A Mixed Model-Based Variance Estimator for Marginal Model Analyses of Cluster Randomized Trials

Thomas M. Braun*

Department of Biostatistics, University of Michigan, 1420 Washington Heights, M4063 SPH II, Ann Arbor, MI 48109, USA

Received 15 December 2005, revised 24 April 2006, accepted 18 July 2006

Summary

Generalized estimating equations (GEE) are used in the analysis of cluster randomized trials (CRTs) because: 1) the resulting intervention effect estimate has the desired marginal or population-averaged interpretation, and 2) most statistical packages contain programs for GEE. However, GEE tends to underestimate the standard error of the intervention effect estimate in CRTs. In contrast, penalized quasi-likelihood (PQL) estimates the standard error of the intervention effect in CRTs much better than GEE but is used less frequently because: 1) it generates an intervention effect estimate with a conditional, or cluster-specific, interpretation, and 2) PQL is not a part of most statistical packages. We propose taking the variance estimator from PQL and re-expressing it as a sandwich-type estimator that could be easily incorporated into existing GEE packages, thereby making GEE useful for the analysis of CRTs. Using numerical examples and data from an actual CRT, we compare the performance of this variance estimator to others proposed in the literature, and we find that our variance estimator performs as well as or better than its competitors.

Key words: Clustered data; Correlated data; GEE; Group randomized trial; PQL.

1 Introduction

1.1 Cluster randomized trials

Cluster randomized trials (CRT) are clinical trials in which each member of a particular group is randomized to the same treatment arm. Although CRTs can be politically motivated or used to reduce bias [4], the genesis of many CRTs is to address a typical public health question: can group-administered interventions lead to improved health to the members of those groups on average? For example, the Community Intervention Trial for Smoking Cessation (COMMIT) was designed to test a community intervention aimed at encouraging smokers to stop smoking (COMMIT Research Group, 1995). The intervention used a wide array of channels, such as media campaigns and public education, health care providers, and cessation resources. The rationale was that these programs, when used in tandem, might effectively reach smokers and induce them to quit.

CRTs offer a huge potential public health impact, but they typically cost many millions of dollars to conduct, making it more cost-effective to include few clusters, but include as many members of each cluster as possible. For example, COMMIT randomized 11 pairs with an average cluster size of over 2800 subjects, and the total cost of the study was $45 million US (COMMIT Research Group, 1991). Furthermore, due to unmeasurable cluster characteristics, outcomes from subjects in the same cluster tend to be associated with each other more than outcomes from subjects in different clusters.

* Corresponding author: e-mail: tombraun@umich.edu, Phone: +1 734 936 9844, Fax: +1 173 763 2215

Thus, when assessing the effect of the intervention, the induced intra-cluster correlation (ICC), denoted ρ, requires investigators to use correlated data methods. Fortunately, there is a growing amount of literature available to readers interested in the statistical issues underlying the design and analysis of CRTs (Gail et al., 1996; Donner and Klar, 2000; Murray, 1998; Feng et al., 2001).

1.2 Marginal models versus mixed models

Generalized estimating equations (GEE) (Liang and Zeger, 1986) are commonly used in the analysis of CRTs because: 1) the intervention effect parameter describes the intervention's impact on a randomly selected subject in the population, rather than a randomly selected subject from a given cluster, and 2) programs for GEE have been developed for most statistical packages. In GEE, the intervention effect is estimated using generalized weighted least-squares, with the weight matrix for each cluster reflecting the possible covariance structure of the cluster's outcomes. However, knowing the correct variance structure of the data is not necessary for estimating the intervention effect because the intervention effect estimate remains consistent even when the weight matrix is misspecified (Liang and Zeger, 1986). Furthermore, the model-based variance estimator (i.e. inverse of the information matrix) can be replaced by an empirical estimator, the so-called "sandwich" estimator, which has been claimed to be robust to an incorrectly specified weight matrix (Liang and Zeger, 1986).

A competing approach to GEE is penalized quasi-likelihood (PQL), which is based upon a conditional or mixed model (Breslow and Glayton, 1993). As applied to CRTs, mixed models incorporate not only the intervention effect, but also a random cluster effect, into the mean of each observation, making the interpretation of the intervention effect conditional upon the value of the (unobservable) cluster effect. PQL assumes the cluster effects are independent, each with an identical normal distribution around zero with unknown variance. By averaging over this normal distribution, the intervention effect of a marginal model can be viewed as the average conditional intervention effect across all cluster effects. PQL and other conditional model approaches are used less frequently in the analysis of CRTs because: 1) the intervention effect parameter does not have the desired marginal interpretation of that in GEE, and 2) few software packages include conditional model approaches for non-normal outcomes.

However, GEE and PQL lead to quantitatively identical intervention effect estimates with Gaussian and Poisson data. With binary data, the ratio of the marginal model intervention effect to the mixed model intervention effect is approximately $[1 + 16\theta\sqrt{3}/(15\pi)]^{-1/2}$, where θ is the variance of the cluster effects (Zeger et al., 1988). This ratio is very close to 1.0 when θ is small (i.e. small variation between clusters), exactly the setting of most CRTs. Thus, when analyzing CRTs with binary outcomes, marginal and mixed models give very similar results when estimating the intervention effect.

