of individuals from within and between clusters. We concluded from our simulation study that the resampling-based estimates approximate the population ICC more precisely than the analysis of variance and method of moments techniques for different event rates, varying number of clusters, and cluster sizes.

ARTICLE HISTORY
Received 24 April 2013
Accepted 25 November 2013

KEYWORDS
Resampling; Intra-cluster correlation; Clustered binary data.

MATHEMATICS SUBJECT CLASSIFICATION
93A30

1. Background

A cluster usually refers to any naturally occurring group of persons, animals, or other elements that are the object of a study. Cluster randomized trials (CRTs) are increasingly being used to evaluate community interventions where the randomization unit is not an individual, but rather a cluster. When the subjects within a cluster are correlated, the resulting data are correlated, and researchers need to account for the correlation during sample size calculation and analysis. If the researcher uses standard sample size estimation methods, the power will likely be far too low. Increasing the average cluster size will increase statistical power, but only to a certain point (Chakraborty and Ray, 2015; Donner, 1998), after which, the increase in power is negligible.

The intra-cluster correlation coefficient (ICC) measures the degree to which responses within the same cluster are correlated. It is important to obtain a precise estimate of ICC to maintain the power of a study (Chakraborty et al., 2009a; Chakraborty et al., 2009b). The ICC estimate depends on the event rate, event rate variability between clusters, cluster size, cluster size variability, and number of clusters (Chakraborty et al., 2009a; Donner, 1998; Chakraborty, 2008). Researchers have developed different methods of calculating ICC. The most popular is the analysis of variance (ANOVA) method, derived by Fleiss, which uses mean square values from a one-way ANOVA (Fleiss, 1981). Other estimators have been developed based on direct probabilistic methods (Fleiss and Cuzick, 1979; Mak, 1988), direct computation of correlation (Donner, 1986; Schouten, 1986; Karlin et al., 1981; Lipsitz et al., 1994), the method of moments (MM) estimation (Kleinman, 1973; Tamura and Young, 1987; Williams, 1982; Yamamoto and

CONTACT Hrishikesh Chakraborty rishic@mailbox.sc.edu Department of Epidemiology and Biostatistics, Arnold School of Public Health, The University of South Carolina, 800 Sumter Street, Suite 208-D, Columbia, SC 29208, USA.

Yanagimoto, 1992), the use of quadratic estimating equations (Crowder, 1978), beta-binomial maximum likelihood estimation (Lunn and Davies, 1998; Mak, 1988), quasi-likelihood estimation (Donner, 1986; Moore and Tsiatis, 1991), and the extended quasi-likelihood estimator (Carroll and Ruppert, 1988; McCullagh and Nelder, 1989; Nelder and Pregibon, 1987). Ridout et al. (1999) used a simulation exercise to compare several of these methods for binary clustered outcomes, concluding that the best estimators were the ANOVA estimators, a few of the MM estimators, and an estimator with a direct probabilistic interpretation. The "best" in this case means they were accurate, were the least biased, and had smaller standard deviations.

In this article, we present a newly developed method to estimate ICC and its 95% confidence interval (CI) for binary variables using a resampling method (Efron and Tibshirani, 1993) and *U*-statistics (Lee, 1990). We compared the simulated ICC estimates of this new method with the ANOVA and MM estimators (which are the most popular methods), for different cluster sizes, event rates, and numbers of clusters.

2. Methodology

Suppose that there are k clusters of individuals and the pth cluster contains n_p individuals ($n_p \geq 2$). Each individual X_{pq} ($p = 1, \ldots \ldots k$, $q = 1, \ldots \ldots n_p$) has a binary response. Let $n = \sum_{p=1}^{k} n_p$ denote the total number of individuals in a study. We assume that the probability of success, or event rate, is the same for all individuals across clusters; therefore $\mathrm{P}\{X_{pq} = 1\} = \alpha$ for all p and q.

