Statistics
in Medicine

Research Article

Received 14 March 2015, Accepted 28 October 2015 Published online 19 November 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.6817

Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples

Ming Wang,[a*†] Lan Kong,[a] Zheng Li[a] and Lijun Zhang[b,c]

Generalized estimating equations (GEE) is a general statistical method to fit marginal models for longitudinal data in biomedical studies. The variance–covariance matrix of the regression parameter coefficients is usually estimated by a robust "sandwich" variance estimator, which does not perform satisfactorily when the sample size is small. To reduce the downward bias and improve the efficiency, several modified variance estimators have been proposed for bias-correction or efficiency improvement. In this paper, we provide a comprehensive review on recent developments of modified variance estimators and compare their small-sample performance theoretically and numerically through simulation and real data examples. In particular, Wald tests and *t*-tests based on different variance estimators are used for hypothesis testing, and the guideline on appropriate sample sizes for each estimator is provided for preserving type I error in general cases based on numerical results. Moreover, we develop a user-friendly R package "geesmv" incorporating all of these variance estimators for public usage in practice.

Keywords: generalized estimating equations; longitudinal data; variance estimator; small sample size; Type I error; hypothesis testing

1. Introduction

Longitudinal data are commonly encountered in biomedical studies [1–4]. For example, in a diabetes study, repeated primary efficacy measures on HbA1c were taken over time (baseline and follow-up visits after treatment) for each patient, and the question of interest was to investigate the trend of HbA1c changing over time or the insulin treatment effect on HbA1c [5]. For such a situation, the responses from the same individual turn to be "more alike"; thus, incorporating within-subject correlation and between-subject variations into model fitting is necessary to improve the efficiency of the estimation and enhance the power.

To analyze repeated measures, several simple traditional approaches exist (i.e., repeated measure analysis of variance) [6]. However, mixed-effect models [2] and generalized estimating equations (GEE) [7] are popularly applied. Of note, the mixed-effects model is an individual-level approach that is able to adopt random effects to capture the correlation among observations from the same subject [8]; GEE is a population-level model based on the quasi-likelihood function [9, 10]. In this paper, we focus on GEE, which holds several defining features as follows: (i) Under mild regularity conditions, the parameter estimates are consistent and asymptotically normal even under the misspecified "working" correlation structure of the responses; (ii) When the inference is intended to be population-based, for instance the overall treatment effect, GEE treats the variance–covariance matrix of the responses as nuisance parameters [9]; and (iii) GEE relaxes the distribution assumption and only requires correct specification of

[a]*Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, U.S.A.*
[b]*Department of Biochemistry and Molecular Biology, Penn State College of Medicine, Hershey, PA, U.S.A.*
[c]*Institute for Personalized Medicine, Penn State College of Medicine, Hershey, PA, U.S.A.*
[*]*Correspondence to: Ming Wang, Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA, U.S.A.*
[†]*E-mail: mwang@phs.psu.edu*

marginal mean and variance as well as the link function between the mean and covariates of interest. GEE has been implemented in statistical software (i.e., SAS, R) and can be directly adopted for analysis.

It is known that the variance estimators of parameters of interest are utilized in hypothesis testing; thus, its accuracy is important for valid inference. Under some specific conditions such as small-sample size, the traditional GEE with the classic "sandwich" variance estimator does not perform satisfactorily, and considerable downward bias is exhibited [11–14], in turn leading to inflated type I errors and lower coverage rates of the resulting confidence intervals [15, 16]. Until now, several remedy strategies on modifications of variance estimators have been proposed to improve the finite small-sample performance [13, 17]. To our knowledge, few studies exist to cover various variance estimators including the most recently developed ones on GEE with small samples for comprehensive comparisons, and there is a lack of a guideline on the adequate sample size for preserving type I error. Note that the recent paper related to this area was discussed by Li and Redden [18] with emphasis on the small-sample performance of bias-corrected sandwich estimators, particularly for cluster-randomized trials with binary outcomes. However, this work has several limitations as follows: (i) only the scenarios with binary outcomes were considered; (ii) the influence of misspecified correlation structure was not explored; (iii) the degrees of freedom of the approximate t-distribution did not take the variability of the variance estimator into account [16, 19, 20], but only depended on the number of clusters; and (iv) limited variance estimators implemented by SAS were considered, but the most recent ones were not. In this paper, we attempt to address these issues and provide a more comprehensive and accurate comparison of different modified variance estimators. Furthermore, we develop a user-friendly R package including functions for calculating the modified variance estimators as well as the degrees of freedom defined as a function of the variance of the estimator for t-tests [16, 20].

The remainder of the paper is organized as follows. In Section 2, we introduce the notations and provide nine variance estimators of GEE as well as their theoretical comparisons. In addition, two types of hypothesis testing, Wald tests, and t-tests, are emphasized. Later, in Section 3, we provide extensive simulation to compare the performance of different variance estimators and identify the suitable sample size for each to ensure their satisfactory performance in controlling type I error. Importantly, we develop an R package "geesmv" for public use with small samples. In addition, we illustrate the application of our R package via a real data example in Section 4. The conclusion with a brief discussion and future work is shown in Section 5.

2. Method

2.1. Notation and generalized estimating equations

Given longitudinal data consisting of K subjects, denote Y_{ij} as the j^{th} response for the i^{th} subject with n_i observations, $i = 1, 2, ..., K, j = 1, \dots, n_i$, and $\boldsymbol{X}_{ij}$ is a $p \times 1$ vector of covariates. $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{in_i})'$ denotes the response vector with the mean vector noted by $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{in_i})'$ where μ_{ij} is the corresponding j^{th} mean for subject i. There exists within-subject correlation, but the observations across subjects are assumed to be independent. In addition, the marginal model specifying an association between μ_{ij} and the covariates of interest $\boldsymbol{X}_{ij}$ is given by

$$g\left(\mu_{ij}\right) = \boldsymbol{X}_{ij}'\boldsymbol{\beta} \tag{1}$$

with g as a known link function and $\boldsymbol{\beta}$ an unknown $p \times 1$ vector of regression coefficients. The conditional variance of Y_{ij} given $\boldsymbol{X}_{ij}$ is $Var(Y_{ij}|\boldsymbol{X}_{ij}) = \nu(\mu_{ij})\phi$ with ν as a known variance function of μ_{ij} and ϕ a scale parameter, which may need to be estimated. Of note, ν and ϕ depend on the distributions of outcomes. For example, if Y_{ij} is a continuous variable, $\nu(\mu_{ij})$ is 1, and ϕ represents the error variance; if Y_{ij} is a count variable, $\nu(\mu_{ij}) = \mu_{ij}$, and ϕ is equal to 1. Moreover, the variance–covariance matrix for $\boldsymbol{Y}_i$ is noted by $\boldsymbol{V}_i = \phi \boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$, where $\boldsymbol{A}_i = Diag\{\nu(\mu_{i1}), \cdots, \nu(\mu_{in_i})\}$ and the "working" correlation structure $\boldsymbol{R}_i(\boldsymbol{\alpha})$ describes the correlation pattern of observations within-subject with $\boldsymbol{\alpha}$ as a vector of association parameters depending on the correlation structure. Several types of "working" correlation structures including independent $\left(\alpha = 0, Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}\right)$, exchangeable $\left(\alpha = \rho, Corr(Y_{ij}, Y_{ik}) = \begin{cases} 1 & j = k \\ \rho & j \neq k \end{cases}\right)$, autoregressive $\left(\alpha = \rho, Corr(Y_{ij}, Y_{i,j+m}) = \rho^m, j + m \leqslant n_i\right)$,

Toeplitz $\left(\boldsymbol{\alpha}=(\theta,\theta_1,\theta_2,\ldots,\theta_{n_i-1})^T, Corr(Y_{ij},Y_{i,j+m})=\left\{\begin{array}{l}1\ m=0\\ \theta_m\ 0<m\leqslant n_i-j\end{array}\right.\right)$ and also unstructured $\left(\boldsymbol{\alpha}=(\theta_{12},\theta_{13},\ldots,\theta_{21},\theta_{23},\ldots,\theta_{n_{i-1}n_i})^T, Corr(Y_{ij},Y_{ik})=\left\{\begin{array}{l}1\ j=k\\ \theta_{jk}\ j\neq k\end{array}\right.\right)$ ones are commonly used. The estimation of $\boldsymbol{\alpha}$ is based on an iterative fitting process using the Pearson residual $e_{ij}=(y_{ij}-\hat{\mu}_{ij})/\sqrt{v(\hat{\mu}_{ij})}$ given the current value of $\boldsymbol{\beta}$; In addition, the scale parameter ϕ is estimated by $\hat{\phi}=\frac{1}{N-p}\sum_{i=1}^{K}\sum_{j=1}^{n_i}e_{ij}^2$ with the total number of observations $N=\sum_{i=1}^{K}n_i$.