However, the two approaches differ drastically in their abilities to estimate the variance of the intervention effect estimate. GEE was designed for use in longitudinal studies, which typically have a large number of clusters (subjects), each of which consists of a small number of observations (repeated measures). Thus, the asymptotic properties of correlated data methods, which are driven by the number of clusters, are meaningful when applied to longitudinal studies, but are often unrealized when applied to CRTs because of the insufficient number of clusters. As a result, the sandwich estimator of GEE underestimates the variance of an estimated cluster-level fixed effect when there are a small number of clusters (Sharples and Breslow, 1992; Emrich and Piedmonte, 1992; Park, 1993; Feng et al., 1996; Braun and Feng, 2001). Such downward bias results from the sandwich variance estimate converging to the true variance at a rate that depends upon the number of clusters, regardless of the number of observations in each cluster. As a result, both traditional Wald tests and generalized marginal score tests (Rotnitzky and Jewell, 1990) have inflated Type I error rates, leading some authors to suggest that GEE should not be used in the analysis of CRTs (Feng et al., 2001) and propose a generalized linear mixed model application for CRTs (Yasui et al., 2004). However, we feel that GEE should be used estimate the intervention effect parameter due to the marginal, rather than conditional, interpretation of the estimator. It is the variance estimator of GEE that we wish to modify.

In contrast to GEE, the variance estimator from PQL for a cluster-level fixed effect estimate is much more accurate. As shown by Bellamy et al., (2005), the variance estimator of PQL converges to the true variance at a rate that depends upon the total number of observations across all clusters. As a result, even with few clusters, PQL will generate a much more accurate variance estimate than GEE if each cluster has many observations, which is specifically the setting of CRTs. Empirical evidence of this fact can be found in Bellamy et al. (2000) and Ten Have and Localio (1999). Therefore, we seek to express the PQL variance estimator as a sandwich-type variance estimator that can replace the traditional robust variance estimator of GEE. In Section 2, we give a more thorough description of GEE and PQL that leads to the motivation for our variance estimator, and we also describe existing alternatives to our variance estimator. Section 3 contains a simulation-based comparison the variance estimators described in Section 2, and Section 4 compares the variance estimators in an actual application. Section 5 contains our final comments.

2 New Variance Estimator & Existing Alternatives

2.1 Marginal models & GEE

We have a CRT in which there are M clusters and the i-th cluster, $i = 1, 2, \ldots, M$ consists of n_i individuals. Each cluster is randomized to receive either an intervention or control, and we let T_i be an arm assignment indicator ($0 =$ control; $1 =$ intervention).

We let Y_{ij} and $\boldsymbol{X}_{ij}^t = [1, \boldsymbol{Z}_{ij}, T_i]$ be the outcome and $(1 \times p)$ design vector, respectively, for the j-th individual in the i-th cluster, where $\boldsymbol{Z}_{ij}$ is a vector of covariates other than arm assignment. We define $\boldsymbol{X}_i^t$ to be the $(n_i \times p)$ design matrix for cluster i by stacking the vectors $\boldsymbol{X}_{1j}^t, \boldsymbol{X}_{2j}^t, \ldots, \boldsymbol{X}_{n_ij}^t$ upon each other. Under the assumptions of a marginal model, the outcome of each individual in a cluster has mean μ_{ij} and variance $\phi v(\mu_{ij})$, where ϕ is a known constant usually denoted σ^2 for normal outcomes and equal to 1 for binary and count data. We relate μ_{ij} to $\boldsymbol{X}_{ij}$ via the function $\eta_{ij} = \eta(\mu_{ij}) = \boldsymbol{X}_{ij}^t \boldsymbol{\beta}$, in which $\boldsymbol{\beta}^t = [\beta_0, \beta_1, \ldots, \beta_{p-1}]$ is a vector of parameters. We assume that η is a canonical link function, i.e. log with Poisson outcomes or logit with binary outcomes.

Via GEE, the regression-based estimate of $\boldsymbol{\beta}$, denoted $\hat{\boldsymbol{\beta}}_G$, is the solution to $\sum_i \boldsymbol{D}_i^t \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = \boldsymbol{0}$, where $\boldsymbol{S}_i = (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, $\boldsymbol{Y}_i$ is the $n_i \times 1$ vector of outcomes for cluster i, and $\boldsymbol{\mu}_i$ is a corresponding vector of means. $\boldsymbol{D}_i$ is an $n_i \times p$ matrix with column k equal to $\partial \boldsymbol{\mu}_i / \partial \beta_k$ and $\boldsymbol{V}_i^{-1}$ is the inverse of an $n_i \times n_i$ weight matrix that reflects the variance and *assumed* correlation structure of $\boldsymbol{Y}_i$. If we assume that $\boldsymbol{V}_i$ is identical to $\boldsymbol{\Sigma}_i$, the actual covariance matrix of $\boldsymbol{Y}_i$, the model-based (naive) variance estimate of $\hat{\boldsymbol{\beta}}_G$ is $\widehat{\boldsymbol{\Sigma}}_m = \left(\sum_i \widehat{\boldsymbol{D}}_i^t \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{D}}_i \right)^{-1}$, in which the hats denote estimates of the true values. If we do not wish to assume that $\boldsymbol{V}_i = \boldsymbol{\Sigma}_i$, we use the vector of residuals $\widehat{\boldsymbol{S}}_i = \boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i$ to compute an estimate $\widehat{\boldsymbol{\Sigma}}_i$ of $\boldsymbol{\Sigma}_i$, and instead of using $\widehat{\boldsymbol{\Sigma}}_m$, we use the "sandwich" variance estimate of $\hat{\boldsymbol{\beta}}_G$, equal to $\widehat{\boldsymbol{\Sigma}}_s = \widehat{\boldsymbol{\Sigma}}_m \widehat{\boldsymbol{\Sigma}}_0 \widehat{\boldsymbol{\Sigma}}_m$, where $\widehat{\boldsymbol{\Sigma}}_0 = \sum_i \widehat{\boldsymbol{D}}_i^t \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{\Sigma}}_i \widehat{\boldsymbol{V}}_i^{-1} \widehat{\boldsymbol{D}}_i$.