The ANOVA estimator of ICC for the binary data X_{pq} is given by

$$\hat{\rho}_{\mathrm{ANOVA}} = \frac{\mathrm{MSB} - \mathrm{MSW}}{\mathrm{MSB} + (n_0 - 1)\mathrm{MSW}}, \quad \text{where } n_0 = \frac{1}{k-1}\left[n - \sum_{p=1}^{k} \frac{n_p^2}{n}\right]$$

And the MM ICC estimate is defined as

$$\hat{\rho} = \frac{S_w - \tilde{\pi}_w(1-\tilde{\pi}_w)\sum_{p=1}^{k} \frac{w_p(1-w_p)}{n_p}}{\tilde{\pi}_w(1-\tilde{\pi}_w)\left[\sum_{p=1}^{k} w_p(1-w_p) - \sum_{p=1}^{k} \frac{w_p(1-w_p)}{n_p}\right]}.$$

The observed proportion of successes in the pth group is denoted as $\tilde{\pi}_p$, $\tilde{\pi}_w = \sum_{p=1}^{k} w_p \tilde{\pi}_p$, *and* $S_w = \sum_{p=1}^{k} w_p(\tilde{\pi}_p - \tilde{\pi}_w)^2$, where the w_p is the weight satisfying $\sum_{p=1}^{k} w_p = 1$. By equating $\tilde{\pi}_w$ and S_w to their expected values under the assumption of common correlation, a series of weighted moment estimators were derived.

To develop the resampling ICC method, we made three assumptions: first, that the responses of individuals from two different clusters are independent (which is the usual assumption used in designing CRTs to minimize contamination). Second, we assumed that the responses of two individuals from the same cluster are correlated, which is the usual assumption when the clusters are defined as distinct geographical areas, physically separate, and without a common border (Carlo et al., 2010). This correlation is assumed to exist because the residents share common factors that may influence the outcome, such as the same genetics, households, doctors, or resources in a geographical area. Third, we assumed that there is no higher order correlation among individuals within clusters. Therefore, the average correlation among a pair of individuals (X_{ij}, X_{il}) ($j \neq l$) from within cluster is ρ, also called the ICC. The correlation among pairs of individuals who are members of different clusters is zero because

we assume that the clusters are independent. Let a U-statistic, $U_1 = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} I\left(x_{ij}=1\right)$, estimate the overall probability of success, where I defines the indicator function. The mean of U_1 is $E\left(U_1\right) = E[\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} I(x_{ij}=1)] = \alpha = \theta_1$ and the variance of U_1 is

$$V\left(U_1\right) = V\left[\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} I\left(x_{ij}=1\right)\right] = \alpha\left(1-\alpha\right)\left[\frac{1}{n}+\frac{\rho}{n^2}\sum_{i=1}^{k} n_i\left(n_i-1\right)\right] \tag{1}$$

From the properties of U-statistics (Lee, 1990), we know that $\mathrm{Z}_1 = \sqrt{n}\left(U_1-\theta_1\right)$ $N\left\{0, V\left(U_1\right)\right\}$

$$\text{i.e., } U_1 = \frac{Z_1}{\sqrt{n}} + \theta_1 \tag{2}$$

If we draw two samples from a cluster with replacement and define a U-statistic T_{W} as the overall probability of drawing a given within-cluster pair, the within-cluster pair samples with replacement will generate a total of $\sum_{i=1}^{k} n_i(n_i-1)$ pairs of samples, with possible outcomes of $(x_{ij}=1, x_{ij'}=1)$, $(x_{ij}=1, x_{ij'}=0), (x_{ij}=0, x_{ij'}=1)$, and $(x_{ij}=0, x_{ij'}=0)$.

$$T_{\mathrm{W}} = \left[\sum_{i=1}^{k} n_i(n_i-1)\right]^{-1} \times \left[\begin{array}{l} \sum_{i=1}^{k}\sum_{j,j'=1;\, j\neq j'}^{n_i} I\left(x_{ij}=1 \text{ and } x_{ij'}=1\right) + \sum_{i=1}^{k}\sum_{j,j'=1;\, j\neq j'}^{n_i} I\left(x_{ij}=0 \text{ and } x_{ij'}=0\right) \\ -\sum_{i=1}^{k}\sum_{j,j'=1;\, j\neq j'}^{n_i} I\left(x_{ij}=1 \text{ and } x_{ij'}=0\right) - \sum_{i=1}^{k}\sum_{j,j'=1;\, j\neq j'}^{n_i} I\left(x_{ij}=0 \text{ and } x_{ij'}=1\right) \end{array}\right]$$

