Published in final edited form as:
Stat Med. 2010 February 20; 29(4): 464–473. doi:10.1002/sim.3776.

Optimal Combination of Estimating Equations in the Analysis of Multilevel Nested Correlated Data

J. A. Stoner[1,*,†], **B. G. Leroux**[2], and **M. Puumala**[3]

[1] Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, OK, U.S.A.

[2] Departments of Biostatistics and Dental Public Health Sciences, University of Washington Seattle, WA, U.S.A.

[3] Office of Measurement Services and Minnesota Statewide Testing Program, University of Minnesota, Minneapolis, MN, U.S.A.

SUMMARY

Multilevel nested, correlated data often arise in biomedical research. Examples include teeth nested within quadrants in a mouth or students nested within classrooms in schools. In some settings, cluster sizes may be large relative to the number of independent clusters and the degree of correlation may vary across clusters. When cluster sizes are large, fitting marginal regression models using Generalized Estimating Equations with flexible correlation structures that reflect the nested structure may fail to converge and result in unstable covariance estimates. Also, the use of patterned, nested working correlation structures may not be efficient when correlation varies across clusters. This paper describes a flexible marginal regression modeling approach based on an optimal combination of estimating equations. Particular within-cluster and between-cluster data contrasts are used without specification of the working covariance structure and without estimation of covariance parameters. The method involves estimation of the covariance matrix only for the vector of component estimating equations (which is typically of small dimension) rather than the covariance matrix of the observations within a cluster (which may be of large dimension). In settings where the number of clusters is large relative to the cluster size, the method is stable and is highly efficient, while maintaining appropriate coverage levels. Performance of the method is investigated with simulation studies and an application to a periodontal study.

Keywords

Between-cluster Effect; Estimation Efficiency; Generalized Estimating Equations; Multilevel Nested Correlated Data; Within-cluster Effect

1. INTRODUCTION

Multilevel (nested) correlated data often arise in biomedical research through cluster-sampling studies, such as teeth within quadrants in a mouth or students within classrooms in schools. The analysis of multilevel correlated data often focusses on mixed, or cluster-specific, modeling [1,2,3,4]. In some settings, inference may focus on marginal regression

*Correspondence to: Julie A. Stoner, Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, 801 NE 13th Street, CHB 309, PO Box 26901, Oklahoma City, Oklahoma 73104, U.S.A..
†julie-stoner@ouhsc.edu

model parameters, which is the topic of this paper. Liang and Zeger [5] proposed the method of generalized estimating equations (GEE) for correlated data analysis that separates the specification of a model for the marginal mean response from the modeling of the correlation structure, which allows valid inferences about mean parameters without correct specification of the correlation structure.

Independence Estimating Equations (IEE) protect against bias due to certain forms of model misspecification [6], while providing equal efficiency compared with optimal estimating equations in special cases [7]. On the other hand, large gains in efficiency are sometimes possible through the use of working correlation structures other than independence [7,8,9,10].

The working covariance structures that are commonly implemented in software packages do not reflect multi-level nested correlation structures, where we may expect observations within a sub-cluster to be more highly correlated than observations from different sub-clusters within the same cluster. In practice, often only a single level of nesting is accounted for when using GEE. Examples include Kadohira et al. [11], Sashegyi et al. [12], and Morrow et al. [13] where within-farm correlation was considered while ignoring within-area and within-district correlation, within-student correlation was considered while ignoring random school effects, and within-subject correlation among repeated longitudinal measures was considered while ignoring residential clusters, respectively. A further limitation of the GEE approach is that large cluster sizes and complicated data structures that arise in applied research, including spatial correlation among measurements within a mouth in dental research, may result in intractable estimators involving empirical covariance matrices, such as GEE with an unstructured working correlation matrix, particularly when the number of independent clusters is limited.

Chao [14] proposed an extension of GEE to address hierarchical (multilevel nested) correlated data settings that involves specification of multiblock and multilayer correlation structures. The method can be applied in settings with large clusters and in settings with heterogeneous correlation across clusters. Under the method, patterned correlation matrices are specified and corresponding correlation parameters are estimated using a Fisher scoring equation. Simulation studies demonstrate possible efficiency gains relative to standard working structures [14]. Under this method, there is a potential for over-parametrization and consequent convergence problems in settings with a small number of clusters.

Alternate methods for fitting marginal regression models have been proposed. Stoner and Leroux [15] proposed the method of Optimal Combined Estimating Equations (OCEE) that empirically weights and combines specific data contrasts, such as between-cluster and within-cluster contrasts, and showed that increased efficiency over GEE with standard working correlation structures, while maintaining coverage levels, is possible without the need to specify a working correlation structure. A related method based on quadratic inference functions was developed by Qu [16]. Under this method, an extended score function is defined based on the working correlation structures typically used under GEE, which may result in a score vector of high dimensions and requires the specification of a working correlation matrix [16]. Qu, Lindsay, and Li [17] have further developed their approach to address situations where inversion of the estimated covariance matrix is difficult.

This paper develops the extension of the OCEE method to multilevel correlated data settings with large cluster sizes and compares the performance of OCEE to GEE using standard working correlation assumptions and the multiblock and multilayer matrices proposed by Chao [14]. Section 2 defines the notation and presents a summary of the OCEE method. The choice of contrast matrices in a multilevel data setting also is discussed in Section 2. Section 3 presents simulation results comparing the performance of OCEE and GEE. An application of the methodology is given in Section 4. Concluding remarks are given in Section 5.

