Published in final edited form as:
Stat Med. 2015 January 30; 34(2): 281–296. doi:10.1002/sim.6344.

Small Sample Performance of Bias-corrected Sandwich Estimators for Cluster-Randomized Trials with Binary Outcomes

Peng Li* and **David T. Redden**
Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294

SUMMARY

The sandwich estimator in generalized estimating equations (GEE) approach underestimates the true variance in small samples and consequently results in inflated type I error rates in hypothesis testing. This fact limits the application of the GEE in cluster-randomized trials (CRTs) with few clusters. Under various CRT scenarios with correlated binary outcomes, we evaluate the small sample properties of the GEE Wald tests using bias-corrected sandwich estimators. Our results suggest that the GEE Wald *z* test should be avoided in the analyses of CRTs with few clusters even when bias-corrected sandwich estimators are used. With *t*-distribution approximation, the Kauermann and Carroll (KC)-correction can keep the test size to nominal levels even when the number of clusters is as low as 10, and is robust to the moderate variation of the cluster sizes. However, in cases with large variations in cluster sizes, the Fay and Graubard (FG)-correction should be used instead. Furthermore, we derive a formula to calculate the power and minimum total number of clusters one needs using the *t* test and KC-correction for the CRTs with binary outcomes. The power levels as predicted by the proposed formula agree well with the empirical powers from the simulations. The proposed methods are illustrated using real CRT data. We conclude that with appropriate control of type I error rates under small sample sizes, we recommend the use of GEE approach in CRTs with binary outcomes due to fewer assumptions and robustness to the misspecification of the covariance structure.

Keywords

generalized estimating equations (GEE); correlated data; type I error rates; power; sample size

1. Introduction

Cluster-randomized trials (CRTs), also called group-randomized trials, are widely used in the evaluation of interventions in health services research [1-3]. In CRT design, the identifiable clusters, rather than individuals, are randomly assigned to different intervention conditions so that the units of observation are the individuals nested within both their condition and their cluster. A key property of CRTs is that the inferences are intended to

* Correspondence to: Peng Li, Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294 Tel: 205 975 9188 pli@uab.edu.

Competing interests
The authors declare that they have no competing interests.

apply at the individual level, even though the randomization is at cluster level. Since the clusters are formed not at random but rather through some physical, geographic, or other connection among their members, there is an expectation for a positive intraclass correlation (ICC, noted by ρ) among observations in the same cluster. Even though for most CRTs the ICCs are low (0.001 to 0.05) [4, 5], any statistical tests ignoring the non-independence of participants within clusters will underestimate the variances of the intervention effects and inflate the type I error rates [5]. The design effect (DE) is the ratio of the variance of an outcome measure accounting for cluster correlation over the variance of the outcome measure under person-level randomization. For clusters of equal size m, it can be shown that $DE=1+(m-1)*\rho$ for two-level CRTs [1].

The generalized estimating equations (GEE) method, developed by Liang and Zeger [6] in the context of longitudinal studies, has proven to be very popular for the analysis of correlated data. Given the number of independent clusters is large, for example greater than 40 in CRTs, the GEE approach has several desirable properties. The GEE approach does not require distributional assumptions because the estimation depends only on correctly specifying the relationship between the marginal mean and covariates through a link function, not on the entire joint distribution of observed data and random effects [6]. Under mild regularity conditions [6], the resulting regression coefficient estimator is consistent and asymptotically normal and its variance-covariance can be estimated by the sandwich estimator, which is robust to the misspecification of the covariance structure of the response [6]. However, the sandwich estimator is biased downward when the number of clusters is not large enough, for example below 40 in CRTs [2, 3, 7], and this problem becomes more severe as the number of clusters becomes smaller [2, 8]. Unfortunately, most CRTs do not include 40 clusters, and there is a median of 21 clusters in a review of a random sample of 300 published CRTs [9].

Due to the small sample bias of the sandwich estimator, some bias-corrected sandwich estimators have been proposed to improve the small sample performance of GEE [8, 10-12]. In the following, we briefly review the GEE approach, the sandwich estimator, explain its poor performance for small number of clusters and review five bias-corrected sandwich estimators which are proposed to decrease the bias of original sandwich estimator given few clusters.

Suppose that a dataset from a CRT consists of K clusters and each of the clusters i (i = 1,2, ... , K) has m_i observations with response Y_{ij} and a p-dimensional covariate vector X_{ij}, i = 1,2, ... , K and j = 1,2, ... , m_i. Denote $Y_i=(Y_{i1},\ldots,Y_{im_i})'$, $X_i=(X_{i1},\ldots,X_{im_i})'$. It is assumed that Y_i and $Y_{i'}$ are independent for any $i \neq i'$. The marginal model specifies a relationship between the marginal mean $E(Y_{ij}|X_{ij})=\mu_{ij}$ and the covariate X_{ij} through a generalized model, $g(\mu_{ij})=X_{ij}\beta$, where β is an unknown p-vector of regression coefficients to be estimated, and $g(.)$ is a known link function. The marginal variance is $Var(Y_{ij})=\nu(\mu_{ij})\varphi$, where ν is a known function of μ_{ij}, and φ is an unknown scale parameter which may need to be estimated. The within-cluster correlation matrix $Corr(Y_i)$ is R_0 whose structure is in general unknown. An attractive point of the GEE is that a consistent $\hat{\beta}$ can be obtained without the requirement of specifying R_0 correctly [6]. Let $V_i=\varphi A_i^{1/2} R_w(\alpha) A_i^{1/2}$ define the working covariance matrix for Y_i, where $A_i=\mathrm{diag}[\nu(\mu_{i1}),\ldots,\nu(\mu_{im_i})]$ and $R_w(\alpha)$ is a

working correlation matrix for Y_i. By Liang and Zeger [6], the GEE estimates, $\hat{\beta}$, are given by the solution to the estimating equations

$$\sum_{i=1}^{K} D_i' V_i^{-1} (Y_i - \mu_i) = 0.$$

where $D_i = \frac{\delta \mu_i}{\delta \beta'}$. The variance-covariance of $\hat{\beta}$ can be consistently estimated by

$$V = \Omega \left(\sum_{i=1}^{K} D_i' V_i^{-1} r_i r_i' V_i^{-1} D_i \right) \Omega,$$

where $\Omega = \left(\sum_{i=1}^{K} D_i' V_i^{-1} D_i \right)^{-1}$ and $r_i r_i' = (Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$. This variance-covariance estimator is commonly called as robust sandwich estimator [6].

Because the fitted values, $\hat{\mu}_i$, tend to be closer to the observed values, Y_i, than the true values, $(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)'$ underestimates the $Cov(Y_i)$. Therefore, the robust sandwich estimator will underestimate the covariance of $\hat{\beta}$, especially when the number of clusters is small. Since the Wald test statistics, which asymptotically follow standard normal distribution, are commonly used in GEE for the hypothesis testing, the underestimation of the covariance of $\hat{\beta}$ causes inflated Type I errors. Due to the small sample bias of sandwich estimator, some biascorrected sandwich estimators have been proposed to improve the small sample performance of GEE.

DF-corrected sandwich estimator

The simplest adjustment makes a degrees-of-freedom (DF) correction [13] that inflates variance by multiplying the sandwich estimator by $K/(K-p)$, where K is the number of clusters and p is the number of regression parameters. That is

$$V_{DF}^{*} = \left(\frac{K}{K-p} \right) V.$$

KC-corrected sandwich estimator

Kauermann and Carroll [10] defined

$$V_{KC}^{*} = \Omega \left(\sum_{i=1}^{K} D_i' V_i^{-1} (I_i - H_i)^{-1/2} r_i r_i' \left(I_i - H_i' \right)^{-1/2} V_i^{-1} D_i \right) \Omega,$$

where I_i is the identity matrix with $m_i \times m_i$ dimension and matrix H_i is an expression for the leverage of the i^{th} cluster and $H_i = D_i \Omega D_i' V_i^{-1}$ [8, 10, 14]. Because H_i is between 0 and 1, V_{KC}^{*} is expected to give larger standard errors than V.

