American Journal of Applied Mathematics and Statistics, 2015, Vol. 3, No. 6, 243-251
Available online at http://pubs.sciepub.com/ajams/3/6/5

DOI:10.12691/ajams-3-6-5

Modification of the Sandwich Estimator in Generalized Estimating Equations with Correlated Binary Outcomes in Rare Event and Small Sample Settings

Paul Rogers[1], Julie Stoner[2,*]

[1]Department of Numerical Sciences, Civil Aerospace Medical Institute, Oklahoma City, Ok, USA
[2]Department of Biostatistics and Epidemiology, College of Public Health, University of Oklahoma Health Sciences Center, Oklahoma City, Ok. USA
*Corresponding author: Paul.Rogers@faa.gov

Abstract Regression models for correlated binary outcomes are commonly fit using a Generalized Estimating Equations (GEE) methodology. GEE uses the Liang and Zeger sandwich estimator to produce unbiased standard error estimators for regression coefficients in large sample settings even when the covariance structure is misspecified. The sandwich estimator performs optimally in balanced designs when the number of participants is large, and there are few repeated measurements. The sandwich estimator is not without drawbacks; its asymptotic properties do not hold in small sample settings. In these situations, the sandwich estimator is biased downwards, underestimating the variances. In this project, a modified form for the sandwich estimator is proposed to correct this deficiency. The performance of this new sandwich estimator is compared to the traditional Liang and Zeger estimator as well as alternative forms proposed by Morel, Pan and Mancl and DeRouen. The performance of each estimator was assessed with 95% coverage probabilities for the regression coefficient estimators using simulated data under various combinations of sample sizes and outcome prevalence values with an Independence (IND), Autoregressive (AR) and Compound Symmetry (CS) correlation structure. This research is motivated by investigations involving rare-event outcomes in aviation data.

Keywords: *sandwich estimator, generalized estimating equation, rare event, finite sample, binary outcome, correlated outcome*

Cite This Article: Paul Rogers, and Julie Stoner, "Modification of the Sandwich Estimator in Generalized Estimating Equations with Correlated Binary Outcomes in Rare Event and Small Sample Settings." *American Journal of Applied Mathematics and Statistics*, vol. 3, no. 6 (2015): 243-251. doi: 10.12691/ajams-3-6-5.

1. Introduction

Regression models with binary outcome variables are prevalent in all research disciplines. If the data are independent, then the covariance between two measured values, which is a measure of linear dependence, is zero. If the data are dependent, Generalized Estimating Equations (GEE) can be used to account for the correlation, which is a function of the covariance among repeated or clustered measurements [1]. The GEE framework contains options for the working covariance structure based upon the assumed pattern of correlation within the data. One of the strengths of using GEE is that the sandwich or robust variance estimator produces unbiased standard errors in large sample sizes for the regression coefficients even when the covariance structure is misspecified. This is a tremendous advantage, but the sandwich estimator of variance is not without drawbacks.

It is well known that the GEE methodology has issues with small sample sizes due to the asymptotic properties inherent in the covariance sandwich estimator [2,3]. Fitzmaurice et al. noted that in small or finite sample sizes, Wald tests using the Liang-Zeger sandwich estimator tend to produce p-values that are too small [3]. The sandwich estimator of variance is biased downward; that is, it underestimates the variability of parameter estimators in small sample sizes. Much research has been performed to improve the performance of GEE under these circumstances. This is evidenced in the works of Mancl and DeRouen as well as Pan [4,5]. Rare outcomes pose a problem as well. Even with a large sample size, a rare outcome can be viewed as a small sample problem. That is, the information concerning the event of interest is, by itself, a small sample. Adding records that do not have the outcome of interest gives no additional information to the model. If an event becomes rare enough, it becomes extremely difficult to collect enough information to construct an informative regression model. The problems GEE experiences with finite sample sizes can become exacerbated when coupled with a rare outcome.

Rare events defined as binary outcomes, which have tens of thousands to hundreds of thousands of non-events (zeroes) compared to the outcome of interest (ones), can be a challenge in observational studies or clinical trials. Logistic regression methods for independent data have binary or ordinal outcomes but can produce predicted

probabilities that grossly underestimate the true probability of a rare event [6]. At present, very few methods are available for modeling and analyzing longitudinal rare event data. The methods currently available are models based upon the Poisson distribution and are appropriate when the dependent variable involves count data. In the rare event situations, with dependent data, the variance matrix for the regression coefficients of the standard logistic regression model is biased; the estimated variances are smaller than the true variances. Furthermore, Carroll and colleagues discovered that under rare event conditions the use of the sandwich estimator with the logistic regression model produced under coverage of Wald-type tests. In the case of logistic regression using the sandwich estimator, "an important part of sample size considerations is the number of events" [2]. In other words, decreasing sample sizes with rare outcomes can worsen the bias of the sandwich estimator.

The GEE methods are fairly robust and compensate for correlation among repeated measures or clustered data. However, in rare event and finite sample size settings, the variances and covariances generated by these models are underestimated and lead to erroneous inferences. Other investigators have proposed corrections for rare events and finite sample sizes with correlated data but there is no universally agreed upon solution for dealing with these circumstances. These solutions have resulted in alternative sandwich estimators that still have performance issues.

We propose the use of an improved sandwich estimator that has the ability to produce unbiased estimates of variances and covariances in studies of correlated data with rare event and small sample sizes. Our approach will be to adjust the sandwich estimator to compensate for underestimation in these situations. In general, this adjustment is performed by taking an alternate sandwich estimator, developed by Pan, and improving its performance in small sample size and rare event settings by adding an appropriate inflation factor, while still preserving the asymptotic nature of the sandwich estimator. The performance of this improved sandwich estimator will be evaluated with simulated and real-world datasets.

