Using second-order generalized estimating equations to model heterogeneous intraclass correlation in cluster randomized trials

Catherine M. Crespi[1,*], **Weng Kee Wong**[1], and **Shiraz I. Mishra**[2]

[1]Department of Biostatistics, University of California Los Angeles, California [2]University of Maryland School of Medicine, Baltimore, Maryland

SUMMARY

In cluster randomized trials, it is commonly assumed that the magnitude of the correlation among subjects within a cluster is constant across clusters. However, the correlation may in fact be heterogeneous and depend on cluster characteristics. Accurate modeling of the correlation has the potential to improve inference. We use second-order generalized estimating equations to model heterogeneous correlation in cluster randomized trials. Using simulation studies we show that accurate modeling of heterogeneous correlation can improve inference when the correlation is high or varies by cluster size. We apply the methods to a cluster randomized trial of an intervention to promote breast cancer screening.

1. INTRODUCTION

In cluster randomized trials, social units or "clusters" of individuals are randomized to intervention and control conditions [1,2]. Responses of individuals within the same cluster are correlated, and analytical methods for cluster randomized trials typically assume an exchangeable, or compound symmetric, correlation structure within cluster. The usual measure of correlation is the Pearson correlation coefficient between pairs of responses in the same cluster, referred to as the intraclass correlation coefficient or ICC, and frequently denoted ρ.

When analyzing data from cluster randomized trials it is typically further assumed that the ICC is constant across all clusters in the study. However, in many studies there is diversity in cluster characteristics that could lead to differing levels of correlation in different clusters. In some cases, heterogeneity may occur by design, such as when one arm involves independent subjects and the other consists of clustered subjects [3]. In other cases, there may be variability in the effect of the intervention from cluster to cluster, leading to a higher variance and thus a higher ICC in the intervention group [4,5]. Other characteristics, such as cluster size, type or location, or presence of familial relationships, may also affect the degree of correlation of the responses.

Heterogeneity in the ICC is especially a concern for group-based interventions, in which subjects participate in educational, psychosocial or behavioral programs delivered in group settings. Such group-based interventions are common in health and medical research; examples include exercise interventions in the elderly [6,7], kidney disease management [8], smoking cessation programs [9], and HIV risk reduction programs [10]. The groups in these trials are likely to have differing degrees of cohesion *a priori* due to social connections or lack thereof,

*Correspondence to: Catherine M. Crespi, Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA 90095-1772. Email: ccrespi@ucla.edu.

and differing degrees of interaction among participants, either by design or due to differences in delivery, may further increase heterogeneity in the correlation of responses.

Our research was motivated by a group-based intervention study, the Breast Cancer Education Program for Samoan Women [11], a cluster randomized trial designed to increase rates of mammogram usage in women of Samoan ancestry funded by the California Breast Cancer Research Program. In the trial, Samoan churches were randomized to intervention and control conditions. At churches allocated to the intervention arm, women participated in group educational sessions intended to create a group dynamic leading to consensus about mammography usage; the control condition was usual care, with no group activities. The outcome was self-reported receipt of a mammogram at follow-up. Fitting the conventional exchangeable correlation model with constant ρ to the data set using generalized estimating equations (GEE) gave an estimated ICC of 0.19. However, separate analyses of the two intervention arms yielded estimated ICCs of 0.06 for the control arm and 0.34 for the intervention arm, indicating that the ICC varied with treatment assignment.

Thus there is reason to believe that in many cluster randomized trials, the ICC may vary among clusters, contrary to the common assumption of a single ρ common to all clusters. The existence of heterogeneity in the ICC raises the questions of whether and how the correlation should be modeled. A number of advantages may accrue from modeling heterogeneous correlation. First, accurate modeling of correlation structure can improve statistical inference on mean parameters through, most notably, gains in efficiency [12,13,14]. Increased efficiency is especially important for cluster randomized trials, which can be expensive to conduct and are less efficient than individually-randomized trials due to the design effect [15]. In many studies, the correlation structure itself may have scientific importance; for example, the degree of correlation among group-therapy participants may be an outcome of interest. A further advantage is that more nuanced estimates of ICC obtained from correlation modeling can be used to improve the design of future studies.

The greatest need for research on modeling correlated data is in the area of generalized linear models for nonnormal response data, where methods are less well developed than in the normal linear model case. A popular method of analyzing correlated response data in the generalized linear model context is GEE [16]. First-order GEE or GEE1 models parametrize the marginal mean and account for the dependence among units in a cluster by specifying a working correlation structure for the observed responses. GEE1 is designed to focus on the marginal mean, treating the correlation structure as a nuisance. Correspondingly, implementations of GEE1 typically allow a limited number of prespecified working correlation structures (e.g., autoregressive(1), exchangeable, independent, m-dependent, unstructured and user-specified fixed in SAS Version 9.1), and these structures are constant across all clusters. An advantage of GEE1 is that the mean parameters are consistently estimated regardless of whether the correlation structure is correctly specified, so long as the mean is correctly specified [16]. However, efficiency can be lost when the working correlation structure is not correct [16,17, 18,19,20,21].