MD-corrected sandwich estimator

Mancl and DeRouen [8] proposed reducing the bias of the sandwich estimator, by defining

$$V^*_{Md}=\Omega\left(\sum_{i=1}^{K}D_i'V_i^{-1}(I_i-H_i)^{-1}r_ir_i'\left(I_i-H_i'\right)^{-1}V_i^{-1}D_i\right)\Omega,$$

which further inflates V^*_{KC}.

FG-corrected sandwich estimator

Fay and Graubard [12] attempted to correct the bias by setting a scale factor to the working variance in the sandwich estimator, so that

$$V^*_{FG}=\Omega\left(\sum_{i=1}^{K}c_iD_i'V_i^{-1}r_ir_i'V_i^{-1}D_ic_i\right)\Omega,$$

where $c_i=\{1-\min(b,\{Q\}_{jj})\}^{-1/2}$ and $Q=D_i'V_i^{-1}D_i\Omega$. $b<1$ is a constant bound defined by the user to prevent extreme adjustments when the jj^{th} element of Q is very close to 1. Fay and Graubard's results suggest that the bound of b is rarely reached and can be arbitrarily set (0.75 by default) without affecting the results [12].

MBN-corrected sandwich estimator

Morel, Bokossa and Neerchal [11] suggested a bias correction of the sandwich estimator that rested on an additive correction of the residual cross-products and a sample size correction.

$$V^*_{MBN}=\Omega\left(\sum_{i=1}^{K}D_i'V_i^{-1}\left(cr_ir_i'+\delta_m\phi V_i\right)V_i^{-1}D_i\right)\Omega,$$

where $c=\frac{f-1}{f-p}\times\frac{K}{K-1}$, $f=\sum_{i=1}^{K}m_i$ is the total observations;

$\delta_m=\begin{cases}\frac{p}{K-p} & if\quad K>(d+1)\,p\\ \frac{1}{d} & otherwise\end{cases}$; $\phi=max\left\{l,trace\left[\Omega\left(\sum_{i=1}^{K}D_i'V_i^{-1}r_ir_i'V_i^{-1}D_i\right)\right]/p\right\}$, $0\le l\le 1$. the term $\delta_m\varphi V_i$ is added for the small sample correction, in which φ is the estimate of design effect [15] and δ_m is a function, not involving parameter estimates, of order K^{-1}. It should be noted that l is the lower bound of the design effect. Morel et al. suggested that r was set to be 1, and the upper bound on δ_m was arbitrarily set to be 0.5 ($d=2$) , which rarely came into play in practice [11]. The performance of the MBN-correction may depend on the choice of d and l. Within this paper, we followed the recommendations of the authors and set $d=2$ and $l=1$. For large K, the term $\delta_m\varphi V_i$ vanishes and hence, the V^*_{MBN} gives an approximation of the original sandwich estimator. The sample size correction factor c should provide additional reduction to the small sample bias associated with the empirical covariance estimate.

Other bias-corrected sandwich estimators include Jackknife estimator [16], which is very close to V^*_{MD} in simulations [8]. Pan et al. [17] and Wang et al. [18] also propose bias corrected sandwich estimators but their modifications assume a common correlation structure across all clusters and this assumption may limit their application in CRTs with the heterogeneous clusters.

Although many simulation studies [8, 10-12, 19] suggest that the Wald tests with above bias-corrected sandwich estimators show type I errors closer to nominal levels than the Wald test with original sandwich estimator, their results are hard to apply directly to the analyses of CRTs. This difficulty lies in the fact that: i) the corrections are shown to be applicable only in clusters with small size and high ICC, which is typical in longitudinal data but not CRT data; ii) furthermore, most of the simulations from the above studies only consider equal cluster sizes, while cluster sizes may vary greatly in real practices [20, 21]; iii) finally, the situation of an extremely small number of clusters, such as fewer than 10 clusters in each intervention condition, is not well explored in above studies. Lu et al. [19] compared the performances of V^*_{KC} and V^*_{MD} under some CRT scenarios, but only under balanced design. More importantly, the inferences from the asymptotic *z*-test in Lu et al. [19] is not optimal for small cluster numbers, even when the bias-corrected sandwich estimators are used. The *t*-distribution approximation of the Wald statistics may perform better.

The first objective of this study is to evaluate the small sample performance of GEE involving the bias-corrected sandwich estimators including V^*_{DF}, V^*_{KC}, V^*_{MD}, V^*_{FG} and V^*_{MBN} under CRT scenarios with few clusters, especially the effects of the unbalanced cluster sizes on the small sample performance of GEE. Both the *z*-distribution and the *t*-distribution approximations of Wald statistics are used for the hypothesis testing. The second objective of this study is to derive a formula to calculate the power and minimum total cluster number needed in CRTs given that the bias-corrected sandwich estimators are required in the analysis. The current GEE-based power and sample size are calculated with the original sandwich estimator and are only applicable to the large number of clusters (>40) in CRTs [22-24]. It can be expected that a sample size calculation based on the original sandwich estimator followed by an analysis with a bias-corrected sandwich estimator will result in the loss of precision and hence power if only small number of clusters are available when designing a CRT. Once a bias-corrected sandwich estimator is chosen in the GEE analysis of CRT to preserve the Type I errors, intuitively and necessarily, we can use it for a better power and sample size calculation for the CRTs with few clusters. We restrict our study on correlated binary data; however, we expect the similar results on continuous data but definitely we need further confirmation in the future study. The type I error rates and empirical powers will be calculated based on Monte Carlo simulations and the valid tests with type I error rates close to nominal are illustrated using the real data from a trial performed to investigate the effect of an intervention on breast screening uptake among women in Central and East London [4].

2. Test Sizes of Small Sample GEE Wald Statistics on Correlated Binary Outcomes

2.1 Data simulation

Correlated binary responses are generated using a Beta-binomial method [25]. The Beta-binomial distribution is derived as a mixture distribution in which the proportion of event in a cluster is a random draw from a beta distribution with parameters a and b. The marginal proportion of event in a cluster is defined as $\mu=\frac{a}{a+b}$ and the ICC can be shown to be $\frac{1}{1+a+b}$. By selecting a and b, we can simulate the marginal proportion and ICC for each cluster. We mimic the CRTs containing 10, 20, or 30 clusters (K) with 10, 20, 50, 100, or 150 observations in average per cluster ($\overline{m}$). The exact number of observations, m_i, for each cluster $i = 1,..., K$, is randomly drawn and rounded from normal distributions with the mean equal to $\overline{m}$ in and variance equal to σ^2. The variation of cluster sizes can be measured by the coefficient of variation (cv), which is the ratio of standard deviation of the cluster sizes over the mean of the cluster sizes. So, we let $\sigma^2 = \overline{m}^2(cv)^2$. In our simulations, cv is at the range of 0 to 1. To avoid the impossible situation that the number of observations in a cluster is negative or zero, we bounded the smallest cluster size to 1. Considering the potential departure of the mean and variance caused by this process, we carefully check the simulated datasets and calculate the true mean, variance of the cluster size, as well as the coefficient of variation, and find that the departure is minor and will not affect our conclusion. The intraclass correlation (ρ) among observations within a cluster is set as 0.001, 0.01 or 0.05, which may reflect the actual ICC in realistic practices [5]. In each scenario, 3000 independent replicates are generated. Due to the small value of K, it can be expected that the mean and variance of cluster sizes drawn from a normal distribution may be slightly different from the expected values, which are part of the simulation errors. These errors can be reduced by the large number of simulation replicates (3000 in our simulation study). For simplicity but without the loss of generalizability, we assume two intervention arms containing equal clusters and no covariates are included. A logistic regression model is used for the marginal mean of y_{ij}

$$logit\left(\mu_{ij}|X_i\right)=\alpha+\beta * X_i$$

where X_i is vector containing cluster-level binary predictor indicating the intervention arms ($X_i = 0$ for control and $X_i = 1$ for active intervention), $i = 1,..., K$ and $j = 1, ..., m_i$. The marginal proportion is set as $\{\mu_i|X_i = 0\} = 0.25$. For the type I error estimation, we set $\beta = 0$ and test the null hypothesis H_0: $\beta = 0$. The simulated data are analyzed using various GEE approaches assuming a compound symmetric working correlation matrix. All simulations and analyses were conducted using SAS 9.3 (Cary, NC). The SAS code will be provided as request.

