STATISTICS IN MEDICINE
*Statist. Med.* 2005; **24**:1245–1261
Published online 6 December 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2023

# A comparison of goodness of fit tests for the logistic GEE model

Scott Evans[*,†] and Lingling Li[‡]

*Department of Biostatistics, Center for Biostatistics in AIDS Research, Harvard School of Public Health,*
*651 Huntington Avenue, Boston, MA 02115, U.S.A.*

## SUMMARY

Generalized estimating equations have become a popular regression method for analysing clustered binary data. Methods to assess the goodness of fit of the fitted models have recently been developed. However, evaluations and comparisons of these methods are limited. We discuss these methods and develop two additional statistics to evaluate goodness of fit. We evaluate the performance of each of the statistics with respect to type I error rates and power in a simulation study. Guidance is provided regarding appropriate use of the statistics under various scenarios. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: goodness of fit; GEE; clustered binary data; logistic regression; type I error; power

## 1. INTRODUCTION

The logistic generalized estimating equations (GEE) model has become a widely used method for analysis for clustered binary data. The model is easily fit with readily available software. Output is easily interpreted and can be used to estimate probabilities and/or odds ratios.

With this increase in application has been an increase in the development of methods to assess the adequacy of the fitted model. Assessment of the adequacy of the fitted GEE models has been a challenging area of research since no likelihood exists and residuals within cluster are correlated. Recently several goodness of fit (GOF) statistics have been proposed, however a comparison and evaluation of these methods is lacking.

[*] Correspondence to: Scott Evans, Department of Biostatistics, Center for Biostatistics in AIDS Research, Harvard School of Public Health, 651 Huntington Avenue, Boston, MA 02115, U.S.A.
[†] E-mail: evans@sdac.harvard.edu
[‡] E-mail: lingling@sdac.harvard.edu

Contract/grant sponsor: National Institutes of Health Statistical and Data Management Centre; contract/grant number: 5 U01 AI038855-08
Contract/grant sponsor: Neurologic AIDS Research Consortium; contract/grant number: 2 U01 NS0322228-08A1

*Received October 2003*
*Accepted July 2004*

Similar to GOF test statistics for ordinary logistic regression (OLR) models, proposed GOF test statistics for logistic GEE models fit into three major categories: statistics based upon partitioning of the covariate space, statistics using groups based on the ranked estimated probabilities, and statistics based on comparing observed versus predicted values (i.e. residuals). We also propose two hybrid statistics combining the ideas of covariate partitioning and partitioning based on ranked estimated probabilities. In addition, a kappa-like classification statistic has been previously proposed.

Hosmer *et al.* [1] compared and evaluated summary measures of GOF for OLR with respect to type I error rate and power. The objective of this paper is to similarly evaluate the performance of GOF statistics for the logistic GEE model.

We introduce the logistic GEE model in Section 2. We then introduce: statistics based on covariate partitioning in Section 3.1, a statistic based on ranked estimated probabilities in Section 3.2, statistics based on residuals in Section 3.3, a classification statistic in Section 3.4, and newly developed hybrid statistics in Section 3.5. In Section 4, we present the results of a simulation study investigating type I error (Section 4.1) and power (Section 4.2). An example is provided in Section 5. A summary and discussion is presented in Section 6.

## 2. THE LOGISTIC GEE MODEL

Liang and Zeger [2] proposed an extension of the generalized linear model for the analysis of longitudinal or multivariate data when regression is the desired method of analysis. This model describes how the average response across subjects (or 'clusters') changes with the covariates. Interest focuses on the relationship between the covariates and the probability of response while response correlation is treated as a nuisance parameter. This model is the natural multivariate analogue of ordinary logistic regression in the independence case. The model allows for adjustment of observational (or 'time-varying') and cluster level covariates. (Additional details may be found in Reference [3]). Note that we use the term 'time-varying' to characterize covariates that can change within cluster and are not limited to longitudinal studies.

The logistic GEE model for the marginal distribution of a binary outcome for observation $j$ in cluster $i$, $Y_{ij}$, assumes that the

$$\operatorname{logit}[\pi(\mathbf{x}_{ij})] = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_{ij}$$

where the

$$\operatorname{logit}[\pi(\mathbf{x}_{ij})] = \ln[\pi(\mathbf{x}_{ij})/\{1-\pi(\mathbf{x}_{ij})\}]$$

Here, $\pi(\mathbf{x}_{ij}) = E[Y_{ij} \mid \mathbf{x}_{ij}]$ for $i = 1, 2, \ldots, K$ and $j = 1, 2, \ldots, n_i$, where $\mathbf{x}_{ij}$ is the $p \times 1$ vector of covariates for observation $j$ in cluster $i$, $\beta_0$ is the population averaged intercept term and $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')'$ is the $(p+1) \times 1$ vector of population averaged (or marginal) coefficients.

Each cluster's data consist of a $n_i \times 1$ response vector $\mathbf{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]'$. The GEE estimator of $\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$, is found by solving the following estimating equations:

$$\mathbf{u}_\beta(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{K} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} [\mathbf{Y}_i - \boldsymbol{\pi}_i(\hat{\boldsymbol{\beta}})] = 0$$

where $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{in_i}]', \mathbf{D}_i = \partial\boldsymbol{\pi}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\mathbf{V}_i$ is the $n_i \times n_i$ 'working' covariance matrix of $\mathbf{Y}_i$. $\mathbf{V}_i = \mathbf{A}_i^{1/2}\mathbf{R}_i(\boldsymbol{\alpha})\mathbf{A}_i^{1/2}$, where $\mathbf{A}_i$ is a $n_i \times n_i$ diagonal matrix with $\text{var}(Y_{ij}\,|\,x_{ij}) = \pi_{ij}(1 - \pi_{ij})$ as the $i$th diagonal entry. $\mathbf{R}_i(\boldsymbol{\alpha})$, a $n_i \times n_i$ 'working' correlation matrix, is fully specified by an $s \times 1$ vector of unknown parameters $\boldsymbol{\alpha}$. Parameter estimates from these equations are consistent and asymptotically normal even when the correlation matrix is mispecified as long as the mean model and the link function are correctly specified [4].

If one assumes that the mean model and the working correlation matrix are correctly specified, then a model-based estimator of covariance matrix of the estimated parameters is given by

$$\left[\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}$$

This is the GEE version of the inverse of the Fisher information matrix that is often used in logistic regression as an estimator of the covariance of the maximum likelihood estimator of $\boldsymbol{\beta}$. This estimator is often termed a 'naive' estimator and is consistent if the structure is correctly specified.

A more robust estimate of the covariance can be made without the assumption of a correctly specified within-cluster structure. Liang and Zeger use the fact that responses in different clusters are independent and linearize the GEE's to obtain the 'sandwich' formula for a robust (or empirical) estimator of the covariance matrix of the estimated coefficients as follows:

$$\text{Cov}(\hat{\boldsymbol{\beta}}) = \left[\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1} \left[\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\text{Cov}(\mathbf{Y}_i)\mathbf{V}_i^{-1}\mathbf{D}_i\right] \left[\sum_{i=1}^{K} \mathbf{D}_i'\mathbf{V}_i^{-1}\mathbf{D}_i\right]^{-1}$$

The outer pieces of the sandwich correspond to the values of the covariance matrix of the estimated coefficients obtained if the working correlation assumptions are correct and the centre terms depend on the true correlation of responses. If the outer terms of the sandwich are replaced by an estimate of $\text{cov}(\mathbf{Y})$ which only requires independence between clusters, then consistent estimates of the covariance matrix of the estimated coefficients that are robust to misspecification of the correlation structure are obtained [5].

## 3. GOODNESS OF FIT STATISTICS

### *3.1. Statistics based on covariate partitioning*

Barnhart and Williamson [6] developed model-based and robust GOF statistics, using model-based and sandwich estimators of $\text{cov}(\mathbf{Y}_i)$, respectively, based on a partitioning of the covariate space into distinct regions and forming score statistics. The statistics are asymptotically distributed as chi-square random variables and can be considered an extension to the test of Tsiatis [7] for OLR.