The expected value of T_{W} is

$$E\left(T_{\mathrm{W}}\right) = \alpha^2 + (1-\alpha)^2 - 2\alpha(1-\alpha) + 4\rho\alpha(1-\alpha) = \theta_2(\text{say}) \tag{3}$$

and the variance of T_{W} is

$$V\left(T_{\mathrm{W}}\right) = \frac{\alpha\left(1-\alpha\right)}{\sum_{i=1}^{k} n_i^2 - n} \times \left[\begin{array}{l} (1+\alpha-\rho\alpha)\left\{\alpha+\rho\left(1-\alpha\right)\right\} + (1-\alpha+\rho\alpha)\left\{2-\alpha-\rho(1-\alpha)\right\} \\ +2\left(1+\rho\right)\left\{1-\alpha(1-\alpha)\left(1+\rho\right)\right\} \end{array}\right] \tag{4}$$

From the properties of U-statistics, we can say $\mathrm{Z}_2 = \sqrt{n}\left(T_{\mathrm{w}}-\theta_2\right) \xrightarrow{d} N\left\{0, V\left(T_{\mathrm{w}}\right)\right\}$

$$\text{i.e., } T_{\mathrm{W}} = \frac{Z_2}{\sqrt{n}} + \theta_2 \tag{5}$$

Now, we draw two samples from two different clusters with replacement and define a U-statistics T_{B} as the overall probability of drawing a given between-cluster pair. The between-cluster pairs sampled with replacement will generate total of $\{n(n-1) - \sum_{i=1}^{k} n_i(n_i-1)\}$pairs of samples, with possible outcomes of $(x_{i'j}=1, x_{i'j}=1)$, $(x_{i'j}=1, x_{i'j}=0)$, $(x_{i'j}=0, x_{i'j}=1)$, *and* $(x_{i'j}=0, x_{i'j}=0)$. We can write T_{B} as

$$T_{\mathrm{B}} = \left[n(n-1) - \sum_{i=1}^{k} n_i(n_i-1)\right]^{-1}$$

$$\times \left[\begin{array}{l} \sum_{i,i'=1;i\neq i'}^{k} \sum_{j,j'=1}^{n_i} I\left(x_{ij}=1 \text{and} x_{i'j'}=1\right) + \sum_{i,i'=1;i\neq i'}^{k} \sum_{j,j'=1}^{n_i} I\left(x_{ij}=0 \text{and} x_{i'j'}=0\right) - \\ \sum_{i,i'=1;i\neq i'}^{k} \sum_{j,j'=1}^{n_i} I\left(x_{ij}=1 \text{and} x_{i'j'}=0\right) - \sum_{i,i'=1;i\neq i'}^{k} \sum_{j,j'=1}^{n_i} I\left(x_{ij}=0 \text{and} x_{i'j'}=1\right) \end{array} \right]$$

The expected value of T_B is

$$E(T_B) = \alpha^2 + (1-\alpha)^2 - 2\alpha(1-\alpha) = \theta_3 (\text{say}) \tag{6}$$

and the variance of T_B is

$$V(T_B) = \frac{1}{n^2 - \sum_{i=1}^{k} n_i^2} + \{2\alpha(1-\alpha)+1\}^2 \tag{7}$$

From the properties of U-statistics, we can write $Z_3 = \sqrt{n}\,(T_B - \theta_3) \xrightarrow{d} N\{0, V(T_B)\}$

$$\text{i.e., } T_B = \frac{Z_3}{\sqrt{n}} + \theta_3 \tag{8}$$

If all the within- and between-cluster individuals are independent, then any pair of samples will have the same probability based on their success or failure status and, therefore, $E(T_B)$ should be equal to $E(T_W)$. The difference between $E(T_B)$ and $E(T_W)$ are due to within-cluster correlated observations. Using (3) and (6), we have$E(T_W) - E(T_B) = 4\alpha(1-\alpha)\rho$. Therefore, the estimate of ICC is $\hat{\rho} = \frac{T_W - T_B}{4U_1(1-U_1)}$. The expected value of $\hat{\rho}$ is $E(\hat{\rho}) = \rho$. Therefore, $\hat{\rho}$ is an unbiased estimator of ρ.

It is important to estimate the CI of ICC estimates to aid in designing studies where investigators want to balance power and expense. Using the CI for ICC, one can easily calculate the ranges of sample sizes required to have a sufficiently powerful study. We provided a formula to calculate the variance of our estimated ICC to facilitate the CI estimation. We showed that $\hat{\rho}$ is a function of U_1, T_W, and T_B and using (2), (5), and (8), the function can be defined as $\hat{\rho} = g(U_1, T_W, T_B) = g\left(\theta_1 + \frac{Z_1}{\sqrt{n}}, \theta_2 + \frac{Z_2}{\sqrt{n}}, \theta_3 + \frac{Z_3}{\sqrt{n}}\right)$.