2. Generalized Estimating Equations and the Sandwich Covariance Estimator

In general, if Y_i is a response variable and X_i is a covariate of interest for $i = 1,.....,\mathrm{K}$ subjects, a regression model can be utilized to describe their relationship. In the case of longitudinal data, j is the index for the number of observations within a given subject. The number of repeated measurements on an individual will be represented as n_i with n_{ij} being the measurement at the j^{th} interval for the i^{th} subject. Marginal models are based on quasi-likelihood and are similar in form to the Generalized Linear Model (GLM) in that a link function (g), is used to specify a mathematical relationship, involving regression coefficients (β), between a marginal mean response (μ_{ij}), and one or more independent variables $\left(X_{ij}\right)$.

$$g(\mu_{ij}) = \eta_{ij} = X_{ij}^T \beta$$

Regarding the GEE methodology, if $\boldsymbol{\mu}_i$ is a vector of predicted means for the i^{th} individual and p is the number of regression coefficients, then $\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_h}$ where $h = 1,..., p$ will be used to represent the partial derivatives of the vector of predicted means with respect to the vector of regression coefficients $(\boldsymbol{\beta})$. Then $\mathbf{D}_i$ is an n_i x p matrix of these partial derivatives and appears as follows:

$$\mathbf{D}_i = \begin{pmatrix} \partial\mu_{i1}/\partial\beta_1 & \cdots & \partial\mu_{i1}/\partial\beta_p \\ \vdots & \ddots & \vdots \\ \partial\mu_{in_i}/\partial\beta_1 & \cdots & \partial\mu_{in_i}/\partial\beta_p \end{pmatrix}.$$

The variance (v_i) of the dependent variable (y_i) in the quasi-likelihood method, just as it is in GLM, can be expressed as a function (h) of the mean. Phi (ϕ) is a scale parameter estimated from the data and is sometimes referred to as a *nuisance* parameter, as it is typically not of primary interest.

$$v_i = \phi h(\mu_i)$$

If $\mathbf{Y}_i$ is used to indicate the n_i x 1 vector of outcomes for individual i, then let $\mathbf{v}_i$ be the vector of variances for these effects. $\mathbf{A}_i$ is a diagonal matrix that has taken on the values of the vector $\mathbf{v}_i$. Let α represent the correlation within the clustered measurements then $\mathbf{R}_i(\alpha)$ is the working correlation matrix for these same quantities. In this study, it is assumed that there is a correlation structure $\mathbf{R}_i(\alpha)$ common to all subjects. If $\mathbf{A}_i$ is an $n_i \times n_i$ matrix with the variances of $\mathbf{Y}_{ij}$ on the diagonal, then let $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\alpha)\mathbf{A}_i^{1/2}/\phi$ indicate the working covariance matrix for these same measurements; $\mathbf{V}_i$ depends on the correlation structure $\mathbf{R}_i(\alpha)$.

In the GEE method, when the dependent variable comes from the exponential family, the following are the score equations for the regression coefficients $(\boldsymbol{\beta}'s)$:

$$\mathbf{S}_h = \sum_{i=1}^{K} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}_h} \mathbf{V}_i^{-1}(\mathbf{Y}_i - \boldsymbol{\mu}_i) = 0 \text{ where : } h = (1,..., p)$$

Liang and Zeger (1986) demonstrated that as the number of subjects or clusters (K) increased in size, that $\hat{\boldsymbol{\beta}}$ is a consistent estimator for $\boldsymbol{\beta}$. That is, as $K->\infty, K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically multivariate Gaussian with zero mean and covariance matrix $(\mathbf{V}_{LZ})$ estimated as follows.

$$\mathbf{V}_{LZ} = \lim_{K->\infty} K\left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \left\{\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)^{\mathbf{T}} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right\} \left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \quad (1)$$

When estimates of $\boldsymbol{\beta}$ and α are inserted, $\mathbf{V}_{LZ}$ is referred to as the empirical-based,or robust sandwich, variance matrix.

3. Summary of Small-Sample Covariance Estimators

The Liang-Zeger sandwich estimator ($\mathbf{V}_{LZ}$) is used frequently in GEE since it produces valid standard errors asymptotically, even if the covariance structure is misspecified. The degree of bias of the sandwich estimator is an asymptotic property that is reduced as the sample size, or number of independent clusters, increases.

The problems caused by rare outcomes relative to the use of GEE models were first noted by Gunsolley while exploring the performance of GEEs with binary outcomes using a compound symmetry covariance structure[7].

3.1. Pan Estimator

Pan argued that the covariance calculated within the sandwich estimator is not an optimal estimator of $Cov(Y_i)$ because it is based on data from the i^{th} subject and is neither efficient nor consistent [5]. Pan proposed an improvement to the sandwich estimator by using a pooled, or averaged, covariance based upon all subjects. This enhancement depends on two assumptions to preserve the asymptotic nature of Pan's estimator:

Assumption 1. The marginal variance of y_{ij} needs to be modeled correctly.

Assumption 2. There is a common correlation structure across all subjects.

In reference to the sandwich estimator proposed by Liang and Zeger in equation (1), Pan proposedreplacing the Cov(Y_i) with $\mathbf{W}_i$:

$$\begin{aligned}\mathbf{W}_i &= \phi\mathbf{A}_i^{1/2}\mathbf{R}_u\mathbf{A}_i^{1/2} \\ &= \mathbf{A}_i^{1/2}\left(\sum_{i=1}^{K}\mathbf{A}_i^{-1/2}\left(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i\right)^{\mathbf{T}}\mathbf{A}_i^{-1/2}/K\right)\mathbf{A}_i^{1/2}\end{aligned}$$

That is, $\mathbf{W}_i = Cov(Y_i)$ and $\mathbf{R}_u$ is a correlation matrix obtained without any parametric specification (α).

