University of North Carolina at Chapel Hill
The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series

Year 2016 *Paper* 46

Prevalence Estimation at the Cluster Level for Correlated Binary Data Using Random Partial-Cluster Sampling

Rujin Wang* John S. Preisser†

*University of North Carolina at Chapel Hill, rujin@email.unc.edu

†University of North Carolina at Chapel Hill, jpreisse@bios.unc.edu

Prevalence Estimation at the Cluster Level for Correlated Binary Data Using Random Partial-Cluster Sampling

Rujin Wang and John S. Preisser

Abstract

For clustered data in the medical sciences, disease is present when one or more of the observations in the cluster has the disease condition. This paper focuses on estimation of periodontal disease prevalence defined as the probability that one or more tooth sites have disease in a randomly selected subject. The prohibitive exam time and monetary cost of the full-mouth examination makes partial-mouth recording protocols attractive alternative methods to assess chronic periodontitis. In particular, Beck et al. (2006) proposed the random site selection method (RSSM), which pre-specifies a fixed number of tooth sites to be selected randomly from each subject. RSSM could reduce the examination time, but standard estimators that define an individual's disease status solely in terms of selected sites tend to underestimate disease prevalence. We define each mouth as a cluster and disease status (presence or absence) at each tooth site as a binary variable. We describe a prevalence estimator based on the conditional linear family (CLF) of correlated binary distributions under the working assumptions of equal site-level means and exchangeable pairwise correlation for all within-cluster pairs of sites. We derive a variance estimator for the CLF-RSSM prevalence estimator by the delta method. Using simulated data, our prevalence estimator and its variance estimator have small to negligible bias and confidence intervals for prevalence have coverage near the 95% nominal level when the working model is correct. Taking missing teeth into consideration, the CLF-RSSM prevalence estimator has approximately 90% coverage in our simulations. Given a more realistic unequal means and dental correlation structure, the CLF-RSSM prevalence and its standard deviation estimator do not perform well under model misspecification. While the overall approach to the estimation of disease prevalence at the cluster level using

partial cluster sampling is promising, new estimators that incorporate more realistic distributional assumptions of correlated binary data (e.g. tooth surfaces in a mouth) may be needed according to the application.

Prevalence Estimation at the Cluster Level for Correlated Binary Data Using Random Partial-Cluster Sampling

by

Rujin Wang and John S. Preisser

Department of Biostatistics, University of North Carolina at Chapel Hill

Abstract

For clustered data in the medical sciences, disease is present when one or more of the observations in the cluster has the disease condition. This paper focuses on estimation of periodontal disease prevalence defined as the probability that one or more tooth sites have disease in a randomly selected subject. The prohibitive exam time and monetary cost of the full-mouth examination makes partial-mouth recording protocols attractive alternative methods to assess chronic periodontitis. In particular, Beck *et al.*(2006) proposed the random site selection method (RSSM), which pre-specifies a fixed number of tooth sites to be selected randomly from each subject. RSSM could reduce the examination time, but standard estimators that define an individual's disease status solely in terms of selected sites tend to underestimate disease prevalence. We define each mouth as a cluster and disease status (presence or absence) at each tooth site as a binary variable. We describe a prevalence estimator based on the conditional linear family (CLF) of correlated binary distributions under the working assumptions of equal site-level means and exchangeable pairwise correlation for all within-cluster pairs of sites. We derive a variance estimator for the CLF-RSSM prevalence estimator by the delta method. Using simulated data, our prevalence estimator and its variance estimator have small to negligible bias and confidence intervals for prevalence have coverage near the 95% nominal level when the working model is correct. Taking missing teeth into consideration, the CLF-RSSM prevalence estimator has approximately 90% coverage in our simulations. Given a more realistic unequal means and dental correlation structure, the CLF-RSSM prevalence and its standard deviation estimator do not perform well under model misspecification. While the overall approach to the estimation of disease prevalence at the cluster level using partial cluster sampling is promising, new estimators that incorporate more realistic distributional assumptions of correlated binary data (e.g. tooth surfaces in a mouth) may be needed according to the application.

Keywords: clinical attachment level, partial-recording protocol, periodontitis, pocket depth.

1 Introduction

The objective of this technical report is to examine the estimation of periodontal disease prevalence by using random partial-mouth recording protocols. The full-mouth examination is the gold standard method for assessing chronic periodontitis and estimating prevalence of periodontal disease. A full-mouth examination requires probing a maximum 168 sites (six sites on 28 teeth) if third molars are excluded, which are time consuming and infeasible in many large epidemiological research studies. An alternative method is the partial-mouth recording protocol (PRP), which only examines a subset of sites or teeth. Two major types of the partial sampling methods are fixed-site selection method (FSSM) and random-site selection method (RSSM). For FSSM, specific sites or teeth are chosen. Instead of holding the set of sites fixed, Beck *et al.*(2006) proposed the RSSM that pre-specifies a fixed number of sites to be drawn randomly. Use of partial-mouth recording protocols could reduce examination times. However, examination on a subset of sites tends to result in biased and underestimated prevalence in a population with the use of a standard estimator that is based on classifying individuals using only each person's observed sites.

Defining periodontitis as one or more sites affected, Marks (2015) proposed an estimator of chronic periodontitis prevalence of random-site selection method (RSSM) using the conditional linear family (CLF) of correlated binary distributions. Under the assumption of a common site-level probability of disease and an exchangeable pairwise correlation between sites, the CLF-RSSM estimator relying on a simple working model for correlated binary data performs substantially better than the standard estimator. This report extends previous work of Marks (2015) for RSSMs and applies the delta method to derive a large-sample variance formula for the prevalence estimator.