A Taylor series expansion of the function $g(U_1, T_W, T_B)$ about the values of $\{E(U_1), E(T_W), E(T_B)\}$is

$$\begin{aligned} \hat{\rho} = \rho &+ \frac{Z_1}{\sqrt{n}} \frac{1}{4\theta_1(1-\theta_1)} - \frac{Z_2}{\sqrt{n}} \frac{1}{4\theta_1(1-\theta_1)} + \frac{Z_3}{\sqrt{n}} \frac{(\theta_2-\theta_3)(2\theta_1-1)}{4\{\theta_1(1-\theta_1)\}^2} \\ &+ O_p\left(n^{-1}\right) = \rho + \underset{\sim}{a}' \underset{\sim}{Z} + O_p\left(n^{-1}\right) \end{aligned} \tag{9}$$

where $\underset{\sim}{a}$ is the coefficient matrix and

$$\underset{\sim}{Z} = \begin{bmatrix} Z_1 \\ Z_2 \\ Z_3 \end{bmatrix} \xrightarrow{d} N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{Bmatrix} V(Z_1) & \text{Cov}(Z_1, Z_2) & \text{Cov}(Z_1, Z_3) \\ \text{Cov}(Z_1, Z_2) & V(Z_2) & \text{Cov}(Z_2, Z_3) \\ \text{Cov}(Z_1, Z_3) & \text{Cov}(Z_2, Z_3) & V(Z_3) \end{Bmatrix} \right]$$

To calculate the variance of $\hat{\rho}$, use the second derivatives for expression (9). Therefore,

$$\begin{aligned} V(\hat{\rho}) = &\frac{V(T_W)}{\left\{4U_1(1-U_1)\sqrt{n}\right\}^2} + \frac{V(T_B)}{\left\{4U_1(1-U_1)\sqrt{n}\right\}^2} + \frac{V(U_1)(T_W - T_B)^2(1-2U_1)^2}{16n\{U_1(1-U_1)\}^4} \\ &- \frac{2\text{Cov}(T_W, T_B)}{\left\{4U_1(1-U_1)\sqrt{n}\right\}^2} - \frac{2\text{Cov}(T_W, U_1)(T_W - T_B)(1-2U_1)}{\left\{4U_1(1-U_1)\sqrt{n}\right\}^2 U_1(1-U_1)} \end{aligned}$$

$$+\frac{2\mathrm{Cov}\left(T_{\mathrm{B}}, U_1\right)\left(T_{\mathrm{W}}-T_{\mathrm{B}}\right)\left(1-2U_1\right)}{\left\{4U_1(1-U_1)\sqrt{n}\right\}^2 U_1(1-U_1)}+O_p\left(n^{-1}\right)$$

We assumed that there is no higher order correlation among individuals within and between clusters. Therefore, ignoring the higher order terms, we have

$$V\left(\hat{\rho}\right)=\frac{1}{16nU_1^2(1-U_1)^2}V\left(T_{\mathrm{W}}\right)+\frac{1}{16nU_1^2(1-U_1)^2}V\left(T_{\mathrm{B}}\right)+\frac{\left(T_{\mathrm{W}}-T_{\mathrm{B}}\right)^2\left(1-2U_1\right)^2}{16n\left\{U_1(1-U_1)\right\}^4}V\left(U_1\right)$$

Using results from (1), (4), and (7), we have

$$V\left(\hat{\rho}\right)=\frac{1}{16nU_1^2(1-U_1)^2}\times\left[\begin{array}{l}\frac{1}{n^2-\sum_{i=1}^{k}n_i^2}+\{2\alpha(1-\alpha)+1\}^2\\ +\frac{\alpha(1-\alpha)}{\sum_{i=1}^{k}n_i^2-n}\left[\begin{array}{l}(1+\alpha-\rho\alpha)\{\alpha+\rho(1-\alpha)\}\\ \quad+(1-\alpha+\rho\alpha)\{2-\alpha-\rho(1-\alpha)\}\\ +2(1+\rho)\{1-\alpha(1-\alpha)(1+\rho)\}\end{array}\right]\\ +\frac{\alpha(1-\alpha)\left(T_{\mathrm{W}}-T_{\mathrm{B}}\right)^2\left(1-2U_1\right)^2}{\{U_1(1-U_1)\}^2}\left[\frac{1}{n}+\frac{\rho}{n^2}\sum_{i=1}^{k}n_i\left(n_i-1\right)\right]\end{array}\right]$$

We used normal approximation to calculate the 95% CI of ρ.