Pan's sandwich estimator ($\mathbf{V}_P$), in matrix notation, can then be written as:

$$\begin{aligned}\mathbf{V}_P = &\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} \\ &\left[\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\left\{\mathbf{A}_i^{1/2}\left(\sum_{i=1}^{K}\mathbf{A}_i^{-1/2}\left(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i\right)\left(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i\right)^{\mathbf{T}}\mathbf{A}_i^{-1/2}/K\right)\mathbf{A}_i^{1/2}\right\}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right] \\ &\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\end{aligned} \quad (2)$$

Pan claimed this modified sandwich estimator has greater efficiency than that proposed by Liang and Zeger as the covariance is based upon all subjects. The results of his initial simulations using an exchangeable and independence covariance structure with both a binary and continuous outcome variable support this claim [5].

3.2. Mancl and DeRouen Estimator

Mancl and DeRouen proposed replacing the covariance of Liang and Zeger's (1986) sandwich estimator ($\mathbf{V}_{LZ}$) with one corrected for bias. That is $\mathbf{V}_{LZ}$ from equation (1) becomes the bias-corrected sandwich estimator ($\mathbf{V}_{MD}$) [4]:

$$\begin{aligned}\mathbf{V}_{MD} = &\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} \\ &\left\{\sum_{i=1}^{K}\begin{matrix}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}(\mathbf{I}-\mathbf{H}_i)^{-1} \\ (\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)^T \\ (\mathbf{I}-\mathbf{H}_i^T)^{-1}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\end{matrix}\right\}\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\end{aligned} \quad (3)$$

where $\mathbf{I}$ is an $n_i x n_i$ identity matrix, $\mathbf{V}_N$ is the "naïve" or model-based variance estimator and $\mathbf{H}_i = \hat{\mathbf{D}}_i\mathbf{V}_N\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}$. Mancl and DeRouen justify this correction on the grounds that the true expected value is expressed as $E[Cov(\mathbf{Y}_i)] \approx (\mathbf{I}-\mathbf{H}_i)(\mathbf{Y}_i-\boldsymbol{\mu}_i)(\mathbf{Y}_i-\boldsymbol{\mu}_i)^T(\mathbf{I}-\mathbf{H}_i^T)$ rather than $E[Cov(\mathbf{Y}_i)] = (\mathbf{Y}_i-\boldsymbol{\mu}_i)(\mathbf{Y}_i-\boldsymbol{\mu}_i)^T$.

3.3. Morel Estimator

Morel originally explored the covariance matrix estimate in logistic regression in complex survey designs as a product of the application of a Taylor series expansion [8]. These results were later extended to the sandwich estimator used within the GEE framework [9]. They clearly delineated the source of the bias suffered by the sandwich estimator in small samplesizes. It was demonstrated that most software implementations of the sandwich estimator omit the term $\frac{N-1}{N-p}\frac{K}{K-1}$ where $N = \sum_{j=1}^{K} m_j$ and m_j represent the number of units in the i^{th} cluster $i = 1, 2.....K$. This term is part of the Taylor series estimation of the sandwich estimator. The omission of these terms is less serious when the sample size or number of clusters is large but becomes increasingly significant as the sample size is reduced. Morel (2003) proposed re-introducing these terms to adjust for bias in the sandwich estimator. He also recommended inflating the sandwich estimator by adding a scaled version of the sandwich estimator trace to itself. This concept is unalike those previously proposed in that it applies the adjustment to the entire sandwich estimator. Whereas most adjustments took place inside the calculation of the covariance of the sandwich estimator, this one was applied outside of the summation and not to the individual residuals. Morel's version of the sandwich estimator, adjusted with the trace,is referred to as $\mathbf{V}_{M\_T}$.

$$\mathbf{V}_{M\_T} = \mathbf{V}_{LZ} + \hat{\delta}_n\hat{\phi}\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1} \quad (4)$$

where:

$$\delta_K = \min\left(0.5, \frac{p}{K-p}\right)$$

$$\hat{\phi} = \max\left[1, trace\left\{\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left\{\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\,\mathrm{cov}(\mathbf{Y}_i)\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right\}\right\} \Big/ p\right]$$

Simulation results supported this modified approach because Type I error rates were nominal, even in small sample sizes, unlike the unmodified GEE and model-based covariance estimators ($\mathbf{V}_N$).

$$\mathbf{V}_N = Cov(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}.$$

A variant of Morel's original estimator was included in this comparative study for evaluation purposes. It is identical to the estimator described in equation (4) but was inflated with the determinant rather than the trace. Morel had originally suggested evaluating the performance of the sandwich estimator inflated with the determinant but had never done so. This paper will be the first to evaluate the performance of this variant of the sandwich estimator.

Incorporating the changes proposed by Pan (2001) and Morel (2003), with some additional modifications, a new hybrid sandwich estimator ($\mathbf{V}_R$) will be constructed. Building on a fusion of these concepts, we believe a modified GEE estimator can be constructed that delivers accurate probabilities, nominal Type I error rates, and confidence intervals with proper coverage. The performance of this hybrid sandwich estimator is compared to the estimators of Liang and Zeger (1986), Mancl and DeRouen (2001), Morel(2003), and Pan (2001). A summary of the sandwich estimators is given in Table 1.

Table 1. Summary of sandwich estimators

Sandwich Estimators	
Authors: Liang & Zeger	**Approach: Standard Sandwich**
$\mathbf{V}_{LZ} = \left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left\{\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)^{\mathbf{T}}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right\}\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}$	
Author: Pan	**Approach: Average Covariance**
$\mathbf{V}_P = \left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left[\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\left\{\mathbf{A}_i^{1/2}\left(\sum_{i=1}^{K}\mathbf{A}_i^{-1/2}(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)^{\mathbf{T}}\mathbf{A}_i^{-1/2}/K\right)\mathbf{A}_i^{1/2}\right\}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right]\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}$	
Authors: Mancl & DeRouen	**Approach: Bias Correction**
$\mathbf{V}_{MD} = \left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left\{\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}(\mathbf{I}-\mathbf{H}_i)^{-1}(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\mu}}_i)^T(\mathbf{I}-\mathbf{H}_i^T)^{-1}\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right\}\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}$ where $\mathbf{H}_i = \hat{\mathbf{D}}_i\mathbf{V}_N\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}$	
Author: Morel	**Approach: Inflation with Trace**
$\mathbf{V}_{M\_T} = \mathbf{V}_{LZ} + \hat{\delta}_n\hat{\phi}\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}$ where $\delta_K = \min\left(0.5, \frac{p}{K-p}\right)$ and $\hat{\phi} = \max\left[1, trace\left\{\left(\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right)^{-1}\left\{\sum_{i=1}^{K}\hat{\mathbf{D}}_i^T\hat{\mathbf{V}}_i^{-1}\,\mathrm{cov}(\mathbf{Y}_i)\hat{\mathbf{V}}_i^{-1}\hat{\mathbf{D}}_i\right\}\right\} \Big/ p\right]$	

