Comparison of Permutation Tests and GEE Methods for Group Randomized Trials with Count Data

Ping Xu[1], Brian Leroux[2]
[1]Axio Research LLC, 2601 4th Ave Suite 200, Seattle, WA 98121
[2]Department of Biostatistics, Box 357475, University of Washington, Seattle WA 98195

Abstract
Group-randomized trials (GRT) often have a small number of clusters, creating challenges for identifying a valid analysis method with sufficient power. Previous studies found problems with two popular methods- GEE and permutation tests. GEE suffers from anti-conservative tests even with small-sample corrections, while permutations suffer from conservativeness due to discreteness of test distributions. We compared the performance of small sample adjusted GEE methods with three weighted permutation procedures for GRT with count data. Type I error rates and power were estimated by simulation in 81 scenarios with number of clusters from 10 to 40, and ranges of values for overall mean count, covariate effect, overdispersion parameter, and intracluster correlation coefficient. Permutation tests had valid type I errors in all scenarios and had power at least as high as other tests. Average power to detect moderate treatment effects was 0.794 for weighted permutation test versus 0.787 for the best-performing GEE method. For large treatment effects, power was 0.905 versus 0.901. In conclusion, permutation tests showed a wider range of validity with no loss of power, compared to GEE methods.

Key Words: Cluster- randomized trial, Generalized estimating equations, Weighted permutation test, Correlated count data, Bias-corrected variance estimator

Section 1: Introduction

A group randomization trial is one in which intact social units, or groups of individuals, rather than individuals themselves, are randomized to different intervention groups. Group randomization trials, sometimes called cluster-randomized trials, have become particularly widespread in the evaluation of non-therapeutic interventions, including life-style modification, educational programs and innovations in the provision of health care (Donner & Klar, 2000). The units of randomization in such studies are diverse, ranging from relatively small clusters, such as households or families, to entire neighborhoods or communities, but also including worksites, hospital wards, classrooms and medical or dental practices. Examples of GRTs include the National Cancer Institute's Working Well Trial (group: work site), their 5-A-Day Program (group: work site), the Hutchinson Smoking Prevention Project (group: school district), the National Cancer Institute's COMMIT project (group: community), and Kaiser Family Foundation's Community Health Promotion Grants Program (group: community).

The statistical features of group randomization were first brought wide attention in the health research community by Cornfield (1978). Although Cornfield stated that such allocation schemes are less efficient than designs which randomize individuals to

intervention groups, reasons for adopting a group randomized trial rest on other considerations, such as ethical issues, the desire to control costs or attempts to minimize experimental contamination. Differing from individual randomized trials, group randomization trials have two fundamental features: First, the individual outcomes within the same group are often correlated and this correlation is called intra-class correlation (ICC). The second characteristic of GRTs is that given that interventions are delivered at the group level, the most cost-efficient design is to randomize a relatively small number of groups (e.g. COMMIT had 11 communities per treatment condition, Gail et al, 1996) to each treatment condition, and to have a moderate or large number of participants per group. These two features create several problems in the design and analysis of GRTs and invalidate standard approaches to both the estimation of sample size and the analysis of the trial data.

With more and more attention paid to the specific features of GRTs, several analytic approaches have been identified to provide valid analysis for GRTs. The three primary approaches for evaluating correlated data arising in GRTs include: (1) generalized linear mixed models (GLMM); (2) generalized estimating equations (GEE); and (3) randomization-based inference (i.e. permutation test, jackknife estimation, bootstrap approach).

Generalized linear mixed models are extensions of the generalized linear models (GLM) involving models with random terms in the linear predictor. Differing from GLMs, GLMMs parameterize the correlation within clusters using a random cluster effect, which has its own distribution (usually assumed to be independently normally distributed). The GLMM models are also called conditional models, since they are based on modeling the response conditional on the random effects (Schall, 1991). In order to compute an estimate for the intervention-effect parameter, one must integrate the joint likelihood over all possible values of the random effects. There are a multitude of numerical methods available to approximate the integral, of which the penalized quasi-likelihood (PQL) (Breslow & Clayton, 1993) is a commonly used method.

An alternate large-sample method for analyzing correlated data is generalized estimating equations (Liang & Zeger, 1986). The equations are extensions of those used in quasi-likelihood methods by simply replacing the identity matrix with a more general correlation matrix. In contrast to GLMM, GEE adopts a marginal model and the regression is based upon the marginal means of the observations and it estimates the intervention-effect parameter by solving a score-like “estimating equation”. In addition, a "working" correlation matrix for the observations for each cluster is specified. This set-up leads to the above-mentioned estimating equations which give consistent estimators of the regression coefficients and of the covariance matrix of the estimates. However, the asymptotic distribution of the estimate is usually not appropriate in GRTs which usually have small numbers of clusters. It has been noted in the literature that for small samples, the unadjusted robust Wald tests tend to be too liberal in terms of maintaining the nominal test sizes (Lin & Wei, 1989; Emrich & Piedmonte, 1992; Gunsolley et al, 1995; Fay et al, 1998; Mancl & DeRouen, 2001). Thus several adjustment approaches for small-sample clustered data have been proposed and three main approaches include: (1) direct modification of the sandwich estimator by correcting its downward or upward bias (Mancl et al, 2001; Fay et al, 2001); (2) using a different approximate test (i.e. a Student t- or F-test) instead of the regular chi-squared test as the reference distribution for coefficient estimators (Paik, 1988; Lipsitz et al., 1990; Lipsitz et al., 1994; Qu et al., 1994) and (3) a combined approach of (1) & (2) to account for both the bias and

variability of the sandwich estimator at the same time (Mancl et al, 2001; Fay et al, 2001). These previous studies have confirmed that using the bias-corrected estimators or/and using an F-distribution to compute the critical values would help with the inflation issue and produce tests with sizes substantially closer to the nominal level, thereby acknowledging the small-sample feature of GRTs.