3. Simulation

We used Monte Carlo simulation to create datasets with fixed population ICC values for different numbers of clusters, cluster sizes, and event rates. We first defined the number of clusters, sizes of each cluster, probability of success (also called event rate), and variations in event rate for each cluster. Also, we defined the cluster size variation as the maximum variation in cluster size; for example, if the average cluster size is 100 and the cluster size variation is 25% then the clusters can vary from 75 to 125 members. Similarly, we defined the event rate variations for clusters: if the event rate is 0.10 and the event rate variation is 25% then the individual cluster event rate can be between 0.075 and 0.125. Therefore, each cluster will have a randomly assigned cluster size and event rate based on these values. For each cluster, we generated correlated binary data with the randomly generated cluster size and event rate using the method described by Lunn and Davies (1998). We generated data for scenarios in which there were 10, 20, 30, 60, and 90 clusters. We specified overall event rates 0.2 and 0.4, with 25% variation between clusters, and we specified three different cluster sizes (10, 25, 50, 100, and 200), with 25% cluster size variation. All the combinations of event rates, numbers of clusters, and cluster sizes were evaluated for three different population ICC values (0.05, 0.10, and 0.15). We used ICCs ranging from 0.05 to 0.15 because previous studies have found similar ICC ranges (Campbell et al., 2000a,, 2000b).

We created 54 independent datasets for each unique combination of population ICC, event rate, cluster size, and fixed number of clusters. We repeated this process 5,000 times; therefore, in total, we created 270,000 independent datasets. We used each simulated dataset to estimate ICC using the resampling, ANOVA, MM methods (Kleinman, 1973), and 95% CIs using the resampling method. All simulations and analysis were conducted using SAS version 9.2 (2008).

Table 1. ICC estimates and confidence interval for event rate of 0.2 by resampling method.

			Estimated ICC values		
				95% confidence interval	
Total clusters	Cluster size	Assumed ICC value	Mean	Lower	Upper
10	10	0.05	0.046	0	0.262
		0.10	0.087	0	0.309
		0.15	0.129	0	0.357
	25	0.05	0.048	0	0.182
		0.10	0.090	0	0.227
		0.15	0.134	0	0.274
	50	0.05	0.046	0	0.140
		0.10	0.088	0	0.184
		0.15	0.131	0.031	0.230
	100	0.05	0.050	0	0.116
		0.10	0.094	0.026	0.161
		0.15	0.136	0.066	0.205
	200	0.05	0.050	0.004	0.097
		0.10	0.093	0.045	0.140
		0.15	0.136	0.087	0.184
20	10	0.05	0.050	0	0.197
		0.10	0.096	0	0.245
		0.15	0.141	0	0.291
	25	0.05	0.052	0	0.143
		0.10	0.097	0.005	0.190
		0.15	0.144	0.050	0.237
	50	0.05	0.052	0	0.116
		0.10	0.097	0.031	0.163
		0.15	0.143	0.076	0.210
	100	0.05	0.053	0.008	0.099
		0.10	0.099	0.052	0.145
		0.15	0.144	0.097	0.191
	200	0.05	0.054	0.022	0.086
		0.10	0.100	0.067	0.132
		0.15	0.146	0.113	0.179
30	10	0.05	0.054	0	0.172
		0.10	0.101	0	0.220
		0.15	0.148	0.027	0.268
	25	0.05	0.054	0	0.128
		0.10	0.100	0.026	0.175
		0.15	0.148	0.072	0.223
	50	0.05	0.053	0.001	0.106
		0.10	0.100	0.047	0.153
		0.15	0.147	0.093	0.200
	100	0.05	0.055	0.017	0.092
		0.10	0.101	0.064	0.139
		0.15	0.148	0.111	0.186
	200	0.05	0.055	0.029	0.081
		0.10	0.103	0.076	0.129
		0.15	0.150	0.123	0.177
60	10	0.05	0.054	0	0.137
		0.10	0.103	0.020	0.185
		0.15	0.151	0.068	0.234
	25	0.05	0.056	0.004	0.108
		0.10	0.105	0.052	0.157
		0.15	0.153	0.100	0.206
	50	0.05	0.055	0.018	0.092
		0.10	0.104	0.067	0.141
		0.15	0.152	0.115	0.189
	100	0.05	0.055	0.029	0.081
		0.10	0.104	0.077	0.130
		0.15	0.152	0.126	0.178
	200	0.05	0.057	0.039	0.075
		0.10	0.106	0.087	0.124
		0.15	0.154	0.136	0.173

