Statistics
in Medicine

Research Article

Received 18 July 2012, Accepted 27 November 2012 Published online 16 December 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5709

A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix

Philip M. Westgate[*†]

Generalized estimating equations (GEEs) are routinely used for the marginal analysis of correlated data. The efficiency of GEE depends on how closely the working covariance structure resembles the true structure, and therefore accurate modeling of the working correlation of the data is important. A popular approach is the use of an unstructured working correlation matrix, as it is not as restrictive as simpler structures such as exchangeable and AR-1 and thus can theoretically improve efficiency. However, because of the potential for having to estimate a large number of correlation parameters, variances of regression parameter estimates can be larger than theoretically expected when utilizing the unstructured working correlation matrix. Therefore, standard error estimates can be negatively biased. To account for this additional finite-sample variability, we derive a bias correction that can be applied to typical estimators of the covariance matrix of parameter estimates. Via simulation and in application to a longitudinal study, we show that our proposed correction improves standard error estimation and statistical inference. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: correlation structure; efficiency; generalized estimating equations; standard error; unstructured

1. Introduction

Correlated data arise often in practice and may occur for a variety of reasons. We focus on longitudinal settings in which correlated data arises because of subjects contributing multiple observations over time. For instance, we later utilize a data set, presented and discussed by Thall and Vail [1], that arose from a study in which 59 epileptic subjects contributed four repeated measurements, each at the end of consecutive 2-week periods, after being randomized to a treatment drug or a placebo in addition to standard chemotherapy. Of interest was the impact of the drug on the rate of seizures epileptics experience.

The method of generalized estimating equations (GEEs) [2] is very popular for the marginal analysis of correlated data, as only working mean and covariance structures need to be specified, whereas only the mean needs to be correctly given in order to obtain consistent parameter estimates. However, accurate modeling of the covariance structure, and thus correlation, of the data enhances efficiency. Specifically, utilizing the correct correlation structure is ideal, although it will be unknown in practice. Two popular correlation structures are therefore the exchangeable and AR-1, for example, as they are reasonable choices in many settings that arise in practice and involve the estimation of only one correlation parameter. The unstructured working correlation matrix is a popular alternative, particularly when the timing of repeated measurements is the same for each subject, as it does not use a fixed form for the correlation. Therefore, assuming marginal variances are correctly specified and all subjects have the same correlation structure [3], use of the unstructured correlation matrix theoretically is equally or more efficient than any other working structure. However, when subjects contribute a large number of observations, many

Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY 40536, U.S.A.
**Correspondence to: Philip M. Westgate, Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY 40536, U.S.A.*
†E-mail: philip.westgate@uky.edu

nuisance parameters must be estimated for the unstructured working correlation matrix, potentially leading to larger variances of resulting regression parameter estimates relative to estimates that are obtained from the use of a structured working correlation matrix (for instance, see Wang [4], Hin *et al.* [5], and Zhou and Qu [6]). Similarly, realized variances of the regression parameter estimates can be larger than theoretically expected, via the estimated covariance matrix of the regression parameter estimates, when utilizing the unstructured working correlation matrix. Therefore, use of an unstructured working correlation matrix can potentially result in inflated test sizes, as seen in Overall and Tonidandel [7] for example, and reduced coverage probabilities (CPs) from confidence intervals (CIs) due to the use of negatively biased standard error (SE) estimates.

We explain that the additional variability in the regression parameter estimates that is not taken into account by typical covariance matrix estimators is due to the use of estimated, rather than true, regression parameters inside the residuals that are utilized to estimate each correlation parameter. To obtain improved covariance and SE estimates, we approximate this additional variability by using a Taylor series expansion of these estimated regression parameters within the generalized estimating equations about the true parameters, leading to a correction that can be applied to any covariance matrix estimator.

We note that the use of the unstructured working correlation is not always applicable, as subjects do not always contribute the same number of observations. Furthermore, the timing of these repeated measurements may vary across subjects. In these scenarios, the necessary assumption of a common correlation structure for all subjects may not be reasonable, and parametric alternatives that can take into account both the number of and timing of observations should therefore be used to obtain working correlation matrices for each subject.