2.2 Observed type I error rates of GEE Wald tests in simulations

Both the z-distribution and the t-distribution approximations of Wald statistics are used for hypothesis testing in this study. It has been shown that using t-distribution can improve the performance of the GEE Wald test given small number of clusters [1, 8] and some

Satterthwaite-type degrees of freedom approximations have been proposed [12, 26, 27]. The simplest degrees of freedom approximation is based on the number of independent clusters [5, 8, 23]. In the case of two intervention arms, the degrees of freedom can be approximated as $K - 2$, where K is the independent cluster number. In our simulations, $K - 2$ degrees of freedom approximation is used. The performance of the sandwich covariance estimator and bias-corrected estimators is evaluated by computing the observed fraction of Wald statistics rejecting the null hypothesis when null hypothesis is true. At the nominal 0.05 level and 3000 simulations, we would expect the simulated type I error rate to be between 0.042 and 0.058 (95% confidence interval) and any procedure with type I error rate in this range will be considered as valid and efficient.

2.2.1 Wald z tests vs Wald t test—Most of the simulation studies only consider the equal cluster size, which mimic the most efficient CRT design; however, cluster sizes often vary in practice [20, 21]. It is intuitively obvious that the cluster size imbalance will affect the statistical inference in the way that the estimates from the smaller clusters will be less precise and the estimates from the larger clusters more precise. For the bias-corrected sandwich estimators, V^*_{KC} and V^*_{MD} take the cluster size into account in H_i matrix for the bias reduction; while V^*_{MBN} considers the cluster size in the φ factor. It is expected that the cluster size imbalance will affect the type I error rates of the Wald tests in GEE.

For small cluster size imbalance ($cv = 0.1$), the observed type I error rates or test size, for the Wald z tests or t tests involving different covariance estimators are shown in Table 1. In general, the intraclass correlation values and average cluster sizes have very little effects on type I error rates for both the Wald z tests and the Wald t tests; however, the Wald t tests outperform the Wald z tests with regard to the type I error rates closer to the nominal level. The observed type I error rates are greatly inflated when the sandwich estimator is used given cluster numbers fewer than 30, and the inflation becomes more severe as the cluster numbers decrease. The Wald t test decreases the inflation compared to asymptotic z test, but still not to the nominal level.

Compared to the original sandwich estimator, all the five bias-corrected sandwich estimators give smaller type I error rates, and the MD-corrected sandwich estimator is the most conservative. However, with the Wald z test, only the MD-corrected sandwich estimator has the type I error rates to the nominal level as long as the number of clusters greater than 20. These findings suggest that the asymptotical GEE Wald z test should be avoided in the analyses of CRTs with few clusters even when the bias-corrected sandwich estimators are used and the CRT has the almost balanced design.

The results from Table 1 clearly show that when the DF- or KC-corrected sandwich estimator is used, almost all of the observed type I error rates from GEE Wald t tests lie in the nominal range, regardless of cluster numbers from 10 to 30. However, under the almost balanced design, the Wald t test with MD-, FG- or MBN-corrected sandwich estimator may cause over control, especially when the cluster number is less than 20. The over control of type I error rate will lead the power loss if the null hypothesis is false.

2.2.2 Wald t test in unbalanced CRT designs—

As discussed before, cluster sizes often vary in practice and such imbalance may affect the statistical inference. Our simulations show that the GEE Wald *t* tests with the bias-corrected sandwich estimators maintain the type I error rate for the nearly balanced design; however, their performance needs to be evaluated in more general situations. The performances of the GEE Wald *t* tests with different bias-corrected sandwich estimators under different variation of cluster sizes are shown in Figure 1-5. The results suggest that the increased imbalance of cluster sizes inflates the type I error rate of the GEE Wald *t* tests; and the inflation differs when different bias-corrected sandwich estimators are used.

The Wald *t* test with the DF-corrected sandwich estimator only maintains the type I error rate for the small variation of cluster sizes ($cv < 0.2$) (Figure 1). The imbalance of the cluster sizes inflates the type I error rate quickly since the DF-corrected sandwich estimator does not consider the cluster size in its bias reduction and inflates the estimates of variance-covariance from different clusters with the same scale, $\frac{K}{K-p}$. Since the KC-corrected sandwich estimator places more weight on the larger clusters for the bias reduction, the Wald *t* test with the KC-corrected sandwich estimator is more robust to the variation of cluster sizes and can maintain the type I error rate for small to moderate variation of cluster sizes ($cv < 0.6$) (Figure 2). The Wald *t* test with the MD-corrected sandwich estimator is the most robust test to the variation of cluster sizes (Figure 3). However, it is too conservative, especially when the cluster number becomes very small. The Wald *t* test with the FG-corrected sandwich estimator tends to be conservative when the variation of cluster sizes is not large ($cv < 0.6$), especially for small cluster number (Figure 4). Interestingly, it maintains the type I error rate to nominal level for larger variation of cluster sizes ($cv > 0.6$), even when the cluster number is as low as 10. The Wald *t* test with the MBN-corrected sandwich estimator tends to be conservative when the variation of cluster sizes is small ($cv < 0.4$) and to be liberal when the variation of cluster sizes is large ($cv > 0.8$) (Figure 5).

2.3 Summary of small sample GEE Wald tests

Our simulations suggest that among the bias-corrected sandwich estimators, the GEE Wald *t* tests outperform the Wald *z* tests with regard to maintaining the type I error rates to nominal level; however, no single bias-corrected sandwich estimator is universally superior to the others when the Wald *t* test is used for the analyses of CRTs with few clusters and considerable variation of cluster sizes. Teerenstra et al. [23] recommended the Wald *t* test with KC-corrected sandwich estimator for the analyses of CRTs with few clusters, but it is only valid for the small to moderate variation of cluster sizes. In practice, when cluster number is small and cluster sizes vary, we suggest a rule of thumb that choosing the Wald *t* test with KC-corrected sandwich estimator when the coefficient of variation of cluster size is less than 0.6 and choosing the Wald *t* test with FG-corrected sandwich estimator, otherwise.

3. Statistical Power of Small Sample GEE for Correlated Binary Data

In a GEE model without covariates, the hypothesis of interest is to test H_0: $\beta = 0$ vs H_a: $\beta \neq 0$. The power estimation for a given $\beta = b > 0$ is given by

$$power=1-\mathbf{\Phi}\left(\mathbf{Z}_{\alpha/\mathbf{2}}-\frac{\mathrm{b}\sqrt{K}}{\sqrt{v_R}}\right) \quad (1)$$

where v_R is the (2,2)-element of KV, where K is the number of clusters and V is the assumed true covariance matrix of parameter coefficients [22, 24].