Marks (2015) used oral examination data from 6,793 participants in the Arteriolosclerosis Risk In Communities (ARIC) Study. ARIC is a predominantly biracial community-based prospective cohort study designed to investigate risk factors for the development and progression of atherosclerosis and coronary heart disease in 15,792 adult participants aged 45 to 64 years enrolled starting in 1987 in four U.S. communities. As part of the fourth visit (1996 to 1998), 6,793 cohort members received a full-mouth oral examination as participants in an ancillary study of oral health, Dental ARIC (Beck *et al.*, 2006). To assess bias in the prevalence estimator, she resampled the ARIC cohort and calculated the relative bias of the CLF-RSSM estimator defining the observed prevalence estimated from the full-mouth examination as the true prevalence of chronic periodontitis for ARIC participants. She showed that the CLF-RSSM estimator gives prevalence estimates that are much closer than standard partial-mouth estimates to full-mouth estimates.

This paper extends the research of Marks (2015) by introducing a large-sample variance formula for the prevalence estimator and methods for computing asymptotic confidence intervals for disease prevalence. In simulation studies, correlated binary data for the full mouth from the conditional linear family are generated under different distributional structures. We assess the bias of the CLF-RSSM prevalence estimator and its standard error, defining the true value as inserting pre-specified marginal means and pairwise correlation into the formula for the probability of one or more sites affected.

Section 2 presents the CLF-RSSM estimator and an expression for its large-sample variance. Section 3 describes the simulation studies, section 4 contains the results and section 5 describes limitations as well as possible future developments in the statistical methodology.

2 Methods

Under random partial-cluster sampling, instead of recording disease status for all n observations in a cluster (e.g., all $n = 168$ tooth sites in a mouth excluding third molars), a much smaller number (m) of sites are selected at random from each cluster. Moreover, typically m is fixed at a constant while, in practice, the number of sites selected from the ith mouth (m_i) may be less than m if m sites are not available. Under such a sampling protocol, a formula for the prevalence of a disease or condition at the cluster level is based on the assumption of equal means and exchangeable correlated sites using the CLF as developed by Marks (2015). A case of chronic periodontitis is defined as one or more tooth sites affected according to a criterion such as pocket depth or clinical attachment loss exceeding an established threshold (Beck *et al.* 2006). To obtain the prevalence, recall the fact that the multivariate joint probability distribution of a sequence of correlated binary variables can be factorized as a product of conditional probabilities, and thus introduce new parameters which indicate the corresponding conditional probabilities. We adopt as the working model the conditional linear family (CLF) with equal means μ and exchangeable pairwise correlation ρ for all pairs of sites. Consequently, applying equation (5) of Qaqish (2003), the prevalence can be expressed as

$$\pi = 1 - (1-\mu)\prod_{j=2}^{n}\left(1 - \frac{(1-\rho)\mu}{1+(j-2)\rho}\right) \tag{1}$$

where $n = 168$ is the number of total sites in a mouth (excluding third molars), $\mu_j = \mu$ for $j = 1, ..., n$, and $\rho_{jk} = \rho$ between all sites for $j = 1, ..., n-1$ and $k = j+1, ..., n$.

In order to examine how sensitive the prevalence is to perturbations in the values of μ and ρ, we

compute the prevalence for a grid of values, specifically $\mu = 0.045$ to 0.055 by 0.0025 increments and $\rho = 0.15$ to 0.25 by 0.025 increments.

Table 1: Prevalence π for a grid of values

ρ \ μ	0.045	0.0475	0.05	0.0525	0.055
0.15	0.594	0.614	0.633	0.651	0.668
0.175	0.547	0.567	0.585	0.603	0.621
0.2	0.506	0.525	0.543	0.561	0.578
0.225	0.469	0.487	0.505	0.522	0.539
0.25	0.436	0.454	0.471	0.488	0.504

Results show that, given ρ, prevalence increases as μ increases and, given μ, prevalence decreases as ρ increases (Table 1). The prevalence π is shown to be sensitive to the values of μ and ρ underscoring the importance of their accurate estimation in the intermediate step in the estimation of π.

2.1 The CLF-RSSM prevalence estimator

An estimator $\hat{\pi}$ of prevalence is obtained by plugging in estimators $\hat{\mu}$ and $\hat{\rho}$ into equation (1). An estimator of the common marginal mean μ is

$$\hat{\mu} = \frac{\sum_{i=1}^{K} m_i \hat{Y}_i}{\sum_{i=1}^{K} m_i} \tag{2}$$

where $i = 1, ..., K$ subjects, $\hat{Y}_i$ denotes the proportion of sampled sites in the ith subject that have periodontitis and m_i is the number of tooth sites randomly sampled from the ith mouth. Let $y_{ij} = 1$ if the jth site in the ith subject has disease and 0 otherwise. Next, define $r_{ij} = \frac{y_{ij} - \hat{\mu}}{\sqrt{\hat{\mu}(1-\hat{\mu})}}$ as the standardized residual for the ith individual's jth site under the working model of a common mean $E(y_{ij}) = \mu$ across subjects and sites and an exchangeable correlation ρ for pairs of tooth sites within mouths. A GEE estimator of the exchangeable correlation (Zeger and Liang; 1986) is

$$\hat{\rho} = \frac{\sum_{i=1}^{K} \sum_{j \neq k} s_{ij} r_{ij} s_{ik} r_{ik}}{\sum_{i=1}^{K} m_i (m_i - 1)} \tag{3}$$

where the second summation in the numerator includes products of residuals from examined pairs (those having $s_{ij} = 1$ and $s_{ik} = 1$) across all possible pairings of sites in the full mouth. The numerator consists of all permutations such that both $r_{i1}r_{i2}$ and $r_{i2}r_{i1}$ are included. Within a cluster, there are $m_i(m_i - 1)$ such terms. Importantly, the number of terms in the numerator is equivalent to the number of terms in the denominator, making it an average of within-cluster residual cross-products. For equal cluster sizes ($m_i \equiv m$), the GEE approach of Prentice (1988) without covariates reduces to equation (2) and equation (3). A macro for the method of Prentice (1988) is available from the second author by request. This procedure will provide an estimate of the joint covariance matrix of $\hat{\mu}$ and $\hat{\rho}$ from which an expression of the asymptotic variance of $\hat{\pi}$ will be obtained as described in the next section. Along with estimators for μ, ρ and prevalence, large-sample confidence intervals can be obtained for each of them.