Section 2 introduces notation and discusses GEE, the unstructured working correlation matrix, and typical estimators of the covariance matrix of the regression parameter estimates. In Section 3, we derive our proposed bias correction. Via simulation study and in application to the seizures study, Section 4 demonstrates the necessity and utility of the proposed correction. Section 5 gives the concluding remarks.

2. Notation and estimation

Let N subjects contribute repeated measurements at the same n time points. For simplicity, we assume each subject contributes all n observations. We denote the vector of correlated outcomes from subject i, $i = 1,\ldots,N$, as $\boldsymbol{Y}_i = [Y_{i1},\ldots,Y_{in}]^T$, which has a marginal mean given by $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i)$. For a known link function, f, a vector of covariate values, $\mathbf{x}_{ij} = [1, x_{1ij},\ldots,x_{(p-1)ij}]^T$, and a $p \times 1$ vector of regression parameters, $\boldsymbol{\beta} = [\beta_0, \beta_1,\ldots,\beta_{p-1}]^T$, we have $f(\mu_{ij}) = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta}$, $j = 1,\ldots,n; i = 1,\ldots,N$. Let $\phi\boldsymbol{A}_i$ be a diagonal matrix of working marginal variances for outcomes from the ith subject, such that $\phi v(\mu_{ij})$, $j = 1,\ldots,n$ gives the jth diagonal element. Here, ϕ is assumed to be a common dispersion parameter, and v is a known function. Furthermore, the working correlation matrix is given as $\boldsymbol{R}(\boldsymbol{\alpha})$, in which $\boldsymbol{\alpha}$ denotes the nuisance correlation parameter(s) to be estimated. Therefore, $\boldsymbol{V}_i = \phi\boldsymbol{A}_i^{1/2}\boldsymbol{R}(\boldsymbol{\alpha})\boldsymbol{A}_i^{1/2}$ gives the working covariance structure for subject i.

With GEE, initial estimates of the regression parameters, $\tilde{\boldsymbol{\beta}}$, are specified and the final estimates, $\hat{\boldsymbol{\beta}}$, are found by solving

$$\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{1}$$

in which $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}^T$, using an iterative procedure in which $\tilde{\boldsymbol{\beta}}$ is continuously updated until convergence to $\hat{\boldsymbol{\beta}}$. Denoting the empirical correlation for the ith subject, $i = 1,\ldots,N$, as $\phi^{-1}\boldsymbol{E}_i(\boldsymbol{\beta}) = \phi^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^T = \phi^{-1}\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T\boldsymbol{A}_i^{-1/2}$, and the corresponding sample average as $\phi^{-1}\boldsymbol{E}_N(\boldsymbol{\beta}) = (N\phi)^{-1}\sum_{i=1}^{N}\boldsymbol{E}_i(\boldsymbol{\beta})$, Liang and Zeger [2] proposed the unstructured working correlation matrix, which can be estimated by

$$\boldsymbol{R}(\tilde{\boldsymbol{\alpha}}) = \frac{1}{\phi}\boldsymbol{E}_N(\tilde{\boldsymbol{\beta}}) \tag{2}$$

at any iteration of this procedure. Here, $\tilde{\boldsymbol{\beta}}$ is used in place of $\boldsymbol{\beta}$, and $\tilde{\boldsymbol{\alpha}}$ denotes the resulting estimate of $\boldsymbol{\alpha}$. Alternatively, a similar version of this working correlation matrix could replace the diagonal elements with 1.