4. A New Hybrid Sandwich Estimator

A new hybrid sandwich estimator was created by inflating Pan's estimator with a scaled version of the determinant. The determinant is a physical representation of the area or volume of the variances and covariances of the sandwich estimator [10]. In terms of the volume, the determinant of the sandwich estimator can be expressed as:

$$\det\left(\widehat{Sandwich}\right) = \frac{(Volume)^2}{(N-1)^p}$$

Our recommended solution will use an averaged or pooled covariance, just as Pan (2001) did, and scale these values using the corrections originally proposed by Morel (2003).

An advantage of this strategy is that as long as the model-based estimate of variance is a positive definite matrix, the hybrid sandwich estimator will also be positive definite. Referencing Pan's version of the sandwich estimator from equation (2) the improvements will change the final version of the Rogers hybrid sandwich estimator $\mathbf{V}_R$ to generally appear as:

$$\mathbf{V}_R = \mathbf{V}_P + \left(\frac{p}{K-p}\right) \times \det\left[\left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \left\langle \sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \left\{ \mathbf{A}_i^{1/2} \left(\sum_{i=1}^{K} \begin{pmatrix} \mathbf{A}_i^{-1/2} \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right) \\ \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right)^{\mathbf{T}} \mathbf{A}_i^{-1/2} \end{pmatrix} / K \right) \mathbf{A}_i^{1/2} \right\} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right\rangle \Big/ p \right] \times \left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \tag{5}$$

5. Asymptotic Properties

The asymptotic properties of the hybrid estimator follow directly from the properties of Pan's estimator. As the number of clusters increases, the Pan and Rogers estimators become more similar. To assure the asymptotic validity of his estimator, Pan needed the two assumptions, listed in section 3.1,to hold true [5].

Our estimator is similar to Pan's but with an additional inflation factor; as the number of subjects increases they converge to the same values. This can be demonstrated with the following limit;

Limit 1. As the number of subjects goes to infinity $K \to \infty$ the following will hold true:

$$\lim_{K\to\infty} \mathbf{V}_R = \lim_{K\to\infty} \left[\mathbf{V}_P + \left(\frac{p}{K-p}\right) \det\left[\left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \left\langle \sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \left\{ \mathbf{A}_i^{1/2} \left(\sum_{i=1}^{K} \begin{pmatrix} \mathbf{A}_i^{-1/2} \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right) \\ \left(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i\right)^{\mathbf{T}} \mathbf{A}_i^{-1/2} \end{pmatrix} / K \right) \mathbf{A}_i^{1/2} \right\} \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i \right\rangle \Big/ p \right] \times \left(\sum_{i=1}^{K} \hat{\mathbf{D}}_i^T \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i\right)^{-1} \right] = \mathbf{V}_P$$

If assumption 2 does not hold, as Pan recommended, subjects can be classified into groups such that the $\mathbf{Y}_i$ have the same correlation matrix. Therefore, as the sample size increases and the marginal variance of $\mathbf{Y_i}$ is modeled correctly we expect the values of the Pan and Rogers sandwich estimators to be more similar. If assumptions 1 and 2 hold then with a large enough sample size we expect the differences in $K^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ to be asymptotically multivariate Gaussian with zero mean and covariance matrix $(\mathbf{V})$ under the Pan and Rogers methodologies as well. In addition to these similarities, if the sample size and prevalence are both increased, we expect to see a convergence of similar values and performance in coverage probabilities from all of the sandwich estimators.

6. Simulation Studies

Due to the asymptotic nature of the sandwich estimators, simulations were conducted to assess their performance under varying small sample and rare event conditions. The sandwich estimators compared included the traditional Liang-Zeger $(\mathbf{V}_{LZ})$, Mancl and DeRouen $(\mathbf{V}_{MD})$, Pan $(\mathbf{V}_P)$, Morel $(\mathbf{V}_{M\_T})$, a version of Morel inflated by the determinant rather than the trace, and the Rogers $(\mathbf{V}_R)$ sandwich estimators. A model with one continuous covariate was used for simulation study. The number of clusters, prevalence, and correlation of the outcome variable were varied.

The single covariate model was fit on a series of simulated datasets with outcomes of differing prevalence values (0.01, 0.05, 0.10, 0.30, 0.50). The simulations were run with data sizesof500, 100, 50, 30,and 20 subjects. The various estimators' performance was also compared when the simulated within-cluster correlation structure was either exchangeable or autoregressive, with correlation set at 0.005 or 0.05, and when observations within clusters were simulated to be independent. All simulations involve balanced designs with four observations per subject. These correlation values were selected based on the relationship between the prevalence of the outcome and the correlation among longitudinal measures. That is, the probability of the outcome restricts the range of possible correlation values[3]. Due to this relationship between the prevalence and correlation, it was not practical to simulate all combinations of prevalence values and correlations.

Simulated correlated binary data were generated with the binary SimCLF R-code library, which is based on the work by Qaqish [11]. The correlations were kept low due to the simulation difficulties in generating large numbers of valid outcome vectors with the binary SimCLF library in small sample size and low outcome prevalence conditions. That is to say, as the sample size and outcome prevalence decreased, the binarySimCLF produced a large number of vectors which failed its own validity check.In all simulations, the covariance structure was correctly specified. The total number of configurations, as well as the number of simulations is summarized in Table 2. Eachsimulation configuration was run 1,000 times, reporting the average of the sandwich estimator undergoing testing.