2.2 The asymptotic variance of the prevalence estimator

To estimate the standard errors of prevalence, we apply the delta method to the prevalence estimator to derive a large-sample variance formula. Let

$$\theta = 1 - \pi = (1 - \mu) \prod_{j=2}^{n} \left(1 - \frac{(1 - \rho)\mu}{1 + (j - 2)\rho}\right)$$

Working with the log probability,

$$\begin{aligned} \theta^* &= \ln \theta \\ &= \ln(1 - \mu) + \sum_{j=2}^{n} \ln \left(1 - \frac{(1 - \rho)\mu}{1 + (j - 2)\rho}\right) \end{aligned}$$

Therefore,

$$\pi = 1 - e^{\theta^*}.$$

A standard method of modeling correlated binary data involves the logit link for the marginal mean model and the identity link for the linear correlation model. The logit link is used to restrict the mean in the range of (0, 1). Since $\text{logit}(\mu) = \beta$, we have

$$\mu = \frac{e^{\beta}}{1 + e^{\beta}}.$$

Using the delta method,

$$Var(\hat{\theta^*}) = \phi' \Sigma_{\beta,\rho} \phi \tag{4}$$

$$\phi = \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} = \begin{pmatrix} \frac{\partial \theta^*}{\partial \beta} \\ \frac{\partial \theta^*}{\partial \rho}. \end{pmatrix}$$

The covariance matrix Σ of $\hat{\beta}$ and $\hat{\rho}$ is estimated as a sandwich covariance estimator based upon the GEE procedure of Prentice (1988).

Next, we differentiate θ^* with respect to β, using the chain rule for differentiation, according to which

$$\frac{\partial \theta^*}{\partial \beta} = \frac{\partial \theta^*}{\partial \mu} \frac{\partial \mu}{\partial \beta}$$

where

$$\frac{\partial \mu}{\partial \beta} = \mu(1-\mu)$$

and

$$\frac{\partial \theta^*}{\partial \mu} = -\frac{1}{1-\mu} + \sum_{j=2}^{n} \left(\frac{-\frac{1-\rho}{1+(j-2)}}{1-\frac{(1-\rho)\mu}{1+(j-2)\rho}} \right).$$

The result is

$$\frac{\partial \theta^*}{\partial \beta} = \mu(1-\mu) \left(-\frac{1}{1-\mu} + \sum_{j=2}^{n} \left(\frac{-\frac{1-\rho}{1+(j-2)}}{1-\frac{(1-\rho)\mu}{1+(j-2)\rho}} \right) \right).$$

Similarly,

$$\begin{aligned} \frac{\partial \theta^*}{\partial \rho} &= \sum_{j=2}^{n} \left(\frac{1+(j-2)\rho}{1-\mu+(j-2+\mu)\rho} \cdot \frac{(j-2+\mu)(1+(j-2)\rho) - (j-2)(1-\mu+(j-2+\mu)\rho)}{(1+(j-2)\rho)^2} \right) \\ &= \sum_{j=2}^{n} \left(\frac{\mu(1+(j-2))}{(1+(j-2)\rho)(1-\mu+(j-2+\mu)\rho)} \right). \end{aligned}$$

Using delta method again provides

$$Var(\hat{\pi}) = \left(\frac{\partial \pi}{\partial \theta^*} \right)^2 Var(\hat{\theta^*})$$

where

$$\frac{\partial \pi}{\partial \theta^*} = -e^{\theta^*}.$$

Therefore,

$$Var(\hat{\pi}) = e^{2\theta^*} Var(\hat{\theta^*}). \tag{5}$$

That is to say, we have derived a large-sample formula for the variance of the CLF-RSSM prevalence estimator as a function of the variances of $\hat{\mu}$ and $\hat{\rho}$ and their joint covariance matrix. Likewise, if desired,

the variance of μ is easily derived by the delta method:

$$\begin{aligned} Var(\hat{\mu}) &= \left(\frac{\partial \mu}{\partial \beta}\right)^2 Var(\hat{\beta}) \\ &= \mu^2(1-\mu)^2 Var(\hat{\beta}). \end{aligned} \tag{6}$$

3 Simulation Studies

Three distributional structures are considered for an evaluation of the finite-sample performance of the CLF-RSSM prevalence estimator $\hat{\pi}$. This section describes simulation scenarios to generate correlated binary data from the conditional linear family (CLF) for the full mouth and methods for the evaluation of the finite-sample performance of $\hat{\pi}$ with partial-mouth samples selected from the full mouth. The performance of estimators including bias and confidence interval coverage is reported in section 4. Evaluation of the newly developed estimation methods requires deduction of the true value of prevalence from each data generating model.

3.1 Simulation study designs

3.1.1 Prevalence under the common mean, exchangeable correlation model

Under a model specifying common site-specific means and exchangeable correlation among site pairs, the true prevalence of periodontitis is defined by inserting μ and ρ into formula (1) directly with the values of μ and ρ that are used when generating correlated binary data from the CLF for the full mouth.

3.1.2 Prevalence under the different means, dental correlation model

The human dentition is divided into four quadrants, which are commonly labeled such that 1 indicates the maxillary right quadrant, 2 indicates the maxillary left quadrant, 3 denotes the mandibular left quadrant, and 4 is the mandibular right quadrant (Figure 1). Notably, quadrants 1 and 3 are not spatially adjacent and quadrants 2 and 4 are not adjacent. Excluding third molars, there are 3 anterior teeth and 4 posterior teeth per quadrant. Anterior teeth (typically incisors and canine teeth) tend to have a lower probability of disease than posterior teeth (typically premolars and molars). Moreover, the relationship of sites within pairs of sites may vary according to their distance or relationship to each other. Such considerations lead us to consider a data-generating model involving different site-level means and a "dental" correlation structure.