The GEE method yields asymptotically consistent $\hat{\boldsymbol{\beta}}$, even when the "working" correlation structure ($\boldsymbol{R}_i(\boldsymbol{\alpha})$) is misspecified [7], and $\hat{\boldsymbol{\beta}}$ is obtained by the following estimating equation

$$U(\boldsymbol{\beta})=\sum_{i=1}^{K}\boldsymbol{D_i'V_i^{-1}(Y_i-\mu_i)}=0, \tag{2}$$

where $\boldsymbol{D}_i=\frac{\partial\boldsymbol{\mu}_i}{\partial\boldsymbol{\beta}'}$. Given the true value of $\boldsymbol{\beta}$ as $\boldsymbol{\beta}_t$ and mild regularity conditions, $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with a mean $\boldsymbol{\beta}_t$ and a covariance matrix estimated based on the "sandwich" estimator by

$$V_{LZ}=\left(\sum_{i=1}^{K}\boldsymbol{D_i'V_i^{-1}D_i}\right)^{-1}M_{LZ}\left(\sum_{i=1}^{K}\boldsymbol{D_i'V_i^{-1}D_i}\right)^{-1}, \tag{3}$$

with

$$M_{LZ}=\sum_{i=1}^{K}\boldsymbol{D_i'V_i^{-1}Cov(Y_i)V_i^{-1}D_i} \tag{4}$$

by replacing $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and ϕ with their consistent estimates, where $Cov(\boldsymbol{Y}_i)=\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i'$ with $\hat{\boldsymbol{r}}_i=\boldsymbol{Y}_i-\hat{\boldsymbol{\mu}}_i$ is an estimator of the variance–covariance matrix of $\boldsymbol{Y}_i$ [7, 21]. This "sandwich" estimator is robust in that it is consistent even if the correlation structure is misspecified. Note that if V_i is correctly specified, a consistent estimator for the covariance matrix of $\hat{\boldsymbol{\beta}}$ is given by $\left(\sum_{i=1}^{K}\boldsymbol{D_i'V_i^{-1}D_i}\right)^{-1}$, which is often referred as the model-based variance estimator [7]. Next, we will discuss the small-sample properties of GEE with several modifications on variance estimators and hypotheses testing.

2.2. *Modified variance estimators of generalized estimating equations with small samples*

Because of the fact that the fitted value $\hat{\boldsymbol{\mu}}_i$ tends to be closer to $\boldsymbol{Y}_i$ than the true value $\boldsymbol{\mu}_i$ and when sample size is small, $\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i'$ in V_{LZ} is biased downward for estimating $E(\boldsymbol{e}_i\boldsymbol{e}_i')$, and the bias turns to be larger when the sample is much smaller; meanwhile, a greater variability may arise [16,22]. Therefore, the hypothesis testing using V_{LZ} tends to be liberal, and the resulting confidence interval is narrow. Table I provides a comprehensive summary of the recent modified variance estimators, and the details of the estimators are provided next.

(1)V_{MK} is the degrees-of-freedom corrected "sandwich" variance estimator proposed by MacKinnon [23]. This estimator incorporates the simplest adjustment by adopting an adjustment factor of $\frac{K}{K-p}$, which is shown by

Table I. Summary of eight modified variance estimators for generalized estimating equations with small sample.

Variance estimator	Modification	Reference
V_{MK}	Degrees-of-freedom adjustment	MacKinnon (1985) [23]
V_{KC}	Bias correction	Kauermann and Carroll (2001) [24]
V_{PAN}	Efficiency improvement	Pan (2001) [20]
V_{GST}	Efficiency improvement	Gosho *et al.* (2014) [25]
V_{MD}	Bias correction	Mancl and DeRouen (2001) [22]
V_{FG}	Bias correction	Fay and Graubard (2001) [26]
V_{MBN}	Bias correction	Morel *et al.* (2003) [27]
V_{WL}	Bias correction and efficiency improvement	Wang and Long (2011) [16]

$$V_{MK} = \frac{K}{K-p} V_{LZ}. \tag{5}$$

When $K \to \infty$, $V_{MK} \to_p V_{LZ}$. V_{MK} corrects the bias, but meanwhile increases the variability.

(2) V_{KC} is a bias-corrected "sandwich" variance estimator under the assumption of correctly specified correlation structure proposed by Kauermann and Carroll [24], which is

$$V_{KC} = \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} M_{KC} \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1}, \tag{6}$$

with

$$M_{KC} = \sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left(\boldsymbol{I}_i - \boldsymbol{H}_{ii} \right)^{-1/2} \hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i' \left(\boldsymbol{I}_i - \boldsymbol{H}_{ii}' \right)^{-1/2} \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \tag{7}$$

where $\boldsymbol{I}_i$ is an $n_i \times n_i$ identify matrix and the subject leverage H_{ii} is a diagonal matrix with the leverage of the i^{th} subjects, which can be calculated by

$$\boldsymbol{H}_{ii} = \boldsymbol{D}_i \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}.$$

(3) V_{PAN} was proposed by Pan [20] given two additional assumptions satisfied: (A1) The conditional variance of Y_{ij} given $\boldsymbol{X}_{ij}$ is correctly specified; (A2) A common correlation structure, $\boldsymbol{R}_c$, exists across all subjects. The modified variance estimator is

$$V_{PAN} = \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} M_{PAN} \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1}, \tag{8}$$

with

$$M_{PAN} = \sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left\{ \boldsymbol{A}_i^{1/2} \left(\frac{1}{K} \sum_{i=1}^{K} \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i' \boldsymbol{A}_i^{-1/2} \right) \boldsymbol{A}_i^{1/2} \right\} \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \tag{9}$$

V_{PAN} pools data across all subjects in estimating $Cov(\boldsymbol{Y}_i)$, which performs more efficiently.

(4) V_{GST} made an additional modification on Pan's estimator by incorporating the bias of $\boldsymbol{A}_i^{1/2} \left(\frac{1}{K} \sum_{i=1}^{K} \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i' \boldsymbol{A}_i^{-1/2} \right) \boldsymbol{A}_i^{1/2}$ for small K, which was proposed by Gosho *et al.* [25]. The modified variance estimator is written as

$$V_{GST} = \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} M_{GST} \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1}, \tag{10}$$

with

$$M_{GST} = \sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left\{ \boldsymbol{A}_i^{1/2} \left(\frac{1}{K-p} \sum_{i=1}^{K} \boldsymbol{A}_i^{-1/2} \hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i' \boldsymbol{A}_i^{-1/2} \right) \boldsymbol{A}_i^{1/2} \right\} \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \tag{11}$$

V_{GST} also pools data across all subjects in estimating $Cov(\boldsymbol{Y}_i)$. In particular, when $K \gg p$ and K is large enough, V_{GST} approximately equals to V_{PAN}.

(5) V_{MD} is another bias-corrected "sandwich" variance estimator proposed by Mancl and DeRouen [22]. Unlike V_{KC}, this estimator does not assume a correctly specified correlation structure, and it is written by

$$V_{MD} = \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1} M_{MD} \left(\sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \right)^{-1}, \tag{12}$$

with

$$M_{MD} = \sum_{i=1}^{K} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} (\boldsymbol{I}_i - \boldsymbol{H}_{ii})^{-1} \hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i' (\boldsymbol{I}_i - \boldsymbol{H}_{ii}')^{-1} \boldsymbol{V}_i^{-1} \boldsymbol{D}_i \tag{13}$$

where $\boldsymbol{I}_i$ and $\boldsymbol{H}_{ii}$ are defined as the same as V_{KC}. Note that to correct this bias in finite samples, Mancl and DeRouen [22] relied on the following approximate identity

$$E\left(\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i'\right) \approx (\boldsymbol{I}_i - \boldsymbol{H}_{ii})Cov(\boldsymbol{Y}_i)\left(\boldsymbol{I}_i - \boldsymbol{H}_{ii}\right)',$$

but they ignore one term $\sum_{j\neq i}\boldsymbol{H}_{ij}\boldsymbol{Cov}(\boldsymbol{Y}_i)\boldsymbol{H}_{ij}^{T}$ from its first-order Taylor expansion, leading to overcorrection.

(6) V_{FG} indicated by Fay and Graubard [26] made a further adjustment on V_{MD} by multiplying a scale factor, which is given by

$$V_{FG} = \left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1} M_{FG}\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}, \tag{14}$$

with

$$M_{FG} = \sum_{i=1}^{K}\boldsymbol{\eta}_i^{-1}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\boldsymbol{\eta}_i'^{-1} \tag{15}$$

where $\boldsymbol{\eta}_i = \boldsymbol{I}_p - \boldsymbol{N}_i$. Note that the jj^{th} diagnal value of $\boldsymbol{\eta}_i^{-1/2}$ equals to $\left(1 - min(b, \{\boldsymbol{N}_i\}_{jj})\right)^{-1}$, where $\boldsymbol{N}_i = \boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}$ for a simple bias correction and b is prespecified subjectively to avoid extreme adjustments when $\boldsymbol{N}_i$ is quite close to 1.

(7) V_{MBN} is a bias-corrected estimator recommended by Morel *et al.* [27] by incorporating correlation on the residual cross-products and sample size, provided as

$$V_{MBN} = \left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1} M_{MBN}\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}, \tag{16}$$

with

$$M_{MBN} = \sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\left(k\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i' + \delta_m\xi\boldsymbol{V}_i\right)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i \tag{17}$$

where $k = \frac{N-1}{N-p}\frac{K}{K-1}$, $\delta_m = \begin{cases}\frac{p}{K-p} & K > (d+1)p\\ \frac{1}{d} & otherwise\end{cases}$ and $\xi = max\left(r, \frac{trace\left(\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}M_{LZ}\right)}{p}\right)$ with $0 \leqslant r \leqslant 1$. Note that k is a factor to adjust the bias of empirical variance estimator of $Cov(\boldsymbol{Y}_i)$ and that δ_m given by Morel can be bounded by $1/d$ [27]. The default values for d and r are 2 and 1, respectively, according to Morel *et al.* [27].