Assuming a two-armed CRT with K clusters, qK clusters are assigned to arm 1 and $(1-q)K$ clusters are assigned to arm 2. For the binary outcome, let π_1 and π_2 represent the true population proportion for each arm, respectively. Although in the real data, the number of participants in different clusters varies, in the design stage we may assume $m_i = m$, such that there are same number of participants for all the clusters. Whether the cluster size can be assumed as a constant at the design stage depends on the knowledge of the investigators. In many trials, only a fixed number of individuals in each cluster are planned to be in the trial, then the cluster can be assumed as a constant or nearly consistent [28]. The high variability of cluster size could happen and is not considered in this study. In addition, the structure of within-cluster variance-covariance matrix $Cov(Y_i)$ is in general unknown and different from the working covariance structure; however, for the power calculation before we collect the data, we have to assign a correlation matrix. The compound symmetry structure is a reasonable assignment in the framework of CRT. If we also allow the compound symmetry structure as the working covariance structure, based on the sandwich estimator and some matrix algebra [24], we have

$$v_R=\left(\frac{1+(m-1)\rho}{m}\right)\left[\frac{1}{q\pi_1(1-\pi_1)}+\frac{1}{(1-q)\pi_2(1-\pi_2)}\right].$$

If equal number of clusters is assigned to the control and intervention arms, we have

$$v_R=\left(\frac{1+(m-1)\rho}{m}\right)\left[\frac{2}{\pi_1(1-\pi_1)}+\frac{2}{\pi_2(1-\pi_2)}\right].$$

When we plug the v_R into the above power equation, we can calculate the power and consequently the sample size needed given α level and estimated proportions. Apparently, this power estimation only works under asymptotic conditions so that the v_R derived from the sandwich estimator is unbiased and the asymptotic Wald z test is used.

When a bias-corrected sandwich estimator and the Wald t test ($K-2$ degrees of freedom) are used in the data analyses, we propose a modified power estimation formula as:

$$power=1-T_{K-2}\left(t_{\alpha/2,K-2}-\frac{b\sqrt{K}}{\sqrt{v_R{}^*}}\right) \quad (2)$$

where $v_R{}^*$ is the (2,2)-element of KV^*, where V^* is the assumed true covariance matrix of parameter coefficients. Because we just consider the well-balanced design in the power calculation, according to the rule of thumb we proposed in section 3, the $v_R{}^*$ should be

derived from the KC-corrected sandwich estimator. Note that the diagonal elements of H_i, denoted by hit, corresponds to the amount of leverage of the i^{th} cluster. Intuitively, under the well-balanced design with an equal number of clusters in each arm ($q = 0.5$) and equal number of participants in all clusters ($m_i = m$), each cluster would have the same leverage of the response on the corresponding fitted value so that we would have $h_{it} = p/K$, where $p = rank(X)$. In our power calculations, $p = 2$. Therefore, under the perfect balanced design, $I_i - H_i = I_i - \frac{2}{K} \times I_i = \frac{K-2}{K} \times I_i$, and $V_{KC}^* = \left(\frac{K}{K-2}\right) V$. The $v_R{}^*$ is the (2,2)-element of KV_{KC}^*, and given by

$$v_R^* = \left(\frac{K}{K-2}\right) v_R = \left(\frac{K}{K-2}\right) \left(\frac{1+(m-1)\rho}{m}\right) \left[\frac{2}{\pi_1(1-\pi_1)} + \frac{2}{\pi_2(1-\pi_2)}\right].$$

Hence, for the GEE analyses of CRTs with few clusters,

$$power = 1 - T_{K-2}\left(t_{\alpha/2,K-2} - \frac{b\sqrt{K}}{\sqrt{\left(\frac{K}{K-2}\left(\frac{1+(m-1)\rho}{m}\right)\left[\frac{2}{\pi_1(1-\pi_1)} + \frac{2}{\pi_2(1-\pi_2)}\right]\right)}}\right).$$

Given the nominal a level, the expected power, the expected cluster size m, and the expected proportions, with this formula, the K can be easily calculated with an iteration process until the power exceed a specified level. If K is fixed, we can calculate m directly from the formula. It should be noted that in practice, there could be different numbers of clusters in different intervention arms. For the power and sample size calculation in the design stage, the balanced design is a reasonable assumption. In addition, for well-balanced design, the KC-corrected sandwich estimator is equivalent to the DF-corrected sandwich estimator.

The empirical power of the GEE Wald t test with the KC-corrected sandwich estimator was evaluated by computing the observed fraction of rejections of the null hypothesis when the intervention effect is set as odds ratio equal to 1.5 or 2. Even though the power formula assumes the common cluster size, we allow some variations ($cv = 0.2$) in our simulations to evaluate its robustness. The empirical power (P_{0BS}) from 1000 simulations and the expected power (P_{NEW}) calculated from proposed formula are shown in Table 2. Our results suggest that the empirical power from the Wald t test with KC-corrected sandwich estimator agrees well with the expected power from calculation. Hence, the proposed power formula can be used for power and sample size calculations if GEE Wald t test with the KC-corrected sandwich estimator is used. In Table 2 we also include the power estimation (P_{SHIH}) from power formula (1), which does not consider the small sample adjustments. Clearly, the formula (1) overestimates the power for the CRTs with few clusters.

In CRTs, there are two components of sample size—the number of clusters (K) and the number of subjects per cluster (m_i) —and increasing both numbers can achieve higher power. As the nature of CRTs, it may be very difficult to recruit large numbers of clusters. However, the effects of the number of subjects per cluster on power depend on the intraclass correlation. For example, assuming the odds ratio of 1.5, K=20 and ρ=0.001, the power

increases from about 0.22 at $\overline{m}$ =10 to 0.99 at $\overline{m}$ =150; while assuming the odds ratio of 1.5, K=20 and ρ =0.05, the power increases from about 0.17 at $\overline{m}$ =10 to only 0.36 at $\overline{m}$ =150. Hence, the estimation of ICC is very important before the power calculation in CRTs because enrolling large number of subjects may not increase the power in a cost effective manner if the ICC is relatively high.

Though the proposed power formula assumes the common cluster sizes, the correction allowing unequal cluster sizes could be obtained by replacing the $\frac{1+(m-1)\rho}{m}$ term in the power formula with $\frac{1+\left(\left((cv)^2\frac{(K-1)}{K}+1\right)\overline{m}-1\right)\rho}{\overline{m}}$[29]. But as pointed out by Campbell et al. [1], this modification may be conservative in GEE analysis. In our simulations allowing some considerable variation of cluster sizes (cv = 0.2), the empirical power matches well the expected 1 power calculated from the formula assuming common cluster sizes. Therefore, our proposed formula can still be used even when the cluster sizes have small to moderate variations.

4. Application

An example CRT is the trial in the Newham borough of East London, investigating whether intervention in general practices improved subsequent attendance at breast screening among women who did not respond to their initial invitation [4, 20]. Among the participating practices, 12 were randomized to the intervention group and 14 to the control. The intervention entailed training of the reception staff of the general practice to contact non-attenders for breast screening. Control practices were given no training or advice. A total of 995 women in the intervention practices and 1069 in the control practices were included in the trial. The outcome of interest was the attendance at breast screening among women who did not respond to their initial invitation for routine breast screening. In the intervention practices overall 9 percent of the women subsequently attended breast screening, as compared to 4 percent in the control practices. The intervention practices generally had higher rates of attendance in comparison to those in the control practices, although the attendance rate varied considerably between practices. It should be noted that a key feature of this trial was the small number of clusters (K = 26) with highly variable cluster sizes (cv $\approx$ 0.71).

Omar et al. [20] used the generalized linear mixed model (GLMM) for the analyses of the trial for women's attendance described above and suggested not using the GEE approach because of the small cluster number (K=26). We re-analyse the data with the GEE method by the use of Wald t tests with bias-corrected sandwich estimators. Because the Wald z tests are not suggested by our study, we only show the results from Wald z test with original sandwich estimator for comparison. The intervention results are shown in Table 3. Our results are similar to Omar's results using a GLMM assuming the normal distribution of the random term. The 95% confidence intervals from the Wald t tests with bias-corrected sandwich estimators are clearly more conservative than the Wald z test with the original sandwich estimator. Although the difference of the small sample inferences among the Wald t tests with different bias-corrected sandwich estimators are small in this data, the Wald t test with FG-corrected sandwich estimator is more conservative than the Wald t tests with other

bias-corrected sandwich estimators. Since the variation of cluster sizes is large in this trial ($cv \approx 0.71$), the Wald t test with FG-corrected sandwich estimator should be preferred. We may conclude that the women in intervention practices were estimated to be about 3 times more likely than those in the control practices to attend for breast screening subsequently. It must be noted that the interpretation of the regression parameters differs between GLMMs and the GEE approach. In the GEE context, the interpretation parameter describes the average change of the responses across the population under different intervention groups; in the GLMM context, the interpretation parameter is specific for the given cluster (practice) and describes the difference of responses assuming the same cluster (practice) is assigned to both control and intervention groups [23, 30].