2.2 Mixed models & PQL

In Section 2.1, the marginal mean of each individual's outcome was modeled without regard to the correlation between individuals in the same cluster. In a mixed model, we assume each cluster has a unobserved random cluster effect u, which if observed, would eliminate the within-cluster correlation. Therefore, the conditional mean μ_{ij}^* for each subject's outcome in a cluster is expressed as a linear combination of covariates and the cluster effect $\eta_{ij}^* = \eta(\mu_{ij}^*) = \boldsymbol{X}_{ij}^t \boldsymbol{\beta} + u_i$. This defines $\phi v(\boldsymbol{\mu}_{ij}^*)$ to be the conditional variance of outcomes in a cluster. PQL assumes that the random cluster effects are mutually independent with each having a mean-zero normal distribution with variance θ. Since the variability between clusters is quantified completely by θ, the PQL model with this single component of variation is analogous to the assumption of an exchangeable correlation structure in marginal mod-

els. PQL is based upon a Laplace approximation to the exact conditional likelihood and iterates between two sets of weighted least-squares equations. The first set of equations compute an estimate of $\boldsymbol{\beta}$, which we denote $\hat{\boldsymbol{\beta}}_P$, for given values of θ and the cluster effects $\boldsymbol{u} = [u_1, u_2, \ldots, u_M]$, while the second set of equations compute values for θ and u at the current value of $\hat{\boldsymbol{\beta}}_P$.

Others have shown that PQL tends to produce biased estimates of both fixed and random effect parameters with longitudinal binary data (Breslow and Lin, 1995; Lin and Breslow, 1996). However, Bellamy et al. show that the bias converges to zero at a rate $\mathcal{O}(1/N)$, where $N = \sum_i n_i$ is the total number of observations (Bellamy et al., 2005). As each cluster in CRTs tends to be relatively large, N is much larger in CRTs than it is in longitudinal studies, leading to a substantial reduction in the bias of PQL. This bias reduction is not realized in GEE because the bias vanishes at rate $\mathcal{O}(1/m)$, where m is the number of clusters. Most importantly, a close examination of the results of Bellamy et al. (2005), demonstrates that the PQL variance estimator can be approximated by a sandwich-type form estimator common to marginal model approaches.

As we stated earlier, the small between-cluster variability in CRTs leads to a negligible difference between the GEE-based estimate $\hat{\boldsymbol{\beta}}_G$ and the PQL-based estimate $\hat{\boldsymbol{\beta}}_P$, and we recommend using $\hat{\boldsymbol{\beta}}_G$ because of its marginal interpretation. However, because $\mathrm{Var}\,(\hat{\boldsymbol{\beta}}_G)$ (via the robust GEE estimator) converges at a slower rate than $\mathrm{Var}\,(\hat{\boldsymbol{\beta}}_P)$, we propose replacing the robust variance estimator of GEE with a PQL-based variance estimator, which is easy to incorporate into existing GEE computations due to its sandwich-type form. We now describe this variance estimator in greater detail.

2.3 Proposed variance estimator

In CRTs in which each cluster size $n_i \equiv n$, Bellamy et al. (2005), show that

$$\hat{\boldsymbol{\beta}}_P = \left(\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^t\right)^{-1} \sum_{i=1}^{m} \boldsymbol{X}_i\, g\left(\bar{Y}_i - \frac{\tilde{\boldsymbol{u}}_i}{n\theta}\right),$$

in which $\bar{Y}_i = \sum_{j=1}^{n} Y_{ij}/n$, $\tilde{u}_i$ is the predicted value of the random effect for cluster i, and $g(\cdot)$ is a canonical link function. Using a series of Taylor series expansions, Bellamy et al. (2005) derive an expression for the variance of $\hat{\boldsymbol{\beta}}_P$. By generalizing their findings to reflect unequal cluster sizes, we find that the variance of $\hat{\boldsymbol{\beta}}_P$ is approximately equal to $\boldsymbol{V}_1 + \boldsymbol{V}_2$, in which

$$\begin{aligned}
\boldsymbol{V}_1 &= \boldsymbol{S}_1{}^{-1}\, \boldsymbol{S}_0\, \boldsymbol{S}_1{}^{-1} \\
\boldsymbol{V}_2 &= \theta \left(\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^t\right)^{-1} \\
\boldsymbol{S}_1 &= \left(\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^t / n_i\right) \\
\boldsymbol{S}_0 &= \phi \left(\sum_{i=1}^{m} \boldsymbol{Z}_i \boldsymbol{X}_i^t / n_i\right)
\end{aligned}$$

where the j-th row of $\boldsymbol{Z}_i^t$ is created by multiplying each element of the covariate vector $\boldsymbol{X}_{ij}^t$ by the scalar

$$z_{ij} = E_{u_i}\left\{\frac{1 - u_i v'(\mu_{ij}^*)}{v(\mu_{ij}^*)}\right\}. \tag{1}$$

With normal outcomes, $v'(\mu_{ij}^*) = 0$, so that $\boldsymbol{S}_0 = \phi \boldsymbol{S}_1$. As a result, our variance estimator is exact and equal to $\phi \left(\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^t / n_i\right)^{-1} + \theta \left(\sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^t\right)^{-1}$, a value that reduces to the usual least-squares var-

iance estimator with independent data ($\theta = 0$). With non-normal outcomes, we see that $\boldsymbol{V}_1$ is a sandwich-type estimator in which S_0 serves to account for the non-normality. Although much of the between-cluster variability is accounted for in $\boldsymbol{V}_2$, $\boldsymbol{V}_1$ also reflects the between-cluster variability as $\boldsymbol{S_0}$ is a function of the conditional variance of each observation.