Suppose $n$ observations within each cluster are measured at different times. The authors proposed two GOF statistics by partitioning the covariate space into $M$ distinct regions in $p$-dimensional space. Let $\mathbf{O}_{ij} = [I_{ij1}, \ldots, I_{ijM}]'$ be an $M \times 1$ vector, where $I_{ijm}$ is the indicator variable that equals one if the $j$th observation in $i$th cluster is in the $m$th region and zero otherwise. $\mathbf{O}_i = [\mathbf{O}_{i1}, \ldots, \mathbf{O}_{in}]'$ is a $n \times M$ matrix. Consider the alternative model

$$\text{logit}(\boldsymbol{\pi}_i) = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_n \boldsymbol{\tau} + \mathbf{O}_i \boldsymbol{\sigma} + \mathbf{S}_i \boldsymbol{\rho}$$

where $\mathbf{X}_i$ is the $n \times (p+1)$ design matrix for $i$th cluster (including the intercept), $\mathbf{Z}_n = [\mathbf{0}, \mathbf{I}_{n-1}]'$ is the $n \times (n-1)$ matrix where the first row has entries of zero and the remaining $(n-1)$ rows form a $(n-1) \times (n-1)$ identity matrix, $\mathbf{S}_i = [\mathbf{0}, \text{diag}(\mathbf{O}_{i2}, \ldots, \mathbf{O}_{in})]'$ is a $n \times (n-1)M$ matrix and $\mathbf{0}$ is a $(n-1)M \times 1$ vector of zeros. Note that $\boldsymbol{\tau}$ is the $(n-1) \times 1$ vector of time effects within each cluster (the first observation is the reference time point), $\boldsymbol{\sigma}$ is the $M \times 1$ vector of region effects, and $\boldsymbol{\rho}$ is the $(n-1)M \times 1$ vector of time and region interaction effects (e.g. for the $j$th observation in $i$th cluster, if its covariate belongs to $m$th region, the alternative model will be $\text{logit}(\pi_{ij}) = \mathbf{X}_{ij} \boldsymbol{\beta} + \tau_{j-1} + \sigma_m + \rho_{(j-1)m}$ for $j > 1$. Or $\text{logit}(\pi_{ij}) = \mathbf{X}_{ij} \boldsymbol{\beta} + \sigma_m$ for $j = 1$). Let $\boldsymbol{\theta} = [\boldsymbol{\tau}', \boldsymbol{\sigma}', \boldsymbol{\rho}']'$ be a $J \times 1$ vector with $J = (n-1) + M + (n-1)M$.

Let $\mathbf{U}$ be the $L = (p+1+J) \times 1$ vector with $l$th component

$$U_l = \sum_{i=1}^{K} \hat{\mathbf{D}}_{il}' \hat{\mathbf{V}}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\pi}}_i)$$

for $l = 1, \ldots, L$, where $\hat{\mathbf{D}}_{il} = \partial \hat{\boldsymbol{\pi}}_i / \partial \beta_l$ for $l \leqslant p+1$, $\hat{\mathbf{D}}_{il} = \partial \hat{\boldsymbol{\pi}}_i / \partial \theta_{l-p-1}$ for $l > p+1$, where $\hat{\boldsymbol{\pi}}_i = \text{logit}^{-1}(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{Z}_n \boldsymbol{\tau} + \mathbf{O}_i \boldsymbol{\sigma} + \mathbf{S}_i \boldsymbol{\rho})$, and $\hat{\boldsymbol{\beta}}$ is the GEE estimator obtained from the original mean model without the indicator coefficients. The estimated 'sandwich' covariance matrix of $\mathbf{U}$ can be written as

$$\mathbf{W}_R = \sum_{i=1}^{K} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \text{cov}(\mathbf{Y}_i) \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i$$

where $\hat{\mathbf{D}}_i = [\hat{\mathbf{D}}_{i1}, \ldots, \hat{\mathbf{D}}_{iL}]$ is a $n \times L$ matrix. The $\text{cov}(\mathbf{Y}_i)$ can be consistently estimated by $(\mathbf{Y}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\pi}}_i)'$ and $\mathbf{W}_R$ would reduce to the model-based estimator, $\mathbf{W} = \sum_{i=1}^{K} \hat{\mathbf{D}}_i' \hat{\mathbf{V}}_i^{-1} \hat{\mathbf{D}}_i$, if the correlation matrix $\mathbf{R}_i$ is correctly specified.

Let

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{pmatrix}, \qquad \mathbf{W}_R = \begin{pmatrix} \mathbf{A}_R & \mathbf{B}_R' \\ \mathbf{B}_R & \mathbf{C}_R' \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} \mathbf{A} & \mathbf{B}' \\ \mathbf{B} & \mathbf{C}' \end{pmatrix}$$

where $\mathbf{U}_2$ is the $J \times 1$ vector and $\mathbf{C}_R$ and $\mathbf{C}$ are $J \times J$ matrices. Under $H_0 : \boldsymbol{\theta} = 0$, both the proposed robust GOF test statistic

$$Q_R = \mathbf{U}_2'(\mathbf{C}_R - \mathbf{B}_R \mathbf{A}_R^{-1} \mathbf{B}_R')^{-} \mathbf{U}_2$$

and the proposed model-based GOF test statistic

$$Q = \mathbf{U}_2'(\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-} \mathbf{U}_2$$

are asymptotically distributed as chi-square random variables with

$$\text{d.f} = \text{rank}((\mathbf{C}_R - \mathbf{B}_R \mathbf{A}_R^{-1} \mathbf{B}_R')^{-}) = \text{rank}((\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-}),$$

where $-$ is the generalized inverse.

The authors reported that the statistics had high power for detecting omission of a quadratic term but low power for detecting omission of an interaction term. A disadvantage of these statistics is that test statistics and the degrees of freedom depend upon the subjective choice of covariate partitioning. The authors state that the statistics may be best employed when only discrete covariates are available because there will be no need to subjectively partition the covariate space. If the model contains only continuous covariates, then one may use medians, quartiles, or other percentiles to partition the space depending on sample size. However, in order to ensure the asymptotic properties of the tests, each partitioned region should contain at least 10 clusters; only 25 per cent of the regions should have less than 25 clusters; and the frequency of a responses and non-responses should be non-zero within each region.

### *3.2. A statistic using groups based on ranked estimated probabilities*

Horton *et al.* [8] have developed a statistic using predicted deciles of risk which can be viewed as an extension of the Hosmer–Lemeshow GOF statistic [9] for OLR to logistic GEE models. The statistic has an approximate chi-square distribution when the model is correctly specified and is appropriate for both categorical and continuous covariates.

The authors formed $G$ groups of approximately equal size based on the estimated probabilities:

$$\hat{\pi}_{ij} = \left( \frac{e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'_1 \mathbf{x}_{ij}}}{1 + e^{\hat{\beta}_0 + \hat{\boldsymbol{\beta}}'_1 \mathbf{x}_{ij}}} \right)$$

Define $(G-1)$ group indicators

$$I_{ijg} = \begin{cases} 1 & \text{if } \hat{\pi}_{ij} \text{ is in group } g, \\ 0 & \text{if otherwise,} \end{cases} \qquad g = 1, \ldots, G-1$$

The authors considered the alternative model

$$\text{logit}(\pi_{ij}) = \beta_0 + \boldsymbol{\beta}'_1 \mathbf{x}_{ij} + \gamma_1 I_{ij,1} + \cdots + \gamma_{G-1} I_{ij,G-1}$$

to test GOF

$$H_0 : \gamma_1 = \cdots = \gamma_{G-1} = 0$$

using the score statistic

$$X^2 = \mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)' \{\widehat{\text{var}}[\mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)]\}^{-1} \mathbf{u}_2(\hat{\boldsymbol{\beta}}, 0)$$

$X^2$ is distributed as $\chi^2_{G-1}$ under the null where

$$\mathbf{u}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \begin{bmatrix} \mathbf{u}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) \\ \mathbf{u}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}) \end{bmatrix} = \sum_{i=1}^{K} \begin{bmatrix} \mathbf{D}'_{1i} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \\ \mathbf{D}'_{2i} \mathbf{V}_i^{-1} [\mathbf{Y}_i - \boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})] \end{bmatrix}$$

with

$$\mathbf{D}_{1i} = \partial[\boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]/\partial \boldsymbol{\beta}, \quad \mathbf{D}_{2i} = \partial[\boldsymbol{\pi}_i(\boldsymbol{\beta}, \boldsymbol{\gamma})]/\partial \boldsymbol{\gamma}, \boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_{G-1}]'$$

The $\hat{\boldsymbol{\beta}}$ is obtained from solving $\mathbf{u}_1(\hat{\boldsymbol{\beta}}, 0) = 0$.