Assuming that the working covariance structure for the data is correctly specified, the covariance of $\hat{\boldsymbol{\beta}}$ can be consistently estimated using the model-based estimator, $\widehat{\text{cov}}_{MB}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}$ [8]. This estimator will be biased if the covariance structure is misspecified, and therefore Liang and Zeger [2] proposed the empirical sandwich estimator, $\widehat{\text{cov}}_{E}(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \left[\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i)(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right] \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1}$, which is routinely utilized with the GEE estimation method as it gives a consistent estimate for the true covariance matrix even when the working covariance structure is misspecified. In both covariance estimators, unknown parameters are replaced with corresponding estimates. When there are a small number of subjects, say less than 40 or 50 [8,9], notable negative bias can exist in SE estimates obtained from $\widehat{\text{cov}}_{E}(\hat{\boldsymbol{\beta}})$ due to the use of estimated, rather than true, residuals to estimate $\text{cov}(\boldsymbol{Y}_i), i = 1, \ldots, N$, in the middle section of this estimator. Therefore, Mancl and DeRouen [8] and Kauermann and Carroll [10] derived small sample bias corrections, which replace $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$ with $(\boldsymbol{I}_n - \boldsymbol{H}_i)^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T \left(\boldsymbol{I}_n - \boldsymbol{H}_i^T\right)^{-\kappa}, i = 1, \ldots, N$, inside $\widehat{\text{cov}}_{E}(\hat{\boldsymbol{\beta}})$. Here, $\hat{\boldsymbol{\mu}}_i, i = 1, \ldots, N$, are estimated marginal means using $\hat{\boldsymbol{\beta}}$, $\boldsymbol{I}_n$ is an $n \times n$ identity matrix, and $\boldsymbol{H}_i = \boldsymbol{D}_i \widehat{\text{cov}}_{MB}(\hat{\boldsymbol{\beta}}) \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1}$. For the Mancl and DeRouen [8] correction, $\kappa = 1$, whereas $\kappa = 0$ with the Kauermann and Carroll [10] correction, which assumes the working covariance structure is correctly specified.

3. A bias correction for covariance estimators

When utilizing the unstructured working correlation matrix in practice, $\tilde{\boldsymbol{\beta}}$ replaces $\boldsymbol{\beta}$ in Equation (2) to estimate $\boldsymbol{\alpha}$, which can notably increase the finite-sample variance of each regression parameter estimate. This increase in variance is not accounted for by the four previously mentioned covariance matrix estimators, which therefore can be biased. Although any additional variability will be negligible when the number of correlation parameters is small, it can be notable when N is not arbitrarily large relative to n, as $n(n-1)/2$ correlation parameters must be estimated. If the true regression parameters were known, then for any given sample, the use of true standardized residuals, and thus true empirical correlations, $\phi^{-1} \boldsymbol{E}_i(\boldsymbol{\beta}), i = 1, \ldots, N$, in Equation (2) would be more informative than their expected values, or the true correlations, thus possibly resulting in greater precision for estimating $\boldsymbol{\beta}$. In practice, however, we must replace $\boldsymbol{\beta}$ with $\tilde{\boldsymbol{\beta}}$ in Equation (2), thus increasing estimation variability in Equation (1). This additional variability can be approximated by using first-order Taylor series expansions. First, conditioning on the use of $\tilde{\boldsymbol{\beta}}$ in Equation (2) and then replacing $\boldsymbol{V}_i^{-1}$ with $\boldsymbol{A}_i^{-1/2} \boldsymbol{E}_N^{-1}(\tilde{\boldsymbol{\beta}}) \boldsymbol{A}_i^{-1/2}, i = 1, \ldots, N$, in Equation (1), we have [2,8]

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{E}_N^{-1}(\tilde{\boldsymbol{\beta}}) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i). \tag{3}$$

Second, as $\tilde{\boldsymbol{\beta}}$ varies about $\boldsymbol{\beta}$, we utilize the following expansion of Equation (3):