By assuming normality for the random cluster effects, the computation of z_{ij} is fairly straightforward with Gauss-Hermite quadrature (Liu and Pierce, 1994) or with the K-th order Taylor series around $u_i = 0$

$$z_{ij} \approx v^{-1}(\mu_{ij})\left\{1 + v'(\mu_{ij}) \sum_{k=1}^{K} \frac{2k+1}{(2k)!}\, E(u_i^{2k})\right\}.$$

Note that the Taylor series is evaluated at the marginal mean μ_{ij}, which is equal to the conditional mean μ_{ij}^* when $u_i = 0$. For example, a second-order approximation for z_{ij} with binary outcomes would be

$$z_{ij} \approx \frac{1 + (1 - 2\,\hat{\mu}_{ij})(3\hat{\theta}/2 + 5\,\hat{\theta}^2/8)}{\hat{\mu}_{ij}(1 - \hat{\mu}_{ij})}\,,$$

with hats indicating estimates for the unknown parameters. We subsequently discuss the computation of $\hat{\theta}$.

2.4 Estimating variation between clusters

Marginal model approaches often express the between-cluster variability θ in terms of an intra-cluster correlation coefficient ρ. Therefore, we need to derive an estimate of θ from the estimate of ρ in order to use our variance estimator. As shown in Commenges and Jacqmin (1994), if we write $\mathrm{Var}\,(Y_{ij}) = E\{\mathrm{Var}\,(Y_{ij} \mid u_i)\} + \mathrm{Var}\,\{E(Y_{ij} \mid u_i)\}$, and we interpret correlation as a percentage of total variation, i.e. $\mathrm{Var}\,(Y_{ij})$, explained by the model, i.e. $\mathrm{Var}\,\{E(Y_{ij} \mid u_i)\}$, we express the correlation ρ as

$$\rho = \frac{\mathrm{Var}\,(\mu^*)}{\mathrm{Var}\,(\mu^*) + E[v(\mu^*)]}\,. \tag{2}$$

Using first-order approximations around $u = 0$, we have $\mathrm{Var}\,(\mu^*) \approx \theta v^2(\mu)$ and $E[v(\mu^*)] \approx \phi v(\mu)$, so that $\rho \approx [\theta v(\mu)]/[\theta v(\mu) + \phi]$. Note that this approximation is exact for normal outcomes, as $\rho = \theta/(\theta + \phi)$ and represents the percentage of total variation explained by the variation between clusters. We also examined second-order approximations around $u = 0$ for $\mathrm{Var}\,(\mu^*)$ and $E[v(\mu^*)]$ and found little improvement over the first-order approximations.

Note that for non-normal outcomes, Eq. (2) will not be constant across all subjects. To derive an overall estimate of θ, we average across all subjects, using the marginal model estimate $\hat{\rho}$ and defining

$$\hat{\theta} = \sum_i \sum_j \frac{\hat{\rho}}{1 - \hat{\rho}}\, v^{-1}(\hat{\mu}_{ij})/N\,. \tag{3}$$

2.5 Existing small sample improvements to GEE

Other approaches to improving the variance estimator of GEE exist. Mancl and DeRouen (2001) state that using the residuals from GEE leads to an under-estimate of the actual covariance structure of the data, and they propose a bias-adjusted covariance estimator to use in the sandwich variance estimate; see also Kauermann and Carroll (2001). As this variance estimator can be liberal in some settings, Mancl and DeRouen (2001) also suggest comparing the resulting Wald test statistic to the quantile of an F distribution with 1 and $(M - p)$ degrees of freedom instead of a traditional χ^2 distribution, where M is number of clusters and p is the number of regression parameters.

Fay and Graubard (2001) state that the middle of the sandwich estimator, the sum of the observed cluster scores, should be inflated to reflect how the observed cluster scores are functions of estimated parameters. They also propose replacing the reference χ^2 distribution with an F-distribution, with the denominator degrees of freedom directly estimated from the data. Pan and Wall (2002) also propose a modification to reflect the extra variability of the cluster scores, although their method is based upon the method of moments and is more computationally intensive. Morel et al. (2003) develop a modification to GEE by inflating the sandwich variance estimator by a fraction of the naive variance estimator that is motivated from bias adjustment in sampling methods. The degree of inflation reflects the level of within-cluster correlation yet disappears with an increasing number of clusters. In the next section, we will compare the performance of some of these estimators to our estimator.

3 Numerical Examples

3.1 Description

We have a CRT that seeks to examine whether or not health behavior messages (i.e. exercising more, eating healthier, etc.) are more likely to encourage adoption of those behaviors if the messages are tailored specifically to a subject's familial risk of disease. In other words, we want to determine if knowledge of increased familial risk for a disease impacts a subject's behavior to prevent that disease. To answer our question, we plan to enroll a total of M physician practices and randomize each practice to one of two approaches: 1) convey standard health behavior messages to patients, or 2) convey health behavior messages to patients that vary by whether the patient has a low, moderate, or high familial risk of disease. We constrain our randomization so that $M/2$ practices are randomized to each arm. Each practice has a variable number of patients, but we expect to enroll an average of 100 patients in each practice. In our simulations, we examined $M \in \{20, 30, 40\}$, with the number of subjects in each practice distributed uniformly over the range $[70, 130]$.