Table 2. Simulation design settings for each of the six estimators.

Total Number of Simulations				
Correlation Structure	**Prevalence**	**Cluster Sizes**	**Correlation**	**Number of Simulations**
Autoregressive (AR-1)	0.01, 0.05, 0.10, 0.30, 0.50	20, 30, 50, 100, 500	0.005	25
Autoregressive (AR-1)	0.10, 0.30, 0.50	20, 30, 50, 100	0.05	12
Compound Symmetric	0.01, 0.05, 0.10, 0.30, 0.50	20, 30, 50, 100, 500	0.005	25
Compound Symmetric	0.10, 0.30, 0.50	20, 30, 50, 100	0.05	12
Independent	0.01, 0.05, 0.10, 0.30, 0.50	20, 30, 50, 100, 500	0	25

The true values for the intercept (β_0) and regression coefficient (β_1) were both set to one for all tests. This model consisted of a single, normally distributed covariate (X_1) with a variance of one centered at a mean appropriate to the simulated prevalence $\pi(x)$ of the outcome. The relationship between the outcome prevalence and continuous covariate is given by:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 X_1)}}{1+e^{(\beta_0+\beta_1 X_1)}}$$

The standard deviation and average estimated standard error of the estimated regression coefficients of the betas (β_0, β_1) were calculated and recorded. The performance of each sandwich estimator was assessed primarily by the 95% coverage probabilities for the regression coefficients.

6.1. Demonstration of Bias as a Poor Performance Measure

A measure of performance usually used in evaluating a new statistic is the bias; the difference between the estimator's true variance and its mean estimated variance. Estimators with a positive bias have underestimated the true variance, while those with a negative bias have overestimated the true variance.

Coverage probabilities, which are related to confidence intervals, are an alternative way of assessing performance. These confidence intervals are centered around the estimated regression coefficients $(\hat{\beta})$, which are the same for each covariance estimation method in this simulation study. Therefore, the coverage probability in this study is only a function of the estimated variance.

The simulation environment was designed to reproduce coverage probabilities analogous to a 95% confidence interval. After completion of the simulations, the distributions of the variance estimates created by each sandwich estimator in small samples were skewed. For all covariance structures, as the simulated prevalence and sample size diminish, the distribution of the variances for each sandwich estimator becomes steeper on the lower end and right-skewed, both to a different degree. The implication is that the distribution of the variances is no longer symmetric, and the mean is no longer in the center of the distribution under these extreme conditions. These differences are so great that measures of bias are not adequate performance indicators and therefore, coverage probabilities will be reported as the performance measure.

6.2. Coverage Probabilities

The coverage probability results are only shown for the autoregressive covariance structures as the results were similar among the three types of correlation. Coverage probabilities were similar between β_1 and the intercept (β_0) therefore, figures are only shown for the β_1 regression coefficient. A composite figure is used for outcome prevalence values of 5% through 50% under a .005 correlation. A single graph is dedicated to the 1% prevalence level to highlight the differences that occur under these extreme conditions. When the correlation increases to .05,a composite figure is again used to display prevalence values of 10%, 30%, and 50%. Coverage probabilities are displayed in Figure1 through 3 for the autoregressive covariance structure.

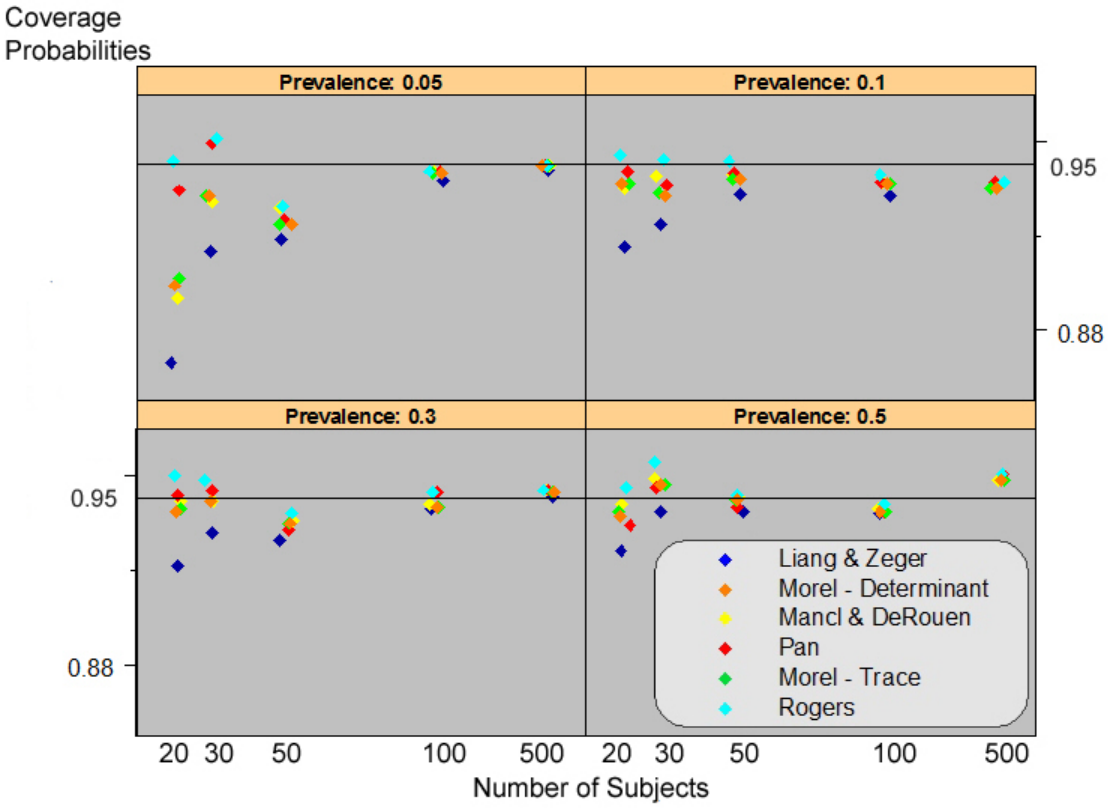

Figure 1. Coverage probabilities when estimating the regression coefficient β_1 under a simulated autoregressive covariance structure for 0.05 through 0.5 prevalence values with 0.005 correlation

The coverage probability of our estimator for the regression coefficient (Figure 1) is very competitive with that of Pan's. These two estimators outperform the remaining estimators at 20 and 30 subjects under a 5% prevalence. At 10%, 30%, and 50% prevalence values, the performance of the estimators begin to cluster and converge with increasing sample size, while the Liang-Zeger estimator lags behind the rest at fewer than 50 subjects.