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{E}_N^{-1}(\boldsymbol{\beta}) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) + \boldsymbol{G}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}), \tag{4}$$

in which the kth column of

$$\boldsymbol{G} = \frac{\partial}{\partial \boldsymbol{\beta}^{*T}} \left[\left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{E}_N^{-1}(\boldsymbol{\beta}^*) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i) \right] |_{\boldsymbol{\beta}^* = \boldsymbol{\beta}}$$

is given by

$$-\left(\sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{D}_i\right)^{-1} \sum_{i=1}^{N} \boldsymbol{D}_i^T \boldsymbol{A}_i^{-1/2} \boldsymbol{E}_N^{-1}(\boldsymbol{\beta}) \frac{\partial \boldsymbol{E}_N(\boldsymbol{\beta})}{\partial \beta_{k-1}} \boldsymbol{E}_N^{-1}(\boldsymbol{\beta}) \boldsymbol{A}_i^{-1/2} (\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

in which

$$\frac{\partial \boldsymbol{E}_N(\boldsymbol{\beta})}{\partial \beta_{k-1}} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \boldsymbol{e}_i}{\partial \beta_{k-1}}\boldsymbol{e}_i^T + \boldsymbol{e}_i\frac{\partial \boldsymbol{e}_i^T}{\partial \beta_{k-1}}\right).$$

Next, we approximate $\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}$ in Equation (4) with $\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)$, giving

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx (\boldsymbol{I}_p + \boldsymbol{G})\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{E}_N^{-1}(\boldsymbol{\beta})\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

in which $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. Therefore, the covariance formula using the derived correction is given as

$$\begin{aligned}\mathrm{cov}(\hat{\boldsymbol{\beta}}) &\approx \mathrm{cov}\left((\boldsymbol{I}_p + \boldsymbol{G})\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i)\right)\\ &\approx (\boldsymbol{I}_p + \boldsymbol{G})\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\mathrm{cov}(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}(\boldsymbol{I}_p + \boldsymbol{G})^T,\end{aligned}$$

in which the typical covariance formula is multiplied by $(\boldsymbol{I}_p + \boldsymbol{G})$ and $(\boldsymbol{I}_p + \boldsymbol{G})^T$ on its left and right sides, respectively, and the unstructured correlation matrix is incorporated in $\boldsymbol{V}_i, i = 1, \ldots, N$. Furthermore, the typical covariance formula can be estimated using one of the four estimators discussed in Section 2. However, because of our derived correction for the approximation of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, the Mancl and DeRouen [8] and Kauermann and Carroll [10] estimators now utilize $\boldsymbol{H}_i = \boldsymbol{D}_i(\boldsymbol{I}_p + \boldsymbol{G})\widehat{\mathrm{cov}}_{MB}(\hat{\boldsymbol{\beta}})\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}$. As an example, the covariance estimator that utilizes both the Mancl and DeRouen [8] correction and our proposed correction is given by $\widehat{\mathrm{cov}}_{CMD}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{I}_p + \boldsymbol{G})\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\left[\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}(\boldsymbol{I}_n - \boldsymbol{H}_i)^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T\left(\boldsymbol{I}_n - \boldsymbol{H}_i^T\right)^{-1}\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right] \times \left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}(\boldsymbol{I}_p + \boldsymbol{G})^T$.

For the unstructured working correlation matrix that has diagonal elements of 1, which therefore are not functions of the regression parameters, our proposed correction can be modified by replacing the diagonal elements of $\partial \boldsymbol{E}_N(\boldsymbol{\beta})/\partial \beta_{k-1}, k = 1, \ldots, p$, with 0. Furthermore, $\boldsymbol{E}_N(\boldsymbol{\beta})$ and $\boldsymbol{E}_N(\hat{\boldsymbol{\beta}})$ are consistent for the average of each subject's true correlation matrix [11]. Therefore, the bias in the uncorrected covariance estimators that is due to the use of the unstructured working correlation matrix will be negligible when N is large relative to n. Consequently, differences in the two versions, one that does and one that does not incorporate our correction, of a specific covariance estimator will be negligible for an arbitrarily large N.

4. Assessing the performance of the proposed covariance correction

4.1. Via simulation study

4.1.1. Study description. We assess the necessity and utility of the proposed covariance correction via simulation study in a variety of settings based upon two distinct scenarios. Settings vary by the number of subjects and the number of repeated measurements each subject contributes. Results are presented in Tables I and II, and each setting was examined via 1000 simulations.