(8) V_{WL} is a combined variance estimator suggested by Wang and Long [16], which considered both the strength of V_{PAN} and V_{MD} for pooling information from all subjects and reducing the bias of the estimate for e_ie_i'. The estimator is as follows

$$V_{WL} = \left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1} M_{WL}\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}, \tag{18}$$

where

$$M_{WL} = \sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{A}_i^{1/2}\left\{\sum_{i=1}^{K}\boldsymbol{A}_i^{-1/2}\left(\boldsymbol{I}_i - \boldsymbol{H}_{ii}\right)^{-1}\hat{\boldsymbol{r}}_i\hat{\boldsymbol{r}}_i'\left(\boldsymbol{I}_i - \boldsymbol{H}_{ii}'\right)^{-1}\boldsymbol{A}_i^{-1/2}/K\right\}\boldsymbol{A}_i^{1/2}\boldsymbol{V}_i^{-1}\boldsymbol{D}_i. \tag{19}$$

This estimator was confirmed to perform as well as or better than V_{PAN} and V_M, but the two additional assumptions specified earlier also need to be satisfied.

We now present theoretical comparisons among those variance estimators. As shown earlier, all variance estimators share the same two outside terms, that is, $\left(\sum_{i=1}^{K}\boldsymbol{D}_i'\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}$. Thus, we focus on assessing and comparing the middle matrix, M, of different variance estimators. The derived covariance

Table II. Covariance matrix of the middle parts from nine variance estimators for generalized estimating equations.

Matrix M	Covariance matrix of $vec(M)$
M_{LZ}	$\sum_{i=1}^{K} \boldsymbol{S}_i \boldsymbol{T}_i \boldsymbol{S}_i'$
M_{MK}	$\sum_{i=1}^{K} \frac{K^2}{(K-p)^2} \boldsymbol{S}_i \boldsymbol{T}_i \boldsymbol{S}_i'$
M_{KC}	$\sum_{i=1}^{K} \boldsymbol{S}_i \boldsymbol{F}_i \boldsymbol{T}_i \boldsymbol{F}_i' \boldsymbol{S}_i'$
M_{PAN}	$\sum_{i=1}^{K} \boldsymbol{S}_i \left[\boldsymbol{E}_i \left(\sum_{j=1}^{K} \frac{1}{K^2} \boldsymbol{E}_j^{-1} \boldsymbol{T}_j \boldsymbol{E}_j^{-1} \right) \boldsymbol{E}_i \right] \boldsymbol{S}_i'$
M_{GST}	$\sum_{i=1}^{K} \boldsymbol{S}_i \left[\boldsymbol{E}_i \left(\sum_{j=1}^{K} \frac{1}{(K-p)^2} \boldsymbol{E}_j^{-1} \boldsymbol{T}_j \boldsymbol{E}_j^{-1} \right) \boldsymbol{E}_i \right] \boldsymbol{S}\vert_i'$
M_{MD}	$\sum_{i=1}^{K} \boldsymbol{S}_i \boldsymbol{G}_i \boldsymbol{T}_i \boldsymbol{G}_i' \boldsymbol{S}_i'$
M_{FG}	$\sum_{i=1}^{K} \boldsymbol{H}_i \boldsymbol{T}_i \boldsymbol{H}_i'$
M_{MBN}	$\sum_{i=1}^{K} \boldsymbol{S}_i \boldsymbol{N}_i \boldsymbol{S}_i'$
M_{WL}	$\sum_{i=1}^{K} \boldsymbol{S}_i \left[\boldsymbol{E}_i \left(\sum_{j=1}^{K} \frac{1}{K^2} \boldsymbol{E}_j^{-1} \boldsymbol{G}_j \boldsymbol{T}_j] \boldsymbol{G}_j' \boldsymbol{E}_j^{-1} \right) \boldsymbol{E}_i \right] \boldsymbol{S}_i'$

$\boldsymbol{T}_i = Cov(vec(\hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i')); \boldsymbol{S}_i = (\boldsymbol{D}_i' \boldsymbol{V}_i^{-1}) \bigotimes (\boldsymbol{D}_i' \boldsymbol{V}_i^{-1}); \boldsymbol{F}_i = (\boldsymbol{I}_i - \boldsymbol{H}_{ii})^{-\frac{1}{2}} \bigotimes (\boldsymbol{I}_i - \boldsymbol{H}_{ii})^{-\frac{1}{2}}; \boldsymbol{G}_i = (\boldsymbol{I}_i - \boldsymbol{H}_{ii})^{-1} \bigotimes (\boldsymbol{I}_i - \boldsymbol{H}_{ii})^{-1}; \boldsymbol{E}_i = \boldsymbol{A}_i^{1/2} \bigotimes \boldsymbol{A}_i^{1/2}; \boldsymbol{H}_i = (\boldsymbol{\eta}_i^{-1} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}) \bigotimes (\boldsymbol{\eta}_i^{-1} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}); \boldsymbol{N}_i = Cov(k vec(\hat{\boldsymbol{r}}_i \hat{\boldsymbol{r}}_i') + vec(\delta_m \xi \boldsymbol{V}_i))$.

matrix for $vec(M)$ are given in Table II. It has been shown by Wang and Long [16] that $Cov\left(vec(M_{LZ})\right) - Cov\left(vec(M_{WL})\right)$ and $Cov\left(vec(M_{MD})\right) - Cov\left(vec(M_{WL})\right)$ are non-negative definite with probability 1, while $Cov\left(vec(M_{PAN})\right) - Cov\left(vec(M_{WL})\right)$ converges to 0 with probability 1 as $K \to \infty$. For comparisons among the other alternatives,

$$Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{MK})\big) = \sum_{i=1}^{K} \left(1 - \frac{K^2}{(K-p)^2}\right) \boldsymbol{S}_i \boldsymbol{T}_i \boldsymbol{S}_i'$$

$$Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{KC})\big) = \sum_{i=1}^{K} \boldsymbol{S}_i \left(\boldsymbol{T}_i - \boldsymbol{F}_i \boldsymbol{T}_i \boldsymbol{F}_i'\right) \boldsymbol{S}_i'$$

$$Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{GST})\big) = \sum_{i=1}^{K} \boldsymbol{S}_i \left(\boldsymbol{T}_i - \boldsymbol{E}_i \left(\sum_{j=1}^{K} \frac{1}{(K-p)^2} \boldsymbol{E}_j^{-1} \boldsymbol{T}_j \boldsymbol{E}_j^{-1}\right) \boldsymbol{E}_i\right) \boldsymbol{S}_i'$$

$$Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{FG})\big) = \sum_{i=1}^{K} \left(1 - \boldsymbol{\eta}_i^{-1} \bigotimes \boldsymbol{\eta}_i^{-1}\right) \boldsymbol{S}_i \boldsymbol{T}_i \boldsymbol{S}_i' \left(1 - \boldsymbol{\eta}_i'^{-1} \bigotimes \boldsymbol{\eta}_i'^{-1}\right)$$

$$Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{MBN})\big) = \sum_{i=1}^{K} \boldsymbol{S}_i \left(\boldsymbol{T}_i - \boldsymbol{N}_i\right) \boldsymbol{S}_i'$$

Based on the aforementioned derivations, under mild conditions, $Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{MK})\big)$, $Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{KC})\big)$, $Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{FG})\big)$, and $Cov\big(vec(M_{LZ})\big) - Cov\big(vec(M_{MBN})\big)$ will converge to 0 with probability 1, while $Cov\left(vec(M_{LZ})\right) - Cov\left(vec(M_{GST})\right)$ is non-negative definite with probability 1 as $K \to \infty$. Hence, these variance estimators are asymptotically equivalent. However, when the sample size is small, V_{LZ} tends to underestimate the variance. Therefore, the modifications through the bias-correction or degrees-of-freedom adjustment are mostly applied (Table I). On the other hand, the efficiency gain by pooling data across all subjects to improve the estimator of $Cov(\boldsymbol{Y}_i)$ instead of only using data from the i^{th} subject, is incorporated in V_{PAN}, V_{GST}, and V_{WL}. Thus, V_{WL} is the only estimator that takes into consideration both bias correction and efficiency improvement. Therefore, it is expected intuitively to outperform the other alternatives if the assumptions (A1) and (A2) are satisfied. In Section 3, extensive numerical comparisons via simulations will be conducted for further investigation.