2. METHODS

A general two-level nested data structure will be considered, although the methods can easily be extended to additional levels of nesting. Let $i = 1, \ldots, C$ index clusters, $j = 1, \ldots, l_i$ index sub-clusters, and $k = 1, \ldots, m_{ij}$ index observations within a sub-cluster. The total cluster size will be denoted by $n_i = \sum_{j=1}^{l_i} m_{ij}$. Let $\mathbf{y}_i$ denote the $n_i \times 1$ vector of outcome values, which depend on an $n_i \times p$ covariate matrix $\mathbf{X}_i$ for $i = 1, \ldots, C$ clusters. We will assume a generalised linear model for the mean response $\mu_{ijk} = E(y_{ijk} \mid \mathbf{X}_i)$ given by $g(\mu_{ijk}) = \mathbf{X}_{ijk}^{\top}\beta$, where g is a known link function and β is the $p \times 1$ coefficient parameter vector of interest. We also assume that the variance is given by $\text{var}(y_{ijk} \mid \mathbf{X}_i) = \varphi v(\mu_{ijk})$, where φ is a dispersion parameter and v is a known variance function. In addition, we assume that observations within sub-clusters and within clusters are correlated while observations from different clusters are independent. Direct sum notation, as presented in Searle et al. [18], will be used where the notation $\oplus_{i=1}^{C} B_i$ indicates a block diagonal matrix with non-zero matrices B_i along the diagonal and 0 values elsewhere.

To motivate the proposed method, we will examine the inherent weighting of between-cluster and within-cluster sources of information by GEE. As discussed by Mancl and Leroux [7] and Stoner and Leroux [15], GEE can be expressed in terms of weighted sums of estimating functions based on particular data contrasts. OCEE extends this observation to involve more flexible data contrasts and optimal combinations of the estimating functions based on different contrasts.

For illustration, as shown in Stoner and Leroux [15], consider a simple linear model given by $y_{ij} = \left(x_{ij} - \bar{x}_{..}\right)^{\top}\beta + \epsilon_{ij}$, where all clusters are of size n and $(x_{ij} - \bar{x}_{..})$ is a single covariate that has been centered at the mean covariate over all observations and only a single level of clustering. The form of generalized estimating equations can be written as

$$D^{\top} V^{-1} (Y - \mu) = 0, \tag{1}$$

where $D = \partial\mu / \partial\beta^{\top}$, $V^{-1} = \oplus_{i=1}^{C} \frac{1}{\phi} A_i^{-1/2} P^{-1} A_i^{-1/2}$, $A_i = \text{diag}\{v(\mu_{ij})\}$ and P is a working correlation matrix.

The inverse of a working exchangeable structure with correlation parameter ρ can be written as

$$V^{-1} = \oplus_{i=1}^{C} \frac{1}{\phi} \left[\frac{1}{\{1 + (n-1)\rho\}} \frac{1}{n} J_i + \frac{1}{1-\rho} \left(I_i - \frac{1}{n} J_i \right) \right],$$

where J_i is an $n \times n$ matrix of 1's and $\mathrm{I_i}$ is an $n \times n$ identity matrix. By inserting this form for V^{-1} into (1) we have

$$GEE_{\text{exch}}=\frac{1}{\phi}\frac{1}{k_B}D^{\top}W_B(Y-\mu)+\frac{1}{\phi}\frac{1}{k_W}D^{\top}W_W(Y-\mu)=0,$$

where $k_B = 1 + (n - 1)\rho$, $k_W = 1 - \rho$, $W_B=\oplus_{i=1}^{C}(\frac{1}{n}J_i)$ and $W_W=\oplus_{i=1}^{C}(I_i-\frac{1}{n}J_i)$. The matrix W_B can be seen as a contrast of between-cluster means and W_W can be seen as a within-cluster contrast of an individual response with the cluster-level mean when GEE_{exch} is rewritten as

$$\begin{aligned} GEE_{\text{exch}} &= \frac{1}{\phi}\frac{1}{k_B}X_B^{\top}\left(Y_B-X_B^{\top}\beta\right)+\frac{1}{\phi}\frac{1}{k_W}X_W^{\top}\left(Y_W-X_W^{\top}\beta\right)\\ &\equiv \frac{1}{\phi}\frac{1}{k_B}G_B+\frac{1}{\phi}\frac{1}{k_W}G_W=0, \end{aligned}$$

where

$$\begin{aligned} X_B^{\top} &= \left[\bar{x}_{1.}-\bar{x}_{..},\cdots,\bar{x}_{1.}-\bar{x}_{..},\cdots,\bar{x}_{C.}-\bar{x}_{..}\right], \quad X_W^{\top}=\left[x_{11}-\bar{x}_{1.},\cdots,x_{1n}-\bar{x}_{1.},\cdots,x_{Cn}-\bar{x}_{C.}\right],\\ Y_B^{\top} &= \left[\bar{y}_{1.}-\bar{y}_{..},\cdots,\bar{y}_{1.}-\bar{y}_{..},\cdots,\bar{y}_{C.}-\bar{y}_{..}\right], \quad Y_W^{\top}=\left[y_{11}-\bar{y}_{1.},\cdots,y_{1n}-\bar{y}_{1.},\cdots,y_{Cn}-\bar{y}_{C.}\right] \end{aligned}$$

and G_B and G_W are estimating functions based on between-cluster and within-cluster information, respectively.

This example demonstrates that generalized estimating equations under particular working covariance structures can be written as weighted sums of estimating equations based on particular data contrasts. The data contrasts and weights corresponding to standard working structures may or may not be optimal depending on the agreement between the working and true covariance structures. In general, more flexible data contrasts and weights, can improve estimation efficiency while maintaining appropriate coverage probability levels.

In Stoner and Leroux [15], we introduced the OCEE method. OCEE weights different sources of information as captured by component estimating functions of the form $D^{\top}W(\mathbf{y}-\mu)=\sum_{i=1}^{C}D_i^{\top}W_i(\mathbf{y}_i-\mu_i)$ where y_i is the response vector for the i^{th} cluster with mean μ_i, $D_i = \partial\mu_i/\partial\beta^{\top}$ is its derivative with respect to the parameter vector β, and W is a block-diagonal weight matrix that reflects a particular source of information. We proposed a general block diagonal form for weight matrices W,

$$W=\oplus_{i=1}^{C}A_i^{-1/2}\left(H_i^{\top}H_i\right)A_i^{-1/2}, \tag{2}$$

where $A_i = diag\{v(\mu_{ijk})\}$ with v a variance function, and H_i corresponds to a particular data contrast.