In addition to the small-sample adjustment GEE methods, there are other studies proposing re-sampling methods for small-sample correlated response data, among which the permutation tests have wider range of applications: permutation tests have been applied in cluster analysis (Hubert & Levin, 1976), Fourier analysis (Freedman & Lane, 1980), multivariate analysis (Arnold, 1964; Mielke, 1986) and single-subject analysis (Kazdin, 1980). A permutation test, also called an exact test, originated from the works of Fisher and Pitman in the 1930s (Fisher, 1935; Pitman, 1938). It is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the treatment assignments on observed data points. Although the permutation test was originally developed for individual-randomized studies, it can be readily applied to GRTs. For GRTs, the treatments assigned to groups are permuted. The first extensive examination of using permutation tests to analyze GRTs was made by Gail et al (1992). This well-known publication described significance tests based on the permutation distribution and the approaches for covariate adjustment. They demonstrated that in GRTs, the permutation test remains valid under almost all practical situations in GRTs, including unbalanced group sizes. Although members of the permutation test family have the appropriate type I error rate, the power of each test depends upon which statistic is used. Gail et al (1996) and Brookmeyer & Chen (1998) empirically examined the power of permutation tests, but there is no comprehensive theoretic development of efficient permutation tests for correlated data regression settings like GRTs. It was Braun & Feng (2001) that systematically developed permutation inference that will have maximal power against a specific alternative, proposed the optimal tests using three different weighted test statistics and also compared the power of these permutation tests to that of GEE and PQL through simulation studies.

Section 2: Methods

In this paper, we examined six tests through simulation studies, including three permutation tests, two small-sample adjustment GEE tests and the unadjusted GEE test. Within the family of permutation tests, we compared one un-weighted permutation test with two weighted permutation tests, in order to examine the effect of weighting on test size and power. We also compared the performance of the three permutation tests with the bias-corrected GEE test and the combined GEE test.

2.1 Un-weighted Permutation Test

The commonly used test statistic for permutation test is the un-weighted one, which uses the un-weighted mean of the residuals of each cluster as the test statistic. This traditional permutation test has been applied in many previous studies including the COMMIT study (Gail et al, 1992). The test statistic is computed by:

$$S_u = \sum_i \frac{1}{ni} \sum_j \boldsymbol{T}_{ij} J_{nij} \left(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij} \right) \qquad (\mathbf{2.1})$$

, where n_i is the cluster size for the *i*th cluster,
$\mathbf{T}_{ij}$ is the treatment assignment for the *i*th cluster and *j*th subject ($\mathbf{T}_{ij}$=+1/-1),
$\mathbf{J}n_{ij}$ is a (1 X n_i) vector of ones,
$\mathbf{Y}_{ij}$ and $\boldsymbol{\mu}_{ij}$ are the corresponding (n_i X 1) vectors of outcomes and their expected values.

2.2 Cluster-size-weighted Permutation Test

Not as commonly used as the un-weighted permutation test, the cluster-size-weighted test has been applied to two group randomized clinical trials, the Hutchinson Smoking Prevention Project (HSPP, Peterson et al 2000) and the Smokeless Tobacco Cessation Intervention Study (Walsh et al, 1999). The test statistic used here is the difference in overall averages between the control and experimental groups and thus is weighted by the cluster size. The formula is given below:

$$S_c = \frac{\sum_i^{Control} \sum_j \boldsymbol{J}_{nij}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij})}{\sum_i^{Control} ni} - \frac{\sum_i^{Test} \sum_j \boldsymbol{J}_{nij}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij})}{\sum_i^{Test} ni} \qquad (\mathbf{2.2})$$

, where $\mathbf{n}_i$, $\mathbf{J}_{nij}$, $\mathbf{Y}_{ij}$, and $\boldsymbol{\mu}_{ij}$ have the same definitions as **2.1**.

2.3 Correctly-weighted Permutation Test

Braun and Feng (2001) developed a permutation test which uses a statistic that is a weighted sum of residuals. In their study, a marginal model was used to motivate this so-called correctly-weighted test statistic. Such a model generates a quasi-score statistic, with the weights based on the cluster sizes and the intra-cluster correlation. This test's optimality was proved by Li (Li, 1993). Besides, Braun demonstrated in his PhD dissertation (Braun, 1999) that this correctly-weighted test statistic is closest to the true score and the corresponding test is locally most powerful (LMP).

The correctly-weighted test statistic is given by:

$$\mathbf{S}_w = \sum_i \sum_j \{\boldsymbol{T}_{ij} D_{ij} \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_{ij})\} \qquad (\mathbf{2.3})$$

, where $\boldsymbol{D}_{ij}$ is the corresponding (1 X n_i) vector of gradients,
$\boldsymbol{V}_i$ is an (n_i X n_i) covariance matrix of the *i*th cluster, with non-zero elements off its diagonal.