For each setting, we compare the empirical means of estimated SEs to the corresponding empirical standard deviation (ESD) of each non-intercept parameter estimate when utilizing the unstructured working correlation matrix with GEE. For simplicity, we only present empirical means for the empirical sandwich SE estimators utilizing only the Mancl and DeRouen [8] correction, denoted by SE_{MD}, or both the Mancl and DeRouen [8] correction and our proposed correction, denoted by SE_{CMD}. We also present empirical CPs from corresponding 95% CIs which utilize critical values obtained from a t-distribution with $N - p$ degrees of freedom, as suggested by Mancl and DeRouen [8]. We do not

Table I. Empirical standard deviations (ESDs) of $\hat{\beta}_1$ and $\hat{\beta}_2$ (in bold) when using the unstructured working correlation matrix with GEE, and respective empirical mean standard error estimates from SE_{MD} and SE_{CMD} with corresponding empirical 95% confidence interval coverage probabilities (CP).

			$\hat{\beta}_1$		$\hat{\beta}_2$	
N	n	SE estimator	Estimate	CP	Estimate	CP
35	5	ESD	**0.47**		**0.14**	
		SE_{MD}	0.42	0.93	0.12	0.91
		SE_{CMD}	0.47	0.95	0.14	0.96
	10	ESD	**0.36**		**0.11**	
		SE_{MD}	0.27	0.88	0.08	0.84
		SE_{CMD}	0.34	0.94	0.10	0.93
50	5	ESD	**0.39**		**0.12**	
		SE_{MD}	0.36	0.93	0.11	0.92
		SE_{CMD}	0.39	0.95	0.12	0.95
	10	ESD	**0.29**		**0.08**	
		SE_{MD}	0.24	0.89	0.07	0.89
		SE_{CMD}	0.28	0.93	0.08	0.95
100	5	ESD	**0.27**		**0.08**	
		SE_{MD}	0.26	0.94	0.08	0.93
		SE_{CMD}	0.27	0.95	0.08	0.95
	10	ESD	**0.19**		**0.06**	
		SE_{MD}	0.17	0.93	0.05	0.91
		SE_{CMD}	0.19	0.95	0.06	0.94

Results are from scenario 1. N, number of independent subjects; n, number of repeated measurements per subject; SE_{MD}, sandwich standard error estimator using only the Mancl and DeRouen [8] correction; SE_{CMD}, sandwich standard error estimator using both the Mancl and DeRouen [8] correction and our proposed correction.

Table II. Empirical standard deviations (ESDs) of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ (in bold) when using the unstructured working correlation matrix with GEE, and respective empirical mean standard error estimates from SE_{MD} and SE_{CMD} with corresponding empirical 95% confidence interval coverage probabilities (CPs).

			$\hat{\beta}_1$		$\hat{\beta}_2$		$\hat{\beta}_3$		$\hat{\beta}_4$	
N	n	SE estimator	Estimate	CP	Estimate	CP	Estimate	CP	Estimate	CP
59	4	ESD	**0.16**		**0.12**		**0.36**		**0.031**	
		SE_{MD}	0.15	0.94	0.11	0.92	0.34	0.93	0.030	0.94
		SE_{CMD}	0.16	0.96	0.12	0.94	0.36	0.95	0.031	0.95
	10	ESD	**0.18**		**0.13**		**0.37**		**0.009**	
		SE_{MD}	0.14	0.89	0.11	0.86	0.31	0.90	0.008	0.89
		SE_{CMD}	0.17	0.93	0.14	0.93	0.39	0.94	0.009	0.93
100	4	ESD	**0.12**		**0.09**		**0.26**		**0.025**	
		SE_{MD}	0.12	0.94	0.09	0.93	0.26	0.94	0.023	0.93
		SE_{CMD}	0.12	0.95	0.09	0.93	0.27	0.95	0.023	0.94
	10	ESD	**0.12**		**0.09**		**0.27**		**0.007**	
		SE_{MD}	0.11	0.94	0.08	0.90	0.24	0.93	0.006	0.92
		SE_{CMD}	0.12	0.95	0.09	0.93	0.27	0.95	0.007	0.95

Results are from scenario 2. N, number of independent subjects; n, number of repeated measurements per subject; SE_{MD}, sandwich standard error estimator using only the Mancl and DeRouen [8] correction; SE_{CMD}, sandwich standard error estimator using both the Mancl and DeRouen [8] correction and our proposed correction.

present results for the unstructured working correlation matrix that has diagonal elements equal to 1 due to its instability in some simulated settings.