A number of extensions to incorporate covariates into covariance structures, termed second-order GEE or GEE2, have been proposed. These include a joint parameter estimation approach [17], a quadratic exponential model approach [22], scale parameter modeling [19], and mean, scale and correlation modeling with separate equations [23]. GEE2 offers the opportunity to specify and estimate more accurate models of correlation structure and thus its application may yield efficiency gains and other benefits for cluster randomized intervention trials. However, cluster randomized intervention trial data is a novel application for GEE2, and work is needed in several areas. First, methods of adapting GEE2 to such applications, including correlation modeling strategies, are needed. Also lacking are studies of the performance of GEE2 in this

context; there is a need to determine whether hypothesized gains in efficiency of estimation of mean parameters do in fact occur, and whether statistical inference is improved. Empirical investigations through real data analysis examples are also needed.

In this paper, we present a method of modeling heterogeneous correlation structure in cluster randomized trials using GEE2. We present simulation studies comparing heterogeneous ICC modeling to the conventional homogeneous ICC model in terms of bias, variance, mean squared error (MSE) and empirical power. We apply the methods to data from our motivating example, the Breast Cancer Education Program for Samoan Women, and conclude with a discussion of the implications of heterogeneous correlation in cluster randomized intervention trials.

Throughout, we focus on binary response data. We use the three-estimating-equation (3EE) GEE2 method of Yan and Fine [23], which has separate estimating equations with separate link functions and linear predictors for the mean, scale and correlation. All models were fit using the R package geepack [24] (downloaded from http://cran.r-project.org/ in February 2007). R code for applying the methods is available from the first author.

2. MODEL

In this section, we present a brief description of second-order generalized estimating equations for generalized linear models and present a method of heterogeneous correlation modeling for cluster randomized trial data. For further details on GEE, the reader is referred to Hardin and Hilbe [25].

Suppose we have K clusters indexed by i, with clusters assumed to be independent. Each cluster has n_i subjects, whose outcomes are collected in the vector $\mathbf{Y}_i^T = (Y_{i1}, \ldots, Y_{in_i})$. Each observation Y_{ij} is associated with a p-dimensional covariate vector $\mathbf{x}_{ij}^T = (x_{ij1}, \ldots, x_{ijp})$. Using a generalized linear model framework [26], the marginal regression model for the mean, $E(Y_{ij}) = \mu_{ij}$, is $g(\mu_{ij}) = \mathbf{x}_{ij}^T \boldsymbol{\beta}$, where g is a known link function and $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients to be estimated. The marginal variance is $\mathrm{Cov}(Y_{ij}) = \phi \upsilon(\mu_{ij})$ where υ is a known function and ϕ is a scale parameter which may need to be estimated or may be specified as fixed.

If observations within clusters were independent, the covariance of responses within the ith cluster, $\mathrm{Cov}(\mathbf{Y}_i)$, could be expressed as

$$\mathrm{Cov}(Y_i) = \mathbf{V}_i = \phi \mathbf{A}_i = \phi \mathbf{A}_i^{1/2} \mathbf{I} \mathbf{A}_i^{1/2}, \tag{1}$$

where the identity matrix $\mathbf{I}$ has dimension $n_i \times n_i$ and $\mathbf{A}_i = \mathrm{diag}(\upsilon(\mu_{i1}), \ldots, \upsilon(\mu_{in_i}))$ is a diagonal matrix of variances of the elements of $\mathbf{Y}_i$. To handle correlated responses, GEE1 replaces the identity matrix with a more general correlation matrix,

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}, \tag{2}$$

where $\mathbf{R}(\alpha)$, of dimension $n_i \times n_i$, is referred to as the working correlation matrix of $\mathbf{Y}_i$. The correlation matrix is estimated through the parameter vector $\boldsymbol{\alpha}$. When the observations within a cluster are assumed to be equally correlated, as is the case for most cluster randomized trial data, $\boldsymbol{\alpha}$ is a scalar and the working correlation matrix is taken as

$$R(\alpha)=\begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix},$$

(3)

an exchangeable or compound symmetric structure.

Expressions (2) and (3) together specify the conventional assumption of an exchangeable correlation structure of constant magnitude across clusters, and such models can be fit using GEE1. GEE2 extends GEE1 to allow covariates in the covariance model, thus allowing correlation structure and/or magnitude to depend on cluster or subject characteristics. When the correlation matrix varies by cluster, the covariance model may be expressed as

$$\mathbf{V}_i=\phi\mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2},$$

(4)

where $\mathbf{R}_i(\boldsymbol{\alpha})$ depends on covariates. A general approach for linking the correlation parameters to a linear predictor has been suggested by Yan and Fine [23]. Take the upper diagonal elements of $\mathbf{R}_i$ and arrange them into a $n_i(n_i-1)/2 \times 1$ vector $\boldsymbol{\rho}_i$ of pairwise correlations between the elements of $\mathbf{Y}_i$, with $\boldsymbol{\rho}_i^T = (\rho_{i,12}, \rho_{i,13}, \ldots, \rho_{i,1n_i}, \rho_{i,23}, \ldots, \rho_{i,n_i-1n_i})$. A model for $\boldsymbol{\rho}_i$ is

$$h(\boldsymbol{\rho}_i)=\mathbf{W}_i\boldsymbol{\alpha}$$

(5)

where h is a known link function, $\mathbf{W}_i$ is a covariate matrix with dimension $n_i(n_i-1)/2 \times q$ and $\boldsymbol{\alpha}$ is a vector of correlation parameters to be estimated, of length q. When the correlation structure is exchangeable, then all elements of $\boldsymbol{\rho}_i$ are equal and we have

$$\begin{pmatrix} h(\rho_{i1}) \\ h(\rho_{i2}) \\ \vdots \\ h(\rho_{in_i}) \end{pmatrix} = \begin{pmatrix} h(\rho_i) \\ h(\rho_i) \\ \vdots \\ h(\rho_i) \end{pmatrix} = \begin{pmatrix} \mathbf{w}_i'\boldsymbol{\alpha} \\ \mathbf{w}_i'\boldsymbol{\alpha} \\ \vdots \\ \mathbf{w}_i'\boldsymbol{\alpha} \end{pmatrix},$$

(6)

where $\mathbf{w}_i$ is a vector of length q.