In particular, a three-parameter dental working correlation structure is based on the four quad-

rants of the mouth. We assume the highest correlation occurs when two sites are from the same quadrant, the medium level of correlation occurs when two sites are from different and adjacent quadrants, and the lowest level correlation arises when two sites are from different and non-adjacent quadrants. The disease status of each site is denoted as Y_{ijk} for the ith subject, jth quadrant, and kth tooth, where $j = 1, 2, 3, 4$ and $k = 1, ..., 7$. Let

$$\rho(Y_{ijk}, Y_{ijk'}) = \alpha_0$$

$$\rho(Y_{ijk}, Y_{ij'k'}) = \begin{cases} \alpha_1 & \text{different and adjacent quadrants, } \{(j, j') : (1,2), (1,4), (2,3), (3,4)\} \\ \alpha_2 & \text{different and non-adjacent quadrants, } \{\text{(j,j')}:(1,3),(2,4)\} \end{cases}$$

where α_0 is the within-quadrant correlation, α_1 is the between and adjacent quadrant correlation, and α_2 is the between and non-adjacent quadrant correlation. Values are considered such that $\alpha_0 > \alpha_1 > \alpha_2$ and

$$\mu_{\text{xp}} > \mu_{\text{mp}} > \mu_{\text{xa}} > \mu_{\text{ma}}$$

where μ_{xp} is the marginal mean for upper back teeth (maxillary/posterior), μ_{mp} indicates the marginal mean for lower back teeth (mandibular/posterior), μ_{xa} denotes the marginal mean for upper front teeth (maxillary/anterior), and μ_{ma} stands for the marginal mean for lower front teeth (mandibular/anterior).

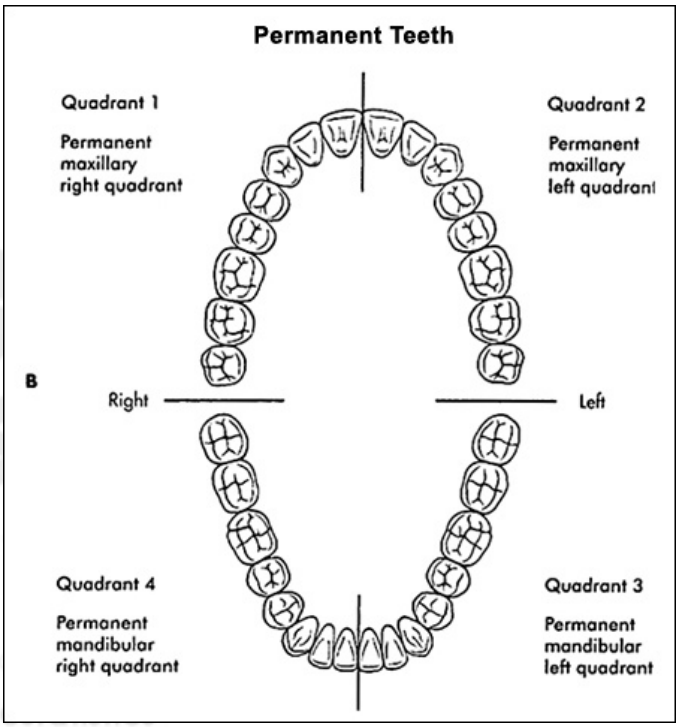

Figure 1: Quadrant numbers

To construct the dental correlation matrix structure, we have

$$R = \begin{array}{c} \\ Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \end{array} \begin{array}{c} \begin{array}{cccc} Q_1 & Q_2 & Q_3 & Q_4 \end{array} \\ \begin{bmatrix} A_0 & A_1 & A_2 & A_1 \\ A_1 & A_0 & A_1 & A_2 \\ A_2 & A_1 & A_0 & A_1 \\ A_1 & A_2 & A_1 & A_0 \end{bmatrix} \end{array}$$

where R is a 168×168 matrix and A_h is a 42×42 matrix for $h = 0, 1, 2$. Specifically,

$$A_0 = \begin{bmatrix} 1 & \alpha_0 & \dots & \alpha_0 \\ \alpha_0 & 1 & \dots & \alpha_0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_0 & \alpha_0 & \dots & 1 \end{bmatrix} \quad A_1 = \begin{bmatrix} \alpha_1 & \alpha_1 & \dots & \alpha_1 \\ \alpha_1 & \alpha_1 & \dots & \alpha_1 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_1 & \alpha_1 & \dots & \alpha_1 \end{bmatrix} \quad A_2 = \begin{bmatrix} \alpha_2 & \alpha_2 & \dots & \alpha_2 \\ \alpha_2 & \alpha_2 & \dots & \alpha_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_2 & \alpha_2 & \dots & \alpha_2 \end{bmatrix}$$

The true prevalence can be determined from the formula for 1 minus prevalence, i.e., the probability under the data-generating model that all 168 site-level indicators for disease are equal to zero.

To derive the general formula for prevalence for arbitrary $(\boldsymbol{\mu}, \boldsymbol{R})$ satisfying the CLF condition (Qaqish, 2003), let

$$\lambda_j = P(Y_j = 1 | Y_1 = 0, ..., Y_{j-1} = 0)$$

From Marks (2015),

$$\begin{aligned} \pi &= P\left(\sum_{j=1}^{n} Y_j \geq 1\right) = 1 - P\left(\sum_{j=1}^{n} Y_j = 0\right) \\ &= 1 - P(Y_1 = 0, ..., Y_n = 0) \\ &= 1 - P(Y_1 = 0)P(Y_2 = 0|Y_1 = 0) \cdots P(Y_n = 0|Y_{n-1} = 0, ..., Y_1 = 0) \\ &= 1 - (1 - \lambda_1)(1 - \lambda_2) \cdots (1 - \lambda_n) \\ &= 1 - \prod_{j=1}^{n} (1 - \lambda_j) \end{aligned}$$

where $\lambda_1 = \mu_1 = P(Y_1 = 1)$. To fully define π, expressions for λ_j are needed.