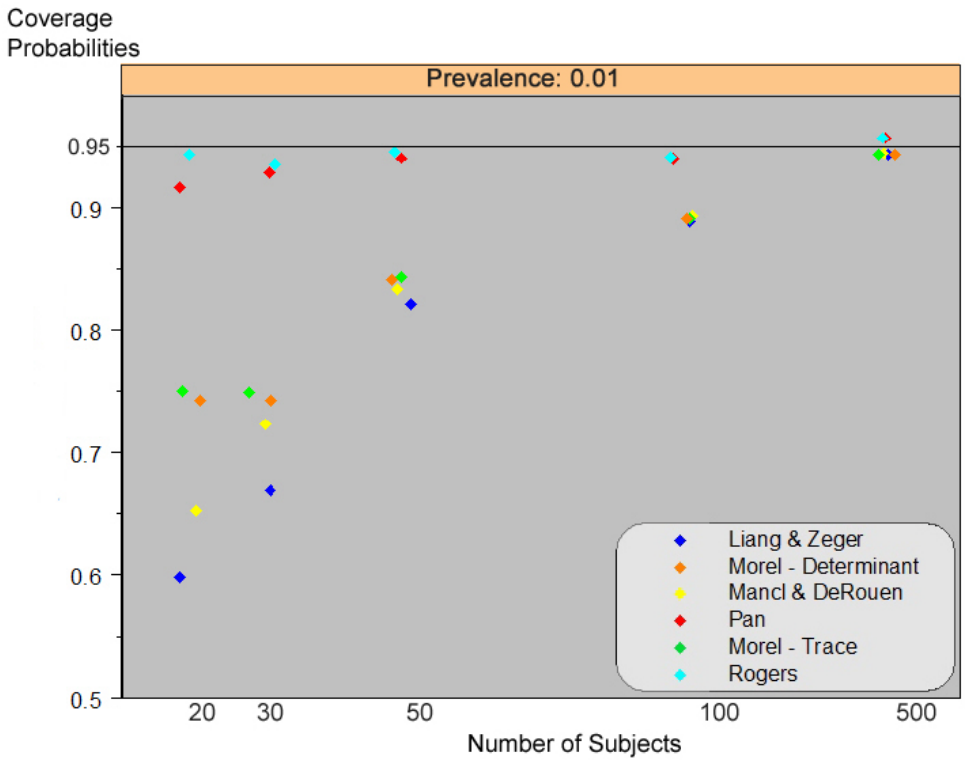

Figure 2. Coverage probabilities when estimating the regression coefficient β_1 under a simulated autoregressive covariance structure for 0.01 prevalence with 0.005 correlation

The Rogers and Pan estimators are very close in their performance in terms of coverage probabilities under a 1% outcome prevalence, with the Rogers estimator slightly edging out Pan on the smaller sample sizes (Figure 2). The remaining estimators performance is poor at 20 and 30 subjects but steadily improves as the sample size increases. At simulated sample sizes of 500 subjects, the estimators have converged to roughly the same coverage probabilities.

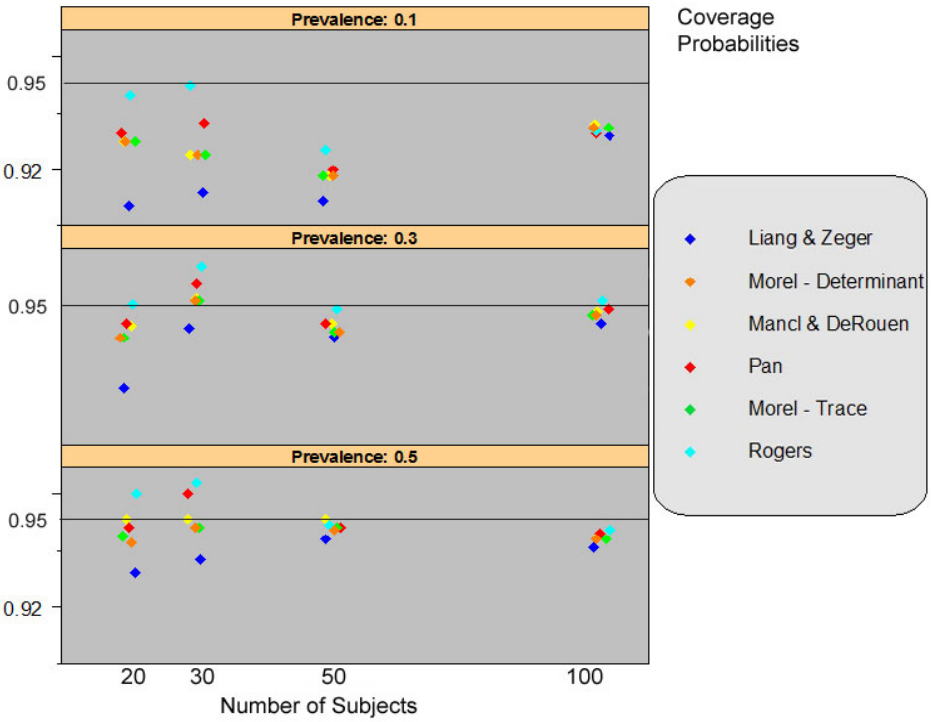

Figure 3. Coverage probabilities when estimating the regression coefficient β_1 under a simulated autoregressive covariance structure for 0.1 through 0.5 prevalence values with 0.05 correlation

When the simulated correlation is .05, a similar trend can be seen for the regression coefficient (Figure 3). As the sample size and prevalence decrease, the estimators begin to diverge from one another in their performance. Typically, the Liang-Zeger estimator lags behind the others as it underestimates the variance, which decreases its coverage probability. At an outcome prevalence of 10%, our proposed estimator performs better than the other estimators. It is interesting to note that when a simulated outcome prevalence is as low as 10% is coupled with a sample size of 100, all the sandwich estimators underestimate the true variance. As the outcome prevalence increases to 50%, our estimator slightly overestimates the variance at sample sizes of 50 subjects or fewer.