The marginal model utilized in scenario 1 is

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \epsilon_{ij}; \quad \epsilon_{ij} \sim N(0,5), \quad j = 1, \ldots, n,$$

where $x_{1ij} = j/n$ and $x_{2ij} \sim N(j/n, 1)$, $i = 1, \ldots, N$, similar to a scenario used by Qu *et al.* [12]. The number of subjects was either 35, 50, or 100, and the number of repeated measurements from each subject was either 5 or 10. The true correlation structure was exchangeable with a parameter value of 0.5 and $\boldsymbol{\beta} = [0, 0.5, 1]^T$. Table I presents the simulation results.

The marginal model utilized in scenario 2 is a representation of our applied example and is

$$log(\mu_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4ij}; \quad j = 1, \ldots, n,$$

where μ_{ij}, $j = 1, \ldots, n$, $i = 1, \ldots, N$ are marginal mean counts. The number of subjects was either 59 or 100, and the number of repeated measurements from each subject was either 4 or 10. For $N = 59$ and $N = 100$, x_{1i} is a subject-level indicator equal to 1 for 31 and 53 subjects, respectively. Furthermore, x_{2i} and x_{3i} were drawn independently across subjects from $N(1.75, 0.5)$ and Uniform$(2.9, 3.7)$, and $x_{4ij} = j$. The true correlation structure was exchangeable with a parameter value of 0.5, the marginal variance for the jth outcome from the ith subject is given as $4\mu_{ij}$ and $\boldsymbol{\beta} = [-2, 0, 1, 0.65, -0.05]^T$. Outcomes were generated using a method presented by Madsen and Dalthorp [13]. Table II presents the simulation results.

4.1.2. Description of results. Results confirm that for a finite number of subjects, with each contributing multiple repeated measurements, typical SE estimators can be negatively biased with the GEE approach that incorporates the unstructured working correlation matrix. Specifically, ratios of the empirical mean of SE_{MD} divided by the corresponding ESD ranged from 0.73 to 1.00 across all settings and parameters. Therefore, empirical CPs were notably less than the nominal value of 0.95 in numerous settings and ranged from 0.84 to 0.94. As expected, in each scenario, the bias in SE_{MD} was most notable in settings with the fewest number of subjects and the largest number of repeated measurements. However, increasing the number of subjects and decreasing the number of repeated measurements led to less bias in the typical uncorrected SE estimators. For instance, when $N = 100$ in either scenario, bias in SE_{MD} was negligible when $n = 4$ or $n = 5$, although small bias could still be evident when $n = 10$.

Use of our proposed correction consistently worked well at approximating the bias in SE_{MD} and therefore improved inference overall. Specifically, SE_{CMD} always led to unbiased or negligibly biased estimates for the SEs of parameter estimates, therefore resulting in near-nominal empirical CPs that ranged from 0.93 to 0.96 across all settings and parameters. Results also confirm that our proposed correction can be utilized even when SE_{MD} contains little or no bias, as the impact of our correction was negligible in such settings. Furthermore, the findings of this study are independent of the true correlation structure. For instance, similar results (not shown) were found when the true correlation structure was AR-1 in the studied settings.

4.2. *Via application*

We now give focus to the longitudinal study in which 59 epileptic subjects, 31 of whom were randomized to receive the drug progabide, contributed data at each of the same four time points [1]. Therefore, we are able to incorporate the unstructured working correlation matrix with GEE to fit the following marginal model, which is the model fit by Song [14] and Song *et al.* [15] with and without an outlying subject, and is the basis for scenario 2 in the simulation study:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4ij}; \quad j = 1, 2, 3, 4.$$

Here, μ_{ij} is the mean number of seizures in the jth 2-week interval for subject i, x_{1i} is the subject-level indicator equal to 1 if subject i received progabide, x_{2i} is the natural log of one quarter of the number of seizures the ith subject had during the 8 weeks preceding randomization, x_{3i} is the natural log of the ith subject's age, and $x_{4ij} = j$ denotes time.