Qaqish (2003) introduced methods for generating correlated binary data that allow general spec-

ifications of the marginal mean and correlation structure. Define $\boldsymbol{Z}_j = (Y_1, ..., Y_{j-1})'$, and

$$\begin{aligned}\lambda_j &= P(Y_j = 1 | \boldsymbol{Z}_j = \boldsymbol{0}) \\ &= \mu_j - \boldsymbol{b}_j' \boldsymbol{\zeta}_j \\ &= \mu_j - \sum_{l=1}^{j-1} b_{jl}\mu_l \quad (j = 2, ..., n)\end{aligned}$$

where $\boldsymbol{\zeta}_j = E(\boldsymbol{Z}_j)$, $\boldsymbol{b}_j = \boldsymbol{G}_j^{-1}\boldsymbol{s}_j$, $\boldsymbol{G}_j = \text{cov}(\boldsymbol{Z}_j)$, $\boldsymbol{s}_j = \text{cov}(\boldsymbol{Z}_j, Y_j)$, and $\boldsymbol{b}_j' = (b_{j1}, b_{j2}, ..., b_{j,j-1})$. Note that

$$\text{cov}(\boldsymbol{Z}_j) = \boldsymbol{A}_j \boldsymbol{R}_j \boldsymbol{A}_j$$

where $\boldsymbol{A}_j = \text{Diag}\left(\mu_1(1-\mu_1), ..., \mu_{j-1}(1-\mu_{j-1})\right)$ and $\boldsymbol{R}_j = \text{Corr}(\boldsymbol{Z}_j)$. Without a formula for $\boldsymbol{R}_j^{-1}$, we cannot simplify the expression further. However, for any known $\boldsymbol{\mu}$ and $\boldsymbol{R}$, we can compute π. We also check to ensure that all λ_j satisfy $0 \leq \lambda_j \leq 1$ to confirm that $(\boldsymbol{\mu}, \boldsymbol{R})$ are compatible with a multivariate binary distribution in the conditional linear family.

3.1.3 Simulation study design under different means, exchangeable correlation structure model, with unequal cluster sizes

More realistically and particularly in older populations, there are almost always missing teeth, which could induce bias. To define data-generating models for correlated binary data with varying cluster sizes due to missing teeth, we first estimate the distribution of the number of missing teeth per participant in the ARIC study. This is done by fitting a negative binomial regression model with only the intercept to the ARIC data. Fixing the estimated $\hat{\mu}$ and $\hat{\phi}$ (the negative binomial dispersion parameter), the count of the number of missing teeth (Z) for an individual is randomly generated i.e., $Z \sim NB(\hat{\mu}, \hat{\phi})$. If the number of missing teeth is greater than 26, we set it as 26. If a tooth is missing, then all six sites on the tooth are missing. If a tooth is observed, then all six sites are observed. We then compute the number of non-missing tooth sites (cluster size) for the $i-$th individual as $n_i = 168 - 6 \times Z$. For given cluster size, we use trial and error to assign the appropriate μ_i based on equation (1) where both ρ and π are fixed to be constant across all considered cluster sizes (and individuals). We generate correlated binary data for individual i from CLF with randomly generated n_i, fixed ρ and calculated μ_i. Finally, the CLF-RSSM prevalence estimator is applied to the generated data. Since $E(Z) = 7$, the expected observed number of sites is 126. Taking an ad-hoc approach to account for missing data, we set $n = 126$ in equation (1) and (5) for estimation.

The rationale for this data generation procedure is as follows. First, as chronic periodontitis for

a person is defined based on the actual teeth in the mouth, missing teeth are irrelevant to the diagnosis of disease. For this reason, the problem of missing teeth is a viewed from the perspective of potentially informative cluster sizes, where n_i is number of tooth sites present, instead of missing data per se. By stratifying the data-generating models according to cluster size and fixing the prevalence, correlated binary data are generated under a non-informative cluster size mechanism.

3.2 Simulation study methods to evaluate the CLF-RSSM estimator

To assess the finite-sample properties of the CLF-RSSM estimator under different scenarios, we conducted simulation studies using SAS 9.4. Specifically, we generated correlated binary data under three different distributional assumptions under the conditional linear family (CLF). Under Model I, correlated binary data are generated from a common mean and exchangeable correlation model with $n = 168$, $\mu = 0.05$ and $\rho = 0.2$. In this case, the true value of prevalence is $\pi = 0.543$. Under Model II, full mouth data are generated from the dental model in section 3.1.2 with $n = 168$, $\alpha_0 = 0.4$, $\alpha_1 = 0.2$, $\alpha_2 = 0.185$, $\mu_{xp} = 0.054$, $\mu_{mp} = 0.051$, $\mu_{xa} = 0.05$, and $\mu_{ma} = 0.048$. Such a parameter setting gives the true value of prevalence as $\pi = 0.542$. In this condition, we only focus on the performance of prevalence estimator and its variance as μ and ρ as estimands are not revelant. Under Model III with different means, exchangeable correlation structure and unequal cluster sizes, we specify $\rho = 0.2$. For given cluster size, we calculate the appropriate μ such that the prevalence is $\pi = 0.543$, and generate data from the CLF distribution for sites from non-missing teeth. Under this model, we focus on the performance of the correlation and prevalence estimators, as well as their variance estimators. We randomly generated data from these three models for each of $K = 500$, $K = 1000$ or $K = 2000$ clusters. We then randomly selected $m = 4$, 6, and 10 sites for each K to assess the performance of the estimators under three cases of partial-cluster sampling resulting in a total of $3^3 = 27$ simulation scenarios. All 6 sites on all teeth excluding third molars have equivalent probability being selected. We used $R = 1000$ replicate simulations for each scenario.

For each simulation replicate, $\hat{\mu}$ and $\hat{\rho}$ were estimated by equations (2) and (3) and their joint empirical covariance matrix is estimated by the GEE method of Prentice(1988). These plug-in estimates are used to calculate $\hat{\pi}$ from equation (1), $\text{Var}(\hat{\mu})$ from equation (6), $\text{Var}(\hat{\rho})$, and $\text{Var}(\hat{\pi})$ from equation (5). Assuming $\hat{\mu}$, $\hat{\rho}$, and $\hat{\pi}$ are approximately normally distributed in large samples, for each simulation replicate, we constructed 95% confidence intervals for π, and for μ and ρ as applicable.