The authors noted that although the statistic is easily interpretable, it may miss important deviations from fit and can only directly test covariates that are in the model. Specifically the statistic could have low power in small samples and low power to detect specific alternatives, but may have broad power to detect an array of general alternatives. A disadvantage of this test is that again the test depends upon subjective partitioning (i.e. choosing the number of groups).

### *3.3. Statistics based on residuals*

Pan [10], Evans [11] and Evans and Hosmer [12] developed two statistics (a Pearson chi-square and an unweighted residual sums of squares) comparing observed versus predicted values (i.e. using residuals). Additional details may be found in Reference [13]. These statistics can be easily applied in practice and are natural extensions of test statistics currently used in OLR. The statistics are theoretically appropriate for ungrouped binary data (i.e. when the number of covariate patterns in the model is larger than the sample size such as when continuous covariates are present). Contrary to other statistics introduced thus far, these statistics do not rely on subjective partitioning.

Let $\mathbf{Y}=[\mathbf{Y}_1',\ldots,\mathbf{Y}_K']', \boldsymbol{\pi}=[\boldsymbol{\pi}_1',\ldots,\boldsymbol{\pi}_K']'$, $\mathbf{A}=\mathrm{diag}(\mathbf{A}_1,\ldots,\mathbf{A}_K), \mathbf{V}=\mathrm{diag}(\mathbf{V}_1,\ldots,\mathbf{V}_K)$, $\mathbf{X}=[(1\ \mathbf{x}_{11}')',(1\ \mathbf{x}_{12}')',\ldots,(1\ \mathbf{x}_{i,j}')',(1\ \mathbf{x}_{i,j+1}')',\ldots,(1\ \mathbf{x}_{K,n}')']'$, the $(nK)\times(p+1)$ design matrix including the intercept, $\mathbf{H}=\mathbf{AX}(\mathbf{X}'\mathbf{AV}^{-\mathbf{1}}\mathbf{AX})^{-1}\mathbf{X}'\mathbf{AV}^{-1}$, and $\mathbf{e}=\mathbf{Y}-\boldsymbol{\pi}$.

The Pearson chi-square statistic is

$$G=\sum_{i=1}^{K}\sum_{j=1}^{n}\frac{(Y_{ij}-\hat{\pi}_{ij})^2}{\hat{\pi}_{ij}(1-\hat{\pi}_{ij})}=Kn+(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}\hat{\mathbf{e}}\approx Kn+(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{e}$$

with $\widehat{E(G)}=Kn$, and $\widehat{\mathrm{var}(G)}=(1-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{A}}^{-1}(\mathbf{I}-\hat{\mathbf{H}})\widehat{\mathrm{cov}(\mathbf{Y})}(\mathbf{I}-\hat{\mathbf{H}}')\hat{\mathbf{A}}^{-1}(1-2\hat{\boldsymbol{\pi}})$. Pan used two estimates of $\mathrm{cov}(\mathbf{Y})$. The first is the empirical covariance estimator: $\widehat{\mathrm{cov}(\mathbf{Y})}_{\mathrm{e}}=\mathrm{diag}(\widehat{\mathrm{cov}(\mathbf{Y}_1)},\ldots,\widehat{\mathrm{cov}(\mathbf{Y}_K)})(G)$, where $\widehat{\mathrm{cov}(\mathbf{Y}_i)}=(\mathbf{Y}_i-\hat{\boldsymbol{\pi}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\pi}}_i)'$. The second is, $\widehat{\mathrm{cov}(\mathbf{Y})}_{\mathrm{u}}=\hat{\mathbf{A}}^{1/2}\mathrm{diag}(\hat{R_{\mathrm{u}}},\ldots,\hat{R_{\mathrm{u}}})\hat{\mathbf{A}}^{1/2}(G2)$, where $\hat{R_{\mathrm{u}}}$ is the unstructured correlation matrix estimate, specifically,

$$\hat{R_{\mathrm{u}}}=\frac{1}{K}\sum_{i=1}^{K}\hat{\mathbf{A}}^{-1/2}(\mathbf{Y}_i-\hat{\boldsymbol{\pi}}_i)(\mathbf{Y}_i-\hat{\boldsymbol{\pi}}_i)'\hat{\mathbf{A}}^{-1/2}$$

The unweighted sums of squares statistic is defined as

$$U=\sum_{i=1}^{K}\sum_{i=1}^{n}(Y_{ij}-\hat{\pi}_{ij})^2=\hat{\boldsymbol{\pi}}'(\mathbf{1}-\hat{\boldsymbol{\pi}})+(\mathbf{1}-2\hat{\boldsymbol{\pi}})'\hat{\mathbf{e}}\approx\hat{\boldsymbol{\pi}}'(\mathbf{1}-\hat{\boldsymbol{\pi}})+(\mathbf{1}-2\hat{\boldsymbol{\pi}})'(\mathbf{I}-\mathbf{H})\mathbf{e}$$

Its mean and variance are approximately

$$\widehat{E(U)}=\hat{\boldsymbol{\pi}}'(\mathbf{1}-\hat{\boldsymbol{\pi}}),\quad \widehat{\mathrm{var}(U)}=(\mathbf{1}-2\hat{\boldsymbol{\pi}})'(\mathbf{I}-\hat{\mathbf{H}})\widehat{\mathrm{cov}(\mathbf{Y})}(\mathbf{I}-\hat{\mathbf{H}}')(\mathbf{1}-2\hat{\boldsymbol{\pi}})$$

and $\mathrm{cov}(\mathbf{Y})$ is estimated using either $\widehat{\mathrm{cov}(\mathbf{Y})}_{\mathrm{e}}(U)$ or $\widehat{\mathrm{cov}(\mathbf{Y})}_{\mathrm{u}}(U2)$ described above. Both $G$ and $U$ have approximately normal distributions.

Pan recommended estimating $\mathrm{cov}(\mathbf{Y})$ using the unstructured correlation matrix $(G2, U2)$ because they are more efficient. He also suggested the use of the working independence model rather than other correlation structures when not knowing the true within cluster correlation structure.

Pan noted that power of these statistics may be limited in practice since the statistics are designed to detect some general model departures. Formulating more specific alternative hypotheses and forming a related test may improve power.

### 3.4. A classification statistic

Williamson *et al.* [14] proposed a kappa-like classification statistic for assessing GOF for logistic GEE models. The summary measure depicts how well binary responses are predicted from the model.

The general expression for the kappa statistic is

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where $P_o$ is the observed proportion of agreement and $P_e$ is the proportion of agreement expected by chance alone. Williamson *et al.* fitted an intercept-only model to estimate $P_e$. For simplicity, assume $n_i = n$ for $i = 1, \ldots, K$. Then

$$\hat{P}_e = (Y/nK)^2 + (1 - Y/nK)^2$$

where $Y = \sum_{i=1}^{K} \sum_{j=1}^{n} Y_{ij}$, the number of observations with a positive outcome. Define $P_{oij}$ to be the probability that the predicted response from the model for $j$th observation in cluster $i$ is equal to the observed response, and $\hat{\zeta}_{ij} = \hat{\pi}_{ij}^{Y_{ij}}(1 - \hat{\pi}_{ij})^{1-Y_{ij}}$. Let $\mathbf{P}_{oi}$ and $\mathbf{U}_i$ denote the $n \times 1$ vectors $[P_{oi1}, \ldots, P_{oin}]'$ and $[\hat{\zeta}_{i1}, \ldots, \hat{\zeta}_{in}]'$. Noting that $P_{oij} = P_e + \kappa(1.0 - P_e)$, they estimate an overall $\kappa$ by solving a set of estimating equations.

$$\mathbf{v}_2(\kappa, \boldsymbol{\beta}) = \sum_{i=1}^{K} \mathbf{C}_i' \mathbf{W}_i^{-1} \{\mathbf{U}_i(\boldsymbol{\beta}) - \mathbf{P}_{oi}(\kappa)\} = 0$$

where $\mathbf{C}_i = \mathrm{d}\mathbf{P}_{oi}/\mathrm{d}\kappa = [1 - \hat{P}_e, \ldots, 1 - \hat{P}_e]'$ and $\mathbf{W}_i \approx \mathrm{var}(\mathbf{U}_i)$ is the $n \times n$ working covariance matrix of $\mathbf{U}_i$. The equations should be solved iteratively if $\mathbf{W}_i$ is different from the identity matrix.