Specifically for the count responses, $\boldsymbol{D}_{ij}$ is the derivative of the mean with respect to the linear predictor, ie, as Braun states that $\mathbf{D}_{ij} = (d\,\eta_{ij}/d\,\mu_{ij})^{-1}$ which is for a log-linear model where $\eta = \log(\mu)$, $(d\,\eta_{ij}/d\,\mu_{ij})^{-1} = (1/\mu_{ij})^{-1} = \mu_{ij}$.

2.4 Un-adjusted GEE Method

The generalized estimating equation methodology (GEE) has become a popular regression method which is applied to analyze correlated responses. The term generalized estimating equations indicates that an estimating equation is not the result of a likelihood-based derivation, but that it is obtained by generalizing another estimating equation. Suppose we have data that consists of correlated observations $\{y_{ij}, x_{ij}\}$, j =1, 2,..., n_i, for each of the i = 1, 2, ..., K subjects, where y_{ij} is the response measure and x_{ij} is a p x 1 vector of covariates. The mean $\mu_{ij} = E(y_{ij} | x_{ij})$ is related to x_{ij} by $\eta(\mu_{ij}(\beta)) = x_{ij}^T\beta$, where η is a known link function and β is a p x 1 vector of unknown regression coefficients.

Liang and Zeger (1986) proposed a sandwich estimator to estimate the covariance matrix of the regression coefficient estimators $\hat{\beta}$:

$$\left(\sum_{i=1}^{K} \boldsymbol{D_i^T V_i^{-1}} \mathbf{D_i}\right)^{-1} \left(\sum_{i=1}^{K} \boldsymbol{D_i^T V_i^{-1}}\, cov(\boldsymbol{y_i}) \boldsymbol{V_i^{-1}} \mathbf{D_i}\right) \left(\sum_{i=1}^{K} \boldsymbol{D_i^T V_i^{-1}} \mathbf{D_i}\right)^{-1}, \quad \textbf{(2.4)}$$

, where $(\mathbf{y_i} - \boldsymbol{\mu_i})\,(\mathbf{y_i} - \boldsymbol{\mu_i})^{T}$ is typically used to estimate cov($\mathbf{y_i}$) and the reader can find detailed definitions of the above parameters in Liang and Zeger (1986).

2.5 Bias-corrected GEE Method

It is well known that the residual estimator of cov ($\mathbf{y_i}$) tend to be too small for the small-sample correlated responses. Thus to reduce the bias of the residual estimator $\mathbf{r_i r_i}^{T}$ and improve the performance of GEE test on small-sample correlated responses, Mancl et al (2001) proposed an alternative robust covariance estimator for $\hat{\beta}$, which is so-called bias-corrected robust covariance estimator. Specifically, the expected value of the residual estimator $\mathbf{r_i r_i}^{T}$ was approximated by multiplying the same matrix $(\boldsymbol{I_i} - \boldsymbol{H_{ii}})^{-1}$ on both sides, where $\boldsymbol{H}_{ii}$ is an expression for the leverage of the *i*th subject (Preisser & Qaqish, 1996) and $\boldsymbol{I}_i$ is an identity matrix of the same dimension as $\boldsymbol{H}_{ii.}$

The following equation gives the bias-corrected covariance estimator:

$$\boldsymbol{Var}^{(\boldsymbol{\beta})}_{\boldsymbol{bias-corrected}} = \boldsymbol{F}^{-1} \left\{ \sum_{i=1}^{K} \boldsymbol{D_i^T V_i^{-1}} (\boldsymbol{I_i} - \boldsymbol{H_{ii}})^{-1} \mathbf{r_i r_i^T} (\boldsymbol{I_i} - \boldsymbol{H_{ii}})^{-1} \boldsymbol{V_i^{-1}} \mathbf{D_i} \right\} \boldsymbol{F}^{-1} \quad \textbf{(2.5)}$$

, where $\boldsymbol{F} = \sum_{i=1}^{K} \boldsymbol{D_i^T V_i^{-1}} \mathbf{D_i}$,

$$\boldsymbol{H}_{ii} = \mathbf{D_i} \left(\sum_{l=1}^{K} \boldsymbol{D_i^T V_i^{-1}} \mathbf{D_i}\right)^{-1} \boldsymbol{D_i^T V_i^{-1}}.$$

2.6 Combined GEE Method

Some investigators (Pan et al, 2002; Fay et al 2001; Mancl et al 2001) proposed using an approximate t- or F- test that takes account of the variability of the sandwich estimator. Specifically, a Student's t- or F-distribution is used instead of the asymptotic normal (or chi-square) distribution to compute the statistical significance. Since our main interests of this study lie only on the treatment effect, we chose to compute the bias-corrected GEE test statistic and then use the F- distribution with 1 (df_1) and d (df_2) degrees of freedom.

The determination of the degree of freedom (df_2) had been rather arbitrary, but if it is desired to have the test size equal to or less than the specified nominal level, it is reasonable to use the number of clusters minus the number of coefficients (m) in the regression model as the denominator degree of freedom, which is m-3 in all the following simulation scenarios.