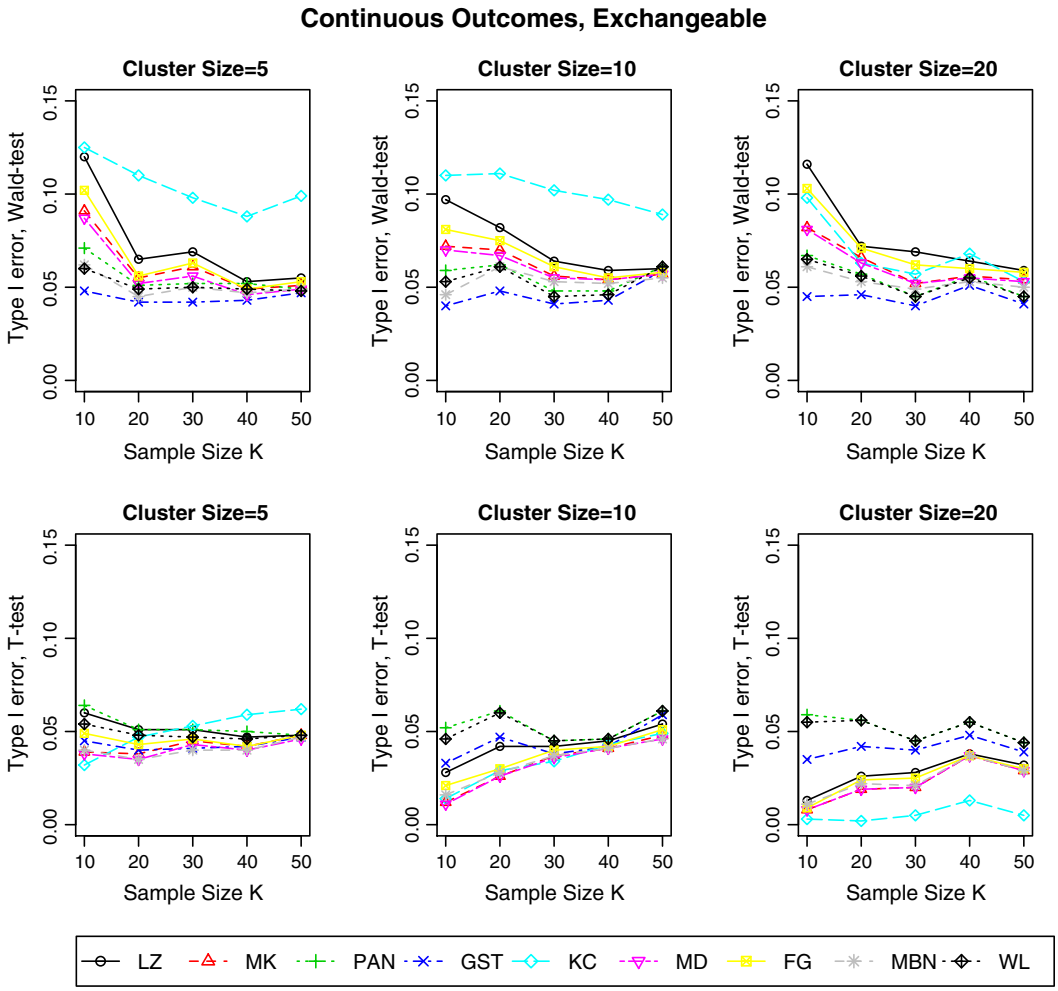

Figure 1. Type I errors based on Wald and t-tests for continuous outcomes with the true correlation structure as exchangeable.

2.3. *Hypotheses testing*

For tests of hypotheses in GEE, the Wald test and score test have been popularly applied [16, 20, 28]. However, when the sample size is small, the Wald test leads to inflated type I error, which seems too liberal [20, 28], and score test has smaller test size than the prespecified nominal level [28]. Therefore, several modifications have been proposed to obtain improved finite performance for GEE, that is, t-test and modified score test. According to Guo *et al.* [28], the score test was modified under the context of small sample, which was shown to be less conservative than the t-test via simulation. Currently, we consider t-tests when the sample size is small, and the brief derivation is shown next.

Without generality, a simple univariate hypothesis testing is taken as an example. Suppose in a clinical trial, the mean model is specified as $\mu_{ij} = \alpha + \beta \times$ treatment. The hypothesis of interest for the treatment parameter β is given by $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$. Thus, the test statistic for the Wald test is $z = \frac{\hat{\beta}}{\sqrt{\hat{V}(\hat{\beta})}}$, where $\hat{V}(\hat{\beta})$ can be replaced by any estimator mentioned earlier. For small samples, the t-test was proposed by Pan [20] and was also extensively studied by Wang and Long [16]. Denote κ and ν as the estimated mean and variance of $V(\hat{\beta})$. It follows the derivation based on vec operator that the distribution of $\frac{\hat{V}(\hat{\beta})}{c}$ is approximated with a chi-square distribution χ^2_d, where the scale parameter $c = \frac{\nu}{2\kappa}$ and the degrees of freedom $d = \frac{2\kappa^2}{\nu}$. The test statistic for t-test is $t = \frac{\hat{\beta}/\sqrt{\kappa}}{\sqrt{\hat{V}(\hat{\beta})/cd}}$, which is the same as the Wald test statistic with the degrees of freedom $d \approx 2\hat{V}(\hat{\beta})^2/\widehat{Var}(\hat{V}(\hat{\beta}))$ [16, 20]. This satterthwaite-type degrees-of-freedom approximation incorporates the variability of the variance estimator and thus performs better compared

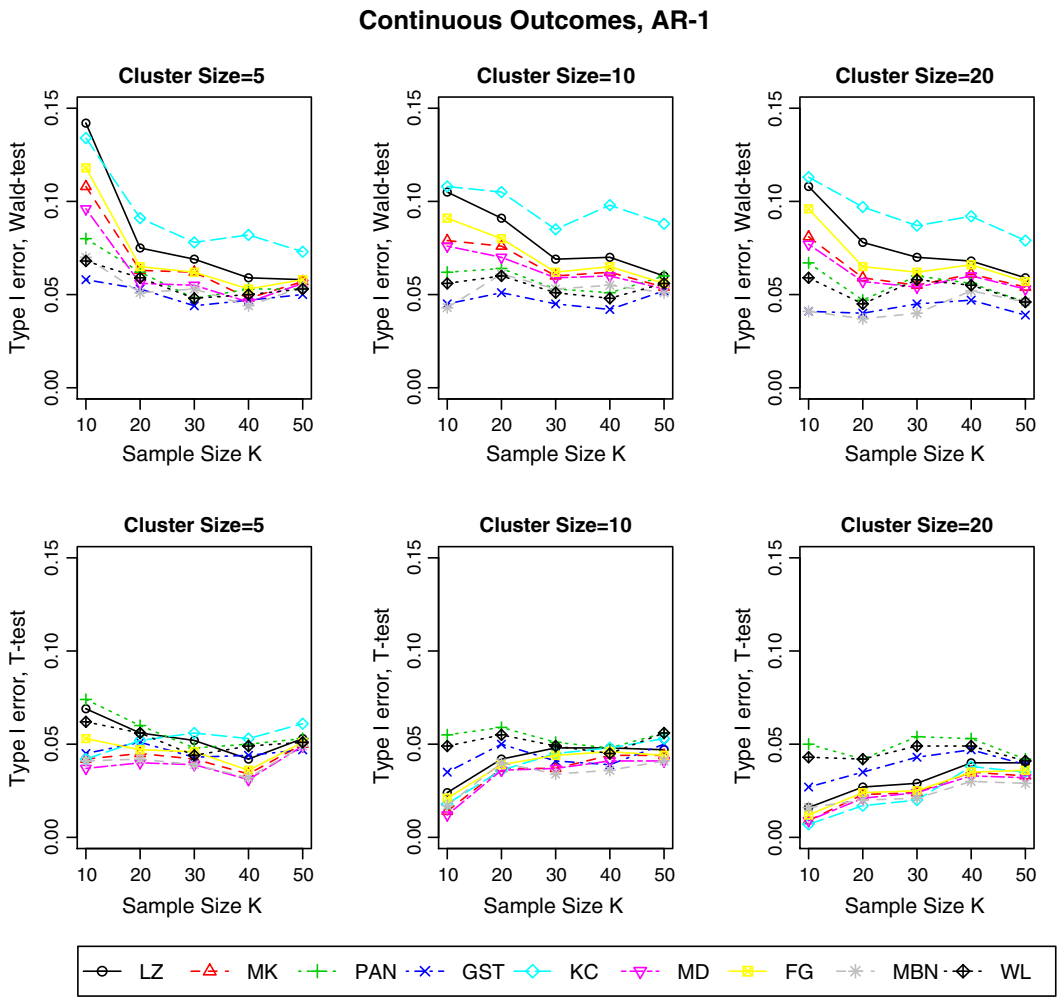

Figure 2. Type I errors based on Wald and *t*-tests for continuous outcomes with the true correlation structure as AR-1.

with depending only on the number of clusters in Li and Redden [18]. The outperformance of *t*-test over the Wald test was identified in the settings with small sample [16, 29].