7. Practical Application

In this section, we demonstrate the application of our sandwich estimator in a practical setting. The two datasets used are random samples of size 500 and 30 airmen, sampled independently, from the Federal Aviation Administration's Decision Support Systems (DSS) and constructed as a longitudinal dataset, as described by Peterman [12]. Airmen undergo a flight physical from an Aviation Medical Examiner (AME) and must meet certain physical requirements to hold a Class I, II, or III medical certificate. The random samples taken from the DSS were restricted to airmen who took a Class I, II, or III flight physical in each of the four years over 2002-2005.

The outcome of interest, expressed as a binary variable, concerns the occurrence of an Accident and Incident Data System (AIDS) event, which can include anything from a major aircraft accident to a minor incident with only slight damage. The covariate of interest, a continuous variable, indicates the number of flight hours over the last six months self-reported by the airmen at the time of their last medical exam. The results should give us insight as to whether the number of accumulated flight hours over the last six months is associated with the occurrence of an AIDS event. The question of interest is represented in equation (6), where Y represents a binary outcome, with a one and zero indicating the occurrence or lack of an AIDS event, respectively.

$$\ln\left\{\frac{\Pr\left(Y_{ij}=1\right)}{\Pr\left(Y_{ij}=0\right)}\right\}=\beta_0+\beta_1[FlightTime_{ij}] \qquad (6)$$

The outcome's prevalence, slightly under 0.5%, is lower than those investigated in our simulation study which was 1% or higher. Among all years for the sample of 500 pilots, the median flight time was 26.62 hours (inter quartile range: 24.43-29.16 hours). In the sample size of 500 subjects, one subject reporteda flight time of 20,750 hours. This outlier was omitted and imputed as the average of the three previously reported flight hours for the previous six months (300 hours).

An autoregressive correlation structure reflects correlation decay with increasing intervals of time between measurements. Use of an autoregressive structure was reasonable, given the design of the study. The analytical results for the 500 and 30 subjects are displayed in Table 3 and Table 4, respectively.

Table 3. Estimated regression coefficients, odds ratios (OR), 95% confidence intervals (CI) and sandwich estimators from analysis of AIDS events in self-reported 100 flight hour changes for last six months for 500 airmen

Estimated Regression Coefficients Under an Autoregressive Correlation Structure						
Method	$\hat{\beta}_1$	$\widehat{SE}(\hat{\beta}_1)$	OR	95% CI for OR	Z-Score	p-Value
Liang-Zeger	0.0094	0.2163	1.0094	0.6607, 1.5422	0.0433	0.9655
Mancl & DeRouen	0.0094	0.2172	1.0094	0.6594, 1.5452	0.0432	0.9655
Morel - Trace	0.0094	0.2167	1.0094	0.6601, 1.5435	0.0433	0.9655
Pan	0.0094	0.2029	1.0094	0.6782, 1.5024	0.0462	0.9632
Morel-Determinant	0.0094	0.2167	1.0094	0.6602, 1.5434	0.0433	0.9655
Rogers	0.0094	0.2033	1.0094	0.6776, 1.5037	0.0461	0.9632

Table 4. Estimated regression coefficients, odds ratios (OR), 95% confidence intervals (CI) and sandwich estimators from analysis of AIDS events in self-reported 100 flight hour changes for last six months for 300 airmen

Estimated Regression Coefficients Under an Autoregressive Correlation Structure						
Method	$\hat{\beta}_1$	$\widehat{SE}(\hat{\beta}_1)$	OR	95% CI for OR	Z-Score	p-Value
Liang-Zeger	0.1541	0.0599	1.1666	1.0374, 1.3120	2.5728	0.0101
Mancl & DeRouen	0.1541	0.9010	1.1666	0.1995, 6.8221	0.1711	0.8641
Morel - Trace	0.1541	0.0767	1.1666	1.0038, 1.3559	2.0099	0.0444
Pan	0.1541	0.3319	1.1666	0.6087, 2.2360	0.4644	0.6424
Morel-Determinant	0.1541	0.0767	1.1666	1.0038, 1.3559	2.0099	0.0444
Rogers	0.1541	0.3379	1.1666	0.6016, 2.2625	0.4561	0.6484

In our analysis of 500 subjects, the differences among the variances of the sandwich estimators for the covariate of interest (β_1) from equation (6) are very small. This is not surprising as the simulation results reported that the sandwich estimator's coverage probabilities converge to the same values in large sample sizes, even with outcome prevalence values as low as 1%. When we analyze the sample of 30 subjects, the variability of the sandwich estimators' variance is even larger, as reflected in their values differing from one another by several magnitudes of 10. When performing statistical hypothesis testing in a situation where the outcome is of low prevalence and the sample size is small, the choice of sandwich estimator affects the outcome of hypothesis testing concerning the regression coefficients. The estimated odds ratio for a100-hour increase of flight time ($\hat{\beta}_1$) in the sample of 30 subjectsis 1.1664. The associated 95% confidence intervals for the Liang-Zeger and Rogers sandwich

estimators are (1.0374, 1.3120) and (0.6016, 2.2625), respectively. For the purposes of this question, the use of the Liang-Zeger or Rogers sandwich estimator impacted the statistical significance of the covariate of interest.