Examples of the general form (2) are between-cluster and within-cluster weight matrices, $W_B=\oplus_{i=1}^{C}A_i^{-\frac{1}{2}}\left(\frac{1}{n_i}J_i\right)A_i^{-\frac{1}{2}}$ and $W_W=\oplus_{i=1}^{C}A_i^{-\frac{1}{2}}\left(I_i-\frac{1}{n_i}J_i\right)A_i^{-\frac{1}{2}}$, which are motivated by the form of the inverse of an exchangeable correlation matrix. (J_i is an $n_i \times n_i$ matrix of 1's and I_i an $n_i \times n_i$ identity matrix, with n_i the cluster size.) To derive the OCEE estimating equation, we applied Heyde's [19] optimal linear combination of estimating functions to a set of estimating functions of the general form $D^{\top}W_k(Y-\mu)$ with different weight matrics W_k ($k = 1, \ldots, r$), with the result

$$\text{OCEE} \quad : \quad \begin{bmatrix} \Sigma_{i=1}^{C} D_i^{\top} W_{1i} D_i \\ \vdots \\ \Sigma_{i=1}^{C} D_i^{\top} W_{ri} D_i \end{bmatrix}^{\top} \left[\sum_{i=1}^{C} \begin{pmatrix} D_i^{\top} W_{1i} \\ \vdots \\ D_i^{\top} W_{ri} \end{pmatrix} cov\,(y_i) \begin{pmatrix} D_i^{\top} W_{1i} \\ \vdots \\ D_i^{\top} W_{ri} \end{pmatrix}^{\top} \right]^{-1} \times \left[\sum_{i=1}^{C} \begin{pmatrix} D_i^{\top} W_{1i} (Y_i - \mu_i) \\ \vdots \\ D_i^{\prime} W_{ri} (Y_i - \mu_i) \end{pmatrix} \right] = 0. \tag{3}$$

Replacing cov(y_i) with $(y_i - \mu_i)(y_i - \mu_i)^{\top}$ and solving by iterated weighted least squares (IWLS) gives the OCEE estimators. Note that there is no working correlation structure to specify, nor any correlation parameters to estimate, which avoids issues raised by Crowder [20] regarding the breakdown of the asymptotic properties of GEE. We showed that OCEE estimators are asymptotically multivariate normal and developed a consistent sandwich-type covariance estimate [15].

The contrast matrices of interest in a multilevel nested setting can be motivated by decomposing the inverse of symmetric, patterned correlation matrices. For example, the inverse of a patterned exchangeable correlation matrix for 2 sub-clusters each of size 2 nested within a single cluster where observations within the same sub-cluster are correlated with coefficient ρ_1 and observations within the same cluster from different sub-clusters are correlated with coefficient ρ_2 can be written as

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_2 \\ \rho_1 & 1 & \rho_2 & \rho_2 \\ \rho_2 & \rho_2 & 1 & \rho_1 \\ \rho_2 & \rho_2 & \rho_1 & 1 \end{bmatrix}^{-1} = c_1 J + c_2 \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} + c_3 I \tag{4}$$

where $c_1 = \frac{-\rho_2}{\rho_1^2 + 2\rho_1 + 1 - 4\rho_2^2}$, $c_2 = \frac{\rho_1^2 + \rho_1 - 2\rho_2^2 + \rho_1\rho_2 - \rho_2}{(\rho_1 - 1)\left(\rho_1^2 + 2\rho_1 + 1 - 4\rho_2^2\right)}$, $c_3 = \frac{-\rho_1^2 - 2\rho_1 + 4\rho_2^2 - 1}{(\rho_1 - 1)\left(\rho_1^2 + 2\rho_1 + 1 - 4\rho_2^2\right)}$, I is a 4 × 4 identity matrix, and J is a 4 × 4 matrix of 1's. The matrix decomposition suggests that data contrasts among cluster-level means, sub-cluster-level means and individual observations may be useful. Even when sub-clusters are of differing size or the within sub-cluster correlation differs across sub-clusters, the decomposition presented in (4) still holds where the scalers c_2 and c_3 are replaced by matrices.

In a 2-level nesting situation, where we have observations nested within sub-clusters nested within clusters, we propose to use three different contrast matrices. The first is a within-sub-cluster contrast between individual observations and the sub-cluster mean denoted by $H_{ai} = \left(I_i - \oplus_{j=1}^{l_i} \frac{1}{m_{ij}} J_{ij}\right)$ where J_{ij} is an $m_{ij} \times m_{ij}$ matrix of 1's and I_i an $n_i \times n_i$ identity matrix. Second, a between-sub-cluster contrast between sub-cluster means and the cluster-level mean denoted by $H_{bi} = \left(\oplus_{j=1}^{l_i} \frac{1}{m_{ij}} J_{ij} - \frac{1}{n_i} J_i\right)$ where J_i is an $n_i \times n_i$ matrix of 1's. Finally, a between-cluster contrast of cluster-level means denoted by $H_{ci} = \frac{1}{n_i} J_i$. For the example presented earlier, the contrast matrices would be:

NIH-PA Author Manuscript
NIH-PA Author Manuscript

$$H_a = I - \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix},$$

$$H_b = \begin{bmatrix} 1/2 & 1/2 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 1/2 \end{bmatrix} - \frac{1}{4}J, \quad \text{and} \quad H_c = \frac{1}{4}J$$

where I is a 4×4 identity matrix and J is a 4×4 matrix of 1's. The set of contrast matrices can be extended to reflect additional levels of nesting by inserting matrices that reflect sub-sub-cluster means and contrasting means at consecutive levels (observation level, sub-sub-cluster level, sub-cluster level, and cluster level).