The value of $\kappa$ is one if prediction is perfect and zero if the fitted model predicts no better than chance. The kappa-like statistic is a more appropriate indicator of how well the model predicts responses at the cluster level as opposed to how well the model fits the data at the group level (e.g. treatment group). The measure is intuitive, estimating the probability of being correctly predicted by the fitted model and this probability is corrected for chance. An advantage of this statistic is that no subjective decisions need to be made concerning partitioning. However, interpretation of the statistic is not trivial since no distribution of the statistic is given. Similar to Landis and Koch's [15] labelling of kappa values, the authors recommended that: a $\kappa$ value from 0.00 to 0.20 represents poor fit, a value from 0.21 to 0.40 represents fair fit, a value from 0.41 to 0.60 represents good fit, and a value from 0.61 to 1.00 represents excellent fit.

### 3.5. Hybrid statistics

Using a similar idea to Pulkstenis and Robinson [16] in OLR, we develop hybrid statistics which combine the ideas of covariate partitioning and using groups based on ranked estimated probabilities. However, our approach differs from Pulkstenis and Robinson in that we utilize score statistics whereas Pulkstenis and Robinson use deviance and chi-square statistics.

We first partition the covariate space into $M$ distinct regions using the categorical covariates. We then partition each region into two parts based on the predicted probabilities within each region. Each observation then belongs to one of $2M$ distinct regions. Let $\mathbf{O}_{ij} = [I_{ij1}, \ldots, I_{ij,2M-1}]'$ be an $(2M-1)\times 1$ vector, where $I_{ijm}$ is the indicator variable that equals one if the $j$th observation in $i$th cluster is in the $m$th region and zero otherwise using the last region as a reference. $\mathbf{O}_i = [\mathbf{O}_{i1}, \ldots, \mathbf{O}_{in}]'$ is a $n\times(2M-1)$ matrix. Consider the alternative model:

$$\text{logit}(\boldsymbol{\pi}_i) = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{O}_i\boldsymbol{\theta}$$

where $\mathbf{X}_i$ is the $n\times(p+1)$ design matrix for $i$th cluster (including the intercept), $\theta$ is a $(2M-1)\times 1$ vector, and use the score statistic to test $H_0 : \boldsymbol{\theta} = 0$. We now have exactly the same definition for $U$, $W$, and $W_R$ as in Section 3.1. The hybrid robust score statistic is

$$N_R = \mathbf{U}_2'(\mathbf{C}_R - \mathbf{B}_R\mathbf{A}_R^{-1}\mathbf{B}_R')^{-}\mathbf{U}_2$$

and the hybrid model-based score statistic is

$$N_m = \mathbf{U}_2'(\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-}\mathbf{U}_2$$

Since both hybrid statistics are score statistics and we use the same variance estimators as in Section 3.1, we may use the asymptotic property of the score test for $M$-estimators and the fact that the GEE parameter estimates are consistent, to obtain the asymptotic distribution of the statistics. Both statistics are asymptotically distributed as chi-square random variables with

$$\text{d.f.} = \text{rank}((\mathbf{C}_R - \mathbf{B}_R\mathbf{A}_R^{-1}\mathbf{B}_R')^{-}) = \text{rank}((\mathbf{C} - \mathbf{B}\mathbf{A}^{-1}\mathbf{B}')^{-})$$

Note that the partitioning of the covariate space is again subjective and using only categorical covariates is preferred. However if the sample size is large and there are a small number of categorical covariates, then using the median or quartiles of continuous covariates to help partition the covariate space is acceptable. However to ensure the asymptotic property of the statistics, sparse cell counts should be avoided. Using exploratory simulations and based upon similar recommendations from Barnhart and Williamson [6], we recommend that each partitioned covariate region should contain at least 10 clusters; only 25 per cent of the regions contain less than 25 clusters; and the frequency of responses and non-responses in each region should be non-zero.

## 4. SIMULATION RESULTS

A simulation study evaluated type I error and power of each of the statistics.

Note that for the statistics based on covariate partitioning and the hybrid statistics which need to partition the covariate space into regions, we used only the categorical covariates to partition the covariate space when categorical covariates were in the model. If the model only contained continuous covariates, then we used medians of the continuous variables to partition the covariate space.

### *4.1. Null distribution*

We consider a number of different models to examine the performance of the statistics when the GEE model is correct. Table I provides a list of the investigated models and simulation study factors. These scenarios provide the opportunity to assess the effect of several factors including: the magnitude of the correlation, cluster-level versus time-varying covariates, covariate distributions, the number of covariates, the number of clusters, the number of observations within a cluster, and the correlation structure. Data were generated using the methods of Qaqish [17]. For each model, correlated binary variables with a specified mean vector were generated such that the intercept was zero and all logistic coefficients were 0.8. After fitting the correct model, each statistic was calculated to evaluate model inadequacy. This process was replicated 1000 times and the type I error rate was aggregately obtained. Table II displays the resulting type I error rates.

With higher magnitude of correlation, the type I error rates of the four statistics based on residuals are inflated. Covariate level does not have much influence on any of the statistics except $N_R$ which has an inflated type I error rate with only cluster level covariates. $G$ displays inflated type I error rates with more covariates. A larger number of clusters tends to decrease the rejection rate of $N_m$ and $N_R$. $N_R$ and $Q_R$ tend to have higher rejection rates than $N_m$ and $Q$, respectively, under most scenarios. Compared with an exchangeable correlation structure, AR(1) tends to slightly decrease the type I error rates of $X^2$ and slightly increase that of other statistics. $X^2$ is conservative with small sample sizes while $Q_R$ and $N_R$ tend to have inflated type I error rates because the empirical variance estimator is known to be highly variable with small sample sizes.

We note that the performance of $G$, $G2$, $U$, and $U2$ is relatively robust to the assumption that number of covariate patterns should be close to sample size. Note that these statistics behave reasonably well when the model only contains categorical covariates. We further note that the hybrid statistics do surprisingly well even with only continuous covariates in the model.

$G$, $G2$, and $X^2$ are extremely conservative with skewed covariate distributions. $X^2$ displays inflated type I error rates in model 17 with six continuous covariates. None of the statistics are robust to the inclusion of two chi-square distributed covariates as in model 13; note that $U$, $U2$, $Q$ and $N_m$ all have type I error rates significantly higher than target levels.

$Q_R$ and $N_R$ have greatly inflated type I error rates in some models; this may be because in those models the probability of $Y_{ij}$ equals 1 is very high in some covariate regions; thus the low frequency of $Y_{ij}=0$ causes problems in the estimation of the robust variance. When the sample size is very large (e.g. 500 clusters), the frequency of $Y_{ij}=0$ is moderate even with low probability, and thus the two statistics have reasonable performance.

$\kappa$ displays extremely inflated type I error rates in nearly all scenarios when utilizing the recommended ‘poor fit’ cut-off of less than 0.2. The type I error rate is greater than 90 per cent in many scenarios.

### *4.2. Power*

We further investigate the power of the statistics to detect various model departures. Data were generated similarly to that described in the previous section. However, the fitted models were incorrect.