Section 3: Simulation Studies

3.1 Model and Simulation Scenarios

A standard "log" link was used in the regression model: $\log(\mu_{ij}) = \alpha + X_{ij}\,\beta + W_{ij}\,\gamma$, where μ_{ij} is the specified mean of count data (Y_{ij}), α is the intercept, X_{ij} is the binary indicator for treatment groups (0 or 1) for the *j*th subject in the *i*th community (but note that for a

GRT the treatment is the same for each subject in the same community), β is the treatment effect, W_{ij} is an individual level covariate adjusted in the model and γ is the covariate effect. In addition, the total number of clusters, m, the correlation within each cluster ρ (we assume equal intra-class correlation for all the clusters) and an overdispersion parameter φ were also specified ($\varphi=\text{var}(Y_i)/\mu_{Yi}$). Various values (ranging from 2 to 15) were examined for the over-dispersion parameter **φ** and they all gave similar results for the test size and the power. Thus a value of 4.0 was chosen for all the simulation scenarios, which was closest to the Pearson dispersion estimate of data in a group randomized dental trial that motivated this work (Harrison et al., 2010).
In total, 81 different scenarios were simulated based on combinations of the above parameters: (α, γ, ρ, m), with 3 values for each parameter. The values of the regression parameters specified in the simulations were chosen to correspond with the dental trial, but alternatives were allowed with **α**=0.2/1.3/3.0 and **γ**=0/0.05/0.10. The values of correlations were chosen so that the correlation **ρ** between subjects within each cluster was 0/0.05/0.3. All simulation sets were run with the exchangeable correlation. We selected values for the parameters that were reasonable based on scientific rationale of dental clinical studies. Lastly, 3 different numbers of clusters were considered with **m**=10/20/40. For each data configuration, 2000 simulations were generated and simulations were performed in the statistical package R 2.11.1.

3.2 Data Generation

The correlated count data were simulated by using a lognormal-Poisson hierarchy (L-P method) by Madsen (Madsen & Dalthorp, 2007). The L-P method is a simple and fast way to simulate count-valued random n-vectors $\boldsymbol{Y}$ with specified mean and correlation structure. A vector of correlated normal $\boldsymbol{Z}$ is generated and transformed to a vector of lognormals $\boldsymbol{X}$. Then, $\boldsymbol{Y}$ is generated as conditionally independent Poissons with means $\boldsymbol{X_i}$.
We simulated GRTs of 10, 20, 40 clusters. Each GRT consisted of equal number of control clusters and intervention (treated) clusters with equal numbers of subjects, whose outcomes were correlated count data. For all the simulation scenarios, the total number of all subjects was forced to be 400 (i.e., the number of subjects per cluster was 40, 20, or 10 corresponding to number of clusters 10, 20, or 40, repectively).

3.3 Results from the Simulation Study

3.3.1 Simulated Test Sizes

We first examined the test sizes and power of the permutation tests based upon 200, 500, 1000, 2000 and 10,000 random permutations. All these significance tests examined were applied to identical simulated data sets, thus any difference in size or power estimates seen from test to test is free of variations due to simulation. Simulation results showed that test sizes and power for all 81 simulation scenarios are similar for the five settings and the largest differences between 2000 and 10,000 random permutations were 0.002 and 0.005 for the test sizes and power, respectively (data not shown). The results thus suggest that 2000 random permutations are feasible and accurate enough. Such an approximation is also discussed in Dwass (1957) and Braun & Feng (2001).

We then investigated the performances of the six tests under 81 different simulation scenarios. **Figure 1** summarizes the test sizes of the six tests. As expected, the Wald test in GEE (G1) based on a sandwich standard error estimate is too liberal, a result also supported by the simulation study of Emrich & Piedmonte (1992). Since the actual test size has a 95% confidence interval of (0.040, 0.060) based upon an estimate of 0.05, estimates of test size that are outside of (0.040, 0.060) indicate that the test size is not

equal to 0.05. As a result, the bias-corrected GEE method (G2) did reasonably well in maintaining the nominal test size in most of the scenarios except that it gave conservative results when the correlation ρ is big and the total cluster number m is small. However, the bias-corrected GEE covariance estimator combined with F-distribution (G3) appears to overcorrect the sandwich covariance estimator in most cases, and this overcorrection may be the source of the conservativeness of the test sizes. Satisfyingly, none of the three permutation tests have sizes significantly larger than 0.05 for all scenarios. And it is worthy to note that the cluster-size-weighted and un-weighted permutation tests gave the same results (denoted by P1) in these balanced cases (of equal cluster sizes) and the correctly-weighted permutation test (denoted P2) did equally well. Additionally, the test sizes remain similar for all the six tests when the covariate effect (γ), correlation (ρ) and intercepts (α) change their values. This suggests that the performance of the permutation test and other tests remains stable and are not influenced by the covariate effect, intra-cluster correlation or the intercept. But, the bias-corrected method and the combined method appear to give more conservative test sizes when there are larger covariate effects and the total cluster numbers are small.