3. SIMULATION STUDIES

Twelve different simulation settings, patterned after the periodontal clinical trial application in Section 4, under two different models were considered where the cluster sizes and covariance parameters were constant or non-constant across clusters in a 2 × 2 factorial design and data sets of 25, 50, and 200 clusters were generated. In the constant cluster size setting, data were generated for 4 sub-clusters, each of size 6, resulting in a cluster size of 24. In the non-constant cluster size cases, there were either 3 or 4 sub-clusters per cluster with sub-cluster sizes of 3 or 6, resulting in cluster sizes of 12 to 24.

Under the first model, the response variable $\mathbf{y}_i$ for cluster i was generated from a linear model where $\mathbf{y}_i = \beta_0 + \beta_1 \mathbf{x1}_i + \beta_2 \mathbf{x2}_i + \beta_3 \mathbf{x3}_i + \varepsilon_i$ where $\beta_0 = 1$, $\beta_1 = 0.5$, $\beta_2 = 0.5$, and $\beta_3 = 0.5$ and ε_i, a vector of length equal to the cluster size, was generated from a multivariate normal distribution with mean 0 and a nested, exchangeable correlation structure where observations in the same sub-cluster were correlated with coefficient 0.4 in the constant correlation case, and with a varying coefficient between 0.2 to 0.4 in the non-constant correlation case, and observations from different sub-clusters in the same cluster were correlated with coefficient 0.2, and with a varying coefficient between 0.1 to 0.2 in the non-constant correlation case. The variance parameter was 5, and ranged from 1 to 5 across clusters in the non-constant case.

The covariate **x1** was a dichotomous covariate with a constant value in some clusters and varied within clusters and within sub-clusters for other clusters, resulting in approximately half of the variation in the covariate to be due to between-cluster variation. Under settings of non-constant covariance parameters, clusters with higher correlation and lower variance parameters had more within-cluster covariate variation. The covariate **x2** was a cluster-level covariate, generated independently from **x1** using a normal distribution with mean 58 and standard deviation 6 (reflective of the age of the subjects in the periodontal clinical trial application that was used to motivate the design of the simulation study). Finally, covariate **x3** was a dichotomous covariate that varied only within clusters, having a mean value of 0.5 for each cluster, and was generated independently of **x1** and **x2**.

Results are summarized for β_1 over 1000 generated data sets, for sets with 50 independent clusters in Table I and then for sets with 200 independent clusters in Table II. Results from data sets with 25 clusters are briefly described and are included in the graphical summary in Figure 1 that summarizes the performance of OCEE relative to the GEE approaches according to the simulation study design factors. The analysis methods being compared are generalized estimating equations with an independence (GEE_i), exchangeable (GEE_e) or

NIH-PA Author Manuscript

NIH-PA Author Manuscript

nested (GEE_n) working structure using the method of Chao [14] to estimate the nested structure parameters. Two variations of the optimal combined estimating equations method are used: 1) using between- and within-cluster contrasts ($OCEE_{bw}$) involving the cluster-level mean and the difference between individual observations and the cluster-level mean (r=2); and 2) using between-cluster, between sub-cluster, and within sub-cluster contrasts ($OCEE_n$) (r=3) as presented in Section 2.

For β_1, bias was small across all methods, with average coefficient estimates within 0.019 of the true parameter when 50 clusters were considered and within 0.006 when 200 clusters were considered. Coverage probabilities of the 95% confidence intervals were as low as 0.88 for the OCEE methods when considering settings with 50 clusters, but improved to values no less than 0.93 when 200 clusters were considered. Estimation efficiency, defined as the ratio of the calculated variance of $\widehat{\beta}$ under GEE_n relative to GEE_e, tended to be higher, by as much as 22%, under the GEE_n method relative to the GEE_e method when covariance was constant and 200 clusters were considered. Estimation efficiency relative to $OCEE_n$ was as low as 0.66 (a 34% loss in efficiency relative to $OCEE_n$) when the covariance was not constant across clusters and was as high as 1.24 (a 24% gain in efficiency relative to $OCEE_n$) under the GEE_n method when the cluster size was constant and 50 clusters were considered, but was reduced to values no greater than 1.02 when 200 clusters was considered. Efficiency gains were most notable for the covariate **x1** in the settings where the covariance parameters were non-constant, but were also seen relative to GEE_i and relative to GEE_e, with a large number of clusters, when the covariance was constant across clusters. Relative efficiency calculations suggest as much as a 19% loss in efficiency for $OCEE_{bw}$ relative to $OCEE_n$.

Solution of the OCEE involves estimation of the empirical covariance matrix for the vector of estimating functions. The OCEE estimates failed to converge only 1 out of 1000 times for the $OCEE_n$ method when constant covariance parameters and constant cluster sizes or non-constant covariance parameters and constant cluster sizes were considered in settings with 50 clusters (these data sets are not included in the data summaries presented in Table I for any of the methods). Non-convergence occurred in fewer than 2.4% of the 1000 replicates in each of the 4 simulated data settings (constant or non-constant cluster size or covariance) when only 25 clusters were considered. This compares to a non-convergence rate of greater than 95% in each of the 4 simulated data settings for GEE with an unstructured working covariance structure when only 25 clusters were considered (data not shown). In addition to non-convergence issues, OCEE resulted in low coverage probabilities ranging from 0.79 to 0.91 across the 4 simulation settings with data sets of 25 clusters.