3. Simulation studies

In this section, we conduct simulation studies to numerically compare the finite small-sample performance of nine types of variance estimators including the original "sandwich" variance estimator under different settings. Moreover, we focus on the Wald test and *t*-test for hypothesis testing to calculate the type I error rate for each estimator and further provide the recommendation on suitable sample size for each one to ensure test sizes at the nominal levels. In particular, three scenarios with continuous, count, and binary repeated outcomes are considered. The models for data generation are

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1 \times x_{ij} + b_i + \epsilon_{ij} \\ log(u_{ij}|b_i) &= \beta_0 + \beta_1 \times x_{ij} + b_i \\ logit(u_{ij}|b_i) &= \beta_0 + \beta_1 \times x_{ij} + b_i \end{aligned} \tag{20}$$

where $\beta_0 = 0$ and $\beta_1 = 0$, $i = 1, \ldots, K$ with sample size $K = 10, 20, 30, 40, 50$ and $j = 1, \ldots, n$ with equal number of observations within-subject (i.e., cluster size) $n = 5, 10, 20$. The covariate x_{ij} is independent and identical distributed (i.i.d) from a standard normal distribution $N(0, 1)$. The subject-level random effects b_i's are i.i.d. from $N(0, \sigma_b^2)$ with $\sigma_b^2 = 0.25$, and the random error ϵ_{ij}'s are i.i.d. from $N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon^2 = 0.8$. The details for each scenarios are listed as follows: (1) For the case with continuous outcomes,

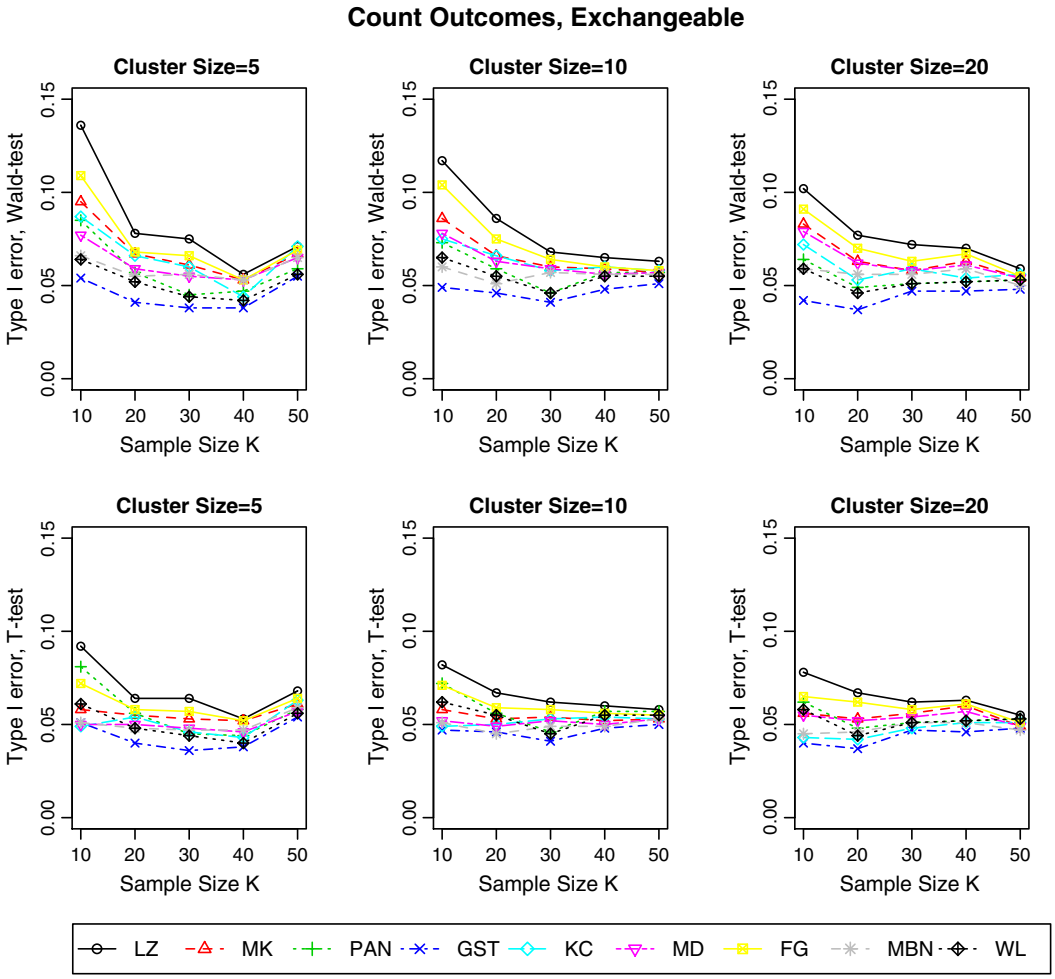

Figure 3. Type I errors based on Wald and *t*-tests for count outcomes with the true correlation structure as exchangeable.

b_i and ϵ_{ij} are independent with each other, leading to the true exchangeable correlation structure with the correlation parameter as $\alpha = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2) \approx 0.2$; (ii) For the case with count outcomes, based on the derivation by Guo *et al.* [28], the correlation parameter $\alpha \approx \sigma_b^2/(1+\sigma_b^2) \approx 0.3$; and (iii) For the case with binary outcomes, the correlation parameter $\alpha \approx \frac{\sigma_b^2/16}{E(\frac{1}{1+exp(-b_i)})[1-E(\frac{1}{1+exp(-b_i)})]} \approx 0.1$ according to Guo *et al.* [28].

In particular, 1000 Monte Carlo data sets are generated for each scenario, where the parameter estimate $\hat{\beta}_1$ along with nine variance estimates are calculated. For each set-up, three types of "working" correlation structures are used: independence, exchangeable, and AR-1. The Wald and *t*-tests are both applied for hypotheses testing, and type I error is calculated given the significance level of 0.05. Note that the degrees of freedom for *t*-distribution vary across different variance estimators. For example, the average degrees of freedom for the first scenario with continuous outcomes, $K = 10$ and $n = 5$, are rounded by 13, 13, 69, 69, 11, 14, 14, 22, 54, respectively, indicating the variability influence of variance estimators on statistical inference.

The partial results are shown in Figures 1–6, are summarized as follows: (i) The results based on Wald tests show that the use of robust variance estimator V_{LZ} always leads to inflated type I error when the sample size is small (i.e., $\leqslant 50$), which is consistent with our expectation; however, the tests using the other estimators also have inflation to some extent, but the degrees of freedom are smaller with V_{WL} performing the best; (ii) *t*-tests for hypotheses testing attain better performance than Wald tests in terms of the control of type I error across all estimators. The estimator V_{LZ} still leads to some degree of inflation. Interestingly, when the "working" correlation structure is specified correctly, V_{LZ} achieves satisfactory performance even though the sample size is as small as 10; (iii) For *t*-tests, the performance of variance estimators is substantially influenced by sample size K, while larger cluster size n leads to more con-

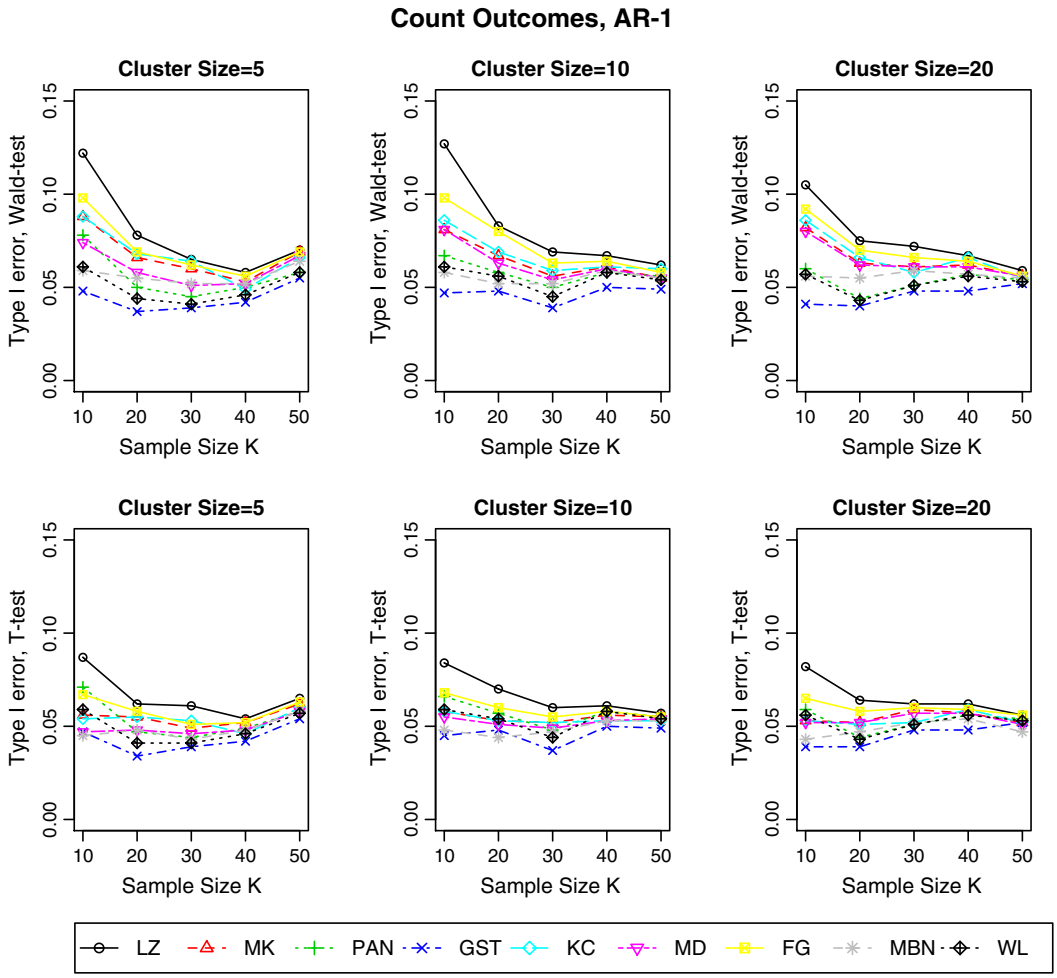

Figure 4. Type I errors based on Wald and *t*-tests for count outcomes with the true correlation structure as AR-1.

servative results; (iv) Note that V_{KC} performs worse than V_{LZ} based on Wald tests as indicated by larger inflation on type I error, but improves with increasing cluster size; in addition, some estimators, such as V_{GST} and V_{MBN}, perform conservatively for small samples; and (v) Among all nine variance estimators, V_{WL} has superior performance consistently across a variety of setups. Thus, it is a preferable estimator for GEE even when the sample size is as small as 10. Note that the results on the independent "working" correlation structure are not provided due to the similar trend as AR-1. In the end, according to our current numerical studies as well as literature [16, 28, 29], we recommend the sample size requirements to preserve type I error for all variance estimators as follows: $V_{LZ}(\geqslant 50)$, $V_{MK}(\geqslant 40)$, $V_{KC}(\geqslant 50)$, $V_{PAN}(\geqslant 30)$, $V_{GST}(\geqslant 20)$, $V_{MD}(\geqslant 30)$, $V_{FG}(\geqslant 40)$, $V_{MBN}(\geqslant 50)$, and $V_{WL}(\geqslant 10)$. Note that we also investigated the effect of cluster sizes via additional simulations (results available upon request) and found out that the higher cluster size n can somewhat improve the performance in preserving type I error, but the effect is not as substantial as the sample size K. Because of the fact that in most practical longitudinal designs, the cluster size (i.e., the number of observations within-subject) is usually less than 30 [30, 31], thus our recommendation can be applied in general cases (i.e., $n \geqslant 5$) based on current extensive simulations.