8. Conclusion

This research explored a novel way of building a hybrid sandwich estimator that would achieve superior performance over that of the standard Liang-Zeger sandwich estimator in settings with low outcome prevalence and reduced sample sizes. The performance of this estimator was also compared with other sandwich estimators adjusted for improved performance in small sample sizes. As the outcome prevalence dropped below 30% and the sample size below that of 50 subjects, the choice of estimators matters, and one should consider using an alternative to the Liang-Zeger estimator. In our limited simulation settings, the Rogers sandwich estimator outperformed the Liang-Zeger and typically outperformed all other estimators as the prevalence and sample size both dropped. The Rogers estimator is an extension of the Pan estimator, which also performed very well in these simulations. The performance of the Rogers estimator is dependent on the determinant calculated in the inflation factor. It is possible that the performance of the Rogers estimator may be inferior in comparison to the Pan estimator under different correlation settings. The performance of the Mancl and DeRouen sandwich estimator deteriorated to coverage probabilities only slightly better than that of the Liang-Zeger in prevalence values of 1% and 5% in sample sizes of 20 and 30 subjects. The Morel sandwich estimators, at the 1% outcome prevalence level, performed better than that of Mancl and DeRouen but not as well as the Pan or Rogers' estimators. Overall, it is wise to select any of these other estimators, if available, over the Liang-Zeger in a situation involving low sample size or low outcome prevalence.

The true or simulated covariance structure had little bearing on the estimators' performance. Mirrored performances were observed by all of the sandwich estimators among the three different covariance structures. This result was also observed by Mancl and DeRouen [4]. It is likely that the simulated covariance structure played no role in the estimators' performance due to the low correlation values used in the simulations. The correlations were kept low due to the simulation difficulties in generating large numbers of valid outcome vectors with the binary SimCLF library in small sample size and low outcome prevalence conditions. It is possible that the performance of the sandwich estimators may differ under simulation conditions using greater correlation values than were used in this project.

A similar performance was observed in our simulations to that of Pan's, in terms of coverage probabilities, for the Liang-Zeger and Pan sandwich estimators under both the independence and compound symmetry structures [5]. The results of Gunsolley et al.'s 1995 simulation study exploring the performance of the Liang-Zeger sandwich estimator were similar to ours: as the outcome prevalence or sample size increased, the performance of the Liang-Zeger improved, as well in terms of Type I error rates [7].

In summary, the performance of the Liang-Zeger sandwich estimator suffers as the sample sizes dropped below 50 subjects, and the outcome prevalence values were less than 30%. This drop off in performance is further exacerbated at the lower outcome prevalence values and smaller sample sizes. Under these extreme conditions, the Rogers and Pan estimators would be good choices for variance estimators followed by any of the two estimators proposed by Morel. The Mancl and DeRouen estimator outperformed the Liang-Zeger estimator under all outcome prevalence values as the sample size dropped below 50 subjects. With outcome prevalence values of 30% or higher and sample sizes less than 50 subjects, the Liang-Zeger estimator still consistently underestimated the coefficient variances even in these nominal conditions.

Future work will be conducted to evaluate the performances of the various sandwich estimators with higher correlations in moderate sample sizes. We will also include different numbers and types of covariates in the assessment of sandwich estimator performances.

Acknowledgements

Partial funding provided by National Institutes of Health, National Institute of General Medical Sciences [grant 1 U54GM104938].

Competing Interest

The authors declare no competing financial interests.

References

[1] Liang, K.-Y. and S.L. Zeger, "Longitudinal Data Analysis Using Generalized Linear Models",*Biometrika*, 1986. **7**3(1): p. 13-22.

[2] Carroll, R.J., Wang, S., Simpson, D. G., Stromberg, A. J., and Ruppert, D.,*The Sandwich (Robust Covariance Matrix) Estimator*. 1998; Available from: http://www.stat.tamu.edu/ftp/pub/rjcarroll/sandwich.pdf.

[3] Fitzmaurice, G.M., N.M. Laird, and J.H. Ware, *Applied Longitudinal Analysis*. Wiley series in probability and statistics. 2004, Hoboken, N.J.: Wiley-Interscience. 506 p.

[4] Mancl, L.A. and T.A. DeRouen, "A Covariance Estimator for GEE with Improved Small-Sample Properties",*Biometrics*, 2001. 57(1): p. 126-134.

[5] Pan, W., "On the Robust Variance Estimator in Generalised Estimating Equations",*Biometrika*, 2001. 88(3): p. 901-906.

[6] King, G. and L. Zeng, "Logistic Regression in Rare Events Data",*Political Analysis*, 2001. 9: p. 137-163.

[7] Gunsolley, J.C., C. Getchell, and V.M. Chinchilli, "Small Sample Characteristics of Generalized Estimating Equations", *Communications in Statistics: Simulation and Computation*, 1995. 24: p. 869-78.

[8] Morel, J.G., "Logistic Regression Under Complex Survey Designs",*Survey Methodology*, 1989. 15(2): p. 203-223.

[9] Morel, J.G., M.C. Bokossa, and N.K. Neerchal, "Small Sample Correction for the Variance of GEE Estimators",*Biometrical Journal*, 2003. 45(4): p. 395-409.

[10] Johnson, R.A. and D.W. Wichern, *Applied Multivariate Statistical Analysis*. 6th ed. 2007, Upper Saddle River, N.J.: Pearson Prentice Hall. 773 p.

[11] Qaqish, B.F., "A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables with Specified Marginal Means and Correlations",*Biometrika*, 2003. 92: p. 455-463.

[12] Peterman, C.L., Rogers, P. B., Veronneau, S. J. H., and Whinnery, J. E., "Development of an Aeromedical Scientific Information System for Aviation Safety", *Office of Aerospace Medicine* 2008.Report No. DOT/FAA/AM-08/01.