Results of $R = 1000$ replicate simulations are aggregated for each scenario. For data generated

under Model I, we calculated coverage of 95% confidence interval for μ, ρ, and π, indicating the proportion of times CI contained the true parameter value. We also calculated the average as well as the percent relative bias of point estimates for each RSSM sample size as follows:

$$\% \text{ Relative Bias } (\hat{\mu}) = \left(\frac{\bar{\hat{\mu}}-\mu}{\mu}\right) \times 100\%$$
$$\% \text{ Relative Bias } (\hat{\rho}) = \left(\frac{\bar{\hat{\rho}}-\rho}{\rho}\right) \times 100\%$$
$$\% \text{ Relative Bias } (\hat{\pi}) = \left(\frac{\bar{\hat{\pi}}-\pi}{\pi}\right) \times 100\%$$

where $\bar{\hat{\mu}}$ =ave$(\hat{\mu}) = \frac{\sum_{r=1}^{R}\hat{\mu}_r}{R}$, $\bar{\hat{\rho}}$ =ave$(\hat{\rho}) = \frac{\sum_{r=1}^{R}\hat{\rho}_r}{R}$, $\bar{\hat{\pi}}$ =ave$(\hat{\pi}) = \frac{\sum_{r=1}^{R}\hat{\pi}_r}{R}$, R is the number of replicate simulation times, and μ, ρ and π are true values. Next, the percent relative bias of the standard error (s.e.) estimates of the $\hat{\mu}$, $\hat{\rho}$ and $\hat{\pi}$ is reported. "True" values of s.e.$(\hat{\mu})$, s.e.$(\hat{\rho})$ and s.e.$(\hat{\pi})$ can be taken by the respective monte carlo standard deviations of $\hat{\mu}$, $\hat{\rho}$ and $\hat{\pi}$.

$$\% \text{ Relative Bias (s.e.}(\hat{\mu})) = \left(\frac{\text{ave s.e.}(\hat{\mu}) - \text{mc s.d.}(\hat{\mu})}{\text{mc s.d.}(\hat{\mu})}\right) \times 100\%$$
$$\% \text{ Relative Bias (s.e.}(\hat{\rho})) = \left(\frac{\text{ave s.e.}(\hat{\rho}) - \text{mc s.d.}(\hat{\rho})}{\text{mc s.d.}(\hat{\rho})}\right) \times 100\%$$
$$\% \text{ Relative Bias (s.e.}(\hat{\pi})) = \left(\frac{\text{ave s.e.}(\hat{\pi}) - \text{mc s.d.}(\hat{\pi})}{\text{mc s.d.}(\hat{\pi})}\right) \times 100\%$$

where Monte Carlo standard deviation for μ is mc s.d.$(\hat{\mu}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\mu}_r - \bar{\mu})^2}{R-1}}$; Monte Carlo standard deviation for ρ is mc s.d.$(\hat{\rho}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\rho}_r - \bar{\rho})^2}{R-1}}$; Monte Carlo standard deviation for prevalence is mc s.d.$(\hat{\pi}) = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\pi}_r - \bar{\pi})^2}{R-1}}$; and $R = 1000$ is the number of replicate simulations. For Model II, results are reported for π only. For Model III, results are reported for only π and ρ.

4 Results

4.1 Common mean, exchangeable correlation

When the working model assumptions of common site-level mean and exchangeable correlation are satisfied, the GEE estimators for μ and ρ have negligible bias and so does the CLF-RSSM estimator $\hat{\pi}$ (Table 3). Moreover, the CLF-RSSM estimators of mean, correlation and prevalence perform well given different random selection sample size m. Generally, the standard errors of mean, correlation, and prevalence estimators perform well with only trivial biases (Table 4). Coverage of 95% confidence intervals for the mean, correlation, and prevalence improves with an increasing number of clusters K while there is slight undercoverage of 95% confidence intervals for π.

Table 2: Average estimates using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has equal mean ($\mu = 0.05$) and exchangeable correlation ($\rho = 0.2$) under the conditional linear family such that the true prevalence is $\pi = 0.543$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$
RSSM4	0.050	0.197	0.552	0.050	0.199	0.547	0.050	0.200	0.544
RSSM6	0.050	0.198	0.550	0.050	0.199	0.546	0.050	0.201	0.543
RSSM10	0.050	0.197	0.548	0.050	0.198	0.546	0.050	0.200	0.544

Table 3: Percent relative bias of estimates using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has equal mean ($\mu = 0.05$) and exchangeable correlation ($\rho = 0.2$) under the conditional linear family such that the true prevalence is $\pi = 0.543$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$
RSSM4	−0.33	−1.61	1.64	−0.19	−0.61	0.68	−0.20	−0.22	0.19
RSSM6	0.10	−1.15	1.21	−0.25	−0.73	0.44	−0.03	0.32	−0.10
RSSM10	−0.07	−1.32	0.90	−0.10	−0.84	0.51	−0.01	−0.15	0.13

Table 4: Percent relative bias of standard deviation estimates using the large sample CLF-RSSM variance estimator for 1000 simulation replicates when the true distribution for the correlated sites has equal mean ($\mu = 0.05$) and exchangeable correlation ($\rho = 0.2$) under the conditional linear family such that the true prevalence is $\pi = 0.543$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	$\hat{\sigma}_\mu$	$\hat{\sigma}_\rho$	$\hat{\sigma}_\pi$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\rho$	$\hat{\sigma}_\pi$	$\hat{\sigma}_\mu$	$\hat{\sigma}_\rho$	$\hat{\sigma}_\pi$
RSSM4	0.14	−4.74	−6.56	−0.59	−2.65	−5.80	−2.88	1.32	−1.16
RSSM6	−1.21	−3.55	−8.30	−2.23	2.64	−1.30	0.28	0.15	−3.31
RSSM10	−3.04	−1.73	−5.62	2.96	3.91	0.51	2.19	2.35	−4.53

Table 5: Coverage of 95% confidence interval for estimates using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has equal mean ($\mu = 0.05$) and exchangeable correlation ($\rho = 0.2$) under the conditional linear family such that the true prevalence is $\pi = 0.543$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	μ	ρ	π	μ	ρ	π	μ	ρ	π
RSSM4	94.4	91.7	91.3	94.2	93.1	92.7	93.9	95.0	94.3
RSSM6	94.7	91.2	92.2	93.9	94.3	94.5	95.0	95.6	93.5
RSSM10	93.1	92.3	92.2	96.2	94.8	94.5	95.1	95.3	92.9

4.2 Different means, dental correlation structure

When the working Model I assumptions of the CLF-RSSM estimator are false such that the true structure of the correlated site-level binary data within a mouth is characterized by different means and dental correlation matrix (Model II), the percent relative bias of $\hat{\pi}$ is about -10% (Table 6). Compared with the relative bias of the prevalence standard error estimator under Model I, the percent relative bias of true standard error of the prevalence estimator under Model II is substantially more severe (Table 7). Also, under Model II, the coverage of 95% CI for prevalence decreases remarkably relative to its performance under Model I. As the number of sampled sites increases, the 95% confidence interval coverage also decreases.