Taking h as the identity function would allow ρ_i to range from $-\infty$ to $+\infty$. A mathematically tractable transformation that restricts correlation coefficients to the interval $(-1, 1)$ is the Fisher transformation, $h(\rho)=\log(\frac{1+\rho}{1-\rho})$, which makes a convenient link function. Using the inverse link function, we obtain

$$\rho_i=\frac{\exp(\mathbf{w}_i'\boldsymbol{\alpha})-1}{\exp(\mathbf{w}_i'\boldsymbol{\alpha})+1}.$$

(7)

This approach can be viewed as a generalization of the homogeneous ICC and independence models. Homogeneous ICC is a special case in which α is a scalar and $\mathbf{w}_i = 1$ for all i. The ICC would then be given by the scalar quantity $\rho=\frac{e^\alpha-1}{e^\alpha+1}$, constant across clusters. Independence of responses within a cluster can be viewed as a special case with $\rho_i = 0$, and can be specified

by coding $\mathbf{w}_i$ as a zero vector. With appropriate coding of $\mathbf{w}_i$, it is possible to specify mixed correlation structures with, for example, independence in some clusters and covariate-dependent compound symmetry in others.

The scale parameters $\boldsymbol{\phi}$ can also be modeled as a function of covariates with a separate link using 3EE GEE2. In our application, we model binary data and set the scale parameter to unity.

Yan and Fine [23] present estimating equations and an alternating Fisher scoring method for the 3EE GEE2 approach. They show that $\{K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T, K^{1/2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})^T, K^{1/2}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi})^T\}$ is asymptotically normal under regularity conditions with mean zero and covariance matrix consistently estimated. Thus this approach provides standard errors for the mean, correlation and scale parameters. Options for the variance estimator include the robust sandwich estimator, analogous to the sandwich estimator used for GEE1, and jackknife estimators that have better performance for small samples ($K \leq 30$).

3. SIMULATION STUDIES

We investigated the performance of heterogeneous ICC modeling for cluster randomized trial data using simulation experiments, in which we generated data sets with heterogeneous correlation structure and compared the performance of the conventional homogeneous ICC model to the performance of heterogeneous ICC modeling in terms of statistical inference. All simulation scenarios were designed as two-arm cluster randomized trials with a binary outcome variable. We presumed that correlation modeling is more likely to be applied when correlation levels are high and there are relatively large numbers of clusters, and selected simulation settings accordingly. The method of Emrich and Piedmonte [27] was used to generate the high-dimensional multivariate binary variables.

We conducted two sets of simulation scenarios. Set I simulated clusters of equal size, with 30 subjects per cluster and 30 clusters per arm. For mean and correlation parameter specification, we used a 4×3 factorial design, with four levels of correlation, all heterogeneous: very high (each arm has 10 clusters with $\rho = 0.20$, 10 with $\rho = 0.40$ and 10 with $\rho = 0.80$), high (each arm has 10 clusters with $\rho = 0.10$, 10 with $\rho = 0.20$ and 10 with $\rho = 0.40$), moderate (each arm has 10 clusters with $\rho = 0.05$, 10 with $\rho = 0.10$ and 10 with $\rho = 0.20$), and low (each arm has 10 clusters with $\rho = 0.02$, 10 with $\rho = 0.04$ and 10 with $\rho = 0.08$), crossed with three levels of (π_c, π_t), the true proportions of successes in the control and treatment arms, specified as (0.40, 0.60), (0.25, 0.45) and (0.10, 0.30).

The second set of simulation scenarios was motivated by our data application and other similar trials, which have clusters of varying sizes. Set II simulated clusters of unequal size, with each arm of the trial having 1, 2, 3, 4, 5 and 10 clusters of sizes 50, 30, 20, 15, 10 and 5, respectively. The true proportions of successes in the control and treatment arms were 0.25 and 0.45, respectively, in all scenarios. The ICC was varied as a function of treatment arm (treatment vs. control) or cluster size (small, defined as 5–15 subjects per cluster, vs. large, defined as 20–50 subjects per cluster). In particular, we specified four scenarios, with

$$\begin{pmatrix} \rho_{c,sm} & \rho_{t,sm} \\ \rho_{c,lg} & \rho_{t,lg} \end{pmatrix} \tag{8}$$

equal to

$$\begin{pmatrix} 0.05 & 0.35 \\ 0.05 & 0.35 \end{pmatrix}, \begin{pmatrix} 0.35 & 0.05 \\ 0.35 & 0.05 \end{pmatrix}, \begin{pmatrix} 0.05 & 0.05 \\ 0.35 & 0.35 \end{pmatrix} \text{ and } \begin{pmatrix} 0.35 & 0.35 \\ 0.05 & 0.05 \end{pmatrix}. \tag{9}$$

Pilot studies showed that the standard deviations in the simulations were in the range of 0.15 to 0.30 for the various scenarios. On this basis, we chose a Monte Carlo sample size of 500, which would lead to standard errors of means on the order of 0.007 to 0.013. Thus we generated 500 simulated data sets for each scenario.