We expect the probability of a subject's adopting the recommended behaviors will vary by the age and gender of the subject; therefore, we designed our analysis to adjust for these two subject-level covariates. Age and gender were generated so that the mean age of all participants was 40 and the average male-female ratio was 50/50. Correlation within a cluster was created by including a random cluster effect into the linear predictor of each observation. These random effects were generated from a normal distribution with mean 0 and variance θ, the value of which was selected so that each cluster had intra-class correlation ρ. When including individual-level covariates, binomial and Poisson outcomes cannot be simulated to have exact correlation ρ. Thus, in the following simulations, all parameters were selected so that the correlation within cluster was ρ on average for each cluster. Note that our data were generated using a mixed model approach, but will be analyzed with a marginal model approach with a canonical link (logit for binary and log for Poisson) function. We also simulated binary data from a marginal model per the methods of Oman and Zucker (2001), as well as with random effects from a gamma distribution with the correct variance and shifted left to have mean 0. Neither of those two simulation approaches produced results that differed significantly from those presented in Table 2.

Under each setting, we simulated data from 1000 hypothetical CRTs; in each CRT, we computed the intervention effect estimate $\hat{\beta}_{\text{int}}$ using GEE with a working independence correlation structure. We also computed the sample standard deviation of the 1000 intervention effect estimates to serve as a reference for all of the following variance estimators. We first computed the "sandwich" or "robust" variance estimator of GEE. We computed our proposed variance estimator with the true value of the between-cluster variance as well as with an estimate of the between-cluster variance using Eq. (3) and the moment-based estimator of ρ computed by GEE. We present our results when using Gauss-Hermite quadrature to compute the scalar z_{ij} in Eq. (1) as we found using a second-order Taylor series approximation for z_{ij} gave nearly identical results.

We also estimated the variance of $\hat{\beta}_{int}$ using the methods of Mancl and DeRouen (2001), Fay and Graubard (2001), and Morel et al. (2003). The variance estimator of Fay and Graubard, which we denote $\hat{v}_{FG}$, was designed for use with a statistic, which we denote $\widehat{F}_{FG}$, whose reference distribution was an F-distribution with 1 and d^* degrees of freedom. Therefore, we had to derive a scaled form of the Fay and Graubard variance estimator that was comparable to the other methods that relied upon a reference χ^2-distribution. Specifically, if $\sqrt{\widehat{F}_{FG}}$ is the q-th quantile of a t-distribution with d^* degrees of freedom, then the scaled Fay and Graubard variance estimator is $[\hat{\beta}_{int}/\Phi(q)]^2$, where $\Phi(\cdot)$ is the standard normal CDF. When examining the size of Fay and Graubard's F-test, we computed $\tilde{d}$, the denominator degrees of freedom, as outlined in their manuscript (Fay and Graubard, 2001).

3.2 Poisson outcomes

We expect subjects receiving standard health messages will practice the negative health behavior an average of $\lambda_0 = 3$ times per month. We anticipate that $\lambda_1 < \lambda_0$, where λ_1 represents the average times per month the negative health behavior is practiced by subjects receiving health messages tailored to their familial disease history. Table 1 compares all the variance estimators under the null hypothesis $\lambda_0 = \lambda_1$ for values of the intra-cluster correlation $\rho \in \{0.00, 0.05, 0.10\}$ which correspond to between-cluster variances of $\theta \in \{0.000, 0.018, 0.037\}$, respectively. For each value of θ, we computed three properties of each estimator: 1) the average standard error value across all 1000 simulations, 2) the mean-squared error (MSE) of each estimator in reference to the empirical standard error, and 3) the size of the hypothesis test corresponding to each variance estimator, computed as the percentage of simulations in which the null hypothesis was incorrectly rejected.

For all values of θ, our estimator produces consistent standard error estimates regardless of the number of clusters and whether θ is known or estimated. The interesting result is that the Wald test, although of correct size using the true value of θ, shows a slightly inflated Type I error rate with θ estimated. This result suggests that we have two options: (1) select another approach for estimating θ, of which there are many (Ridout et al., 1999), or (2) select a longer-tailed reference distribution for the Wald statistic. We hesitate to explore option (1) since our variance estimator is already accurate with our selected method of estimating θ. With regard to suggestion (2), we note that all the approaches demonstrate how poorly the χ^2 distribution approximates the distribution of the Wald statistic in small samples. As proposed in Mancl and DeRouen (2001), we could replace the reference χ^2 distribution of the Wald test with an F-distribution with 1 and $M - p$ degrees of freedom to create a Type I error rate closer to a nominal rate of 0.05.

The estimator of Mancl and DeRouen produces slightly overestimated standard errors, yet leads to a Wald test with nominal size when $M > 20$. In fact when $M = 20$, the Mancl and DeRouen estimator leads to the largest standard error estimate, yet creates a hypothesis test of smallest size. It appears that the increased value and MSE of the standard error from the Mancl and DeRouen estimator compensates for the poor approximation of the χ^2 distribution when $M = 20$. The estimator of Morel tends to create consistent standard error estimates in all settings, although the corresponding Wald test is liberal for non-zero values of θ and for values of $M < 40$. The variance estimator of Fay and Graubard tends to produce underestimated standard errors and an overly liberal F-test and performs similarly to the robust variance estimator of GEE.