In summary, when the number of clusters was 25 and clusters were large (24 observations per cluster), non-convergence rates of approximately 2.5% and coverage probabilities for 95% confidence intervals from 79% to 91% suggest that more structured approaches, such as GEE with a nested or exchangeable working correlation structure, are preferred over OCEE. In settings with 50 clusters of large size (24 observations per cluster), OCEE with between-, within-subcluster, and within-cluster weights results in increased estimation efficiency, while maintaining coverage probabilities that were greater than 90%, relative to GEE with exchangeable or nested working structures, when the covariance parameters are non-constant. In this same cluster size setting, GEE with a nested working structure results in increased estimation efficiency relative to GEE with an exchangeable working structure or OCEE methods when the covariance parameters are constant. Coverage probabilities are still somewhat low for OCEE using nested sub-cluster contrasts when considering 50 clusters, but are similar to the other methods when considering 200 clusters of large size. When 200 clusters are considered, estimation efficiency is greatest among the GEE and OCEE nested approaches in settings with constant covariance parameters and is greatest for the OCEE nested method when the covariance parameters are non-constant. When covariance parameters are non-constant, performance of GEE using an exchangeable working structure is similar to GEE using a nested working structure.

When considering β_2, the coefficient for the cluster-level covariate, the performance was similar among the GEE and OCEE methods. Coverage probabilities were at least 0.90 for all methods when 50 clusters were considered and at least 0.91 when 200 clusters were considered. Relative efficiency of the methods to $OCEE_n$ ranged from 0.83 under GEE_i to 1.14 under GEE_n when covariance parameters were constant and ranged from 0.90 under GEE_i to 1.07 under GEE_n when covariance parameters were non-constant in settings with 50 clusters. In settings with 200 clusters, relative efficiency of the methods to $OCEE_n$ ranged from 0.76 under GEE_i to 1.09 under GEE_n when covariance parameters were constant and ranged from 0.92 under GEE_i to 1.07 under GEE_n when covariance parameters were non-constant. When considering the coefficient for the covariate that varied within clusters but was constant on average across clusters, β_3, the performance again was similar among the GEE and OCEE methods. Coverage probabilities were at least 0.91 and 0.92 for all methods when 50 and 200 clusters were considered, respectively. Relative efficiency of the methods to $OCEE_n$ ranged from 1.03 under GEE_i to 1.17 under GEE_n when covariance parameters were constant and ranged from 1.03 under GEE_i to 1.12 under GEE_n when covariance parameters were non-constant in settings with 50 clusters. In settings with 200 clusters, relative efficiency of the methods to $OCEE_n$ ranged from 0.99 to 1.04 across all simulation settings.

Under a second model, the response variable $\mathbf{y}_i$ for cluster i was generated from a linear model where $\mathbf{y}_i = \beta_0 + \beta_1\mathbf{x1}_i + \beta_2\mathbf{x2}_i + \varepsilon_i$ where $\beta_0 = 1$, $\beta_1 = 0.5$, and $\beta_2 = -0.7$ and ε_i, a vector of length equal to the cluster size, was generated from a multivariate normal distribution with mean 0 and the same nested, exchangeable covariance structure as was used under the first model. The covariate **x1** was generated as under the first model. Covariate **x2** was generated, independently from **x1**, using a multivariate normal distribution with a random mean, generated from a normal distribution with mean 9 and standard deviation of 1, and a nested, exchangeable covariance structure where observations from the same sub-cluster were correlated with coefficient 0.6 and measures from different sub-clusters within the same mouth were correlated with coefficient 0.3 and variance parameter of 1.5.

Results for β_1 over 1000 generated data sets were very similar to the results for β_1 under the first model (data not shown). When considering β_2, the performance was similar between the GEE_n and $OCEE_n$ methods, but was improved under $OCEE_n$ relative to GEE_i and GEE_e methods. Coverage probabilities were at least 0.91 and 0.94 for all methods when 50 and 200 clusters were considered, respectively. Relative efficiency of the methods to $OCEE_n$ ranged from 0.43 and 0.88 under GEE_i and GEE_e, respectively, to 1.13 under GEE_n when covariance parameters were constant and ranged from 0.60 and 0.99 under GEE_i and GEE_e, respectively, to 1.08 under GEE_n when covariance parameters were non-constant in settings with 50 clusters. In settings with 200 clusters, relative efficiency of the methods to $OCEE_n$ ranged from 0.41 and 0.77 under GEE_i and GEE_e, respectively, to 1.04 under GEE_n when covariance parameters were constant and ranged from 0.64 and 0.95 under GEE_i and GEE_e, respectively, to 1.04 under GEE_n when covariance parameters were non-constant.

4. APPLICATIONS

Data from a two-center placebo-controlled, double-blind randomized clinical trial of subantimicrobial dose doxycycline (SDD; 20 mg bid) for treatment of alveolar bone density loss and periodontitis over a 2-year period (R01DE012872, Dr. J. Payne PI) [22] were

NIH-PA Author Manuscript

NIH-PA Author Manuscript

analyzed. The target population consisted of postmenopausal, osteopenic, estrogen-deficient women on periodontal maintenance therapy for moderate to advanced chronic periodontitis. This application focuses on the association between 12-month changes in alveolar bone density and 24-month changes in alveolar bone height among 61 subjects randomized to receive placebo. It was hypothesized that changes in alveolar bone density were positively associated with subsequent changes in bone height. Alveolar bone density changes at the 12-month visit relative to baseline were determined using the CADIA method [23]. The CADIA measures were each coded into 3 categories of change (gain in bone density, no change, loss of bone density) using defined thresholds [22]. Alveolar bone height measurements between the fixed reference point (cementoenamel junction or restoration margin) and the alveolar crest were made for baseline and 24-month radiographs. Increases in the alveolar bone height measurements correspond to loss of alveolar bone height. Up to 6 site-level measurements in each of 4 posterior sextants was made per subject. The number of measurements within a mouth ranged from 8 to 24 with a median cluster size of 18. Measurements were made on the mandible and maxilla (lower and upper jaws), considered in this example as the sub-cluster, where sub-cluster sizes ranged from 1 to 12 with a median sub-cluster size of 10.