Table I. Simulation study factors for the null distribution investigation.

| | | | | Correlation | | |
|---|---|---|---|---|---|---|
| Model | Covariate distribution | Covariate level* | Dimension† | Magnitude | Structure | Comparison model and factor evaluated |
| 1 | U[−1,1], U[−1,1] | T, T | 100 × 2 | 0.2 | Exchangeable | Anchor |
| 2 | U[−1,1], U[−1,1] | T, T | 100 × 2 | 0.6 | Exchangeable | #1 Magnitude of correlation |
| 3 | B(0.5),B(0.5) | T, T | 100 × 2 | 0.2 | Exchangeable | #1 Covariate distribution |
| 4 | B(0.2), B(0.2) | T, T | 100 × 2 | 0.2 | Exchangeable | #1,#3 Covariate distribution |
| 5 | U[−1,1], U[−1,1] | C, C | 100 × 2 | 0.2 | Exchangeable | #1 Covariate level |
| 6 | U[−1,1], B(0.5) | T, T | 100 × 2 | 0.2 | Exchangeable | #1 More covariates |
| | U[−1,1], B(0.5) | C, C | | | | |
| 7 | U[−1,1], B(0.2) | T, T | 100 × 2 | 0.2 | Exchangeable | #1 More covariates |
| | U[−1,1], B(0.2) | C, C | | | | #6 Covariate distribution |
| 8 | U[−1,1], U[−1,1] | T, T | 250 × 2 | 0.2 | Exchangeable | #1 Large number of clusters |
| 9 | U[−1,1], U[−1,1] | T, T | 100 × 5 | 0.2 | Exchangeable | #1 Large cluster size |
| 10 | U[−1,1], U[−1,1] | T, T | 100 × 5 | 0.2 | AR(1) | #1 Correlation structure |
| 11 | U[−1,1], U[−1,1] | T, T | 100 × 5 | 0.6 | AR(1) | #1, #10 Correlation structure and magnitude |
| 12 | N(0,1), $\chi_3$ | T, T | 100 × 2 | 0.2 | Exchangeable | #1 Covariate distribution |
| 13 | $\chi_3$, $\chi_3$ | T, T | 100 × 2 | 0.2 | Exchangeable | #1, #12 Covariate distribution |
| 14 | U[−1,1], U[−1,1] | T, T | 25 × 2 | 0.2 | Exchangeable | #1 Small sample size |
| 15 | U[−1,1], U[−1,1] | T, T | 50 × 2 | 0.2 | Exchangeable | #1, #14 Small sample size |
| 16 | U[−1,1], B(0.5) | T, T | 500 × 2 | 0.2 | Exchangeable | #1, #6 Large sample size |
| | U[−1,1], B(0.5) | C, C | | | | |
| 17 | U(−1,1), U(−3,3), N(0,1) | T, C, T | 700 × 2 | 0.2 | Exchangeable | #1 Six continuous covariates |
| | N(0,2), $N(0,1)^2$,$U(-1,1)^2$ | T, T, T | | | | |

*Cluster-level($C$) versus time-varying($T$).

†Number of clusters by number of observations within cluster.

Table II. Type I error rates at the 5/10 per cent (poor for $\kappa$) levels.

| | Statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G^*$ | $G2^\dagger$ | $U^\ddagger$ | $U2$ | $X^{2\S}$ | $Q^{\P}$ | $Q_R$ | $N_m{}^{\|\|}$ | $N_R$ | $\kappa^{**}$ |
| 1 | 4.1/9 | 4.2/9.1 | **3.1**/9.3 | 4.3/8.6 | 6.0/10.6 | 4.8/9.1 | 5.0/11.1 | 4.1/8.9 | 4.9/11.0 | 98 |
| 2 | **6.8**/**12.2** | **7.2**/**12.9** | **7.4**/**13.3** | **7.2**/**13.6** | 3.8/9.0 | 5.1/9.0 | 6.2/11.4 | 4.3/10.2 | **6.8**/11.4 | 98.3 |
| 3 | 5.1/10.1 | 5.0/10.3 | 5.4/10.6 | 6.1/11.0 | 6.0/**12.0** | 4.6/9.8 | **8.2**/**13.6** | 4.1/8.5 | 4.7/11.0 | 100 |
| 4 | 5.8/**12.2** | 4.8/9.3 | 5.3/11.3 | 5.3/10.2 | 4.6/10.3 | 5.6/11.0 | **14.3**/**24.2** | 5.2/11.1 | 5.6/**12.6** | 100 |
| 5 | 5.2/10.9 | 4.6/9.3 | 4.8/10.9 | 4.3/9.1 | 4.1/9.9 | 4.9/9.7 | 5.8/11.6 | 5.3/11.5 | **6.5**/**14.5** | 97.8 |
| 6 | **6.7**/**13.5** | 3.85/**7.6** | 6.0/11.4 | 5.9/11.3 | 5.3/**12.0** | 4.8/11.3 | **8.4**/**15.1** | **6.9**/**13.2** | **11.2**/**20.7** | 81.5 |
| 7 | 5.4/**12.8** | 4.9/8.5 | 5.0/10.0 | 4.6/10.3 | 5.8/**12.5** | 5.2/9.9 | **10.2**/**18.4** | 5.2/10.1 | **7.4**/**14.9** | 88.7 |
| 8 | 4.3/8.8 | 4.0/9.0 | 4.9/10.1 | 5.1/9.7 | 4.5/9.8 | 4.8/9.0 | 4.6/10.2 | **3.2**/**7.6** | **3.5**/8.9 | 100 |
| 9 | 4.9/9.6 | 5.1/9.7 | 5.0/10.2 | 5.1/9.9 | 4.2/9.8 | 5.1/11.6 | 5.4/11.0 | 4.3/10.5 | 5.0/11.2 | 100 |
| 10 | 4.8/8.7 | 4.9/9.3 | 5.2/9.9 | 5.1/10.4 | 3.9/9.6 | 6.2/10.8 | 5.4/**12.0** | 4.7/10.1 | 4.2/10.8 | 100 |
| 11 | 6.3/11.8 | 6.3/11.5 | 6.3/**12.5** | **6.4**/12.5 | **3.6**/8.9 | **7.1**/**14.9** | 5.2/**12.8** | 4.0/8.9 | 4.5/10.9 | 100 |
| 12 | **0.2**/**3.0** | **0.7**/**0.9** | 5.0/10.8 | 5.2/10.8 | **2.9**/**7.7** | 5.7/10.8 | **63.3**/**71.2** | 4.6/10.6 | **14.7**/**23.7** | 30.2 |
| 13 | **0**/**0.8** | **0.0**/**0.0** | **6.6**/**15.3** | **13.0**/**19.5** | **1.0**/**1.0** | **10.5**/**13.5** | **97.5**/**98.4** | **8.1**/**13.2** | **83.0**/**89.7** | 83.0 |
| 14 | 5.0/11.1 | 4.6/8.9 | 5.8/**12.0** | 6.1/11.6 | **1.2**/**4.8** | 5.8/9.7 | **6.5**/11.3 | 4.9/9.5 | **7.5**/**14.4** | 81.2 |
| 15 | 5.2/9.6 | 5.9/9.4 | 5.5/9.9 | 5.3/9.9 | 4.8/10.2 | 4.9/10.7 | 5.9/11.8 | **3.5**/**7.8** | 4.6/10.6 | 92.3 |
| 16 | **7.9**/11.3 | 4.5/9.4 | 4.9/9.5 | 4.7/9.5 | 4.4/10.2 | 3.9/9.9 | 5.2/10.8 | 5.4/8.9 | 6.1/10.3 | 99.8 |
| 17 | **19.7**/**27.5** | **0.9**/**1.1** | 5.5/10.9 | 5.8/10.7 | **23.9**/36.0 | 4.8/10.0 | **29.6**/**38.6** | 6.1/11.3 | **15.1**/**21.8** | 0.0 |

*Note*: approximate 95 per cent confidence intervals are obtained using ±1.35/1.86, and numbers falling out of CI are bolded.

*Pearson chi-square statistics based on residuals.

†$G$ and $U$ use empirical covariance estimator, $G2$ and $U2$ use unstructured correlation matrix estimate.

‡Statistics based on an unweighted sum of residual squares.

§Statistic based on ranked estimated probabilities.

¶Statistics based on covariate partitioning. $Q$ is model-based, $Q_R$ is robust.

||Hybrid statistics, $N_m$ is model-based, $N_R$ is robust.