3.3.2 Simulated Power

Figure 2 summarizes the power of the six tests with moderate treatment effect of $\beta = 0.2$. **Figure 3** summarizes the power of the six tests with larger treatment effect of 0.3. As shown in both Figure 2 and Figure 3, the power goes up with the increase of the cluster number m, which suggests that higher power would be achieved by having more clusters and the cluster number is more influential than the cluster size. The results also show that the power is higher for larger covariate effect (γ). However, it is important to keep in mind that a larger covariate effect comes with a larger mean and thus the power of testing a specific value of treatment effect for larger means is going to be higher. When we compare the six tests in the same simulation scenario, we found that the powers of the five adjusted methods are close to each other and are slightly lower than the un-adjusted GEE method. Although GEE test appears to be more powerful, this is due to its liberal size. Compared with the permutation tests, the bias-corrected GEE combined with the F-distribution test (G3) has lower power and is actually the least powerful test among the six tests. Considering its conservative property and the lower power, we would not suggest using the bias-corrected GEE combined with F-test for GRTs with small cluster numbers ($m<20$).

Comparing Figure 2 with Figure 3, it is noted that the power goes up when the treatment effect is larger (bigger β) for all the six tests. We also compared the 5 adjusted methods only considering the tests with valid type I errors. The average power comparing cluster-size-weighted permutation test with un-weighted permutation, correctly-weighted permutation, bias-corrected GEE and bias-corrected GEE with F test are: 0.776 vs. 0.776 vs. 0.794 vs. 0.787 vs. 0.765 for treatment effect of 0.2. As for a larger treatment effect of 0.3, the average power comparing cluster-size-weighted permutation test with un-weighted permutation, correctly-weighted permutation, bias-corrected GEE and bias-corrected GEE with F test are: 0.895 vs. 0.895 vs. 0.905 vs. 0.901 vs. 0.887. Therefore, the correctly weighted permutation test yields slightly higher power than the other tests in the situations examined in this study.

In summary, the randomized cluster number has more influence on the test power than the cluster size. Within a certain amount of resource and study costs, we suggest the

researchers to try to increase the total cluster number as much as possible, rather than spend more effort on increasing the cluster sizes.

Section 4: Discussion

In our simulation studies, we found that all three permutation tests have very close to nominal test sizes and are generally performing better than the bias-corrected GEE method (G2). They gave very similar power for almost all the simulation scenarios. Thus in general, we will recommend either using one of the three permutation tests or using the bias-corrected GEE method when the cluster size is equal and the number of cluster is the same in each arm. However, the bias-corrected GEE method should not be used in situations where the covariate effect or/and the intra-class correlation is large, since conservative test sizes would be produced. We also examined a different setting of small GRTs ($m<10$) with unequal numbers of clusters in control and treatment groups (data not shown). Based on the comparison results with regard to test sizes and power, we highly recommend having at least 8 total clusters in designing a group randomized trial and maintain the same number of clusters in each arm of the trial.

As far as the authors are aware, this paper is the first study that compares the use of un-weighted and weighted permutation tests with the small-sample adjustment GEE methods to analyze group randomized trials with correlated count-valued responses. Previous studies only addressed continuous or binary responses by applying either permutation test (s) (Gail et al, 1992; Peterson et al, 2000; Braun, 1999) or small-sample adjustment GEE method (s) (Mancl et al, 2001; Fay et al, 2001; Pan et al, 2002) in the context of GRTs or other settings involving clustered data. It is not uncommon to see count outcomes in group randomized trials (i.e. number of cavities, blood cell counts, hospitalization days, etc) and sometimes results from Gaussian or binary distribution might not carry over to skewed distributions.. Also, we were able to compare the performances of six tests in the simulation studies, including the un-weighted, cluster-size-weighted and correctly-weighted permutation tests, two small-sample adjustment GEE methods and the unadjusted GEE method as the reference. Thus these findings would deliver much more comprehensive understanding in terms of which approach works best in certain situation.

There are several limitations in this study. Firstly, we used a lognormal-Poisson hierarchy (L-P method) for simulating count-valued random n-vectors Y with specified mean and correlation structure. One strength of this technique is its simplicity; once the desired moments of the simulated vector Y have been determined, the generation of the normal vector (Z) is nearly instantaneous for moderate n. However, the L-P method fails in situations where underdispersed random variables or strongly correlated random variables with similar means need to be simulated. As mentioned in Madsen's paper (Madsen et al, 2007), for simulating samples from populations with small means and high correlations, an alternative method, overlapping sums of independent variables is available to address some of the shortcomings of the L-P method. Another limitation of this study lies in estimation of covariate effects. We included a covariate effect (γ) in all the simulation scenarios and did not examine cases where there is no covariate effect. However, whether or not bias in the covariate estimates will cause very large intervention effects to be detected with less power is unknown and underestimated covariate effects will directly impact power and lead to an inefficient permutation test. In fact, it may be more powerful to not adjust for covariates at all. Therefore, it would be good to compare the power of these tests which adjust for small or larger covariate effect with the power of

tests that do not adjust for any covariate. Lastly, findings and conclusions of our study are limited to the simulation scenarios that were used in this study and might not be generalizable to other scenarios.

References

Arnold, H.J. (1964). Permutation support for multivariate techniques. Biometrika, 51: 65-70.

Braun, T. M. (1999). PhD dissertation: Optimal analysis of group randomized trials with permutation tests.

Braun, T. M. and Feng, Z. (2001). Optimal permutation tests for the analysis of group randomized trials. Journal of the American Statistical Association (2001), Vol.96, No.456.

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

Brookmeyer, R., and Chen, Y.-Q. (1998), "Person-Time Analysis of Paired Community Intervention Trials When the Number of Communities Is Small," Statistics in Medicine, 17, 2121–2132.