5. Discussion

We evaluated the small sample properties of GEE with the sandwich estimator and five alternative bias-corrected sandwich estimators in analysing binary outcomes under the CRT scenarios with low ICC (0.001, 0.01 and 0.05), unequal cluster sizes, and small cluster numbers. The GEE sandwich estimator underestimates the variance for small number of clusters, and the Wald z or t tests have substantially greater type I error rates than the nominal level. Our simulations suggest that no single bias-corrected sandwich estimator is universally superior to the others when the Wald t test is used for the analyses of CRTs with few clusters and considerable variation of cluster sizes. The variation of cluster sizes must be considered in the choice of bias-corrected sandwich estimators. A rule of thumb is that choosing the Wald t test with KC-corrected sandwich estimator when the coefficient of variation is less than 0.6 and choosing the Wald t test with FG-corrected sandwich estimator, otherwise. Even though Lu et al. [19] recommended the Wald z test with MD-corrected sandwich estimator in the GEE analyses to deal with the small number of clusters, our results suggest that the MD-corrected sandwich estimator should be used with caution because the Wald z test maintains the type I error rate only when the number of clusters is greater than 20, while the Wald t test is too conservative when the number of clusters is smaller than 20. Teerenstra et al. [23] also found that the Wald t test with MD-corrected sandwich estimator tends to be conservative and suggested using the Wald t test with KC-corrected sandwich estimator instead. However, Teerenstra et al. only evaluated the Wald t test with KC-corrected sandwich estimator under well-balanced trial design. Our results show that Wald t test with KC-corrected sandwich estimator could only maintain the type I error rates to nominal level for the CRTs with small to moderate variation of cluster sizes ($cv < 0.6$). For the large variation of cluster sizes, as in the trial discussed above [20], the Wald t test with FG-corrected sandwich estimator should be used. Regarding Type I error control, the performance of the MBN-correction may depend on the choice of d and l. Within this paper, we followed the recommendations of the authors and set $d = 2$ and $l = 1$. However, future research is warranted to determine if other choices of d l might improve Type I error control.

A number of previous studies have discussed the sample size and statistical power for CRT s with regard to linear mixed models [31-33]. GEE-based sample size and power calculation have also been proposed for the correlated data [22, 34, 35]. However, these GEE-based methods assume the asymptotic conditions, utilize the original sandwich estimator, and

assume the standard normal distribution approximation. Many CRTs may not be able to recruit large number of clusters in reality [5, 9], and the sandwich estimator underestimates the variances of regression parameters for small samples and consequently inflates type I error rates. To improve the small sample performance of GEE, bias-corrected sandwich estimators [8, 10-12] and the Wald *t* test [8, 23, 26, 27] are proposed to replace the sandwich estimator and the asymptotical Wald *z* test when the cluster number is small. Teerenstra et al. [23] develop a formula to calculate the sample size and power for GEE analyses of three-level CRTs and suggest the *t* distribution approximation in the calculation. Even though Teerenstra et al. [23] recommended the Wald *t* test with KC-corrected sandwich estimator for the GEE analyses, they use the original sandwich estimator for the power calculation. It can be expected that a power calculation based on the sandwich estimator followed by an analysis using the KC-corrected sandwich estimator would result in the loss of precision. Considering the Wald *t* test with the KC-corrected sandwich estimator for both data analysis and power calculation, we derive a power formula for CRTs with the expectation of small cluster number. The expected powers based on this formula agree well with the empirical powers from simulations. Kauermann and Carroll [10] reduce the small sample bias of the sandwich estimator by weighting the estimate of $Cov(Y_i)$ in each cluster with a function of the leverage of the cluster, i.e., placing more weight on the cluster with larger leverage in the variance-covariance estimation. For the well-balanced CRT design with equal cluster numbers in different arms and equal numbers of participants in all clusters, we can expect that each cluster has the same leverage and weights the same in the variance-covariance estimation. Our power calculation assumes the well-balanced design, which is reasonable and commonly used in power estimations. In practice, however, it is hard to have the same number of participants in all clusters, and the cluster with more participants is expected to have larger leverage. Theoretically, our power calculation based on the averaged leverage may lose some precision when the exact leverages are used in the analyses. However, our simulation study suggests that the proposed power calculation is robust to some variation of cluster sizes ($cv < 0.2$). The impact of varying cluster size on the power or sample size estimation in CRTs has been investigated [29, 36, 37]. Eldridge et al. [29] show that for $c < 0.23$, the impact of the variation of cluster sizes is very small, which is also confirmed by our simulations. Given the large variation of cluster sizes, our power estimation could lose precision. Eldridge et al. [29] suggest an adjustment by replacing the $\frac{1+(m-1)\rho}{m}$ term in the power or sample size formula with $\frac{1+\left(\left((cv)^2\frac{(K-1)}{K}+1\right)\bar{m}-1\right)\rho}{\bar{m}}$. We expect this adjustment also works in our case, but is a subject for future research.

Our power and sample size estimation only considers the compound symmetry correlation structure, which is commonly assumed in practice. For those CRTs with a more complicated correlation structure, the proposed formulas may not work well. Another limitation is that we do not consider covariates other than the treatment, which may be included in the real data analyses. The addition of covariates will change the variance estimation and the degrees of freedom, so that the impact could be complicated, especially when the covariates are correlated with treatment. Further research is warranted when individual/patient level covariates affect the log odds of event.

In conclusion, we have compared and contrasted various approaches to maintaining appropriate Type I error control for GEE analyses when the number of clusters is small. Our results indicate that when the variation in cluster size is small to moderate (cv <0.6), the Wald t-test with the KC-corrected sandwich estimator maintains appropriate Type I error control. If the variation of cluster size is large (cv > 0.6), then the Wald t-test with the FG-corrected sandwich estimator maintains appropriate Type I error control. Because adequate Type I error control can be maintained even when a small number of clusters are analysed, we recommend the use of GEE method in the analyses of CRTs due to fewer assumptions and the robustness to the misspecification of the covariance structure. Finally we have provided a convenient power and sample size calculation for the GEE analyses of CRTs based upon the appropriate corrected sandwich estimators.

Acknowledgements

This research was supported by NIH grant T32HL079888 (PL), P60AR048095 (DTR), P60AR064172 (DTR) and UL1TR000165(DTR).

References

1. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and Statistics in Medicine. Stat Med. 2007; 26:2–19. [PubMed: 17136746]
2. Murray, DM. Design and Analysis of Group-Randomized Trials. Oxford University Press Inc; New York, NY: 1998.
3. Feng ZD, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. Annual Review of Public Health. 2001; 22:167–187.
4. Turner RM, Omar RZ, Thompson SG. Bayesian methods of analysis for cluster randomized trials with binary outcome data. Stat Med. 2001; 20:453–472. [PubMed: 11180313]
5. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. Am J Public Health. 2004; 94:423–432. [PubMed: 14998806]
6. Liang K-Y, Zeger SL. Longitudinal Data Analysis Using Generalized Linear Models. Biometrika. 1986; 73:10.
7. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. Stat Med. 1996; 15:1793–1806. [PubMed: 8870161]
8. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001; 57:126–134. [PubMed: 11252587]
9. Taljaard M, McRae AD, Weijer C, Bennett C, Dixon S, Taleban J, Skea Z, Eccles MP, Brehaut JC, Donner A, Saginur R, Boruch RF, Grimshaw JM. Inadequate reporting of research ethics review and informed consent in cluster randomised trials: review of random sample of published trials. British Medical Journal. 2011; 342:d2496. [PubMed: 21562003]
10. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association. 2001; 96:1387–1396.
11. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. Biometrical Journal. 2003; 45:395–409.
12. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics. 2001; 57:1198–1206. [PubMed: 11764261]
13. MacKinnon JG WH. Some heteroskedasticityconsistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics. 1985; 29:11.
14. Preisser JS, Qaqish BF. Deletion diagnostics for generalised estimating equations. Biometrika. 1996; 83:551–562.
15. Morel JG. Logistic Regression Under Complex Survey Designs. Survey Methodology. 1989; 15:21.