We fit two correlation models to each data set using 3EE GEE2, the conventional constant ρ model, designated C, which entails the correlation model $h(\rho_i) = \alpha$, and a model with properly specified covariates in the correlation model, designated M. For example, for ICC varying by treatment arm in Set II, Model M specified $h(\rho_i) = \alpha_0 + \alpha_1 T_i$, where T_i is the indicator function having value 1 if the subject is assigned to the intervention group and 0 otherwise. For all models, we used a logit link for the mean, with $\log[p_i/(1 - p_i)] = \beta_0 + \beta_1 T_i$. The proportions (0.40, 0.60), (0.25, 0.45) and (0.10, 0.30) correspond to β_1 of 0.81, 0.90 and 1.35, respectively. The variance function was specified as binomial, and the scale was fixed at the value 1.

We estimated bias as the mean of $\hat{\beta}_1 - \beta_1$ over the 500 replications and relative bias as the mean of $(\hat{\beta}_1 - \beta_1)/\beta_1$. The dispersion of bias was characterized by the 2.5th and 97.5th percentiles. The empirical variance of $\hat{\beta}_1$, $\hat{V}ar(\hat{\beta}_1)$, was used to estimate efficiency and MSE. Relative efficiency was estimated as the mean of $[\hat{V}ar(\hat{\beta}_1)$ for M$]/[\hat{V}ar(\hat{\beta}_1)$ for C$]$. Relative MSE was estimated as the mean of $[\hat{V}ar(\hat{\beta}_1) + (\hat{\beta}_1 - \beta_1)^2$ for M$]/[\hat{V}ar(\hat{\beta}_1) + (\hat{\beta}_1 - \beta_1)^2$ for C$]$. The empirical coverage probability (ECP) for 95% confidence intervals for β_1 was also calculated using the empirical variance. The power to reject the null hypothesis that β_1 equals 0 was estimated as $\frac{1}{500}\sum I(\widehat{\beta}_1/\widehat{SE} > 1.96)$ where I is the indicator function having value 1 if $\widehat{\beta}_1/\widehat{SE} > 1.96$ and 0 otherwise, and $\widehat{SE}$ is the robust sandwich estimator.

Table I presents the results for the first set of simulation scenarios. Bias tended to be higher under correlation modeling (M) compared to the conventional model (C), and was highest for scenarios with $\beta_1 = 1.35$. Increases in efficiency under correlation modeling were evident for all very high and high correlation scenarios, as well as the moderate correlation scenario with $\beta_1 = 0.81$. The greatest efficiency gain was achieved under very high correlation with $\beta_1 = 0.81$, which had a decrease in empirical variance of 22%. Most of these scenarios also showed a reduction in MSE; however, these gains were more modest due to the increase in bias.

For the very high and high correlation scenarios, the null hypothesis was correctly rejected at slightly higher rates when the correct correlation model was used. The very high correlation scenario with $\beta_1 = 0.81$ had the largest increase in power, with an 11% difference.

For the moderate and low correlation scenarios, there appeared to be little advantage to using the correct correlation model compared to the conventional model in terms of statistical inference; in fact, for the low correlation scenarios, MSE was 5–10% higher when using the correct correlation model.

Empirical coverage probabilities were close to the nominal level for all scenarios. For all scenarios, the 2.5th and 97.5th percentiles of bias were similar for the correct and the conventional models. The very high correlation scenarios had the most dispersed distributions of bias (2.5th and 97.5th percentiles on the order of −0.8 and 0.8, respectively), while the low correlation scenarios had the least dispersed (2.5th and 97.5th percentiles on the order of −0.3 and 0.3, respectively).

Table II presents the results for the second set of simulation scenarios. In the scenarios in which the ICC varied by treatment arm, there were no differences in performance of the two modeling strategies. However, in the scenarios in which the ICC varied by cluster size, the correlation modeling strategy was superior in terms of efficiency, MSE and power. Relative efficiencies

were 0.81 and 0.83, and the correlation modeling approach correctly rejected the null hypothesis at rates that were 7–11% higher compared to the conventional modeling approach. Empirical coverage probabilities were close to the nominal level for all scenarios. The 2.5th and 97.5th percentiles of bias were similar for M and C for all scenarios, and were on the order of −0.6 and 0.7, respectively.

These simulation studies are necessarily limited since only a finite number of scenarios can be reasonably explored. However, the results suggest that modeling heterogeneous ICC can have an impact on statistical inference for cluster randomized trials under some circumstances. In particular, benefits may accrue when levels of correlation are high and/or correlation varies by cluster size. On the other hand, there are circumstances in which improvements in inference do not occur. We did not see improvements in statistical inference when levels of correlation were low or correlation varied only with treatment assignment, in the context of the particular parameter specifications in our simulations. In fact, MSE was higher in low correlation scenarios when the correct correlation model was used.