Cornfield, J. (1978). Randomization by group: a formal analysis. American Journal of Epidemiology 108, 100-102.

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

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. Annals of Mathematical Statistics 28, 181-187.

Edgington, E. S. (1995). Randomization tests, 3rd edition. New York, Marel-Dekker.

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.P., Graubard, B.I., et al (1998). Conditional Logistic Regression with Sandwich Estimators: Application to a Meta-Analysis. Biometrics 54, 195-208.

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

Fisher, R. A. (1935). The design of experiment. Oliver and Boyd, Edinburgh.

Freedman, D., Lane, D. (1980). The empirical distribution of Fourier coefficients. Annal Statist; 8: 1244-1251.

Gail, M. H, Byar, D. P., Pechacek, T. F., and Corle, D. K. (1992). Aspects of statistical design for the community intervention trial for smoking cessation (COMMIT). Controlled Clinical Trials 13, 6-21.

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

Gunsolley, J.C. Getchell, C. and Chinchilli, V.M. (1995). "Small Sample Characteristics of Generalized Estimating Equations," Communications in Statistics, Part B - Simulation and Computation, V24, 869-878.

Harrison R., Veronneau J. and Leroux B. (2010). Design and implementation of a dental caries prevention trial in remote Canadian Aboriginal communities. Trials 2010, 11:54.

Hubert, L. J., Levin, J.R. (1976). General statistical framework for assessing categorical clustering in free recall. Psych Bull, 83:1072-1080.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experiments. Journal of Educational Statist, 5: 253-260.

Li, B. (1993). A deviance function for the quasi-likelihood method. Biometrika 80, 741-753.
Liang, K. Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 13-22.
Lin, D.Y. and Wei, L. J. (1989). The robust inference for the cox proportional hazards model. Journal of the American Statistical Association 84, 1074-1078.
Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1990). Using the jackknife to estimate the variance of regression estimators from repeated measures studies. Communications in Statistics, Part A. Theory and Methods 19, 821-845.
Lipsitz, S. R., Fitzmaurice, G. M., Orav, E. J., and Laird, N. M. (1994). Performance of generalized estimating equations in practical situations. Biometrics 50, 270- 278.
Madsen, L. and Dalthorp, D. (2007). Simulating correlated count data. Environ Ecol Stat (2007) 14:129-148.
Mancl, L. A., and DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. Biometrics 57, 126-134, 2001.
Mielke, P.W.Jr. (1986). Non-metric statistical analysis: Some metric alernatives. Journal of Statist Plan Inference, 13: 377-387.
Pan W. and Melanie M. Wall (2002). Small sample adjustments in using the sandwich variance estimator in generalized estimating equations. Statist. Med. 2002; 21:1429–1441.
Paik, M. C. (1988). Repeated measurement analysis for nonnormal data in small samples. Communications in Statistics: Simulations 17, 1155-1171.
Peterson AV, Kealey KA, Mann SL, Marek PM, Sarason IG (2000): Hutchinson Smoking Prevention Project: long-term randomized trial in school based tobacco use prevention--results on smoking. Natl Cancer Inst 2000, 92(24):1979-1991.
Pitman, E. J. G. (1938). Significance tests which may be applied to samples from any population. Part III, the analysis of variance test. Biometrika, 29, 322-335.
Preisser, J. S. and Qaqish, B. F. (1996). Deletion diagnostics for generalized estimating equations. Biometrika 83, 551-562.
Qu, Y., Piedmonte, M. R., and Williams, G. W. (1994). Small sample validity of latent variable models for correlated binary data. Communications in Statistics: Simulations 23, 243-269.
Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. Biometrika, 40, 719-727.
Small, D. S., Have, T.R.Ten, Rosenbaum, P.R. (2008). Randomization inference in a group-randomized trial of treatments for depression. Journal of Amer Statistical Assoc. V103, Issue: 481, 271-279.
Veronneau, J. (2002). Report on Cree dental health in Eeyou Istchee (Northern Quebec). Cree Board of Health and Social Service of James Bay.
Walsh, M. W., Hilton, J. F., Masouredis, C. M., Gee, L., Chesney, M.A. and Ernster, V.L. (1999). Smokeless tobacco cessation intervention for college athletes: results after one year. American Journal of Public Health 89, 228-234.
Zeger, S. L., Liang, K-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. Biometrics 44, 1049-1060.