16. Lipsitz SR, Dear KB, Zhao L. Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. Biometrics. 1994; 50:842–846. [PubMed: 7981404]
17. Pan W. On the robust variance estimator in generalised estimating equations. Biometrika. 2001; 88:901–906.
18. Wang M, Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. Stat Med. 2011; 30:1278–1291. [PubMed: 21538453]
19. Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala S, Wolfson M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. Biometrics. 2007; 63:935–941. [PubMed: 17825023]
20. Omar RZ, Thompson SG. Analysis of a cluster randomized trial with binary outcome data using a multi-level model. Stat Med. 2000; 19:2675–2688. [PubMed: 10986541]
21. Eldridge SM, Ashby D, Feder GS, Rudnicka AR, Ukoumunne OC. Lessons for cluster randomized trials in the twenty-first century: a systematic review of trials in primary care. Clin Trials. 2004; 1:80–90. [PubMed: 16281464]
22. Shih WJ. Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. Biometrical Journal. 1997; 39:899–908.
23. Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample Size Considerations for GEE Analyses of Three-Level Cluster Randomized Trials. Biometrics. 2010; 66:1230–1237. [PubMed: 20070297]
24. Pan W. Sample size and power calculations with correlated binary data. Control Clin Trials. 2001; 22:211–227. [PubMed: 11384786]
25. Lee EW, Dubin N. Estimation and Sample-Size Considerations for Clustered Binary Responses. Stat Med. 1994; 13:1241–1252. [PubMed: 7973205]
26. Fai AHT, Cornelius PL. Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. Journal of Statistical Computation and Simulation. 1996; 54:363–378.
27. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. Stat Med. 2002; 21:1429–1441. [PubMed: 12185894]
28. Campbell MJ. Cluster randomized trials in general (family) practice research. Statistical Methods in Medical Research. 2000; 9:81–94. [PubMed: 10946428]
29. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. International Journal of Epidemiology. 2006; 35:1292–1300. [PubMed: 16943232]
30. Hubbard AE, Ahern J, Fleischer NL, Van der Laan M, Lippman SA, Jewell N, Bruckner T, Satariano WA. To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. Epidemiology. 2010; 21:467–474. [PubMed: 20220526]
31. Donner A. Sample-Size Requirements for Stratified Cluster Randomization Designs. Stat Med. 1992; 11:743–750. [PubMed: 1594813]
32. Lipsitz SR, Fitzmaurice GM. Sample-Size for Repeated-Measures Studies with Binary Responses. Stat Med. 1994; 13:1233–1239. [PubMed: 7973204]
33. Heo M, Leon AC. Statistical Power and Sample Size Requirements for Three Level Hierarchical Cluster Randomized Trials. Biometrics. 2008; 64:1256–1262. [PubMed: 18266889]
34. Liu G, Liang KY. Sample size calculations for studies with correlated observations. Biometrics. 1997; 53:937–947. [PubMed: 9290224]
35. Jung SH, Ahn CW. Sample size for a two-group comparison of repeated binary measurements using GEE. Stat Med. 2005; 24:2583–2596. [PubMed: 16118812]
36. Manatunga AK, Hudgens MG, Chen SD. Sample size estimation in cluster randomized studies with varying cluster size. Biometrical Journal. 2001; 43:75–86.
37. van Breukelen GJP, Candel MJJM, Berger MPF. Relative efficiency of unequal versus equal cluster sizes in cluster randomized and multicentre trials. Stat Med. 2007; 26:2589–2603. [PubMed: 17094074]

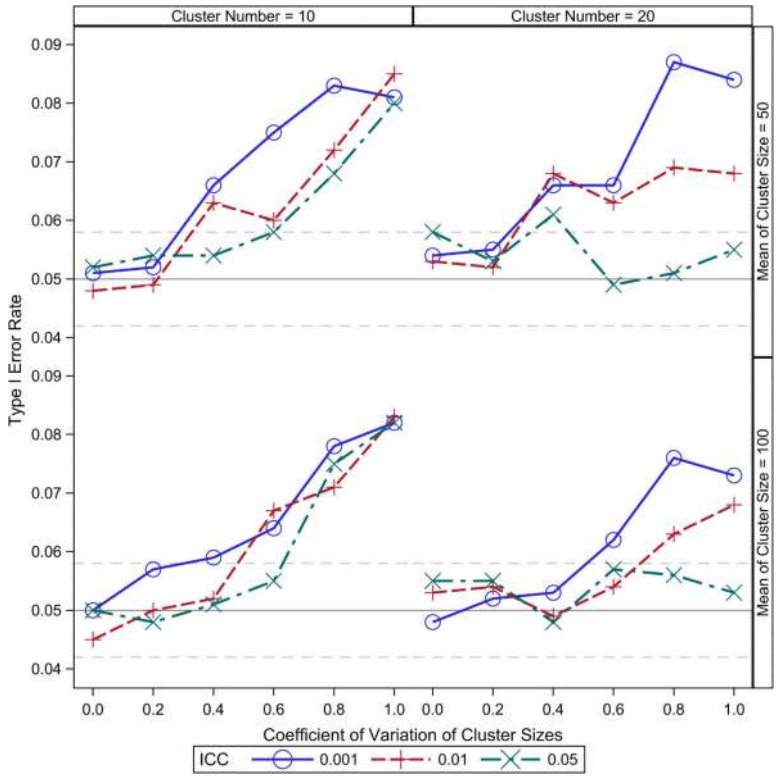

Figure 1.

Observed type I error rates of GEE Wald t test with DF-corrected sandwich estimator. The type I error rates are calculated from 3000 independent replications.

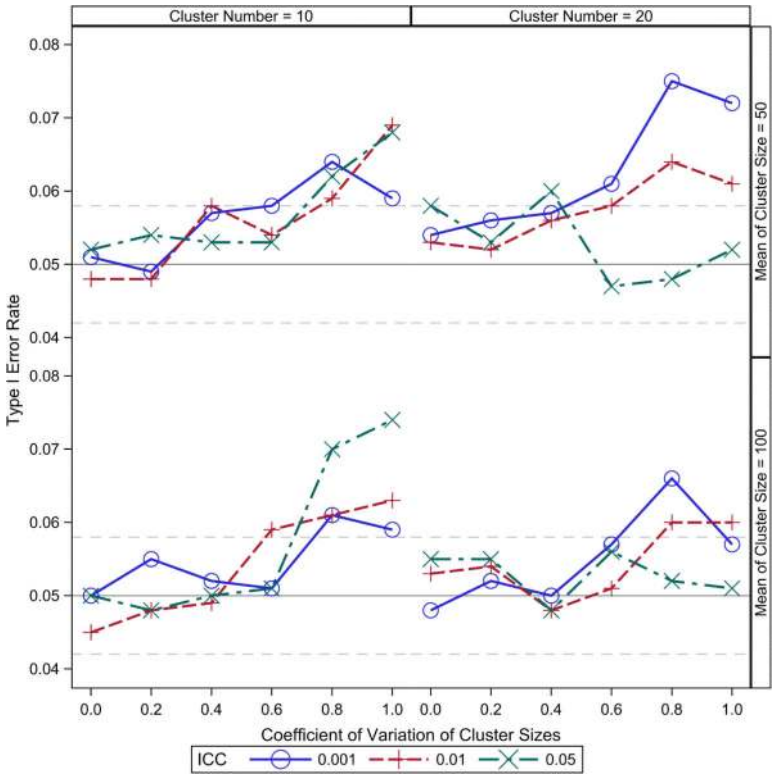

Figure 2.