4. APPLICATION

To illustrate the methods in a real data analysis context, we apply heterogeneous ICC modeling to our motivating application, the Breast Cancer Education Program for Samoan Women. This study, conducted between July 1998 and June 2001 by the National Office of Samoan Affairs, the University of California Irvine and the University of California Los Angeles, was designed to test the effectiveness of a culturally appropriate breast cancer education program tailored to women with Samoan ancestry, who were found in previous studies to have low rates of mammogram use [28]. In the trial, 61 Samoan churches in southern California were randomized to intervention or control conditions. Subjects from churches in the intervention arm participated in a series of culturally-tailored interactive group discussion sessions with a health educator; the control condition was usual care, with subjects receiving educational materials after the follow-up survey. The primary outcome was self-reported receipt of a mammogram between the baseline and follow-up surveys, which were eight months apart.

The data set consisted of 776 subjects, with the number of subjects per church ranging from 1 to 42. The median cluster size was 13, with comparable distributions of cluster sizes in each arm. Rates of self-reported receipt of mammography were 38.7% in the control arm and 47.3% in the intervention arm. As previously discussed, fitting a GEE1 model with an exchangeable working correlation structure to the entire data set gave an estimated ICC of 0.19, whereas fitting models to the control and intervention arms separately gave estimated ICCs of 0.06 and 0.34, respectively, motivating the application of correlation modeling.

We fit correlation models to the data using 3EE GEE2, assuming compound symmetry within each cluster but allowing the magnitude of the correlation to depend on cluster-level covariates. To identify covariates potentially affecting correlation magnitude, we considered the nature of the intervention and the implications of variation in cluster size. Based on our arm-specific analyses and the fact that the intervention entailed extensive interactions among the participants which were lacking in the control arm, we considered treatment arm, coded as T_i equal to 0 for control and 1 for intervention, as a covariate. We also considered cluster size as a covariate, since in this church-based trial, congregations of different size had differing levels of resources and leadership involvement, which may have affected within-group cohesion. In addition, the study investigators had observed based on past experience that different group sizes produce different group dynamics. We used the dichotomous variable S_i to identify clusters as large (defined as over 15 subjects from the church electing to participate, coded as $S_i = 1$) or small (15 or fewer subjects, $S_i = 0$). We also considered an interaction between treatment and cluster

size, based on the concern that the degree of consensus engendered by the intervention might vary with cluster size. In sum, we evaluated the following correlation models:

Model 1. Compound symmetry, constant across clusters:

$$h(\rho_i) = \alpha_0 \tag{10}$$

Model 2. Compound symmetry with magnitude depending on treatment arm:

$$h(\rho_i) = \alpha_0 + \alpha_1 T_i \tag{11}$$

Model 3. Compound symmetry with magnitude depending on cluster size category:

$$h(\rho_i) = \alpha_0 + \alpha_2 S_i \tag{12}$$

Model 4. Compound symmetry with magnitude depending on treatment arm and cluster size category:

$$h(\rho_i) = \alpha_0 + \alpha_1 T_i + \alpha_2 S_i \tag{13}$$

Model 5. Model 4 with a treatment $\times$ cluster size interaction:

$$h(\rho_i) = \alpha_0 + \alpha_1 T_i + \alpha_2 S_i + \alpha_3 T_i S_i \tag{14}$$

The marginal mean was modeled using the logit link, $\ln[p_i/(1 - p_i)] = \beta_0 + \beta_1 T_i$. Thus $\hat{\beta}_1$ gives the estimated intervention effect on the log odds scale. We found that estimates of ρ_i obtained using the linear link $h(\rho_i) = \rho_i$ and using the inverse Fisher transformation (7) were essentially identical and the choice of link function did not affect the model fitting; thus we used the linear link for the correlation model. The scale parameter was fixed at 1. Standard errors were obtained using robust sandwich variance estimators, which are provided for the mean, correlation and (when allowed to vary) scale parameter estimates when using 3EE GEE2.

The results are presented in Table III. The difference in overall unadjusted response rates in the study was a modest 8.6%; hence it is not surprising that the intervention effect was not significantly different from zero under any model. However, there is a suggestion of a trend toward a slightly stronger intervention effect estimate (increasing $\hat{\beta}_1$ and $\widehat{OR}$), slightly smaller standard error, and smaller *p*-value with greater allowance for heterogeneity in the correlation model.

The intraclass correlation coefficient estimates are informative about patterns of correlation in the data. As expected, there is evidence of heterogeneity of the ICC by treatment arm. Model 2 yielded estimates of 0.06 and 0.34 for the control and intervention arms, respectively; these are the same as the estimates obtained using separate analyses for each arm. Model 1, which is the conventional model assuming a ρ of constant magnitude across all clusters, yielded an estimated ICC of 0.19, which can be viewed as an average across the treatment arms which masks the difference between the arms.

Coefficient estimates for the cluster size category indicator S_i in Model 3 through 5 suggest correlation may have been higher in larger clusters. The difference in correlation between small and large clusters was not statistically significant at the 0.05 level in any model, but unless the correlation is quite high or the sample size is large, there is clearly a strong risk of Type II error in using Wald statistics for inference about correlation parameters.

There was no evidence of a significant interaction between treatment and cluster size (Model 5), and the estimates from Models 4 and 5 are very close, suggesting that the interaction term can be safely dropped.

In Models 2, 4 and 5, Wald tests would fail to reject the null hypothesis that the ICC is zero in the clusters in the control arm. Removing statistically insignificant terms would imply zero correlation, which is generally an untenable assumption for cluster randomized trials. Thus models that assume correlation in all clusters will generally be preferred. For this data analysis, considering the correlation parameter estimates in the context of theoretical expectations, we would consider Models 2 or 4 to be suitable models for the correlation structure of the data.