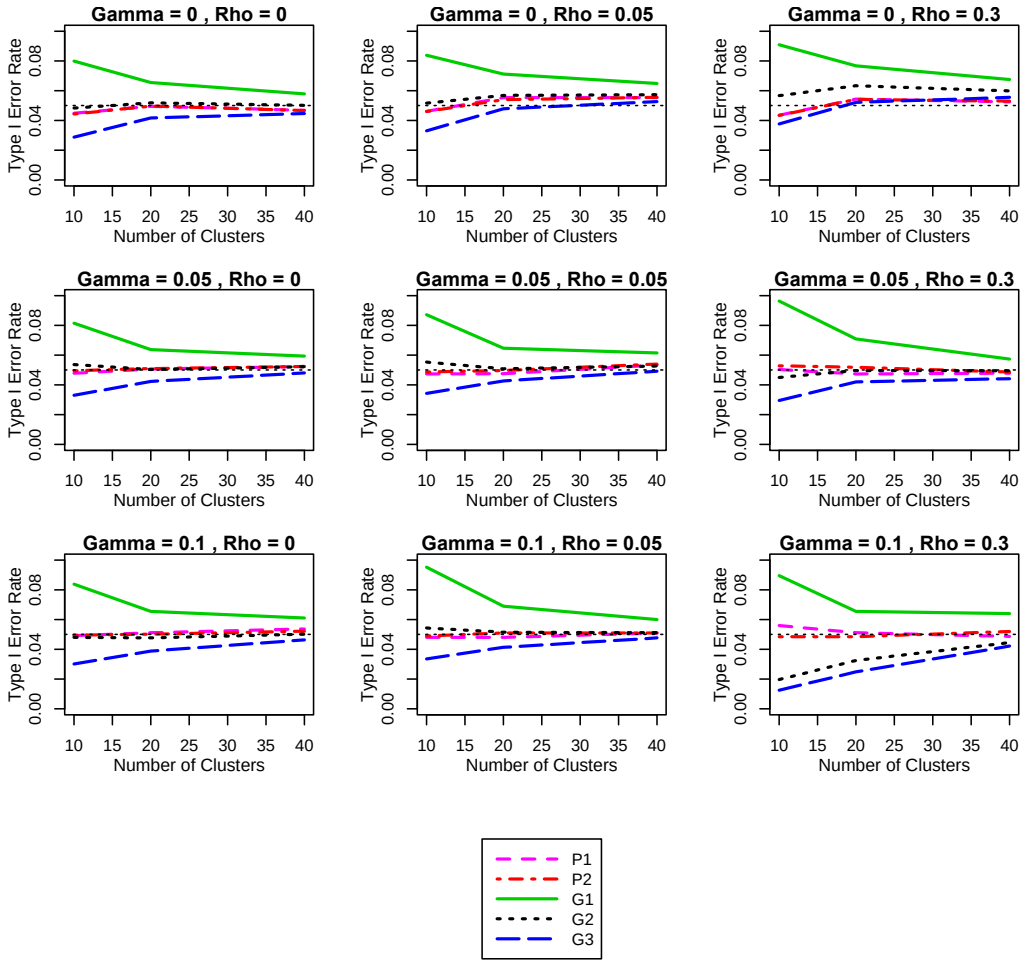

Figure 1: Simulated test sizes for 5 tests as a function of number of clusters (m) with m/2 clusters per treatment group and a total of 400 subjects. Results are based upon 2000 simulations of log-linear models as described in section 3.1 with overdispersion parameter φ=4.0. Results for three different values of the intercept are averaged and plotted. The tests considered are: P1= cluster-size-weighted permutation test (equivalent to the un-weighted permutation test in this setting); P2 = correctly-weighted permutation test; G1= unadjusted GEE; G2= bias-corrected adjusted GEE; G3= bias-corrected GEE combined with F test.