Observed type I error rates of GEE Wald t test with KC-corrected sandwich estimator. The type I error rates are calculated from 3000 independent replications.

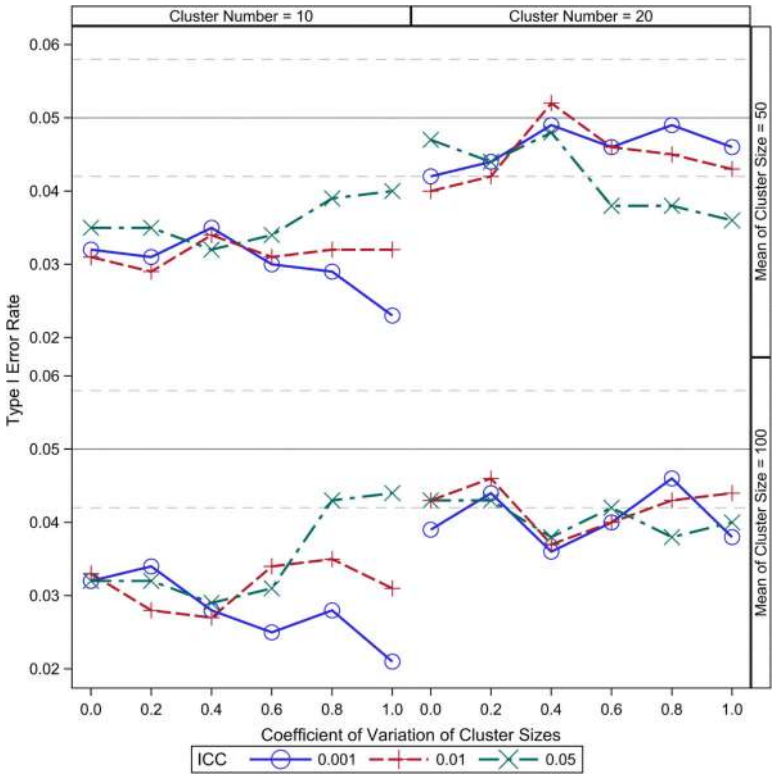

Figure 3.

Observed type I error rates of GEE Wald t test with MD-corrected sandwich estimator. The type I error rates are calculated from 3000 independent replications.

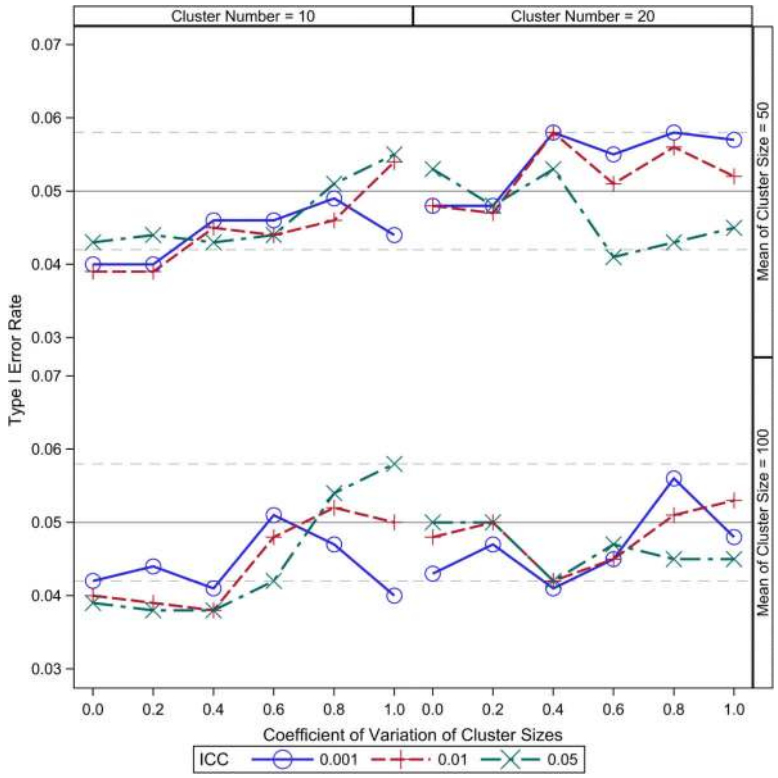

Figure 4.

Observed type I error rates of GEE Wald t test with FG-corrected sandwich estimator. The type I error rates are calculated from 3000 independent replications.

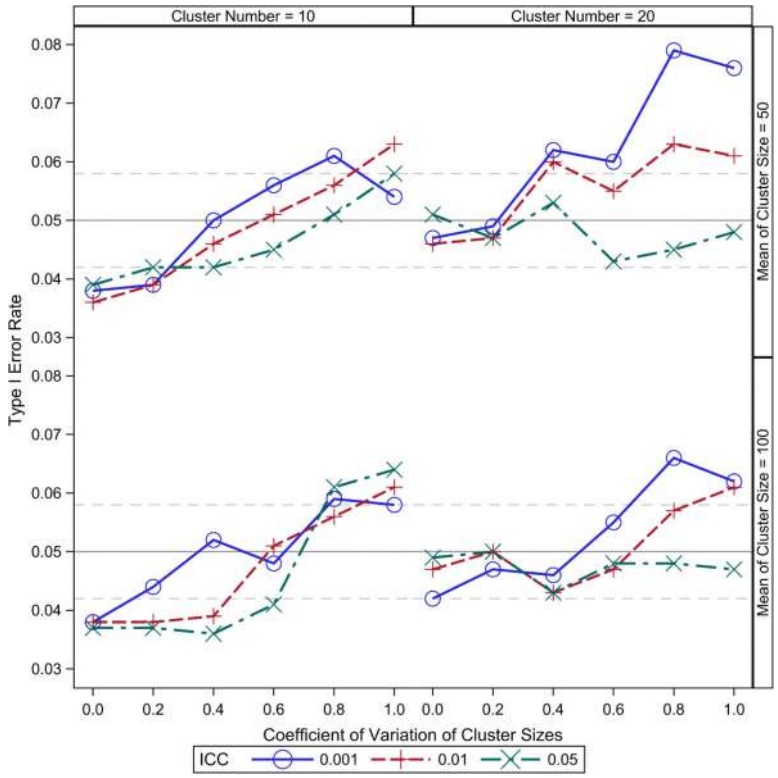

Figure 5.

Observed type I error rates of GEE Wald t test with MBN-corrected sandwich estimator. The type I error rates are calculated from 3000 independent replications.

Table 1

Observed type I error rates of GEE Wald tests. The tests are based on two-sided z-distribution approximation or t-distribution approximation with $K - 2$ degrees of freedom, at a nominal 0.05 significance level. Results are calculated from 3000 independent replications.