In this real data application, the ICC clearly varies by treatment arm, and parameter estimates are slightly different when correlation modeling is used. In Set II of the simulation studies, we did not find a difference in inference when correlation modeling was applied to simulated data with ICC varying by treatment arm, for the particular settings used in the simulation. Since the true parameter values in the real data application are unknown and surely differ from the settings used in the simulation, it is difficult to ascertain whether or not the differences seen with correlation modeling reflect true improvements in inference.

Overall, the exercise of fitting and comparing candidate correlation models for this data suggests that correlation modeling in the context of cluster randomized trials may be beneficial in several respects. Inference on the mean parameters, in particular the intervention effect, did not appear to be negatively impacted and may have been slightly sharpened when the correlation structure was more accurately modeled. In addition, correlation modeling was helpful in identifying predictors of correlation. These predictors are useful in interpreting the impact of the intervention and for designing future studies. For example, if a future study were to be restricted to large clusters only, the sample size would need to be increased to achieve adequate power.

5. DISCUSSION

Current approaches to analyzing data from cluster randomized trials typically assume that the clusters are homogeneous with respect to the intraclass correlation coefficient. Here we present a different view, that the sample may contain subsets of clusters with correlations of different magnitude. Modeling and estimating the correlation structure can yield several benefits, including increased efficiency and power in some cases, and better understanding of the determinants of intraclass correlation.

Our findings suggest that any increases in efficiency and power are likely to be modest, and are more likely to occur when correlation is high or varies by cluster size. Improvements in statistical inference will not accrue uniformly across all scenarios. In our low correlation scenarios, MSE was slightly increased when the correct correlation model was fit. However, there may be other benefits of using a more accurate correlation model. Modeling the correlation can provide insights into the intervention mechanism. In our data application, correlation modeling revealed that treatment assignment was a strong predictor of higher correlation magnitude. This finding is consistent with the behavior-theoretic framework of the intervention, which was designed to create a group dynamic which was not fostered in the control condition. From another point of view, this may be interpreted as variability of the intervention effect from cluster to cluster. In contrast, fitting the conventional constant ρ model yielded an estimate of the ICC that was a compromise between the control and intervention groups. This estimate did not characterize either group nor provide insights about the intervention.

Another application of correlation modeling for cluster randomized trials, suggested by a reviewer, is as part of a sensitivity analysis. In particular, one could use these methods to assess whether different assumptions about the correlation affect inference on the mean parameters. Our analyses in Section 4 can be viewed as an illustration of this type of application.

The estimates and predictors of ICC derived from correlation modeling are also useful for the design of future studies. A persistent difficulty in designing cluster randomized trials has been the scarcity of reliable estimates of ρ, which are necessary in order to determine the variance inflation factor measuring the amount by which one should increase a variance estimate to allow for the clustering effect [2]. Few studies report ρ, perhaps in large part because it is regarded as a nuisance. Expressly modeling the ICC gives this important parameter more prominence. The confidence intervals for the ICC such as are derived from 3EE GEE2 may also improve study design; Turner et al. [29] have shown that averaging power across uncertainty in the ICC may be superior to using a point estimate.

If the ICC is expected to vary by treatment arm or other covariates, and will be modeled as such in the final analysis, then prudence dictates that this variation in ICC should be accounted for at the design stage of the study when the sample size requirements are estimated. Sample size formulae for the case in which ρ varies by group, developed in the context of case-control/family sampling designs, are presented by [30] and may be useful for this purpose.

Selection of correlation models in GEE2, while not a primary focus of our research, is an important topic and an active area of research. 3EE GEE2 provides Wald statistics which may be used for variable selection for the linear predictor of the correlation. However, unless the number of clusters is quite large, there is a danger that important predictors will be discarded as insignificant. In addition, Wald test statistics have been shown to behave in an aberrant manner under some circumstances [31]. Another alternative for correlation model selection is the quasilikelihood information criterion, proposed by Pan [32]. Other goodness of fit criteria which could be explored further are discussed by Zheng [33]. The choice of correlation model should also be guided by clinical or theoretical expectations.

Our work has focused on group-based intervention trials, which typically involve large numbers of small clusters. Other types of cluster randomized trials include community intervention trials, which typically involve a relatively small number of communities each enrolling a large number of subjects. These trials tend to have much lower ICCs, often on the order of 0.01 to 0.001 [2]. These trials may also be prone to heterogeneity in the ICC given the inherent diversity among communities in characteristics which may be associated with correlation. Furthermore, when the number of randomized units is small, randomization may fail to achieve balance across treatment arms. For these reasons, correlation modeling may be considered for community-randomized studies. Such studies may call for somewhat different modeling strategies and will often require different variance estimation approaches (e.g., jackknife variance estimation when the number of clusters is less than or equal to 30 [23]), and present an area for future research.