**Percentage of time $\kappa \leqslant 0.2$.

Table III. Simulation study factors for the alternative distribution investigation.

| Model | Correlation | Dimension | Model | Correlation | Dimension |
|---|---|---|---|---|---|
| 1 | 0.2 | 50 × 2 | 6 | 0.6 | 50 × 2 |
| 2 | 0.2 | 100 × 2 | 7 | 0.6 | 100 × 2 |
| 3 | 0.2 | 250 × 2 | 8 | 0.6 | 250 × 2 |
| 4 | 0.2 | 100 × 5 | 9 | 0.6 | 100 × 5 |
| 5 | 0.2 | 100 × 20 | 10 | 0.6 | 100 × 20 |

We examined the power of the tests of fit to detect four particular types of model departures. The situations studied were: (1) the omission of a covariate, (2) the omission of an interaction term, (3) the omission of quadratic term, and (4) an incorrectly specified link function. For each situation, we compared results of different correlations and dimensions to access the effect of following factors: the magnitude of the correlation, number of clusters, and the number of observations within a cluster. Exploratory simulations using different correlation structures revealed that correlation structure had little effect on the power of the statistics. Table III displays the simulation study schema for the power investigation. In all situations, $X_{i,1}$

Table IV. Power for omitted covariate at the $\alpha = 5/10$ per cent (poor/fair for $\kappa$) levels.

| | Statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G$ | $G2$ | $U$ | $U2$ | $X^2$ | $Q$ | $Q_R$ | $N_m$ | $N_R$ | $\kappa$ |
| 1 | 6.5/13.5 | 4.8/9.9 | 4.9/10.0 | 5.5/10.5 | 2.6/8.1 | 4.8/11.1 | 7.2/13.2 | 5.4/10.4 | 7.0/14.0 | 14.6/**89.1** |
| 2 | 9.5/16.8 | 5.0/10.9 | 7.6/13.2 | 6.2/12.5 | 6.5/13.3 | 3.8/8.1 | 10.4/16.2 | 6.3/13.3 | 6.1/12.3 | 9.1/**97.5** |
| 3 | 8.1/12.4 | 5.4/10.4 | 4.9/10.0 | 4.1/9.7 | 5.6/11.8 | 4.1/7.9 | 5.9/9.7 | 4.7/10.2 | 5.0/10.3 | 2.1/**99.9** |
| 4 | 6.2/11.7 | 6.2/11.5 | 5.5/10.3 | 6.2/10.6 | 5.5/11.7 | 6.3/11.8 | 10.1/17.7 | 5.6/12.5 | 7.0/15.4 | 3.6/**100.0** |
| 5 | 9.4/16.3 | 9.8/17.4 | 9.6/16.3 | 10.2/18.1 | 3.9/8.5 | 10.3/17.9 | 0.0/0.1 | 6.0/11.7 | 6.2/12.9 | 0.0/**100.0** |
| 6 | 7.6/13.3 | 6.8/11.7 | 6.6/11.7 | 6.6/12.8 | 2.8/7.9 | 5.8/10.7 | 6.4/12.3 | 5.5/11.1 | 8.1/14.1 | 18.8/**91.4** |
| 7 | 9.7/15.4 | 5.5/10.6 | 7.7/13.8 | 5.3/12.0 | 5.8/12.9 | 3.0/6.4 | 10.9/17.5 | 11.9/20.1 | 7.7/14.0 | 13.7/**98.3** |
| 8 | 8.8/14.3 | 8.0/12.2 | 6.6/12.6 | 7.4/13.8 | 6.2/11.4 | 3.7/9.1 | 6.0/11.8 | 5.0/10.5 | 5.5/11.1 | 5.6/**100.0** |
| 9 | 7.6/12.9 | 8.0/13.8 | 7.4/13.6 | 7.9/14.5 | 4.4/10.0 | 6.6/12.3 | 8.2/14.3 | 6.8/12.0 | 7.8/12.6 | 15.9/**100.0** |
| 10 | 9.9/18.1 | 13.5/21.6 | 10.0/18.5 | 13.4/21.7 | 3.4/9.3 | 17.5/26.2 | 0.0/0.0 | 7.1/13.3 | 5.5/9.5 | 6.8/**100.0** |

*Note*: moderate or high power are bolded (over 50 per cent).

is a cluster-level covariate distributed as Bernoulli(0.5), and $X_{ij,2}$ is a time-varying covariate distributed as U$(-3, 3)$.

We defined the power to be: ‘high’ if it is over 75 per cent, ‘moderate’ if less or equal to 75 per cent but greater than 50 per cent, ‘low’ if less or equal to 50 per cent but greater than 25 per cent, and ‘very low’ if less or equal to 25 per cent.

*4.2.1. Power for detecting omitted covariate.* We generated data using the following model: $\text{logit}[\pi(\mathbf{x}_{ij})] = 0.8X_{i,1} + 0.8X_{ij,2} + 0.8X_{ij,3}$, where $X_{ij,3}$ is distributed as N$(0, 1)$. We tested the power for omission of $X_{ij,3}$. Table IV displays the results.

All statistics have low power in nearly all of the scenarios considered. Only $Q_R$, $N_m$, and $\kappa$ display power above 10 per cent (but still less than 20 per cent) in some limited scenarios.

*4.2.2. Power for detecting omitted interaction term.* We generated data using the following model: $\text{logit}[\pi(\mathbf{x}_{ij})] = 0.8X_{i,1} + 0.8X_{ij,2} + 0.8X_{i,1}X_{ij,2}$. We tested the power for omission of an interaction term $X_{i,1}X_{ij,2}$. Table V displays the results.

The statistics based on covariate partitioning, $Q$ and $Q_R$ have moderate power under some scenarios considered. More clusters increases power. $G$ and $G2$ have moderate power with large cluster size and high magnitude of correlation. The power for the other statistics is consistently very low across scenarios. Larger cluster sizes also improves power for $Q$, $G$, $G2$ but not for $Q_R$.

*4.2.3. Power for detecting omitted quadratic term.* We generated data using the following model: $\text{logit}[\pi(\mathbf{x}_{ij})] = 0.8X_{i,1} + 0.8X_{ij,2} + 0.8X_{ij,2}^2$. We tested the power for omission of a quadratic term $X_{ij,2}^2$. Table VI displays the results.

$X^2$ has very high power (e.g. at least 85 per cent) when the number of clusters is at least 100. $G$, $G2$, $U$, and $U2$ have moderate or high power when the number of clusters is at least 100 and the cluster size is not too large. In general $G$ and $U$ consistently have higher power than $G2$ and $U2$, respectively.

Table V. Power for omitted interaction term at the $\alpha = 5/10$ per cent (poor/fair for $\kappa$) levels.

| | Statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G$ | $G2$ | $U$ | $U2$ | $X^2$ | $Q$ | $Q_R$ | $N_m$ | $N_R$ | $\kappa$ |
| 1 | 3.9/10.8 | 3.5/7.0 | 4.4/10.1 | 5.7/9.9 | 1.3/5.2 | 5.7/10.4 | 9.0/14.9 | 8.5/14.9 | 15.2/21.8 | 0.3/33.4 |
| 2 | 5.4/11.2 | 2.8/6.4 | 4.3/9.8 | 4.4/9.8 | 4.0/9.2 | 23.2/33.4 | 47.7/**56.1** | 13.3/21.7 | 15.6/22.6 | 0.0/27.1 |
| 3 | 4.4/10.7 | 4.4/8.8 | 4.7/10.3 | 5.8/10.8 | 12.6/21.7 | **63.4**/**74.7** | **71.0**/**80.5** | 18.7/30.1 | 23.1/33.0 | 0.0/21.5 |
| 4 | 4.7/12.2 | 5.2/10.1 | 6.1/10.7 | 7.1/13.2 | 8.3/19.6 | 44.0/**58.2** | **60.7**/**73.3** | 7.5/13.2 | 8.5/16.3 | 0.0/25.5 |
| 5* | 18.4/26.8 | 22.4/29.5 | 10.2/17.1 | 13.0/19.5 | 13.5/23.7 | **89.3**/**93.1** | 1.57/3.14 | 9.6/15.6 | 5.5/11.3 | 0.0/22.6 |
| 6 | 5.1/11.8 | 6.1/10.5 | 6.3/12.7 | 9.3/15.7 | 1.4/5.7 | 6.1/11.6 | 8.6/14.8 | 7.7/14.0 | 13.0/21.2 | 0.4/40.7 |
| 7 | 4.8/11.1 | 4.4/7.1 | 5.1/10.3 | 5.6/10.9 | 4.3/10.2 | 21.7/32.5 | 46.9/**55.9** | 18.0/26.3 | 14.8/23.6 | 0.0/40.0 |
| 8 | 7.1/12.3 | 5.3/10.0 | 9.2/17.9 | 12.5/21.0 | 10.3/19.2 | **70.4**/**79.3** | **73.7**/**81.6** | 15.9/25.4 | 18.5/28.3 | 0.0/42.3 |
| 9 | 7.3/13.6 | 10.7/18.1 | 8.9/15.1 | 10.1/17.4 | 7.1/15.3 | **57.0**/**67.7** | **53.2**/**69.2** | 7.4/12.9 | 7.7/12.6 | 0.0/**60.0** |
| 10† | 48.7/**53.0** | **52.5**/**56.8** | 16.4/23.6 | 18.7/26.5 | 17.9/31.2 | **93.1**/**97.7** | 16.5/16.9 | 40.4/46.9 | 7.7/14.2 | 0.0/**77.6** |

*Note*: moderate or high power are bolded (over 50 per cent).