Table 6: Average estimates using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has unequal means and dental correlation structure under the conditional linear family such that the true prevalence is $\pi = 0.542$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$
RSSM4	0.051	0.243	0.491	0.051	0.244	0.487	0.051	0.245	0.485
RSSM6	0.051	0.242	0.491	0.051	0.244	0.486	0.051	0.245	0.485
RSSM10	0.051	0.241	0.490	0.051	0.243	0.487	0.051	0.244	0.485

Table 7: Percent relative bias for prevalence, coverage of 95% CI for prevalence, and percent relative bias for standard deviation estimates of prevalence using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has unequal means and dental correlation structure under the conditional linear family such that the true prevalence is $\pi = 0.542$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	% rel bias		Coverage	% rel bias		Coverage	% rel bias		Coverage
	$\hat{\pi}$	$\hat{\sigma}_\pi$	π	$\hat{\pi}$	$\hat{\sigma}_\pi$	π	$\hat{\pi}$	$\hat{\sigma}_\pi$	π
RSSM4	−9.41	−8.77	81.20	−10.15	−6.35	72.80	−10.49	−3.48	54.10
RSSM6	−9.28	−8.64	78.40	−10.20	−6.45	64.90	−10.41	−8.11	37.40
RSSM10	−9.56	−5.39	73.40	−10.03	−7.54	54.60	−10.39	−6.78	24.60

4.3 Different means, exchangeable correlation structure, with unequal cluster sizes

For unequal cluster sizes (Model III), the estimated prevalence is slightly overestimated relative to π (Table 8) where the percent relative bias of prevalence is about 3% (Table 9). Compared with the relative bias of the prevalence standard error estimator under Model I (Table 4), the percent relative bias of standard error of the prevalence estimator under the situation of missing teeth (Table 9) is similar. However, the coverage of 95% CI for the prevalence decreases slightly under Model III relative to when Model I is the true distribution. As the number of sampled sites increases, the 95% confidence interval coverage decreases slightly.

Table 8: Average estimates using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has unequal means and exchangeable ($\rho = 0.2$) correlation structure with missing teeth under the conditional linear family such that the true prevalence is fixed as $\pi = 0.543$

	500 clusters per replicate			1000 clusters per replicate			2000 clusters per replicate		
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$	$\hat{\mu}$	$\hat{\rho}$	$\hat{\pi}$
RSSM4	0.057	0.203	0.564	0.057	0.201	0.564	0.057	0.201	0.561
RSSM6	0.057	0.201	0.564	0.057	0.201	0.562	0.057	0.203	0.558
RSSM10	0.057	0.202	0.560	0.057	0.203	0.559	0.057	0.203	0.558

Table 9: Percent relative bias for prevalence, coverage of 95% CI for prevalence, and percent relative bias for standard deviation estimates of prevalence using the CLF-RSSM estimator for 1000 simulation replicates when the true distribution for the correlated sites has unequal means and exchangeable correlation ($\rho = 0.2$) structure with missing teeth under the conditional linear family such that the true prevalence is fixed as $\pi = 0.543$

	500 clusters per replicate					
	% rel bias				Coverage	
	$\hat{\rho}$	$\hat{\sigma}_\rho$	$\hat{\pi}$	$\hat{\sigma}_\pi$	ρ	π
RSSM4	1.33	−1.32	3.94	−7.66	93.6	91.8
RSSM6	0.41	−6.92	3.95	−7.29	92.4	89.9
RSSM10	0.80	−5.72	3.16	−12.32	92.2	88.4

	1000 clusters per replicate					
	% rel bias				Coverage	
	$\hat{\rho}$	$\hat{\sigma}_\rho$	$\hat{\pi}$	$\hat{\sigma}_\pi$	ρ	π
RSSM4	0.28	−2.46	3.79	−3.85	94.3	92.3
RSSM6	0.48	−2.46	3.41	−5.31	94.1	90.7
RSSM10	1.27	−4.40	2.89	−11.24	93.8	87.5

	2000 clusters per replicate					
	% rel bias				Coverage	
	$\hat{\rho}$	$\hat{\sigma}_\rho$	$\hat{\pi}$	$\hat{\sigma}_\pi$	ρ	π
RSSM4	0.62	−3.27	3.4	−6.57	93.8	90.2
RSSM6	1.38	−2.51	2.85	−5.7	94.2	90.1
RSSM10	1.35	0.88	2.76	−5.3	94.3	85.8

5 Discussion

This report describes the development of a large-sample variance formula for an estimator of the prevalence of a cluster-level binary variable such as the presence of disease derived from correlated binary observations within the cluster. When clusters are partially observed with observations missing completely at random, such as with random-tooth site selection methods proposed by Beck et al (2006) for chronic periodontitis surveillance in humans, a prevalence estimator that is defined from the disease status of observed tooth sites (i.e., incomplete clusters) may be defined under certain distributional model assumptions for the correlated binary data. Our formula defines prevalence as the probability that a randomly selected individual has one more diseased tooth sites in the full mouth (i.e., completely observed cluster) and is derived under a model (Model I) assuming equal site-level means and exchangeable within-cluster correlation from the conditional linear family of correlated binary distributions (Qaqish, 2003). Simulation studies were used to evaluate the finite-sample performance of the prevalence estimator in moderately large samples of individuals characteristic of epidemiological cohort studies. Not surprisingly, the proposed inferential methods performed well when the working model of equal site-level means and exchangeable correlation was correct such that the prevalence estimator and the proposed large-sample variance estimator were consistent estimators of the true prevalence and variance, respectively. However, their performance was not as good when the working assumptions of Model I were violated, such as with periodontal studies of humans characterized by missing teeth, unequal site-level means (probability of disease at the site) and complex spatial correlation structures for the pattern of disease within the mouth.