(*Continued on next page*)

Table 1. ICC estimates and confidence interval for event rate of 0.2 by resampling method. (*Continued*)

			Estimated ICC values		
				95% confidence interval	
Total clusters	Cluster size	Assumed ICC value	Mean	Lower	Upper
90	10	0.05	0.056	0	0.124
		0.10	0.106	0.038	0.173
		0.15	0.155	0.087	0.223
	25	0.05	0.057	0.014	0.099
		0.10	0.106	0.063	0.148
		0.15	0.155	0.112	0.197
	50	0.05	0.056	0.026	0.086
		0.10	0.105	0.075	0.135
		0.15	0.153	0.123	0.184
	100	0.05	0.056	0.035	0.078
		0.10	0.105	0.084	0.127
		0.15	0.154	0.133	0.176
	200	0.05	0.057	0.042	0.072
		0.10	0.106	0.091	0.121
		0.15	0.155	0.140	0.170

Note: Assumed 25% cluster size and event rate variations among clusters.

4. Results and discussion

We calculated the mean of 5,000 estimates for each unique combination of event rate, cluster size, and number of clusters. Table 1 presents the estimated ICC and its 95% CIs using the resampling method for population ICC values of 0.05, 0.10, and 0.15, with an average event rate of 0.2, where the event rate varies from cluster to cluster by 25% for 30, 60, and 90 clusters and the average cluster sizes is 50, 100, and 200 with 25% cluster size variation. We found that the estimated ICC using the resampling method is very close to the population ICC for all different cluster sizes and number of clusters. As the total number of clusters increases, the 95% confident interval of ICC becomes narrower. For example, using the totals for 30, 60, and 90 clusters with cluster size 50, the 95% CI estimates are (0.001–0.106), (0.018–0.092), and (0.026–0.086), respectively, for a population ICC value of 0.05. We observed a coverage rate of 89–98% for our estimated CIs. We observed a similar pattern for cluster size of 100 and 200 for different number of clusters and different population ICC values. As the cluster sizes increased, for a fixed number of clusters, the 95% CI of ICC became narrower. For example, with 60 clusters, when the cluster sizes increase from 50, 100, to 200, the 95% CIs of ICC are (0.018–0.092), (0.029–0.081), and (0.039–0.075), respectively, for a population ICC value of 0.05. This is true for 30 and 90 clusters for different population ICC values. We observed similar trends for different numbers of cluster size variations, event rate variations, and event rate of 0.40.

In Table 2 we present the ICC estimates using the resampling, ANOVA, and MM methods for population ICC values of 0.05, 0.10, and 0.15 with an average event rate of 0.2, where the event rate varies from cluster to cluster by 25% for 30, 60, and 90 total clusters and where the average cluster sizes are 50, 100, and 200, with 25% cluster size variation. We observe that the estimates based on this new resampling method are more precise than the estimates by other two methods for different number of clusters and cluster sizes. We also observed similar trends for different number of cluster size variations, event rate variations, and different event rates.

The main limitation of the resampling method is that it requires an intensive amount of processing time for large cluster sizes. Therefore, much computing power is needed to get the

Table 2. ICC estimates comparisons for event rate of 0.2.