3.3 Binary outcomes

We expect a proportion $p_0 = 0.30$ of subjects receiving standard health messages will perform a specific negative health behavior. We anticipate that $p_1 < p_0$, where p_1 is the proportion of subjects receiving health messages tailored to their familial disease history who will perform a specific negative health behavior. Table 2 compares all the variance estimators under the null hypothesis $p_0 = p_1$ for values of the intra-cluster correlation $\rho \in \{0.00, 0.05, 0.10\}$ which correspond to between-cluster variances of $\theta \in \{0.00, 0.27, 0.61\}$, respectively.

Table 1 Comparison of several standard error estimators, corresponding mean-squared errors (MSE) $(x10^{-4})$, and size of resulting hypothesis test, for an estimate of an intervention's effect upon the number of negative health behaviors (Poisson outcomes). M = number of clusters; θ = between-cluster variation; asterisk indicates value is < 0.001. E = empirical; B_θ = Braun with true value of θ; $B_{\hat{\theta}}$ = Braun with estimate of θ; GEE = GEE robust; MD = Mancl/DeRouen; M = Morel; FG = Fay/Graubard.

Estimator	$\theta = 0.00$			$\theta = 0.018$			$\theta = 0.037$		
	Value	MSE	Size	Value	MSE	Size	Value	MSE	Size
				$M = 20$					
E	0.026	*n/a*	n/a	0.065	*n/a*	*n/a*	0.091	*n/a*	*n/a*
B_θ	0.026	*	0.044	0.065	*	0.054	0.091	*	0.057
$B_{\hat{\theta}}$	0.027	*	0.040	0.063	0.026	0.088	0.091	0.132	0.075
GEE	0.024	*	0.094	0.062	0.024	0.092	0.086	0.093	0.089
MD	0.027	0.001	0.054	0.070	0.039	0.063	0.097	0.155	0.061
M	0.028	0.001	0.040	0.066	0.026	0.085	0.092	0.101	0.081
FG	0.023	0.001	0.086	0.058	0.035	0.086	0.081	0.134	0.090
				$M = 30$					
E	0.021	*n/a*	*n/a*	0.054	*n/a*	*n/a*	0.075	*n/a*	*n/a*
B_θ	0.021	*	0.051	0.053	*	0.053	0.074	0.001	0.053
$B_{\hat{\theta}}$	0.022	*	0.045	0.052	0.008	0.066	0.076	0.041	0.061
GEE	0.020	*	0.078	0.051	0.007	0.071	0.072	0.029	0.074
MD	0.022	*	0.052	0.056	0.009	0.050	0.078	0.037	0.052
M	0.022	*	0.048	0.054	0.007	0.066	0.075	0.028	0.072
FG	0.019	*	0.085	0.049	0.011	0.075	0.068	0.042	0.073
				$M = 40$					
E	0.018	*n/a*	*n/a*	0.046	*n/a*	*n/a*	0.064	*n/a*	*n/a*
B_θ	0.019	*	0.046	0.046	*	0.048	0.064	*	0.051
$B_{\hat{\theta}}$	0.019	*	0.044	0.046	0.003	0.054	0.066	0.019	0.048
GEE	0.018	*	0.068	0.045	0.003	0.057	0.063	0.011	0.061
MD	0.019	*	0.055	0.048	0.004	0.044	0.067	0.016	0.044
M	0.019	*	0.055	0.047	0.003	0.055	0.065	0.012	0.061
FG	0.017	*	0.069	0.043	0.003	0.065	0.060	0.014	0.065

Table 2 contains information analogous to that in Table 1. However, if θ is known, our variance estimator leads to overestimated standard errors with binary data and a Wald test with overly conservative rejection rates. Nonetheless, this bias is alleviated when we substitute θ with its estimate and our variance estimator performs as well as the Mancl and DeRouen and Morel estimators. The Fay and Graubard estimator continues to produce underestimated standard errors and liberal hypothesis tests like the GEE robust variance estimator.

4 Actual Application

We apply the aforementioned variance estimators to data collected from the Working Well Trial (WWT), which is one of the largest randomized worksite health promotion studies ever conducted, with over 20 000 workers participating (Abrams et al., 1994). Two cross-sectional samples were taken

Table 2 Comparison of several variance estimators, corresponding mean-squared errors (MSE) $(x10^{-4})$, and size of resulting hypothesis test, for an estimate of an intervention's effect upon the probability of performing a negative health behavior (Binomial outcomes). M = number of clusters; θ = between-cluster variation. E = empirical; B_θ = Braun with true value of θ; $B_{\hat{\theta}}$ = Braun with estimate of θ; GEE = GEE robust; MD = Mancl/DeRouen; M = Morel; FG = Fay/Graubard. Data simulated with mixed model.