Covariance estimator	K	$\bar{m}$	$\rho = 0.001$		$\rho = 0.01$		$\rho = 0.05$	
			z	t	z	t	z	t
V	10	50	0.110	0.071	0.120	0.074	0.118	0.071
		100	0.110	0.067	0.120	0.079	0.113	0.068
		150	0.118	0.079	0.129	0.067	0.107	0.068
	20	50	0.071	0.055	0.084	0.067	0.077	0.062
		100	0.079	0.066	0.085	0.067	0.076	0.060
		150	0.081	0.063	0.078	0.061	0.086	0.066
	30	50	0.070	0.059	0.063	0.054	0.072	0.060
		100	0.066	0.056	0.071	0.062	0.073	0.059
		150	0.074	0.054	0.077	0.063	0.069	0.056
V^*_{DF}	10	50	0.081	0.045	0.089	0.051	0.085	0.050
		100	0.079	0.048	0.086	0.049	0.081	0.046
		150	0.089	0.054	0.093	0.053	0.078	0.047
	20	50	0.059	0.044	0.071	0.051	0.066	0.049
		100	0.068	0.054	0.070	0.051	0.062	0.050
		150	0.065	0.050	0.065	0.050	0.071	0.057
	30	50	0.060	0.050	0.056	0.046	0.063	0.050
		100	0.057	0.048	0.064	0.054	0.061	0.053
		150	0.056	0.048	0.065	0.052	0.059	0.048
V^*_{KC}	10	50	0.081	0.045	0.092	0.051	0.084	0.049
		100	0.079	0.048	0.086	0.049	0.081	0.046
		150	0.088	0.054	0.093	0.053	0.078	0.047
	20	50	0.059	0.044	0.071	0.051	0.066	0.049
		100	0.068	0.054	0.070	0.051	0.066	0.051
		150	0.065	0.050	0.065	0.050	0.071	0.057
	30	50	0.060	0.050	0.056	0.045	0.063	0.050
		100	0.058	0.048	0.064	0.054	0.061	0.053
		150	0.056	0.058	0.065	0.052	0.059	0.048
V^*_{MD}	10	50	0.055	0.030	0.061	0.033	0.058	0.032
		100	0.056	0.032	0.060	0.028	0.055	0.031
		150	0.064	0.037	0.064	0.034	0.058	0.032
	20	50	0.048	0.037	0.056	0.044	0.053	0.042
		100	0.058	0.031	0.056	0.042	0.054	0.038
		150	0.054	0.038	0.055	0.041	0.059	0.045
	30	50	0.053	0.041	0.048	0.040	0.054	0.041

Covariance estimator	K	$\bar{m}$	$\rho = 0.001$		$\rho = 0.01$		$\rho = 0.05$	
			z	t	z	t	z	t
V^*_{FG}	10	100	0.052	0.041	0.056	0.047	0.054	0.044
		150	0.049	0.042	0.056	0.046	0.051	0.041
		50	0.070	0.036	0.076	0.050	0.070	0.040
		100	0.068	0.040	0.071	0.042	0.068	0.038
		150	0.078	0.044	0.078	0.037	0.067	0.038
	20	50	0.053	0.040	0.063	0.041	0.060	0.046
		100	0.064	0.047	0.064	0.046	0.060	0.044
		150	0.061	0.045	0.061	0.046	0.064	0.049
	30	50	0.057	0.044	0.052	0.042	0.058	0.045
		100	0.055	0.044	0.060	0.050	0.058	0.048
		150	0.052	0.044	0.059	0.049	0.055	0.042
V^*_{MRN}	10	50	0.067	0.035	0.071	0.041	0.068	0.038
		100	0.065	0.038	0.070	0.036	0.064	0.037
		150	0.075	0.044	0.073	0.040	0.066	0.035
	20	50	0.051	0.040	0.063	0.045	0.059	0.045
		100	0.064	0.046	0.063	0.045	0.056	0.044
		150	0.059	0.044	0.060	0.046	0.063	0.050
	30	50	0.057	0.045	0.052	0.042	0.058	0.044
		100	0.055	0.044	0.060	0.049	0.058	0.048
		150	0.053	0.043	0.059	0.049	0.055	0.043

Table 2

Expected and empirical powers for cluster-randomized trials with few clusters.

Odds ratio	K	$\bar{m}$	$\rho = 0.001$			$\rho = 0.01$			$\rho = 0.05$		
			P_{SHIH}	P_{OBS}	P_{NEW}	P_{SHIH}	P_{OBS}	P_{NEW}	P_{SHIH}	P_{OBS}	P_{NEW}
1.5	10	10	0.149	0.118	0.119	0.141	0.120	0.113	0.118	0.091	0.098
		20	0.249	0.200	0.189	0.220	0.170	0.169	0.153	0.123	0.122
		50	0.514	0.406	0.385	0.388	0.339	0.289	0.196	0.161	0.152
		100	0.788	0.701	0.632	0.536	0.447	0.402	0.220	0.178	0.169
		150	0.910	0.836	0.783	0.612	0.537	0.465	0.230	0.205	0.175
	20	10	0.251	0.222	0.221	0.236	0.188	0.208	0.189	0.163	0.168
		20	0.441	0.411	0.387	0.388	0.363	0.340	0.258	0.243	0.227
		50	0.806	0.765	0.740	0.659	0.585	0.589	0.344	0.305	0.301
		100	0.974	0.954	0.949	0.826	0.769	0.762	0.388	0.342	0.340
		150	0.997	0.996	0.990	0.888	0.861	0.833	0.406	0.340	0.356
	30	10	0.351	0.324	0.322	0.329	0.296	0.302	0.260	0.250	0.239
		20	0.602	0.537	0.558	0.537	0.473	0.496	0.361	0.340	0.331
		50	0.933	0.929	0.907	0.827	0.823	0.786	0.479	0.440	0.440
		100	0.998	0.997	0.995	0.944	0.928	0.920	0.537	0.523	0.496
		150	1.000	1.000	1.000	0.973	0.973	0.958	0.560	0.535	0.517
2.0	10	10	0.353	0.254	0.263	0.331	0.240	0.247	0.262	0.205	0.198
		20	0.606	0.466	0.460	0.540	0.409	0.406	0.364	0.289	0.271
		50	0.935	0.836	0.822	0.830	0.726	0.679	0.482	0.377	0.360
		100	0.997	0.987	0.977	0.946	0.889	0.841	0.540	0.476	0.406
		150	1.000	0.995	0.997	0.974	0.935	0.899	0.563	0.460	0.424
	20	10	0.610	0.552	0.542	0.577	0.514	0.511	0.463	0.420	0.407
		20	0.883	0.868	0.827	0.830	0.805	0.767	0.625	0.571	0.556
		50	0.998	0.997	0.995	0.985	0.977	0.967	0.773	0.722	0.704
		100	1.000	1.000	1.000	0.999	0.998	0.996	0.830	0.802	0.767
		150	1.000	1.000	1.000	1.000	1.000	0.999	0.850	0.804	0.789
	30	10	0.783	0.761	0.740	0.751	0.720	0.707	0.628	0.585	0.584
		20	0.971	0.960	0.956	0.946	0.928	0.923	0.797	0.777	0.754

Odds ratio	K	$\bar{m}$	$\rho = 0.001$			$\rho = 0.01$			$\rho = 0.05$		
			P_{SHIH}	P_{OBS}	P_{NEW}	P_{SHIH}	P_{OBS}	P_{NEW}	P_{SHIH}	P_{OBS}	P_{NEW}
		50	1.000	1.000	1.000	0.999	0.997	0.998	0.912	0.887	0.882
		100	1.000	1.000	1.000	1.000	1.000	1.000	0.946	0.911	0.923
		150	1.000	1.000	1.000	1.000	1.000	1.000	0.956	0.952	0.936

Note: P_{SHIH} is the expected power calculated by Shih's method[22] which does not consider the small sample adjustment; P_{OBS} is the observed power from 1000 simulated independent replicates; and P_{NEW} is the expected power calculated from our proposed method with small sample adjustment.

Table 3

Intervention effects on women's attendance at breast screening.

Tests	Intervention	Odds ratio	*P*-value
GLMM[20]	1.04 (0.08, 2.00)	2.83 (1.08, 7.39)	0.030
GEE			
V and z test	1.13 (0.32, 1.95)	3.11 (1.38, 7.01)	0.0062
V^*_{DF} and t_{k-2} test	1.13 (0.24, 2.03)	3.11 (1.28, 7.58)	0.0147
V^*_{KC} and t_{k-2} test	1.13 (0.25, 2.02)	3.11 (1.28, 7.55)	0.0143
V^*_{MD} and t_{k-2} test	1.13 (0.22, 2.05)	3.11 (1.24, 7.80)	0.0176
V^*_{FG} and t_{k-2} test	1.13 (0.20, 2.07)	3.11 (1.22, 7.95)	0.0197
V^*_{MBN} and t_{k-2} test	1.13 (0.21, 2.06)	3.11 (1.24, 7.82)	0.0179