Table III presents the parameter estimates and their corresponding SE estimates, SE_{MD} and SE_{CMD}, with 95% CIs. Results are from analyses of the data set that either include or exclude one subject who had outlying numbers of seizures [14, 15]. When analyzing the full data set and comparing SE_{CMD} with

Table III. Results from analyses, including and excluding observations from the subject with outlying numbers of seizures, of the epileptic seizures data set.

	β_1		β_2		β_3		β_4	
	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
$\hat{\beta}$	−0.07		1.21		0.53		−0.068	
SE_{MD}	0.27	(−0.61, 0.47)	0.27	(0.67, 1.75)	0.31	(−0.10, 1.16)	0.030	(−0.129, −0.007)
SE_{CMD}	0.28	(−0.63, 0.50)	0.30	(0.60, 1.81)	0.32	(−0.12, 1.17)	0.031	(−0.129, −0.006)
$\hat{\beta}^*$	−0.31		0.95		0.76		−0.056	
SE^*_{MD}	0.15	(−0.61, −0.02)	0.08	(0.79, 1.12)	0.27	(0.22, 1.30)	0.033	(−0.122, 0.011)
SE^*_{CMD}	0.15	(−0.62, −0.01)	0.08	(0.78, 1.12)	0.30	(0.16, 1.35)	0.034	(−0.123, 0.012)

The unstructured working correlation matrix was incorporated in GEE. Presented are parameter estimates ($\hat{\beta}$), standard error estimates from SE_{MD} and SE_{CMD} and corresponding 95% confidence intervals (CIs) using critical values from a t-distribution with 54 degrees of freedom. *Analysis excludes observations from the subject with outlying numbers of seizures. SE_{MD}, sandwich standard error estimator using only the Mancl and DeRouen [8] correction; SE_{CMD}, sandwich standard error estimator using both the Mancl and DeRouen [8] correction and our proposed correction.

SE_{MD}, our proposed correction estimates that standard errors of parameter estimates are slightly larger than theoretically estimated, especially for $\hat{\beta}_2$. However, when excluding the observations from the subject with outlying numbers of seizures, differences between SE_{CMD} and SE_{MD} are negligible except with respect to the standard error estimate for $\hat{\beta}_3$. This result supports the findings of our simulation study in scenario 2, as only slight to negligible biases in SE_{MD} were evident in results in which $N = 59$ and $n = 4$. However, simulation results and estimates in the analyses of this data set suggest that our correction should still be used, as it will appropriately correct even small biases. Furthermore, without the proposed correction, we could not estimate how much bias, if any, exists in a standard error estimator because of the need for estimating the nuisance correlation parameter estimates within the unstructured working correlation matrix.

5. Concluding remarks

When incorporating the unstructured working correlation matrix with the GEE approach, the variances, and thus standard errors, of regression parameter estimates can be larger than theoretically estimated by typical standard error estimators when the number of subjects is not large relative to the number of repeated measurements each subject contributes. This variance inflation is due to the necessary estimation of a potentially large number of nuisance correlation parameters. Specifically, the use of estimated, rather than true, regression parameters to obtain the unstructured working correlation matrix creates additional variability in the estimating equations and thus the final regression parameter estimates. We therefore developed a bias correction that can be applied with typical covariance formulas. To our knowledge, the proposed correction and the reasoning behind its derivation have not been given in the GEE literature. As demonstrated in Section 4, the proposed correction results in greatly improved SE estimation and thus improved statistical inference.

Although implementation of the proposed correction is not always necessary when using the unstructured working correlation matrix, we do recommend that the correction be routinely utilized in practice as knowledge of when the non-corrected standard error estimators are non-negligibly biased will be limited. The proposed correction will only negligibly inflate standard error estimates when the correction is not needed and will appropriately adjust estimators in settings in which they would otherwise be notably biased. The proposed correction also allows us to obtain more appropriate standard error estimates for improving inference, rather than just subtracting additional degrees of freedom for estimating numerous nuisance parameters [7].