Estimator	$\theta = 0.00$			$\theta = 0.270$			$\theta = 0.610$		
	Value	MSE	Size	Value	MSE	Size	Value	MSE	Size
				$M = 20$					
E	0.100	*n/a*	*n/a*	0.244	*n/a*	*n/a*	0.336	*n/a*	*n/a*
B_θ	0.100	0.002	0.052	0.259	1.142	0.038	0.374	14.071	0.028
$B_{\hat{\theta}}$	0.103	0.033	0.047	0.240	4.346	0.073	0.340	17.626	0.068
GEE	0.092	0.122	0.086	0.230	4.014	0.081	0.316	13.479	0.079
MD	0.104	0.150	0.051	0.260	6.498	0.058	0.356	20.222	0.052
M	0.105	0.116	0.042	0.247	4.125	0.077	0.338	12.698	0.072
FG	0.087	0.197	0.090	0.217	6.089	0.080	0.297	21.829	0.080
				$M = 30$					
E	0.076	*n/a*	*n/a*	0.202	*n/a*	*n/a*	0.274	*n/a*	*n/a*
B_θ	0.081	0.013	0.046	0.211	0.297	0.044	0.304	6.285	0.028
$B_{\hat{\theta}}$	0.084	0.035	0.038	0.200	1.263	0.073	0.283	5.769	0.053
GEE	0.076	0.024	0.061	0.192	1.240	0.079	0.263	3.667	0.070
MD	0.083	0.053	0.052	0.208	1.411	0.059	0.285	4.911	0.049
M	0.083	0.047	0.046	0.201	1.095	0.078	0.274	3.501	0.067
FG	0.072	0.030	0.067	0.181	2.040	0.081	0.249	6.144	0.076
				$M = 40$					
E	0.067	*n/a*	*n/a*	0.173	*n/a*	*n/a*	0.235	*n/a*	*n/a*
B_θ	0.070	0.003	0.035	0.182	0.244	0.036	0.263	3.965	0.025
$B_{\hat{\theta}}$	0.072	0.013	0.034	0.175	0.532	0.053	0.247	2.941	0.048
GEE	0.067	0.011	0.058	0.168	0.438	0.063	0.230	1.372	0.062
MD	0.071	0.019	0.049	0.178	0.583	0.047	0.244	2.037	0.050
M	0.071	0.016	0.047	0.174	0.444	0.059	0.238	1.466	0.062
FG	0.064	0.014	0.067	0.160	0.660	0.063	0.220	2.017	0.062

for each worksite: a set of baseline measurements (Heimendinger et al., 1995), and a corresponding set of measurements taken three years after baseline (Sorensen et al., 1996).

As quitting smoking is believed to decrease the risk of developing lung and other cancers, one aim of the WWT was to determine if employees of worksites that encouraged smoking cessation actually modified their daily routine and quit smoking. The WWT enrolled employees from a total of 114 worksites within 4 geographic centers: the University of Florida (UF), the Dana-Farber Cancer Institute (DFCI), the MD Anderson Cancer Center (MDACC), and Brown University (BU). As the UF center did not collect information on employees' smoking, our analysis focuses upon the three centers that did collect smoking information, leading to a total of 87 worksites and nearly 4000 employees who were current smokers at baseline included in our analysis.

The outcome of interest was whether or not a smoker had quit smoking by the time of the post-intervention survey, the probability of which was modeled using logistic regression as a function of treatment arm assignment, as well as the age, gender, and race of each employee. We computed the

Table 3 Comparison of several standard error estimates for the estimated intervention effect on the prevalence of smoking in worksites participating in the WWT.

	All Centers	DFCI	MDACC	BU
Number of Worksites	87	24	40	23
Number of Employees	3,778	1,312	1,319	1,147
Intervention Effect Estimate	0.152	0.424	−0.131	0.107
Standard Error Estimates:				
Braun	0.130	0.221	0.194	0.234
Mancl/DeRouen	0.120	0.233	0.159	0.234
Morel	0.119	0.232	0.163	0.233
Fay/Graubard	0.110	0.191	0.136	0.171
GEE Robust	0.115	0.209	0.147	0.204

intervention effect estimate and each of the standard error estimates across all three centers, as well as separately for each of the three centers. The results are displayed in Table 3.

Pooling the data from all three centers gives us an overall intervention effect estimate of 0.152, indicating a higher quit rate in intervention worksites than in control worksites (14% versus 12%, respectively). However, we find that the differential in quit rates was greatest in the DFCI center and that the quit rates in the MDACC center were actually reversed, with intervention worksites having lower quit rates than control worksites.

With regard to the standard error estimates, we see that when pooling the data from all three centers, the variance estimators of Fay and Graubard and GEE robust both tended to produce standard errors lower than the others, confirmed earlier in the simulation results. In addition, our variance estimator produced a larger standard error estimate than those of Mancl and DeRouen and Morel, a trend not observed in our simulations.

When stratifying our analysis across the three centers, we found that the variance estimators of Fay and Graubard and GEE continued to produce underestimated standard errors. For the analysis of the DFCI (MDACC) worksite, we found that our variance estimator tended to produce a standard error less (greater) than that of Mancl and DeRouen or Morel, while with the BU worksite, all three methods gave similar standard errors. As a result, it remains inconclusive as to which of our approach and those of Mancl and DeRouen and Morel will tend to work uniformly best in practice.

5 Concluding Remarks

Our methodology has combined the mean parameter estimation abilities of GEE with the variance estimation abilities of PQL. We still propose estimating the intervention effect in CRTs with marginal model approaches, yet propose a PQL-based variance estimator that has better small sample properties than the robust variance estimator of GEE, performs as well as those proposed by Mancl and DeRouen and Morel et al., and better than that proposed by Fay and Graubard. In our simulations, our variance estimator tended to produce standard error estimates that were as close or closer to the empirical standard error than that of Mancl and DeRouen. However, with a small number of clusters, the corresponding Wald test had a nominal size with the estimator of Mancl and DeRouen more often than with our estimator, suggesting the need for a different reference null distribution. One possible area of research is to compare the performance of our estimator to that of Mancl and DeRouen after generating the null distribution of the Wald statistic using permutation methods similar to that described by Braun and Feng (2001). Another approach would be the use of a modified χ^2 statistic similar to that examined by Jung et al., (2001), although such an approach would first require a generalization for individual-level covariates.

With binary data, our variance estimator had an upward bias with θ known, although this bias was alleviated by replacing θ with its estimate. Although Breslow and Lin (1995) have shown that extending PQL to a second-order Laplace approximation leads to estimators with reduced bias, additional research is needed to determine if including this bias adjustment will be useful for our approach, as the variation between clusters in our examples was outside the range examined in Breslow and Lin (1995).