*Approximately 350 out of 1000 replications did not converge for all statistics except $X^2$ and $\kappa$.

†Approximately 700 out of 1000 replications did not converge for all statistics except $X^2$ and $\kappa$.

Table VI. Power for omitted quadratic term at the $\alpha = 5/10$ per cent (poor for $\kappa$) levels.

| | Statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G$ | $G2$ | $U$ | $U2$ | $X^2$ | $Q$ | $Q_R$ | $N_m$ | $N_R$ | $\kappa$ |
| 1 | 44.0/**53.2** | 25.8/35.7 | 40.7/49.5 | 30.0/38.6 | 22.7/43.3 | 5.5/10.4 | 14.9/20.3 | 5.7/13.9 | 17.9/25.5 | **100.0** |
| 2 | **74.4**/**81.9** | **52.0**/**62.0** | **76.8**/**83.2** | **58.8**/**68.4** | **90.3**/**95.8** | 30.9/43.1 | 48.1/**55.2** | 9.6/18.3 | 13.2/20.4 | **100.0** |
| 3 | **94.9**/**97.2** | **84.0**/**90.0** | **94.6**/**96.6** | **87.8**/**93.3** | **99.9**/**100.0** | **68.2**/**77.7** | **57.9**/**70.2** | 10.4/18.0 | 11.9/20.6 | **100.0** |
| 4 | 44.9/**55.3** | 36.8/47.6 | 12.6/21.1 | 13.1/21.5 | **100.0**/**100.0** | 46.4/**57.5** | **71.2**/**78.9** | 4.5/11.5 | 7.6/14.6 | **100.0** |
| 5 | 11.7/18.3 | 10.1/16.1 | 14.0/23.1 | 14.1/23.2 | **100.0**/**100.0** | **87.2**/**92.5** | 0.1/19.3 | 7.3/13.1 | 7.4/13.2 | **100.0** |
| 6 | 30.3/41.2 | 18.0/26.8 | 23.0/31.3 | 17.8/26.1 | 15.0/32.1 | 4.6/9.9 | 14.2/20.3 | 6.1/12.3 | 23.1/28.4 | **99.9** |
| 7 | **71.9**/**79.2** | **50.1**/**60.3** | **73.9**/**80.4** | **57.1**/**66.7** | **86.2**/**93.8** | 25.9/37.6 | 45.5/**53.4** | 15.2/24.7 | 14.8/23.9 | **100.0** |
| 8 | **88.2**/**92.0** | **73.0**/**83.5** | **60.6**/**70.0** | **54.1**/**65.0** | **100.0**/**100.0** | **67.8**/**77.7** | **56.5**/**69.6** | 8.0/14.1 | 9.4/16.7 | **100.0** |
| 9 | 27.8/37.3 | 21.0/30.5 | 6.8/12.9 | 6.8/12.5 | **99.8**/**100.0** | 47.5/**58.9** | **68.1**/**76.9** | 6.1/12.2 | 10.5/16.0 | **100.0** |
| 10 | 7.2/13.2 | 6.2/11.3 | 12.1/20.5 | 11.4/20.0 | **100.0**/**100.0** | **86.5**/**91.9** | 0.6/31.8 | 8.9/16.7 | 5.7/13.7 | 100.0 |

*Note*: moderate or high power are bolded (over 50 per cent).

$Q$ and $Q_R$ have moderate power under some scenarios. $Q$ has high power when cluster size is 20 but increasing cluster size does not improve power for $G$, $G2$, $U$, $U2$, and $Q_R$. $N_m$ and $N_R$ have very low power in all scenarios. Increasing the number of clusters does improve power for $G$, $G2$, $U$, $U2$, $X^2$, $Q$, and $Q_R$. All statistics have very low to low power when the number of clusters is 50.

*4.2.4. Power for detecting incorrect link function.* We generated data using a GEE model with the complimentary log–log link function such that $\log(-\log(\pi(\mathbf{x}_{ij}))) = 0.8X_{i,1} + 0.8X_{ij,2}$. We tested the power for misspecification of the (logit) link function. Table VII displays the results.

$G$ had moderate to high power when the total number of observations (number of clusters times the number of observations within cluster) is large enough. $Q_R$ has moderate to high

Table VII. Power for incorrect link function at the $\alpha = 5/10$ per cent (poor/fair for $\kappa$) levels.

| | Statistic | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G$ | $G2$ | $U$ | $U2$ | $X^2$ | $Q$ | $Q_R$ | $N_m$ | $N_R$ | $\kappa$ |
| 1 | 2.4/13.6 | 0/0.2 | 7.2/17.1 | 7.7/16.3 | 1.9/4.9 | 5.3/10.0 | 8.1/14.8 | 4.1/10.5 | 9.4/17.2 | 0.0/17.9 |
| 2 | 25.8/**51.5** | 0/1.2 | 25.1/37.4 | 15.4/29.1 | 4.2/10.6 | 3.0/5.5 | **89.5/91.6** | 7.3/12.3 | 8.4/13.4 | 0.0/10.9 |
| 3 | **87.5/93.7** | 1.4/13.2 | 32.0/45.2 | 28.2/40.2 | 38.2/**60.5** | 6.2/11.8 | **64.5/66.8** | 6.0/11.5 | 6.5/12.8 | 0.0/3.2 |
| 4 | **74.6/85.3** | 3.9/16.9 | 8.7/15.8 | 11.1/18.9 | 28.4/48.3 | 3.2/7.7 | **87.2/92.9** | 7.6/13.4 | 6.8/12.0 | 0.0/5.0 |
| 5 | **92.2/94.7** | **65.0/80.4** | 14.8/21.9 | 16.2/23.5 | **98.9/99.7** | 10.2/14.0 | 1.3/3.4 | 14.3/21.1 | 5.4/12.1 | 0.0/0.2 |
| 6 | 4.1/14.2 | 0.1/0.9 | 3.8/10.4 | 6.5/12.8 | 0.9/4.2 | 5.4/9.3 | 9.2/14.6 | 4.5/11.6 | 8.4/16.9 | 0.3/22.3 |
| 7 | 22.3/47.6 | 0.1/1.4 | 22.8/35.2 | 14.8/26.8 | 4.0/10.3 | 1.7/4.6 | **85.0/87.9** | 9.4/16.3 | 9.0/15.3 | 0.0/14.2 |
| 8 | **82.0/90.5** | 6.1/24.4 | 4.4/8.8 | 6.6/12.3 | 32.1/**52.6** | 4.2/10.2 | **56.9/60.5** | 5.4/11.9 | 5.7/12.3 | 0.0/8.6 |
| 9 | **65.3/77.0** | 8.6/22.6 | 5.8/10.8 | 8.4/15.2 | 21.3/38.9 | 3.0/7.3 | **65.8/78.6** | 9.4/15.8 | 6.7/11.1 | 0.0/15.9 |
| 10* | **85.7/88.7** | **51.0/67.6** | 21.3/30.3 | 23.3/32.3 | **96.3/98.7** | 18.1/23.4 | 1.53/2.94 | 22.8/31.0 | 7.6/12.4 | 0.0/5.9 |

*Note*: moderate or high power are bolded (over 50 per cent).

*Approximately 140 out of 1000 replications did not converge for all statistics except $X^2$ and $\kappa$.

power when the number of clusters is at least 100 and the number of observations within cluster is not too large. $X^2$ and $G2$ have low power under most scenarios, but have moderate to high power when cluster size is 20. All statistics have very low power when the number of clusters is 50.