4. Data examples with small samples

In this paper, we present the results using two real data applications to test our R program and compare the finite performances of different variance estimators under finite sample size: one with continuous outcomes and the other one with count outcomes. The first data are from a study of orthodontic measurements on children, which includes 11 girls and 16 boys measured at ages 8, 10, 12, and 14 years [32]. The response is the measurement of the distance (in millimeters) from the center of the pituitary to the pteryomaxillary fissure, and the primary covariates of interest are age (in years) and gender (male;

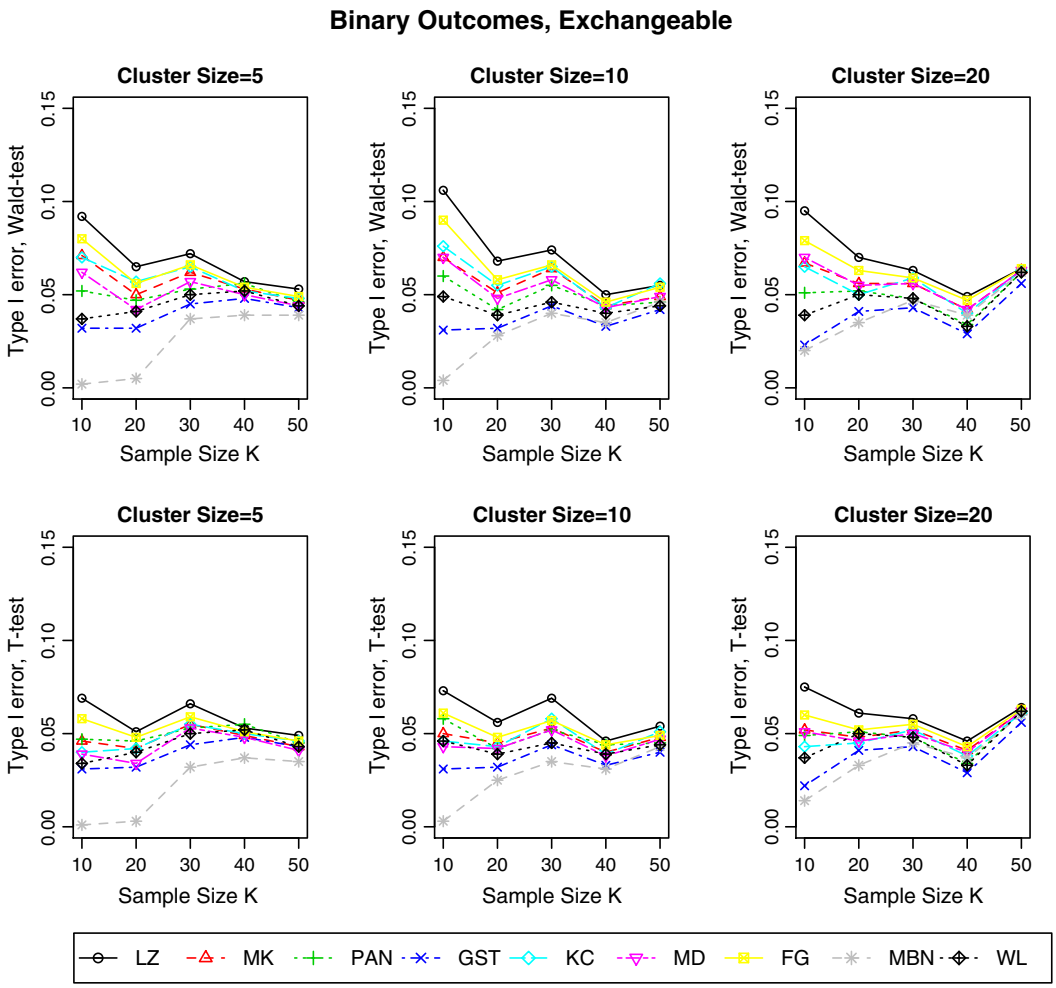

Figure 5. Type I errors based on Wald and *t*-tests for binary outcomes with the true correlation structure as exchangeable.

female). The objective is to investigate whether there exist statistically significant gender differences in dental growth measurements and their temporal trends as age increases. This example has been analyzed by Wang and Long [16] for small-sample properties of several estimators. Here, we conduct comparisons by considering eight types of modified variance estimators in addition to the robust original "sandwich" estimator. Therefore, the mean model of GEE is given by

$$E(y) = \beta_0 + \beta_1 \times \sqrt{Age} + \beta_2 \times Gender \tag{21}$$

The scatter plot of orthodontic measurements is shown in Figure 7, where the black lines are for girls and the red lines are for boys. It turns out that the boys have higher measurements than the girls on average, and the measurements tend to increase with age. GEE analysis results, including parameter estimates and various variance estimators, are shown in Table III. Both Wald and *t*-tests with the significance levels of 0.01 and 0.05 are applied for hypotheses testing. All variance estimators provide comparable results on hypotheses testing of $\sqrt{Age}$ using Wald tests, but when using *t*-tests at the significance level of 0.01, different testing conclusions for gender are obtained, indicating that the choice of different small-sample adjustments in variance estimators may affect the testing results.

The second example is from the randomized trial of progabide consisting of 59 individuals [33]. The subjects were randomly assigned to receive the anti-epileptic treatment (progabide) or placebo (control). The outcome is the number of epileptic seizures in each of four consecutive 2-week intervals, and the variables recorded include age and baseline epileptic seizure counts (in an 8-week interval) prior to the treatment assignment and the indicator for treatment (Trt, 1=progabide; 0=control). In particular, for modeling fitting, the variable *Baseline* is noted by the baseline epileptic seizure count rate per week;

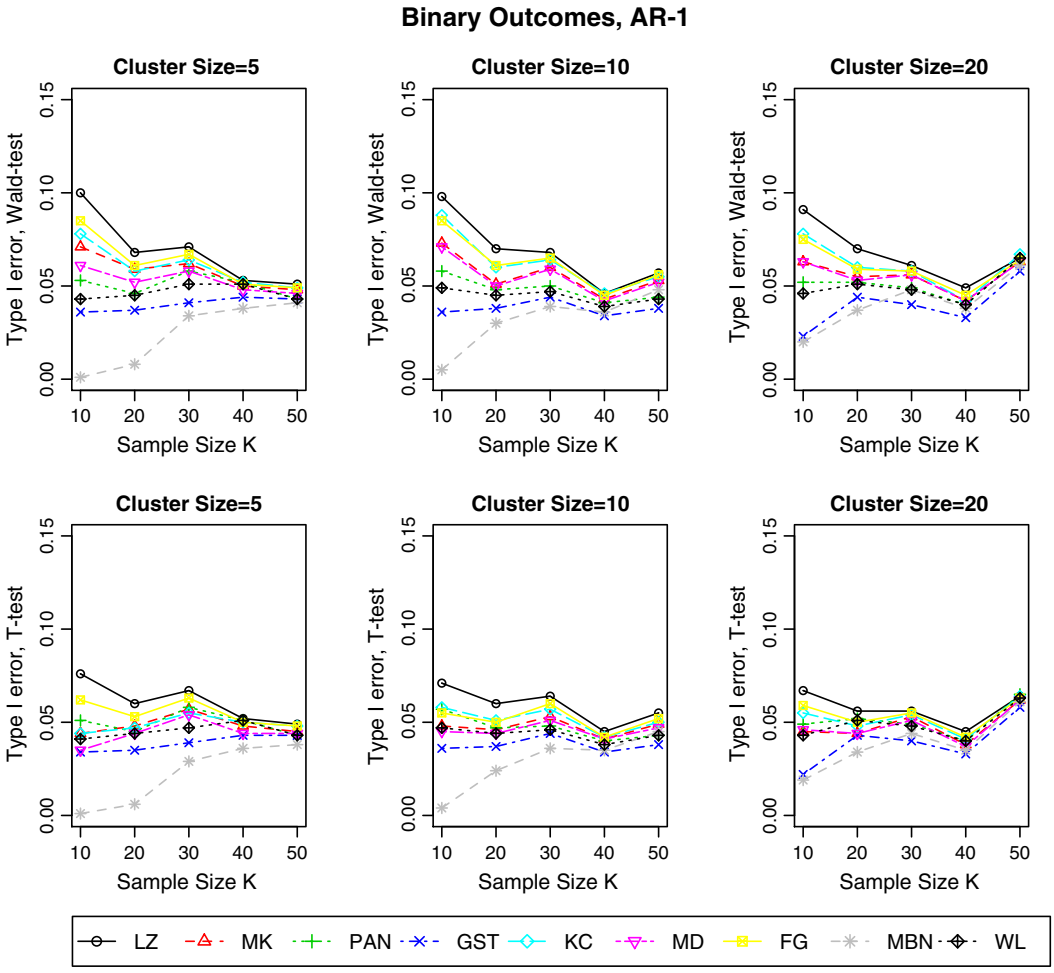

Figure 6. Type I errors based on Wald and *t*-tests for binary outcomes with the true correlation structure as AR-1.