Heagerty (1999) provides an alternate two-model approach to combining marginal and conditional mean model approaches with longitudinal binary data. As applied to our setting, the first model specifies the marginal mean as a function of intervention and other covariates as is traditionally done. The second model specifies the conditional mean as a function of a random cluster effect and unobserved latent effects that address the dependence within-cluster. This approach is contrasted to a strictly conditional model which models the conditional mean on observed covariates rather than unobserved latent effects. As a result, the two-model approach includes a intervention effect estimate that has a marginal interpretation. Nonetheless, this methodology was developed for longitudinal studies, and further research is needed to determine if the resulting variance estimator converges quickly enough to make it a viable approach for analyzing CRTs.

Acknowledgements This research is supported (in part) by the National Institutes of Health through the University of Michigan's Cancer Center Support Grant (5 P30 CA46592).

References

Abrams, D. B., Boutwell, W. B., Grizzle, J., Heimendinger, J., Sorensen, G., and Varnes, J. (1994). Cancer control at the workplace: The Working Well Trial. *Preventive Medicine*, **23**, 15–27.

Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., Lipsitz, S., and Ryan, L. (2000). Analysis of dichotomous outcome data for community intervention trials. *Statistical Methods in Medical Research*, **9**, 135–139.

Bellamy, S. L., Li, Y., Lin, X., and Ryan, L. M. (2005). The bias in penalized quasilikelihood estimator of cluster-level covariate effect in generalized linear mixed models for group-randomized trials. *Statistica Sinica*, **15**, 1015–1032.

Braun, T. M. (1999). *Optimal Analysis of Group Randomized Trials with Permutation Tests*. PhD thesis University of Washington.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. *Journal of the American Statistical Association*, **96**, 1424–1432.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalised linear models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Commenges, D. and Jacqmin, H. (1994). The intraclass correlation coefficient: distribution-free definition and test. *Biometrics*, **50**, 517–526.

COMMIT Research Group (1991). Community intervention trial for smoking cessation (COMMIT): summary of design and intervention. *Journal of the National Cancer Institute*, **83**, 1620–1628.

COMMIT Research Group (1995). Community intervention trial for smoking cessation (COMMIT). *American Journal of Public Health*, **85**, 183–192.

Donner, A. and Klar, N. (2000). *The Design and Analysis of Cluster Randomization Trials in Health Research*. Arnold, London.

Emrich, L. J. and Piedmonte, M. R. (1992). On some small sample properties of generalized estimating equation estimates for multivariate dichotomous outcomes. *Journal of Statistical Computation and Simulation*, **41**, 19–29.

Fay, M. and Graubard, B. (2001). Small-sample adjustment for Wald-type tests using sandwich estimators. *Biometrics*, **57**, 1198–1206.

Feng, Z., Diehr, P., Peterson, A., and McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health*, **22**, 167–187.

Feng, Z., McLerran, D., and Grizzle, J. (1996). A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine*, **15**, 1793–1806.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, **15**, 1069–1092.

Heagerty, P. J. (1999). Marginally specified logistic-normal models for longitudinal binary data. *Biometrics*, **55**, 688–698.

Heimendinger, J., Feng, Z., Emmons, K., Stoddard, A., Kinne, S., Biener, L., Sorensen, G., Abrams, D., Varnes, J., and Boutwell, B. (1995). The Working Well Trial: baseline dietary and smoking behaviors of employees and related worksite characteristics. *Preventive Medicine*, **24**, 180–193.

Jung, S.-H., Ahn, C., and Donner, A. (2001). Evaluation of an adjusted chi-square statistic as applied to observational studies involving clustered binary data. *Statistics in Medicine*, **20**, 2149–2161.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, **96**, 1387–1396.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Lin, X. and Breslow, N. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, **91**, 1007–1016.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, **81**, 624–629.

Mancl, L. and DeRouen, T. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, **57**, 126–134.

Morel, J., Bokossa, M., and Neerchal, N. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal*, **45**, 395–409.

Murray, D. M. (1998). *Design and Analysis of Group-Randomized Trials*. Oxford University Press, New York.

Oman, S. D. and Zucker, D. M. (2001). Modelling and generating correlated binary variables. *Biometrika*, **88**, 287–290.

Pan, W. and Wall, M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, **21**, 1429–1441.

Park, T. (1993). A comparison of the generalized estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine*, **12**, 1723–1732.

Ridout, M. S., Demtétrio, C. G., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, **55**, 137–148.

Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika*, **77**, 485–497.

Sharples, K. and Breslow, N. (1992). Regression analysis of correlated binary data: some small sample results for the estimating equation approach. *Journal of Statistical Computation and Simulation*, **42**, 1–20.

Sorensen, G., Thompson, B., Glanz, K., Feng, Z., Kinne, S., DiClemente, C., Emmons, K., Heimendinger, J., Probart, C., and Lichtenstein, E. (1996). Work site-based cancer prevention: primary results from the Working Well Trial. *American Journal of Public Health*, **86**, 939–947.

Ten Have, T. and Localio, R. (1999). Empirical Bayes estimation of random effects parameters in mixed models. *Biometrics*, **55**, 1022–1029.

Yasui, Y., Feng, Z., Diehr, P., McLerran, D., Beresford, S. A. A., and McCulloch, C. (2004). Evaluation of community-intervention trials via generalized linear mixed models. *Biometrics*, **60**, 1043–1052.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