			Estimated ICC values		
Total clusters	Cluster size	Assumed ICC value	Resampling	ANOVA	Moment
10	10	0.05	0.046	0.055	0.060
		0.10	0.087	0.099	0.105
		0.15	0.129	0.143	0.151
	25	0.05	0.048	0.056	0.056
		0.10	0.090	0.100	0.104
		0.15	0.134	0.146	0.153
	50	0.05	0.046	0.054	0.054
		0.10	0.088	0.098	0.101
		0.15	0.131	0.143	0.148
	100	0.05	0.050	0.055	0.053
		0.10	0.094	0.100	0.101
		0.15	0.136	0.145	0.148
	200	0.05	0.050	0.055	0.052
		0.10	0.093	0.101	0.100
		0.15	0.136	0.147	0.147
20	10	0.05	0.050	0.056	0.058
		0.10	0.096	0.103	0.104
		0.15	0.141	0.150	0.152
	25	0.05	0.052	0.057	0.055
		0.10	0.097	0.104	0.104
		0.15	0.144	0.153	0.154
	50	0.05	0.052	0.055	0.054
		0.10	0.097	0.103	0.102
		0.15	0.143	0.150	0.151
	100	0.05	0.053	0.056	0.054
		0.10	0.099	0.104	0.103
		0.15	0.144	0.151	0.151
	200	0.05	0.054	0.057	0.053
		0.10	0.100	0.105	0.102
		0.15	0.146	0.153	0.151
30	10	0.05	0.054	0.057	0.055
		0.10	0.101	0.105	0.102
		0.15	0.148	0.153	0.152
	25	0.05	0.054	0.057	0.054
		0.10	0.100	0.105	0.104
		0.15	0.148	0.154	0.154
	50	0.05	0.053	0.056	0.057
		0.10	0.100	0.105	0.105
		0.15	0.147	0.153	0.154
	100	0.05	0.055	0.057	0.057
		0.10	0.101	0.105	0.105
		0.15	0.148	0.153	0.154
	200	0.05	0.055	0.057	0.057
		0.10	0.103	0.106	0.106
		0.15	0.150	0.154	0.155
60	10	0.05	0.054	0.057	0.054
		0.10	0.103	0.106	0.103
		0.15	0.151	0.155	0.153
	25	0.05	0.056	0.057	0.055
		0.10	0.105	0.107	0.105
		0.15	0.153	0.156	0.154
	50	0.05	0.055	0.057	0.058
		0.10	0.104	0.106	0.107
		0.15	0.152	0.156	0.156
	100	0.05	0.055	0.057	0.057
		0.10	0.104	0.106	0.107
		0.15	0.152	0.155	0.156
	200	0.05	0.057	0.058	0.058
		0.10	0.106	0.107	0.107
		0.15	0.154	0.156	0.157
90	10	0.05	0.056	0.057	0.054
		0.10	0.106	0.106	0.103

(*Continued on next page*)

Table 2. ICC estimates comparisons for event rate of 0.2. (*Continued*)

Total clusters	Cluster size	Assumed ICC value	Estimated ICC values		
			Resampling	ANOVA	Moment
		0.15	0.155	0.155	0.153
	25	0.05	0.057	0.058	0.054
		0.10	0.106	0.107	0.104
		0.15	0.155	0.157	0.154
	50	0.05	0.056	0.058	0.058
		0.10	0.105	0.107	0.107
		0.15	0.153	0.156	0.156
	100	0.05	0.056	0.057	0.057
		0.10	0.105	0.106	0.106
		0.15	0.154	0.155	0.156
	200	0.05	0.057	0.058	0.058
		0.10	0.106	0.107	0.107
		0.15	0.155	0.156	0.157

Note: Assumed 25% cluster size and event rate variations among clusters.

results, especially as the cluster size and number of clusters increase. Despite this limitation, our resampling method is very useful for estimating the ICC precisely for different event rates, cluster sizes, and numbers of clusters. In this method, we assumed that there are no third or higher order correlations among observations within clusters, though this may not always be true.

Precise ICC estimates are extremely important in designing CRTs because it has been shown that a small change in ICC can significantly increase the required number of clusters, which can drive up the cost (Chakraborty, 2009a; Chakraborty and Ray, 2015). Previous studies comparing ICC estimators found that most of the existing estimators are asymptotically equivalent (Lee, 1990) and the methods to estimate CI for ICC were developed separately (Ridout et al., 1999; Donner and Wells, 1986). Our method presented here will estimate both variance and CIs, and the CIs can be used to estimate the required sample sizes for CRTs. In summary, we conclude that our new resampling method is useful for precisely estimating the ICC and its CIs simultaneously, and it is broadly applicable to the design and analysis of cluster-randomized trials and other studies involving clustered data.

References

Campbell, M.K., Grimshaw, J.M., Steen, I.N., for the Changing Professional Practice in Europe Group. (2000a). Sample size calculations for cluster randomised trials. *J. Health Serv. Res. Policy* 5:12–16.