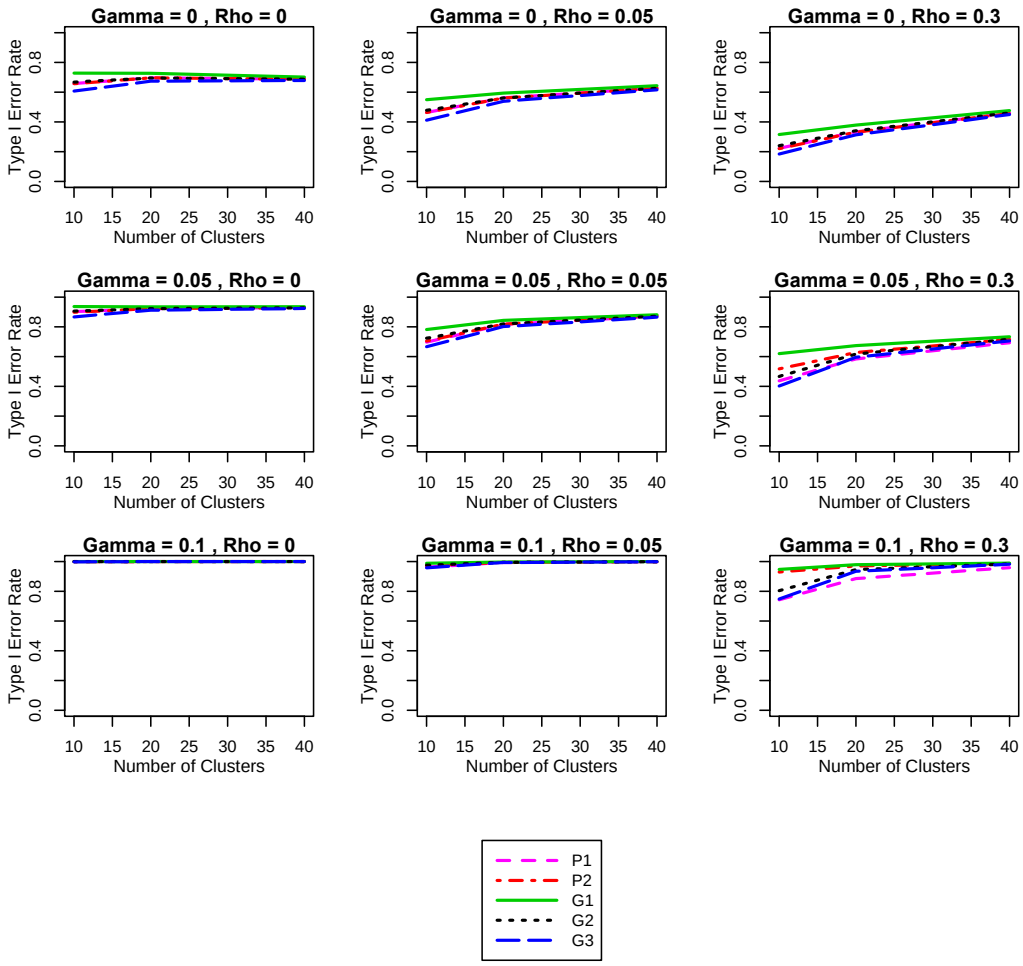

Figure 2: Simulated power for 5 tests as a function of number of clusters (m) with m/2 clusters per treatment group and a total of 400 subjects. Results are based upon 2000 simulations of log-linear models as described in section 3.1 with overdispersion parameter φ=4.0 and treatment effect 0.2. Results for three different values of the intercept are averaged and plotted. The tests considered are: P1= cluster-size-weighted permutation test (equivalent to the un-weighted permutation test in this setting); P2 = correctly-weighted permutation test; G1= unadjusted GEE; G2= bias-corrected adjusted GEE; G3= bias-corrected GEE combined with F test.

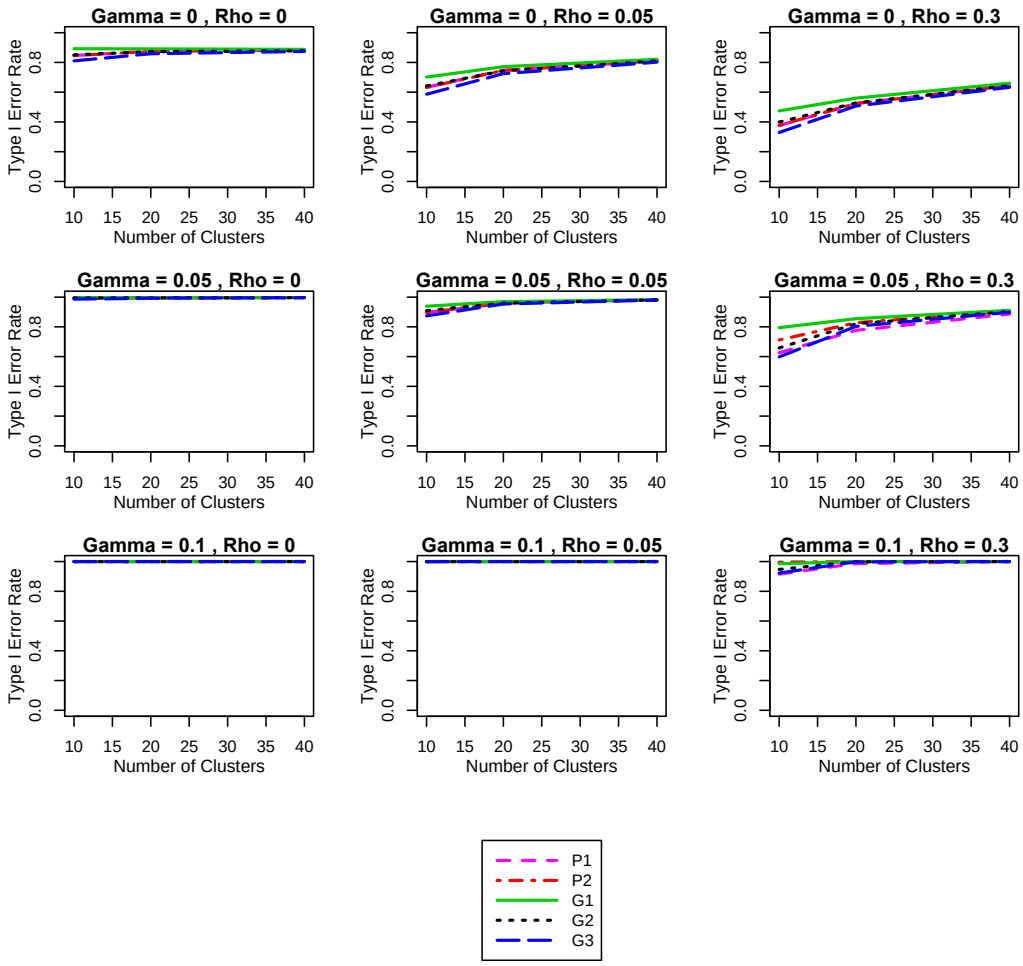

Figure 3: Simulated power for 5 tests as a function of number of clusters (m) with m/2 clusters per treatment group and a total of 400 subjects. Results are based upon 2000 simulations of log-linear models as described in section 3.1 with overdispersion parameter φ=4.0 and treatment effect 0.3. Results for three different values of the intercept are averaged and plotted. The tests considered are: P1= cluster-size-weighted permutation test (equivalent to the un-weighted permutation test in this setting); P2 = correctly-weighted permutation test; G1= unadjusted GEE; G2= bias-corrected adjusted GEE; G3= bias-corrected GEE combined with F test.