We conclude by noting that 3EE GEE2 is not the only available method for modeling heterogeneous intraclass correlation. Alternating logistic regression is another approach to covariate-dependent correlation structure modeling which can be used in the context of generalized estimating equations [34]. Generalized linear mixed models (GLMMs) are an alternative to GEE for fitting correlated data with a nonlinear link function for the mean; an example of a multi-level logistic model for cluster randomized trial data, with two between-cluster variance components, is provided in [4]. In such models, interval estimates of variance parameters can be examined to investigate the extent of overlap, similar to comparison of confidence intervals for the ICCs in 3EE GEE2. Both GLMM and 3EE GEE2 can be

computationally intensive when fitting complex correlation models. Bayesian methods are another alternative; see, for example, [5], who fit a Bayesian version of the model used by [4] to the same data.

Acknowledgments

REFERENCES

1. Murray, DM. Design and Analysis of Group-Randomized Trials. Oxford University Press; 1998.

2. Donner, A.; Klar, N. Design and Analysis of Cluster Randomization Trials in Health Research. Oxford University Press; 2000.

3. Roberts C, Roberts SA. Design and analysis of clinical trials with clustering effects due to treatment. Clinical Trials 2005;2:152–162. [PubMed: 16279137]

4. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multilevel model. Statistics in Medicine 2000;19:2675–2688. [PubMed: 10986541]

5. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Statistics in Medicine 2001;20:453–472. [PubMed: 11180313]

6. Brismee JM, Paige RL, Chyu MC, Boatright JD, Hagar JM, McCaleb JA, Quintela MM, Feng D, Xu KT, Shen CL. Group and home-based tai chi in elderly subjects with knee osteoarthritis: a randomized controlled trial. Clinical Rehabilitation 2007;21:99–111. [PubMed: 17264104]

7. Timonen L, Rantanen T, Makinen E, Timonen TE, Tormakangas T, Sulkava R. Effects of a group-based exercise program on functional abilities in frail older women after hospital discharge. Aging Clinical and Experimental Medicine 2006;18:50–56.

8. Sharp J, Wild MR, Gumley AI, Deighan CJ. A cognitive behavioral group approach to enhance adherence to hemodialysis fluid restrictions: a randomized controlled trial. American Journal of Kidney Diseases 2005;45:1046–1057. [PubMed: 15957134]

9. Romand R, Gourgou S, Sancho-Garnier H. A randomized trial assessing the Five-Day Plan for smoking cessation. Addiction 2005;100:1546–1554. [PubMed: 16185216]

10. Morrison-Beedy D, Carey MP, Kowalski J, Tu X. Group-based HIV risk reduction intervention for adolescent girls: evidence of feasibility and efficacy. Research in Nursing and Health 2005;28:3–15. [PubMed: 15625713]

11. Mishra SI, Bastani R, Crespi CM, Chang LC, Luce PH, Baquet CR. Results of a randomized trial to increase mammogram usage among Samoan women. Cancer Epidemiology, Biomarkers and Prevention 2007;16:2594–2604.

12. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. Biometrics 1995;51:309–317. [PubMed: 7766784]

13. Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika 2003;90:29–41.

14. Ye H, Pan J. Modeling of covariance structure in generalised estimating equations for longitudinal data. Biometrika 2006;93:927–941.

15. Donner A. Some aspects of the design and analysis of cluster randomized trials. Applied Statistics 1998;47:95–113.

16. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika 1986;73:13–22.

17. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics 1988;44:1033–1048. [PubMed: 3233244]

18. Liang KY, Zeger SL, Qaqish B. Multivariate regression analysis for categorical data. Journal of the Royal Statistical Society B 1992;54:3–24.

19. Paik MC. Parametric variance function estimation for nonnormal repeated measurement data. Biometrics 1992;48:19–30. [PubMed: 1581484]

20. Albert PS, McShane LM. A generalized estimating equations approach for spatially correlated binary data: applications to the analysis of neuroimaging data. Biometrics 1995;textbf51:627–638. [PubMed: 7662850]

21. Chao EC. Structured correlation in models for clustered data. Statistics in Medicine 2006;25:2450–2468. [PubMed: 16220520]

22. Prentice RL, Zhao LP. Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. Biometrics 1991;47:825–839. [PubMed: 1742441]

23. Yan J, Fine J. Estimating equations for association structures. Statistics in Medicine 2004;23:859–874. [PubMed: 15027075]

24. Halekoh U, Hojsgaard S, Yan J. The R package geepack for generalized estimating equations. Journal of Statistical Software 2006;15:1–11.

25. Hardin, JW.; Hilbe, JM. Generalized Estimating Equations. Chapman & Hall/CRC; 2003.

26. McCulloch, CE.; Searle, SR. Generalized, Linear and Mixed Models. John Wiley and Sons; 2001.

27. Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. American Statistician 1991;45:302–304.

28. Mishra SI, Luce PH, Hubbell FA. Breast cancer screening among American Samoan women. Preventive Medicine 2001;33:9–17. [PubMed: 11482991]

29. Turner RM, Prevost AT, Thompson SG. Allowing for imprecision of the intraclass correlation coefficient in the design of cluster randomized trials. Statistics in Medicine 2004;23:1195–1214. [PubMed: 15083478]

30. Liang KY, Pulver AE. Analysis of case-control family sampling design. Genetic Epidemiology 1996;13:253–270. [PubMed: 8797008]

31. Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association 1977;72:851–853.

32. Pan W. Akaike's information criterion in generalized estimating equations. Biometrics 2001;57:120–125. [PubMed: 11252586]

33. Zheng B. Summarizing the goodness of fit of generalized linear models for longitudinal data. Statistics in Medicine 2000;19:1265–1275. [PubMed: 10814976]

34. Carey V, Zeger SL, Diggle P. Modelling multivariate binary data with alternating logistic regressions. Biometrika 1993;80:517–526.