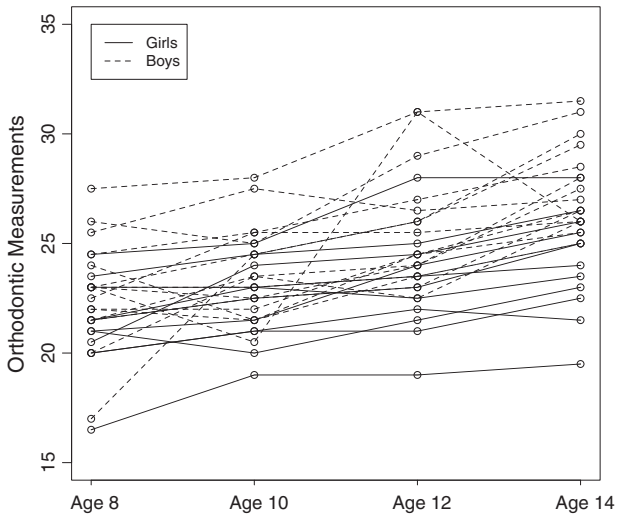

Figure 7. Orthodontic measurements by subject over time.

Table III. Estimation results for the case study of orthodontic measurements.

		$\sqrt{\text{Variance estimator}}$								
	$\hat{\beta}$	$\sqrt{V_{LZ}}$	$\sqrt{V_{MK}}$	$\sqrt{V_{PAN}}$	$\sqrt{V_{GST}}$	$\sqrt{V_{KC}}$	$\sqrt{V_{MD}}$	$\sqrt{V_{FG}}$	$\sqrt{V_{MBN}}$	$\sqrt{V_{WL}}$
					Independence					
$\sqrt{Age}$	4.32	0.461	0.489	0.461	0.489	0.470	0.479	0.486	0.574	0.479
Gender(=M)	2.32	0.750	0.795	0.733	0.777	0.782	0.816	0.836†	0.803	0.793
					Exchangeable					
$\sqrt{Age}$	4.32	0.461	0.489	0.461	0.489	0.470	0.479	0.469	0.498	0.479
Gender(=M)	2.32	0.750	0.795	0.733	0.777	0.782	0.816	0.836‡	0.818	0.793
					AR-1					
$\sqrt{Age}$	4.25	0.480	0.509	0.480	0.509	0.472	0.499	0.525	0.538	0.499
Gender(=M)	2.41	0.754	0.800	0.734	0.778	0.783	0.821	0.841†	0.813	0.795
					Unstructured					
$\sqrt{Age}$	4.27	0.466	0.495	0.466	0.495	0.516	0.484	0.473	0.507	0.484
Gender(=M)	2.22	0.730	0.774‡	0.713	0.756	0.783‡	0.795‡	0.814†	0.794‡	0.772

All are significant at both significance levels of 0.01 and 0.05 using Wald or *t*-tests except the ones with the superscripts.
† Not significant based on either Wald or *t*-tests at the significance level of 0.01.
‡ Significant based on Wald tests but not on *t*-tests at the significance level of 0.01.

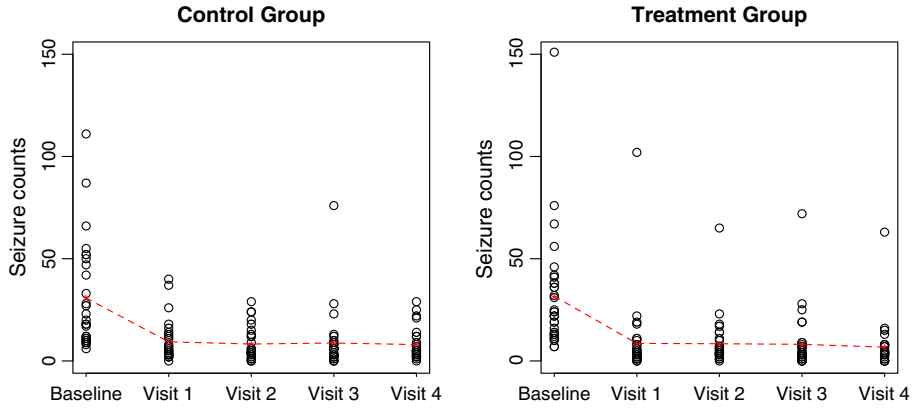

Figure 8. Seizure counts over time for treatment and control groups. The dotted red line is the average number of seizure counts over time.

Time is the number of weeks, which is valued by 2, 4, 6, and 8; $Interval_duration$ is the duration of each interval (i.e., 2 weeks), and $log(Interval\_duration)$ is treated as an offset variable in the model. The goal of this trial is to evaluate whether the anti-epileptic treatment is effective. We use the complete data set of all 59 subjects and a subset of 30 children, which are randomly drawn from the original complete data without replacement, to perform hypotheses testing and evaluate the small-sample properties of different estimators. Note that the interaction term of *Trt* and *Time* is also investigated but is not significant; thus, the final log-linear model for this study is given by

$$log\,(E(y)) = \beta_0 + \beta_1 \times Baseline + \beta_2 \times Trt + \beta_3 \times Time + log(Interval\_duration) \tag{22}$$

The scatter plots of seizure counts by time intervals for progabide and control groups are shown in Figure 8 and indicate that the counts dramatically decrease after the treatment in the first 2 weeks and remain stable afterwards for both groups. The GEE-based parameter estimates as well as the square root of various variance estimates are shown in Table IV. No significant (progabide) treatment effect or temporal trend is detected using either complete data or subset data based on all variance estimators, but *Baseline* has a significant effect on the seizure counts throughout. However, for subset data analysis, only slightly different conclusions of significance on temporal trend are obtained depending on the type of tests and the significance level. For instance, the tests of temporal effect using V_{GST}, V_{MD}, and V_{WL} are significant at the significance level of 0.05 but not at the significance level of 0.01 based on Wald or *t*-tests, but V_{KC} and V_{FG} are significant only based on Wald tests at the significance level of 0.05. This data

Table IV. Estimation results for the case study of epileptic seizures.

		$\sqrt{\text{Variance estimator}}$								
	$\hat{\beta}$	$\sqrt{V_{LZ}}$	$\sqrt{V_{MK}}$	$\sqrt{V_{PAN}}$	$\sqrt{V_{GST}}$	$\sqrt{V_{KC}}$	$\sqrt{V_{MD}}$	$\sqrt{V_{FG}}$	$\sqrt{V_{MBN}}$	$\sqrt{V_{WL}}$
Complete data ($K = 59$)										
					Independence					
Baseline	0.17	0.008	0.009	0.013	0.014	0.009	0.010	0.009	0.009	0.014
Trt(= 1)†	−0.23	0.176	0.182	0.159	0.165	0.181	0.191	0.180	0.182	0.167
Time†	−0.03	0.017	0.018	0.016	0.016	0.018	0.018	0.018	0.019	0.016
					Exchangeable					
Baseline	0.17	0.008	0.009	0.013	0.014	0.009	0.010	0.009	0.009	0.014
Trt(= 1)†	−0.22	0.174	0.180	0.158	0.164	0.182	0.189	0.179	0.181	0.166
Time†	−0.03	0.017	0.018	0.016	0.016	0.019	0.018	0.018	0.018	0.016
					AR-1					
Baseline	0.17	0.008	0.009	0.012	0.013	0.010	0.010	0.009	0.009	0.013
Trt(= 1)†	−0.25	0.166	0.172	0.149	0.155	0.185	0.182	0.171	0.173	0.157
Time†	−0.03	0.017	0.018	0.015	0.016	0.018	0.018	0.017	0.018	0.015
Subset data ($K = 30$)										
					Independence					
Baseline	0.18	0.014	0.015	0.017	0.018	0.014	0.026	0.015	0.015	0.019
Trt(= 1)†	-0.29	0.270	0.290	0.257	0.276	0.293	0.302	0.286	0.289	0.290
Time	-0.05	0.020†	0.021$^{\#}$	0.019$^{\#}$	0.021$^{\#}$	0.023*	0.022$^{\#}$	0.023*	0.025†	0.020$^{\#}$
					Exchangeable					
Baseline	0.18	0.013	0.014	0.017	0.019	0.015	0.025	0.015	0.015	0.020
Trt(= 1)†	-0.27	0.288	0.310	0.261	0.281	0.296	0.325	0.305	0.311	0.295
Time	-0.05	0.020$^{\#}$	0.021$^{\#}$	0.019‡	0.021$^{\#}$	0.021*	0.022$^{\#}$	0.021*	0.022†	0.020$^{\#}$
					AR-1					
Baseline	0.18	0.013	0.014	0.017	0.018	0.020	0.025	0.015	0.015	0.019
Trt(= 1)†	-0.34	0.279	0.299	0.255	0.274	0.296	0.314	0.295	0.299	0.289
Time	-0.05	0.023$^{\#}$	0.025‡	0.021$^{\#}$	0.023$^{\#}$	0.022*	0.026$^{\#}$	0.026*	0.026†	0.022$^{\#}$

Complete data: All are significant at both significance levels of 0.05 and 0.01 using Wald or *t*-tests except the ones with the superscripts.