When the number of missing teeth varied randomly across individuals but prevalence did not vary (Model III), the proposed CLF-RSSM estimator had fair performance with percent relative bias of about 3% and coverage of nominal 95% confidence intervals (CIs) ranging from 0.86 to 0.92. Coverage was poorest when the number of individuals was greatest ($K = 2000$). The undercoverage of CIs may be the result of the underestimation of the standard error of the prevalence estimator or other factors. Several adjustments could be pursued to see whether they improve performance: (i) CI's for π could be constructed based on inverting CIs for $\theta^* = \log(1-\pi)$ instead of using $\text{var}(\hat{\pi})$ directly in CI construction; (ii) the ad-hoc procedure of inserting the mean cluster size for n in the prevalence formula may need modification, such as by estimating prevalence within strata defined by cluster size and then combining the stratum-specific prevalence estimates using weights proportional to the stratum sizes; (iii) use of an estimation procedure other than GEE, such as maximum likelihood under an assumed model, to obtain estimates of μ and ρ and their covariance matrix. For example, a computationally faster algorithm

for computing prevalence under our working model for correlated binary data may be provided by the beta-binomial distribution because CLF under equal means and exchangeable correlation reduces to the beta-binomial distribution.

Finally, the proposed methods performed most poorly, under data generating Model II that assumed equal cluster sizes, unequal means and a "dental" spatial correlation structure for the correlated binary data. These features resulted in severe undercoverage of 95% confidence intervals for prevalence based on the CLF-RSSM estimator's reliance on a simple mis-specified working model. The development of other estimators of prevalence for RSSMs may be needed to achieve good statistical properties under more realistic assumptions. One modification is to define a CLF-RSSM estimator as in equation (1) except that means are allowed to vary by sites. Without a model for the site-level means, all 168 of them could be allowed to have a potentially distinct value. Otherwise, a model for the means such as specified by Model II could be incorporated into the prevalence estimator as a simple modification of equation (1). If application of the delta method proves difficult or prohibitive, bootstrapping could be use to obtain standard errors and/or confidence intervals. Specification of a correlation structure other than exchangeable would give an estimator of a different form than in equation (1).

The simulation study had other limitations. While the CLF was chosen for generating correlated binary data due to its flexibility in terms of allowing unequal site-level means and general correlation structures, other procedures such as that of Emrich and Piedmonte (1993) could have been used. Generally, the generation of correlated binary data with known marginal means and correlations is complicated (Preisser and Qaqish, 2014). A necessary condition for the existence of a correlated binary distribution with known marginal means and correlations is that the correlation matrix be positive definite. Another necessary condition is that certain bounds on the pairwise correlations are satisfied. The so-called Fréchet bounds are functions of the marginal means. In Model II, three correlations need to be within bounds defined by corresponding marginal means. Additionally, the two sets of conditions cited above, while necessary, are not sufficient to guarantee the existence of a correlated binary distribution with the fixed marginal means and pairwise correlations. Particularly, even if a correlated binary distribution exists with the specified marginal means and pairwise correlations, it may not be a member of the conditional linear family (CLF). To guarantee the CLF compatibility and conditional mean within the range [0,1], we had to specify μ_{xp}, μ_{mp}, μ_{xa}, and μ_{ma} with only trivial departures from equal means. Such conditions limit the scope of simulations that may be conducted based on CLF for correlated binary data at least for larger cluster sizes, such as n=168. Preisser and Qaqish (2014) found that the multivariate probit method of Emrich and Piedmonte (1993) had fewer compatibility restrictions than CLF for "exchangeable-type" correlation matrices such as considered here, implying potential avenues

for future statistical research. Conversely, CLF was better suited for generating longitudinal data with auto-correlated within-cluster association.

Studies of chronic periodontitis have varied on what constitutes an individual with the condition (Page and Eke, 2007). In the proposed statistical approach to estimating disease prevalence, the form of the CLF estimator depends upon both an assumed statistical distribution for correlated binary data as well as the definition of disease. The definition employed in this report defines a case based on an individual having one or more sites in the full mouth affected by the condition. Generating other CLF-based prevalence formulae for case definitions where multiple sites (e.g., two or more) must exceed a specific threshold is possible, while various complex definitions of disease that have been proposed in the dental literature could be difficult to handle with the method described in this report.

References

Beck, J.D., Caplan, D.J., Preisser, J.S., and Moss, K. Reducing the Bias of probing depth and attachment level estimates using random partial-mouth recording. *Community Dentistry and Oral Epidemiology* 2006; 34:1-10.

Marks, S. Estimation of disease prevalence using random partial-cluster sampling. *Department of Biostatistics, University of North Carolina at Chapel Hill master's paper.* 2015.

Page, R.C. & Eke, P.I. Case Definitions for Use in Population-Based Surveillance of Periodontitis. *Journal of Periodontology.* 2007; 78:1387-1399.

Preisser, J.S., Lu, Bing and Qaqish, B.F. Finite sample adjustments in estimating equations and covariance estimators for intracluster correlations. *Statistics in Medicine.* 2008; 27:5764-5785.

Preisser, J.S., and Qaqish, B.F. A comparison of methods for simulating correlated binary variables with specified marginal means and correlation. *Journal of Statistical Computation and Simulation.* 2014; 84:2441-2452.

Prentice, R.L. Correlated binary regression with covariates specific to each binary observation. *Biometrics.* 1988; 44:1033-1048.

Qaqish, B.F. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika.* 2003; 90:455-463.

Zeger, S.L., and Liang, K.Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics.* 1986;42:121-130.