The OCEE method was used to fit a model that weighted and combined between-mouth and within-mouth contrasts of the 12-month indicator variables for alveolar bone density change ($OCEE_{bw}$), which were the mouth-level mean values and the difference between a tooth site-level measurement and the mouth-level mean value (r=2), respectively. The OCEE method also was used to fit a model that weighted and combined between-mouths contrasts, between-jaw contrasts, and within-jaw contrasts of the bone density change indicator variables ($OCEE_n$), which were the mouth-level mean values, the difference between the jaw-level means and the mouth-level mean, and the difference between tooth-site level measures and the jaw-level mean measure (r=3), respectively. The working assumptions used under GEE were independence, exchangeable, and a nested exchangeable structure where measures within the maxilla/mandible were correlated with coefficient ρ_1 and measures within a mouth from different halves were correlated with coefficient ρ_2. The coefficient, standard error, and p-value estimates are presented in Table III. Positive coefficient estimates correspond to greater alveolar bone height loss for sites with no change or a decrease in density at 12 months compared to sites that demonstrate an increase in density at 12 months. The coefficient estimates are similar for the GEE_i, $OCEE_{bw}$ and $OCEE_n$ methods. The GEE_i and OCEE methods suggest that sites demonstrating decreases in bone density are associated with an average bone height loss of 0.078 to 0.081 mm beyond the bone height changes seen among sites that demonstrate an increase in density, which is significant at the 0.05 alpha level compared to GEE methods utilizing an exchangeable or nested working correlation structure where the estimated loss in bone height is not significant. Lower significance values for the OCEE methods are a function of increased coefficient estimates and decreased standard error estimates.

5. DISCUSSION

The OCEE method using between-cluster, between-sub-cluster and within-sub-cluster contrasts is a flexible approach to the analysis of multilevel nested data. Under OCEE, no covariance parameters must be directly estimated.

The simulation studies demonstrate that there can be important differences in the performance of OCEE relative to GEE. OCEE can result in increased estimation efficiency relative to the GEE method for continuous response variables, particularly when the covariance structure varies across clusters, and still maintain acceptable coverage probabilities. In cases where the covariance is non-constant and is associated with the

NIH-PA Author Manuscript

NIH-PA Author Manuscript

covariate distribution (β_1 in the simulation studies), the efficiency loss under GEE may be as high as 34% relative to OCEE. In settings with a limited number of clusters relative to the cluster size where the covariance is constant, losses in efficiency and reductions in coverage probabilities may occur under OCEE relative to GEE. Although estimation efficiency in these cases can be recovered by including correlation parameter estimates and cluster sizes in the weight matrices of OCEE, as described in Stoner and Leroux [15], the proposed implementation of OCEE is preferred in order to avoid direct estimation of correlation parameters. Furthermore, efficiency losses are limited, particulary when a large number of clusters is considered. Although coverage probabilities may be low when OCEE is used in settings with a limited number of large clusters, the coverage probabilities are improved when a large number of clusters is considered.

It is interesting to note that performance of GEE with an exchangeable correlation structure was similar, and even slightly better in terms of estimation efficiency, to that of GEE with a more complicated nested correlation structure in a number of settings, particularly when the covariance parameters were non-constant across the clusters.

Additional work on OCEE is needed to identify useful data contrasts in multi-level nested data settings where the expected correlation is more complicated than a patterned exchangeable structure. For example, children nested in classrooms within schools may be followed longitudinally, in which case, contrasts over time within a student would also be of interest. Extensions of OCEE to these more complicated correlated data settings could be compared to alternative methods of analysis including the multi-layer and multi-block GEE methods of Chao [14] and the quadratic inference function methods proposed by Qu, Lindsay, and Li [17].

Software to implement the OCEE method is available at the following website: http://www.coph.ouhsc.edu/coph/bse/documents/ocee_code.html. Software to implement the GEE method, including the multiblock and multilayer working correlation structures proposed by Chao, is available in the Splus Correlated Data library (Copyright 1988, 2005 Insightful Corp.).

Acknowledgments

This research is supported by the National Institutes of Health (NIH) grant R01DE015651. The authors would like to thank Dr. Jeffrey Payne at the University of Nebraska Medical Center College of Dentistry for the use of the periodontal data from his NIH-funded clinical trial (National Institute of Dental and Craniofacial Research grant R01DE012872).

Contract/grant sponsor: National Institutes of Health grants ; contract/grant number: R01DE015651 R01DE012872

REFERENCES

1. Neuhaus JM, Kalbfleisch JD, Hauck WW. A comparison of cluster-specific and population-averaged approaches for analyzing correlated data. International Statistical Review. 1991; 59(1):25–35.
2. Lambert PC, Burton PR, Abrams KR, Brooke AM. The analyiss of peak expiratory flow data using a three-level hierarchical model. Statistics in Medicine. 2004; 23(24):3821–3839. DOI: 10.1002/sim.1951. [PubMed: 15580603]
3. Murray DM. Statistical models appropriate for designs often used in group-randomized trials. Statistics in Medicine. 2001; 20(9):1373–1385. DOI: 10.1002/sim.674. [PubMed: 11343359]
4. Heitjan DF, Sharma D. Modelling repeated-series longitudinal data. Statistics in Medicine. 1997; 16(4):347–355. DOI: 10.1002/(SICI)1097-0258(19970228)16:4<347::AID-SIM423>3.0.CO;2-W. [PubMed: 9044525]

5. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22. DOI: 10.1093/biomet/73.1.13.
6. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. Communications in Statistics—Simulation and Computation. 1994; 23(14):939–951.
7. Mancl LA, Leroux BG. Efficiency of regression estimates for clustered data. Biometics. 1996; 52(2):500–511.
8. Neuhaus JM. Estimation efficiency and tests of covariate effects with clustered binary data. Biometrics. 1993; 49(4):989–996. [PubMed: 8117909]
9. Lipsitz SR, Fitzmaurice GM, Orav EJ, Laird NM. Performance of generalized estimating equations in practical situations. Biometrics. 1994; 50(1):270–278. [PubMed: 8086610]
10. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics. 1995; 51(1):309–317. [PubMed: 7766784]
11. Kadohira M, McDermott JJ, Shoukri MM, Thornburn MA. Assessing infections at multiple levels of aggregation. Preventive Veterinary Medicine. 1997; 29(3):161–177. [PubMed: 9234402]
12. Sashegyi AI, Brown KS, Farrell PJ. Application of a generalized random effects regression model for cluster-correlated longitudinal data to a school-based smoking prevention trial. American Journal of Epidemiology. 2000; 152(12):1192–1200. DOI: 10.1093/aje/152.12.1192. [PubMed: 11130626]
13. Morrow AL, Guerrero ML, Shults J, Calva JJ, Lutter C, Bravo J, Ruiz-Palacios G, Morrow RC, Butterfoss FD. Efficacy of home-based peer counselling to promote exclusive breastfeeding: a randomised controlled trial. Lancet. 1999; 353(9160):1226–1231. DOI: 10.1016/S0140-6736(98)08037-4. [PubMed: 10217083]
14. Chao EC. Structured correlation in models for clustered data. Statistics in Medicine. 2006; 25(14): 2450–2468. DOI: 10.1002/sim.2368. [PubMed: 16220520]
15. Stoner JA, Leroux BG. Analysis of correlated data: A combined estimating equations approach. Biometrika. 2002; 89(3):567–578. DOI:10.1093/biomet/89.3.567.
16. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000; 87(4):823–836. DOI: 10.1093/biomet/87.4.823.
17. Qu A, Lindsay BG. Building adaptive estimating equations when inverse of covariance estimation is difficult. Journal of the Royal Statistical Society Series B. 2003; 65(1):127–142. DOI: 10.1111/1467-9868.00376.
18. Searle, SR.; Casella, G.; McCulloch, CE. Variance Components. John Wiley and Sons, Inc.; 1992. p. 443
19. Heyde, CC. Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation. Springer–Verlag; 1997. p. 209-213.
20. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. Biometrika. 1995; 82:407–410. DOI: 10.1093/biomet/82.2.407.
21. Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. The American Statistician. 1996; 50(4):306–310.
22. Payne JB, Stoner JA, Nummikoski PV, Reinhardt RA, Goren AD, Wolff MS, Lee H, Lynch JC, Valente R, Golub LM. Subantimicrobial dose doxycycline effects on alveolar bone loss in postmenopausal women. Journal of Clinical Periodontology. 2007; 34(9):776–787. DOI: 10.1111/j.1600-051X.2007.01115.x. [PubMed: 17716313]
23. Bragger U, Pasquali L, Rylander H, Carnes D, Kornman KS. Computer-assisted densitometric image analysis in periodontal rediography. A methodological study. Journal of Clinical Periodontology. 1988; 15(1):27–37. DOI: 10.1111/j.1600-051X.1988.tb01551.x. [PubMed: 3276740]

NIH-PA Author Manuscript
NIH-PA Author Manuscript

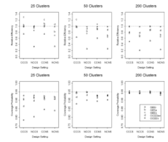

Figure 1.

Summary of simulation study results for β_1, the coefficient for a covariate that varied between and within clusters, under the first model that also included a covariate that varied only between clusters and a covariate that varied only within clusters, across design settings (CCCS: Constant Covariance, Constant Size; NCCS: Non-constant Covariance, Constant Size; CCNS: Constant Covariance, Non-constant Size; NCNS: Non-constant Covariance, Non-constant Size) according to the number of clusters (C=25, 50, or 200). The first row of figures summarizes relative efficiency, calculated as the ratio of the calculated variance of $\widehat{\beta}_1$ under $OCEE_n$ relative to each method, and the second row summarizes the coverage probability of 95% confidence intervals for β_1. Summarized methods include GEE methods, using independence(i), exchangeable(e) and nested(n) working structures, compared to OCEE, combining between- and within-cluster(bw) in addition to nested subcluster(n) data.

Table I

Performance of Generalized Estimating Equations (GEE) and Optimal Combination of Estimating Equations (OCEE) approaches in a dental simulation setting with moderate correlation and cluster sizes. GEE methods, using independence(i), exchangeable(e) and nested(n) working structures, compared to OCEE, combining between- and within-cluster(bw) in addition to nested subcluster(n) data, for a linear model with a dichotomous covariate that varies between and within clusters and sub-clusters ($\mathbf{x_1}$), a continuous covariate that varies only between clusters ($\mathbf{x_2}$) (results not shown), and a dichotomous covariate that varies only within clusters ($\mathbf{x_3}$) (results not shown). The correlation follows a nested, block diagonal, exchangeable structure. Results are summarized over 1000 datasets with 50 independent clusters.

Design	**Method**	**50 Clusters**				
		$\beta_1 = 0.50$				
		Bias[a]	**SE**[b]	**MSE**[c]	**RE**[d]	**Coverage**[e]
Constant	GEE_i	0.019	0.257	0.066	0.659	0.951
Covariance	GEE_e	0.012	0.207	0.043	1.010	0.945
	GEE_n	0.012	0.187	0.035	1.236	0.946
Constant	$OCEE_{bw}$	0.014	0.211	0.045	0.978	0.921
Cluster Size	$OCEE_n$	0.004	0.208	0.043	1	0.882
Non-constant	GEE_i	0.013	0.182	0.033	0.277	0.950
Covariance	GEE_e	0.008	0.104	0.011	0.848	0.951
	GEE_n	0.009	0.109	0.012	0.777	0.950
Constant	$OCEE_{bw}$	0.006	0.099	0.010	0.932	0.933
Cluster Size	$OCEE_n$	0.007	0.096	0.0092	1	0.904
Constant	GEE_i	−0.003	0.288	0.083	0.627	0.932
Covariance	GEE_e	0.010	0.236	0.056	0.931	0.929
	GEE_n	0.004	0.219	0.048	1.082	0.928
Non-constant	$OCEE_{bw}$	0.010	0.238	0.057	0.915	0.918
Cluster Size	$OCEE_n$	0.003	0.228	0.052	1	0.898
Non-constant	GEE_i	−0.005	0.202	0.041	0.279	0.930
Covariance	GEE_e	0.002	0.127	0.016	0.709	0.930
	GEE_n	−0.001	0.131	0.017	0.664	0.932
Non-constant	$OCEE_{bw}$	0.006	0.113	0.013	0.896	0.932
Cluster Size	$OCEE_n$	0.002	0.107	0.011	1	0.906