Subset data: All are significant at both significance levels of 0.05 and 0.01 using Wald or *t*-tests except the ones with the superscripts.

Note that if the notation is put with the variable, the significance result is the same for the whole row.

† Not significant on either Wald or *t*-tests at the significance level of either 0.05 or 0.01.

‡ Significant on Wald test, but not significant on *t*-test at the significance level of 0.01.

$^{\#}$ Significant at the significance level of 0.05 but not at the significance level of 0.01 based on Wald or *t*-tests.

* Significant only based on Wald tests at the significance level of 0.05.

example shows that when the sample size is smaller (i.e., $\leqslant 30$), the validity of hypothesis testing could be influenced by the bias of the variance estimators.

5. Conclusions and discussions

In this paper, we provide a systematic review of recent developments on modified variance estimators for GEE to improve finite small-sample properties, including the formulation of these modifications and their theoretical and numerical comparisons. In addition, to conveniently implement these modifications, we develop the R package "geesmv", which is available at http://cran.r-project.org/web/packages/geesmv/ for free download and public access. We also discuss two main types of hypothesis testing for GEE, Wald and *t*-tests, and evaluate their corresponding type I error when sample size is small. Through extensive simulation studies and two real data examples, we compared the performance of various variance estimators under different scenarios and provide the guidance of the appropriate sample size for controlling type I error. As indicated in our simulation study, in general, *t*-tests based on the variance estimator V_{ML} perform robustly well across different set-ups. In particular, the degrees of freedom for *t*-statistic are more accurately approximated as compared with Li and Redden [18]. However, there are still several limitations

for this work. First, the modifications discussed here for variance estimation are directly focusing on the "sandwich" variance estimator, but some other methods were also proposed but not covered here (i.e., improving the efficiency and robustness of parameter estimates) [28, 29, 34–36]; Second, our recommendation on the appropriate sample sizes for each estimator for preserving type I error is obtained through limited simulation studies under general set-ups (i.e., equal cluster sizes); however, this guideline may not be always applicable, for instance, the cases with unequal cluster sizes; Third, we only evaluate the type I error, but the type II error or the power warrants further investigations. It is also worth pointing out that the selection of an appropriate modification method relies on various aspects of the real application (i.e., study design or intra-subject correlation) [2, 4, 37].

In addition to modifications on variance estimators and test statistics, another important issue, power analysis, tends to be challenging, for example, in a cluster randomized trial with a small number of clusters [5, 38, 39]. Previous studies on sample size/power calculation included Liu and Liang [40], where the generalized score test was utilized to draw statistical inference and the resulting non-central Chi-square distribution of test statistic under the alternative hypothesis was derived. Afterwards, Shih provided an alternative formula on sample size and power calculation, which relied on Wald tests using the estimates of regression parameters and robust variance estimators [41]. This power analysis is valid only when the $V(\hat{\beta})$ is unbiased and asymptotic normality is satisfied. However, when K is small, the estimated power tends to be overestimated. Hence, the modification on the power estimation is necessary to guarantee its unbiasedness, where an approximated t-distribution could be considered. Moreover, the adjustment of power estimation incorporating the variance estimators in Section 2 for efficiency improvement is expected to be advantageous. In the application of GEE method, other issues such as model selection or missing data under the circumstance of small samples or even with informative cluster sizes are also of interest. Thus, novel methodologies are still necessary and urged to develop for GEE to accommodate various data features for valid inference in real applications.

Acknowledgements

The author was supported by a pilot grant and KL2 career grant from the Penn State Clinical and Translational Science Institute (CTSI). The project was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health (NIH), through Grant 5 UL1 RR0330184-04 and Grant 5 KL2 TR 126-4. The content is solely the responsibility of the author and does not represent the views of the NIH. The authors thank Nicholas Sterling for correcting grammar errors during the revision of the manuscript.

References

1. Feng ZD, Diehr P, Peterson A, McLerran D. Selected statistical issues in group randomized trials. *Annual Review of Public Health* 2001; **22**:167–187.
2. Diggle P, Heagerty P, Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford University Press: Oxford, UK, 2002.
3. Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G. *Longitudinal Data Analysis*. Chapman and Hall/CRC: Boca Raton, Florida, 2008.
4. Hedeker D, Gibbons RD. *Analysis of Longitudinal Data*. John Wiley & Sons: New York, 2006.
5. Friedman LM, Furberg CD, DeMets DL. *Fundamentals of Clinical Trials* (3rd edn). Springer: New York, 1989.
6. McCullagh P, Nelder JA. *Generalized Linear Models*. Chapman and Hall: London, 1989.
7. Liang KY, Zeger SL. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrika* 1986; **73**:13–22.
8. Crowder M. On the use of a working correlation matrix in using generalized linear model for repeated measures. *Biometrika* 1995; **82**:407–410.
9. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika* 1974; **61**:439–447.
10. Hardin JW, Hilbe JM. *Generalized Estimating Equations*. Chapman and Hall/CRC Press: Boca Raton, FL, 2003.
11. Paik MC. Repeated measurement analysis for nonnormal data in small samples. *Communications in Statistics: Simulations* 1988; **17**(4):1155–1171.
12. Feng Z, McLerran D, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. *Statistics in Medicine* 1996; **15**:1793–1806.
13. Cox DR, Hinkley DV. *Theoretical Statistics*. Chapman and Hall: London, 1974.
14. Gunsolley JC, Getchell C, Chinchilli VM. Small sample characteristics of generalized estimating equations. *Communications in Statistics-Simulations* 1995; **24**:869–878.
15. Qu Y, Piedmonte MR, Williams GW. Small sample validity of latent variable models for correlated binary data. *Communications in Statistics: Simulations* 1994; **23**:243–269.

16. Wang M, Long Q. Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in Medicine* 2011; **30**(11):1278–1291.
17. Sharples K, Breslow N. Regression analysis of correlated binary data: some small sample results for the estimating equation approach. *Journal of Statistical Computation and Simulation* 1992; **42**:1–20.
18. Li P, Redden DT. Small sample performance of bias-corrected sandwich estimators for cluster-randomized trials with binary outcomes. *Statistics in Medicine* 2015; **34**(2):281–296.
19. Fay MP, Graubard BI, Freedman LS, Midthune DN. Conditional logistic regression with sandwich estimators: application to a meta analysis. *Biometrics* 1998; **54**:195–208.
20. Pan W. On the robust variance estimator in generalized estimating equations. *Biometrika* 2001; **88**:901–906.
21. Qu A, Lindsay B, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**:823–836.
22. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
23. MacKinnon JG. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* 1985; **29**:305–325.
24. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; **96**:1387–1398.
25. Gosho M, Sato Y, Takeuchi H. Robust covariance estimator for small-sample adjustment in the generalized estimating equations: a simulation study. *Science Journal of Applied Mathematics and Statistics* 2014; **2**(1):20–25.
26. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* 2001; **57**:1198–1206.
27. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. *Biometrical Journal* 2003; **45**(4):395–409.
28. Guo X, Pan W, Connett JE, Hannan PJ, French SA. Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in Medicine* 2005; **24**:3479–3495.
29. Pan W, Wall MM. Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine* 2002; **21**:1429–1441.
30. Ma Y, Mazumdar M, Memtsoudis SG. Beyond repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Regional Anesthesia and Pain Medicine* 2012; **37**(1):99–105.
31. Locascio JJ, Atri A. An overview of longitudinal data analysis methods for neurological research. *Dement Geriatr Cogn Discord Extra* 2011; **1**:330–357.
32. Potthoff RF, Roy SW. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* 1964; **51**:313–326.
33. Thall PF, Vail SC. Some covariance models for longitudinal count data with overdispersion. *Biometrics* 1990; **46**:657–671.
34. Lipsitz ST, Laird NM, Harrington DP. Using the jackknife to estimate the variance of regression estimators from measure studies. *Communications in Statistics: Theory and Methods* 1990; **19**:821–845.
35. Sherman M, Le Cessie S. A comparison between bootstrap methods and generalized linear model. *Communications in Statistics: Simulations* 1997; **26**:901–925.
36. Wang L, Zhou J, Qu A. Penalized generalized estimating equations for high-dimensional longitudinal data analysis. *Biometrics* 2011; **68**(2):353–360.
37. Fitzmaurice G, Larid NM, Ware JH. *Applied Longitudinal Data*. John Wiley & Sons: New York, 2004.
38. Shuster JJ. *Practical Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press: Boca Raton FL, 1993.
39. Teerenstra S, Lu B, Preisser JS, van Achterberg T, Borm GF. Sample size considerations of GEE analyses of three-level cluster randomized trials. *Biometrics* 2010; **66**:1230–1237.
40. Liu G, Liang KY. Sample size calculations for studies with correlated observations. *Biometrics* 1997; **53**:937–947.
41. 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.