## 5. EXAMPLE

We use data from the Wisconsin epidemiologic study of diabetic retinopathy [18] as an example. The goal of this study is to determine the risk factors for diabetic retinopathy. This study has 720 individuals (clusters) and each individual has two binary responses (two observations within each cluster) to indicate the presence or absence of diabetic retinopathy in each of two eyes.

We fit two models to the data. The first model includes four main effects: duration of diabetes, glycosylated haemoglobin level, diastolic blood pressure, and body mass index. The second model adds two additional covariates: the square of duration of diabetes and square of body mass index. The four main effects are all continuous covariates, so we used the medians to partition the covariate space into 16 regions to construct the statistics based on covariate partitioning and the hybrid statistics.

Table VIII displays the results. In the first model, all statistics have $p$-values less than 0.01 except for $G$ which has a $p$-value of 0.166 and $\kappa$ which concludes that fit is 'Fair'. There is some comfort in the consistency that most of the statistics imply that there is evidence of lack of fit.

After adding the two quadratic terms and fitting the second model, the $p$-values for nearly all of the statistics are not significant. The exceptions are that the two hybrid statistics $N_m$ and $N_R$ have $p$-values less than 0.01. Pulkstenis and Robinson [16] note that their hybrid statistics, although distinct from the hybrid statistics presented here, are not recommended for use in OLR when only continuous variables are in the model. We also note an inflated type I error rate for $N_R$, particularly for model 17 which contains six continuous covariates as in

Table VIII. $P$-values of GOF test statistics using example data.

| | $P$-values | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | $G$ | $G2$ | $U$ | $U2$ | $X^2$ | $Q$ | $Q_R$ | $N_m$ | $N_R$ | $\kappa$ |
| 1 | 0.166 | $<0.01$ | $<0.01$ | $<0.01$ | $<0.01$ | $<0.01$ | $<0.01$ | $<0.01$ | $<0.01$ | Fair |
| 2 | 0.161 | 0.228 | 0.117 | 0.113 | 0.240 | 0.371 | 0.171 | $<0.01$ | $<0.01$ | Good |

this example. Thus similar concerns for our hybrid statistics exist for GEE. Also note that the $\kappa$ classification is 'Good' in this model, showing that this model is at least better than the first.

## 6. SUMMARY

We note that summary measures of goodness of fit are only one aspect of thorough model development and evaluation. Proper assessment of fit should also include examination of individual components of summary measures such as fit to individual clusters and individual observations with assessment of regression diagnostics. Preisser and Qaqish [19] proposed diagnostics for GEE and suggest ways in which to deal with influential data. The newly proposed hybrid statistics also have the potential to diagnose the location of the lack of fit similar to the statistics proposed by Pulkstenis and Robinson in OLR. In addition it is particularly important that fitted models are biologically supported. Researchers should further be cautioned that a non-significant GOF test does not provide sufficient evidence of model fit. These statistics actually evaluate evidence of 'lack of fit'. The value of summary measures of GOF is that they may indicate lack of fit thus prompting researcher to search for more appropriate models.

Performance of each of the statistics reviewed in this investigation varies under various scenarios. $Q$, $Q_R$, $X^2$, $N_m$, and $N_R$ rely on subjective partitioning, and all except $X^2$ reportedly are most appropriate when the models contain categorical covariates. $X^2$ is appropriate when the models contain continuous and/or categorical covariates. Our simulations show however that even with only continuous covariates, the statistics do maintain appropriate type I error rates under certain scenarios. $Q_R$ and $N_R$ are sensitive when the number of clusters is small. However when the number of clusters is large, there appears to be enough information to estimate the robust variance.

$G$, $G2$, $U$, and $U2$ do not rely on subjective partitioning. However, as noted by Pan [10] and Evans and Hosmer [12], $G$, $G2$, $U$, and $U2$ may be better approximated by chi-square distribution when the observed number of covariate patterns approximately equals the sample size. This requirement is satisfied when the model contains at least one continuous covariate. However, we note that performance of these statistics with respect to the type I error rates appears relatively robust to this assumption.

Several factors can affect the type I error rate including small sample sizes, high correlation, covariates with skewed distributions, and models that contain only one type of covariate. Which statistic is best depends upon the model scenario. For example $Q$ performs better than

$G$, $G2$, $U$, and $U2$ when the model contains only binary covariates. $\kappa$ performs poorly under all scenarios and is not recommended for general use as a definitive test for GOF.

All tests had low power to detect an omitted covariate. Statistics based on covariate partitioning had the highest power to detect an omitted interaction term whereas $X^2$ had the highest power to detect an omitted quadratic term. $G$ and $Q_R$ had the highest power to detect a misspecified link function. The effect of cluster size varies with statistics and alternatives as well. Large cluster sizes improves power for $Q$ but not for $Q_R$.

Due to the varying results in the simulation study, we propose that researchers do not rely on a single goodness of fit statistic but alternatively use the statistics to compliment each other. Researchers may also consider using a significance level of 0.10 as an alternative to the standardly applied 0.05 to assess lack of fit due to the low power of the statistics in some scenarios.

A SAS macro (goflgee.mac) which computes each of the statistics discussed in this manuscript is available from the authors.

## ACKNOWLEDGEMENTS

The authors would like to thank Dave Hosmer, Lu Tian, the reviewers, and the editor for their helpful comments. The authors further wish to thank Nick Horton, Huiman Barnhart, Wei Pan, Garrett Fitzmaurice, David Zucker, Erik Pulkstenis, and John Williamson for their programming examples. This work was supported in part by the National Institutes of Health Statistical and Data Management Centre Grant Number 5 U01 AI038855-08 and the Neurologic AIDS Research Consortium Grant Number 2 U01 NS0322228-08A1.

## REFERENCES

1. Hosmer DW, Hosmer T, le Cessie S, Lemeshow S. A comparison of goodness of fit tests for the logistic regression model. *Statistics in Medicine* 1997; **16**:965–980.
2. Liang K, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
3. Hardin JW, Hilbe JM. *Generalized Estimation Equations*. CRC Press: Boca Raton, FL, 2002.
4. Zeger SL, Liang K. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**: 121–130.
5. Neuhaus JM. Statistical methods for longitudinal and clustered designs with binary responses. *Statistical Methods in Medical Research* 1992; **1**:249–273.
6. Barnhart HX, Williamson JM. Goodness of fit for GEE modeling with binary responses. *Biometrics* 1998; **54**:720–729.
7. Tsiatis AA. A note on a goodness of fit test for the logistic regression model. *Biometrika* 1980; **67**:250–251.
8. Horton NJ, Bebchuk JD, Jones CL, Lipsitz SR, Catalano PJ, Zahner GEP, Fitzmaurice GM. Goodness of fit for GEE: an example with mental health service utilization. *Statistics in Medicine* 1999; **18**:213–222.
9. Hosmer DW, Lemeshow S. A goodness of fit test for the multiple logistic regression model. *Communications in Statistics*, *Part A: Theory and Methods* 1980; **10**:1043–1069.
10. Pan W. Goodness of fit tests for GEE with correlated binary data. *Scandinavian Journal of Statistics* 2002; **29**(1):101–110.
11. Evans SR. Goodness of fit in two models for clustered binary data. *Ph.D. Dissertation*, University of Massachusetts, 1998.
12. Evans SR, Hosmer DW. Goodness of fit tests for logistic GEE models: simulation results. *Communications in Statistics: Simulation and Computation* 2004; **33**(1), to appear.
13. Hosmer DW, Lemeshow S. *Applied Logistic Regression* (2nd edn). Wiley: New York, 2000.
14. Williamson JM, Lin HM, Barnhart HX. A classification statistic for GEE categorical response models. *Journal of Data Science* 2003; **1**:149–165.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
16. Pulkstenis E, Robinson TJ. Two goodness of fit tests for logistic regression models with continuous covariates. *Statistics in Medicine* 2002; **21**:79–93.

17. Qaqish BF. A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika* 2003; **90**(2):455–463.
18. Klein R, Klein BEK, Moss SE, Davis MD, DeMets DL. The Wisconsin epidemiologic study of diabetic retinopathy: II. Prevalence and risk of diabetic retinopathy when age at diagnosis is less than 30 years. *Archives of Ophthalmology* 1984; **102**:520–526.
19. Preisser JS, Qaqish BF. Deletion diagnostics for generalized estimating equations. *Biometrika* 1996; **83**: 551–562.