NIH-PA Author Manuscript
NIH-PA Author Manuscript

[a] Bias $= \left(\overline{\widehat{\beta}} - \beta\right)$

[b] Calculated standard error (SE) of $\widehat{\beta}$

[c] Mean Square Error (MSE)

[d] Relative Efficiency: Ratio of calculated variance of $\widehat{\beta}$ under $OCEE_n$ relative to each method

[e] Coverage probability of 95% confidence intervals

Table II

Performance of Generalized Estimating Equations (GEE) and Optimal Combination of Estimating Equations (OCEE) approaches in a dental simulation setting with moderate correlation and cluster sizes. GEE methods, using independence(i), exchangeable(e) and nested(n) working structures, compared to OCEE, combining between- and within-cluster(bw) in addition to nested subcluster(n) data, for a linear model with a dichotomous covariate that varies between and within clusters and sub-clusters ($\mathbf{x_1}$), a continuous covariate that varies only between clusters ($\mathbf{x_2}$) (results not shown), and a dichotomous covariate that varies only within clusters ($\mathbf{x_3}$) (results not shown). The correlation follows a nested, block diagonal, exchangeable structure. Results are summarized over 1000 datasets with 200 independent clusters.

Design	**Method**	**200 Clusters**				
		$\beta_1 = 0.50$				
		Bias[a]	**SE**[b]	**MSE**[c]	**RE**[d]	**Coverage**[e]
Constant	GEE_i	−0.006	0.129	0.017	0.515	0.957
Covariance	GEE_e	0.000	0.104	0.011	0.791	0.947
	GEE_n	−0.001	0.092	0.009	1.015	0.947
Constant	$OCEE_{bw}$	−0.001	0.103	0.011	0.818	0.945
Cluster Size	$OCEE_n$	−0.001	0.093	0.009	1	0.946
Non-constant	GEE_i	−0.005	0.091	0.008	0.228	0.957
Covariance	GEE_e	−0.001	0.052	0.003	0.708	0.944
	GEE_n	−0.002	0.053	0.003	0.677	0.957
Constant	$OCEE_{bw}$	0.000	0.049	0.002	0.797	0.946
Cluster Size	$OCEE_n$	0.000	0.044	0.002	1	0.943
Constant	GEE_i	−0.002	0.141	0.020	0.542	0.945
Covariance	GEE_e	−0.006	0.110	0.012	0.885	0.955
	GEE_n	−0.006	0.102	0.011	1.021	0.951
Non-constant	$OCEE_{bw}$	−0.005	0.109	0.012	0.898	0.952
Cluster Size	$OCEE_n$	−0.006	0.104	0.011	1	0.950
Non-constant	GEE_i	−0.002	0.101	0.010	0.253	0.943
Covariance	GEE_e	−0.003	0.059	0.004	0.731	0.949
	GEE_n	−0.003	0.062	0.004	0.672	0.947
Non-constant	$OCEE_{bw}$	−0.004	0.054	0.003	0.875	0.944
Cluster Size	$OCEE_n$	−0.004	0.051	0.003	1	0.934

NIH-PA Author Manuscript

NIH-PA Author Manuscript

[a] Bias= $\left(\overline{\widehat{\beta}}-\beta\right)$

[b] Calculated standard error (SE) of $\widehat{\beta}$

[c] Mean Square Error (MSE)

[d] Relative Efficiency: Ratio of calculated variance of $\widehat{\beta}$ under $OCEE_n$ relative to each method

[e] Coverage probability of 95% confidence intervals

Table III

Analysis of Periodontal Data: Comparison of Generalized estimating equation (GEE) methods, using independence(i), exchangeable(e), and nested(n) working structures, and optimal combination of estimating equations (OCEE), combining between- and within-cluster(bw) in addition to nested subcluster(n) data. The regression model estimates the association between the 12-month change in alveolar bone density and the 24-month change in alveolar bone height (mm). Changes in alveolar bone density relative to baseline were categorized as an increase (reference group), no change, or a decrease in density. Summarized estimates include the regression coefficient (COEF), the estimated standard error (SE), and corresponding p-value. The regression coefficient provides an estimate of the difference in mean alveolar bone height change at 24-months relative to baseline (mm) for sites with no change (5^{th} column) or a decrease (8^{th} column) in alveolar bone density at 12 months versus sites with an increase in density.

	Model Term					
	Intercept		No Change in Density		Decrease in Density	
Method	**COEF (SE)**	**P–value**	**COEF (SE)**	**P–value**	**COEF (SE)**	**P–value**
GEE_i	0.038 (0.030)	0.2	0.049 (0.031)	0.1	0.078 (0.040)	0.05
GEE_e	0.053 (0.031)	0.09	0.038 (0.032)	0.2	0.068 (0.039)	0.08
GEE_n	0.053 (0.031)	0.09	0.036 (0.031)	0.3	0.070 (0.041)	0.09
$OCEE_{bw}$	0.037 (0.030)	0.2	0.048 (0.031)	0.1	0.079 (0.038)	0.04
$OCEE_n$	0.036 (0.030)	0.2	0.049 (0.031)	0.1	0.081 (0.038)	0.03