Table I

Simulation results (Set I) comparing correlation modeling (M) with conventional assumption of exchangeable correlation of constant magnitude (C). ECP is empirical coverage probability for 95% confidence intervals for β_1. Details of the simulation settings are provided in the text.

Correlation level	Very high			High			Moderate			Low		
True β_1	0.81	0.90	1.35	0.81	0.90	1.35	0.81	0.90	1.35	0.81	0.90	1.35
Bias (M)	0.01	0.04	0.11	0.05	0.04	0.10	0.01	0.01	0.06	0.01	0.02	0.05
Bias (C)	0.00	0.03	0.10	0.03	0.03	0.03	0.00	0.00	0.02	0.01	0.01	0.02
Rel. Bias (M)	0.01	0.04	0.08	0.06	0.04	0.07	0.01	0.01	0.04	0.01	0.02	0.03
Rel. Bias (C)	0.00	0.03	0.07	0.03	0.04	0.02	0.00	0.00	0.00	0.01	0.01	0.02
Rel. Eff.	0.78	0.84	0.91	0.89	0.87	0.91	0.88	1.00	1.01	1.01	0.99	1.03
Rel. MSE	0.81	0.89	0.95	0.98	0.94	0.99	0.94	1.08	1.06	1.06	1.05	1.10
Emp. power (M)	0.71	0.78	0.93	0.93	0.95	1.00	0.99	1.00	1.00	1.00	1.00	1.00
Emp. power (C)	0.60	0.69	0.86	0.89	0.90	0.98	0.98	0.99	1.00	1.00	1.00	1.00
ECP (M)	0.93	0.95	0.95	0.95	0.95	0.96	0.96	0.96	0.94	0.95	0.94	0.94
ECP (C)	0.95	0.95	0.95	0.94	0.95	0.96	0.95	0.96	0.94	0.95	0.95	0.95

Table II

Simulation results (Set II) comparing correlation modeling (M) with conventional assumption of exchangeable correlation of constant magnitude (C). ECP is empirical coverage probability for 95% confidence intervals for β_1. Details of the simulation settings are provided in the text.

$\begin{pmatrix} \rho_{c,sm} & \rho_{t,sm} \\ \rho_{c,lg} & \rho_{t,lg} \end{pmatrix}$	ρ varying by treatment arm		ρ varying by cluster size	
	$\begin{pmatrix} 0.05 & 0.35 \\ 0.05 & 0.35 \end{pmatrix}$	$\begin{pmatrix} 0.35 & 0.05 \\ 0.35 & 0.05 \end{pmatrix}$	$\begin{pmatrix} 0.05 & 0.05 \\ 0.35 & 0.35 \end{pmatrix}$	$\begin{pmatrix} 0.35 & 0.35 \\ 0.05 & 0.05 \end{pmatrix}$
Bias (M)	0.02	0.07	0.01	0.03
Bias (C)	0.01	0.07	0.02	0.04
Rel. Bias (M)	0.02	0.07	0.01	0.04
Rel. Bias (C)	0.02	0.08	0.02	0.04
Rel. Eff.	0.99	0.98	0.81	0.83
Rel. MSE	1.00	0.98	0.87	0.90
Emp. power (M)	0.82	0.81	0.92	0.88
Emp. power (C)	0.81	0.81	0.85	0.77
ECP (M)	0.96	0.95	0.95	0.95
ECP (C)	0.95	0.94	0.95	0.94

Table III

Results of fitting candidate correlation models to data from the Breast Cancer Education Program for Samoan Women.

$h(\rho)$	Model 1 α_0	Model 2 $\alpha_0 + \alpha_1 T$	Model 3 $\alpha_0 + \alpha_2 S$	Model 4 $\alpha_0 + \alpha_1 T + \alpha_2 S$	Model 5 $\alpha_0 + \alpha_1 T + \alpha_2 S + \alpha_3 TS$
Intervention effect estimates					
β_1	0.20	0.25	0.27	0.30	0.30
$\hat{SE}$	0.28	0.27	0.27	0.26	0.26
$\hat{\delta}_R$	1.22	1.28	1.31	1.35	1.35
95% CI	(0.71, 2.10)	(0.75, 2.17)	(0.77, 2.22)	(0.81, 2.23)	(0.81, 2.25)
p-value	0.47	0.36	0.32	0.25	0.25
Intraclass correlation coefficient estimates (with 95% CIs)					
$\hat{\rho}$	0.19 (0.09,0.29)				
$\hat{\rho}_C$		0.06 (−0.02,0.14)			
$\hat{\rho}_T$		0.34 (0.23,0.45)			
$\hat{\rho}_{sm}$			0.12 (0.02,0.22)		
$\hat{\rho}_{lg}$			0.22 (0.07,0.36)		
$\hat{\rho}_{C,sm}$				−0.01 (−0.11,0.09)	0.01 (−0.06,0.08)
$\hat{\rho}_{C,lg}$				0.08 (−0.02,0.18)	0.08 (−0.03,0.18)
$\hat{\rho}_{T,sm}$				0.27 (0.14,0.41)	0.25 (0.07,0.42)
$\hat{\rho}_{T,lg}$				0.36 (0.24,0.48)	0.37 (0.24,0.50)