Campbell, M.K., Mollison, J., Steen, N., Grimshaw, J.M., Eccles, M. (2000b). Analysis of cluster randomized trials in primary care: a practical approach. *Family Pract.* 17:192–196.

Chakraborty, H., Moore, J., Carlo, A.W., Hartwell, T.D., Wright, L.L. (2009a). A simulation based technique to estimate intracluster correlation for a binary variable. *Contemp. Clin. Trials* 30:71–80.

Chakraborty, H., Moore, J., Hartwell, T.D. (2009b). Intracluster correlation adjustments to maintain power in cluster trials for binary outcomes. *Contemp. Clin. Trials* 30(5):473–480.

Chakraborty, H. (2008). The design and analysis aspects of cluster randomized trials. In: Biswas, A., Datta, S., Fine, J.P., Segal, M., Fine, J., Biswas, A.R., eds. *Statistical Advances in the Bio-Medical Sciences: State of the Art and Future Directions. Clinical Trials, Epidemiology, Survival Analysis, and Bioinformatics*. New York: John Wiley & Sons, Inc.

Chakraborty, H., Ray, G. (2015). Cluster randomized trials: Considerations for design and analysis. *Journal of Statistical Theory and Practice*. 9:685–698.

Carlo, W.A., Goudar, S.S., Jehan, I., Chomba, E., Tshefu, A., Garces, A., Parida, S., Althabe, F., McClure, E.M., Derman, R.J., Goldenberg, R.L., Bose, C., Krebs, N.F., Panigrahi, P., Buekens, P., Chakraborty,

H., Hartwell, T.D., Wright, L.L., the FIRST BREATH Study Group. (2010). Newborn care training and perinatal mortality in communities in developing countries. *N. Engl. J. Med.* 362(7):614–623.
Carroll, R.J., Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman and Hall.
Crowder, M.J. (1978). Beta-binomial ANOVA for proportions. *Appl. Stat.* 27:34–37.
Donner, A. (1998). Some aspects of the design and analysis of cluster randomization trials. *Appl. Stat.* 47:95–113.
Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effect model. *Int. Stat. Rev.* 54:67–82.
Donner, A., Wells, G. (1986). A comparison of confidence interval methods for the intraclass correlation coefficient. *Biometrics* 42:401–412.
Efron, B., Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.
Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: Wiley.
Fleiss, J.L., Cuzick, J. (1979). The reliability of dichotomous judgments: unequal number of judges per subject. *Appl. Psychol. Meas.* 3:537–542.
Karlin, S., Cameron, P.E., Williams, P. (1981). Sibling and parent-offspring correlation with variable family age. *Proc. Natl. Acad. Sci, USA* 78:2664–2668.
Kleinman, J.C. (1973). Proportions with extraneous variance: single and independent samples. *J. Am. Stat. Assoc.* 8:46–54.
Lee, A.J. (1990). *U-Statistics: Theory and Practice*. New York: Marcel Dekker Inc.
Lipsitz, S.R., Laird, N.M., Brennan, T.A. (1994). Simple moment estimates of the k-coefficient and its variance. *Appl. Stat.* 43:309–323.
Lunn, A.D., Davies, S.J. (1998). A note on generating correlated binary variables. *Biometrika* 85: 487–490.
Mak, T.K. (1988). Analyzing intraclass correlation for dichotomous variables. *Appl. Stat.* 37:344–352.
McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.
Moore, D.F., Tsiatis, A. (1991). Robust estimation of the standard error in moment methods for extra-binomial and extra-Poisson variation. *Biometrika* 47:383–401.
Nelder, J.A., Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* 74:221–232.
Ridout, M.S., Demetrio, C.G.B., Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics* 55:137–148.
SAS Institute Inc. (2008). *SAS Release 9.2*. Cary, North Carolina.
Schouten, H.J.A. (1986). Nominal scale agreement among observers. *Psychometrica* 51:453–466.
Tamura, R.N., Young, S.S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biometrics* 43:813–824. Correction: (1994), 50:321.
Williams, D.A. (1982). Extra-binomial variation in logistic linear models. *Appl. Stat.* 31:144–148.
Yamamoto, E., Yanagimoto, T. (1992). Moment estimators for the binomial distribution. *J. Appl. Stat.* 19:273–283.