Choice of the working correlation structure with GEE is a popular research area with respect to improving estimation efficiency. However, use of the unstructured working correlation matrix has received little attention with respect to proposed methods, as many more nuisance parameters must be estimated relative to structures such as the exchangeable and AR-1 [16]. However, we have shown that the impact of estimating these nuisance parameters should be accounted for by inflating the typical estimators of the covariance matrix of the regression parameter estimates. Therefore, future research with respect to choosing a working correlation structure should be carried out with respect to utilizing our proposed correction, thus allowing focus to also be given to the unstructured correlation matrix.

One limitation of our proposed correction is that it requires the unstructured working correlation matrix to be stable. However, when the number of subjects is not large relative to the number of repeated measurements, the unstructured working correlation matrix, and thus regression parameter estimates, can be unstable [17]. For instance, although stable results were observed in our simulation study, results (not shown) from the use of the unstructured working correlation matrix version with diagonal elements equal to 1 were sometimes unstable, leading to ESDs and SE estimates that were very large. However, we found that our correction with this version of the unstructured working correlation matrix also works well when estimates are stable.

The proposed correction, with its theoretical aspects, is similar to work performed for generalized method of moments and the quadratic inference function method [12, 18–21] for the marginal analysis of correlated data. These methods involve the use of empirical covariances within their respective estimating equations, resulting in a similar type of bias and correction method as demonstrated in the present work with GEE and the unstructured correlation matrix. Future work is needed to incorporate this type of correction with the quadratic inference function method when utilizing a working unstructured correlation structure.

References

1. Thall PF, Vail SC. Some covariance models for longitudinal count data with overdispersion. *Biometrics* 1990; **46**:657–671.
2. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
3. Pan W. On the robust variance estimator in generalised estimating equations. *Biometrika* 2001; **88**:901–906.
4. Wang L. GEE analysis of clustered binary data with diverging number of covariates. *The Annals of Statistics* 2011; **39**:389–417.
5. Hin L-Y, Carey VJ, Wang Y-G. Criteria for working correlation structure selection in GEE: assessment via simulation. *The American Statistician* 2007; **61**:360–364.
6. Zhou J, Qu A. Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* 2012; **107**:701–710.
7. Overall JE, Tonidandel S. Robustness of generalized estimating equation (GEE) tests of significance against misspecification of the error structure model. *Biometrical Journal* 2004; **46**:203–213.
8. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. *Biometrics* 2001; **57**:126–134.
9. Lu B, Preisser JS, Qaqish BF, Suchindran C, Bangdiwala SI, Wolfson M. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* 2007; **63**:935–941.
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. Balan RM, Schiopu-Kratina I. Asymptotic results with generalized estimating equations for longitudinal data. *The Annals of Statistics* 2005; **33**:522–541.
12. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; **87**:823–836.
13. Madsen L, Dalthorp D. Simulating correlated count data. *Environmental and Ecological Statistics* 2007; **14**:129–148.
14. Song PX-K. *Correlated Data Analysis: Modeling, Analytics, and Applications*. Springer: New York, 2007.
15. Song PX-K, Jiang Z, Park E, Qu A. Quadratic inference functions in marginal models for longitudinal data. *Statistics in Medicine* 2009; **28**:3683–3696.
16. Hin L-Y, Wang Y-G. Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* 2009; **28**:642–658.
17. Warton DI. Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics* 2011; **67**:116–123.
18. Windmeijer F. A finite sample correction for the variance of linear two-step GMM estimators, Institute for Fiscal Studies Working Paper Series No. W00/19, London, 2000.
19. Windmeijer F. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 2005; **126**:25–51.
20. Westgate PM, Braun TM. The effect of cluster size imbalance and covariates on the estimation performance of quadratic inference functions. *Statistics in Medicine* 2012; **31**:2209–2222.
21. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. *Statistics in Medicine* 2012; **31**:4003–4022.