Research Article

Statistics in Medicine

Received 28 February 2012, Accepted 13 June 2013 Published online 30 July 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5908

On assessing model fit for distribution-free longitudinal models under missing data

P. Wu,[a*†] X. M. Tu[a,b] and J. Kowalski[c]

The generalized estimating equation (GEE), a distribution-free, or semi-parametric, approach for modeling longitudinal data, is used in a wide range of behavioral, psychotherapy, pharmaceutical drug safety, and healthcare-related research studies. Most popular methods for assessing model fit are based on the likelihood function for parametric models, rendering them inappropriate for distribution-free GEE. One rare exception is a score statistic initially proposed by Tsiatis for logistic regression (1980) and later extended by Barnhart and Willamson to GEE (1998). Because GEE only provides valid inference under the missing completely at random assumption and missing values arising in most longitudinal studies do not follow such a restricted mechanism, this GEE-based score test has very limited applications in practice. We propose extensions of this goodness-of-fit test to address missing data under the missing at random assumption, a more realistic model that applies to most studies in practice. We examine the performance of the proposed tests using simulated data and demonstrate the utilities of such tests with data from a real study on geriatric depression and associated medical comorbidities.

Keywords: goodness of fit; missing at random; weighted generalized estimating equations; score test; small-sample adjusted score test

1. Introduction

Longitudinal designs are employed in a wide range of behavioral, psychotherapy, pharmaceutical drug safety, and healthcare-related research studies. One popular approach for modeling data from such repeated assessments is to use latent variables, or random effects, to describe the intraclass correlation across multiple within-subject outcomes, such as the generalized linear mixed-effect model, which are widely used for longitudinal data [1–3]. The main drawback of such random-effect-based models is their dependence on distribution assumptions for inference, which too often is at odds with real study data. The generalized estimating equation (GEE) is a popular alternative for robust inference under such situations, because by virtue of being distribution free, it effectively improves inference validity for a much broad class of data distributions [1, 4, 5].

However, existing methods for assessing model fit are by and large based on likelihood methods, rendering them inappropriate for distribution-free, or semi-parametric, models such as GEE. In recent years, efforts have been made to develop non-likelihood-based methods for broader applications. For example, Evans and Li [6] reviewed and compared five different methods in testing goodness of model fit for binary responses, ranging the popular Hosmer–Lemeshow test to Kappa-like classification statistics to score tests based on covariates partitioning. Only the score test based on covariates partitioning, initially proposed by Tsiatis for logistic regression [7] and later extended by Barnhart and Willamson to GEE [8], has an asymptotic chi-square distribution. Unfortunately, this GEE-based score test only provides valid inference under the very restricted missing completely at random (MCAR) assumption, which requires

[a]*Department of Biostatistics and Computational Biology, Rochester, NY 14623, U.S.A.*
[b]*Department of Psychiatry, University of Rochester, Rochester, NY 14623, U.S.A.*
[c]*Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, U.S.A.*
**Correspondence to: P. Wu, Biostatistics and Computational Biology, University of Rochester, Rochester, NY 14623, U.S.A.*
[†]*E-mail: pan_wu@urmc.rochester.edu*

that the event of missing a response during the longitudinal study does not depend on any of the variables of interest, observed or otherwise [9]. However, missing values arising in most real studies usually follow the rather more general missing at random (MAR) assumption, which posits that the missingness depends only on observed variables, a more plausible mechanism for most longitudinal studies [9].

In this paper, we propose extensions of this score test to address its limitations so that they provide valid inference when used for assessing model fit under MAR. In Section 2, we present the proposed goodness-of-fit test statistics and derive their asymptotic distributions. In addition, we discuss adjustments for small samples when applying the test statistics in practice. Simulation studies presented in Section 3 provide an evaluation of the performance of the test statistics in terms of type I error and power for hypothesis testing. In Section 3.2, we also demonstrate the utility of the proposed model fit tests using data from a real study on geriatric depression and associated medical comorbidities. We conclude with a discussion in Section 4.

2. Test statistics and asymptotic properties

2.1. Generalized linear models and generalized estimating equations

Consider a random sample of n individuals with T assessment times. Let y_{it} denote response and $\mathbf{x}_{it}$ be a $(p+1)$-dimensional vector of covariates from the ith individual at time t $(i = 1, \ldots, n, t = 1, \ldots, T)$. Consider the following generalized linear model:

$$\mu_{it} = E\left(y_{it}|\mathbf{x}_{it}\right) = h\left(\mathbf{x}_{it}^{\top}\beta\right), \quad v\left(\mu_{it}\right) = \mathrm{Var}\left(y_{it}|\mathbf{x}_{it}\right), \quad 1 \leqslant i \leqslant N,\ 1 \leqslant t \leqslant T, \tag{1}$$

or in a more compact form:

$$\begin{aligned}
&\boldsymbol{\mu}_i = E\left(\mathbf{y}_i \mid \mathbf{x}_i\right) = h\left(\mathbf{x}_i\boldsymbol{\beta}\right), \quad v\left(\boldsymbol{\mu}_i\right) = \mathrm{Var}\left(\mathbf{y}_i \mid \mathbf{x}_i\right), \quad 1 \leqslant i \leqslant N, \\
&\mathbf{y}_i = \left(y_{i1}, \ldots, y_{iT}\right)^{\top}, \quad \mathbf{x}_i = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}\right)^{\top}, \quad h\left(\mathbf{x}_i\boldsymbol{\beta}\right) = \left(h\left(\mathbf{x}_{i1}^{\top}\boldsymbol{\beta}\right), \ldots, h\left(\mathbf{x}_{iT}^{\top}\beta\right)\right)^{\top}, \\
&\boldsymbol{\mu}_i = \left(\mu_{i1}, \mu_{i2}, \ldots, \mu_{iT}\right)^{\top}, \quad \mathrm{Var}\left(\mathbf{y}_i \mid \mathbf{x}_i\right) = \left(\mathrm{Var}\left(y_{i1} \mid \mathbf{x}_{i1}\right), \ldots, \mathrm{Var}\left(y_{iT} \mid \mathbf{x}_{iT}\right)\right)^{\top},
\end{aligned} \tag{2}$$

where $h^{-1}(\cdot)$ is a known link function and $v(\cdot)$ a known function. Inference about $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_P)^{\top}$ is obtained by solving the following GEE:

$$\mathbf{w}_n\left(\boldsymbol{\beta}\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{w}_{ni} = \sum_{i=1}^{n} D_i V_i^{-1} S_i = \mathbf{0}, \tag{3}$$

where

$$D_i = \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\mu}_i, \quad S_i = \mathbf{y}_i - \boldsymbol{\mu}_i, \quad V_i = A_i^{\frac{1}{2}} R\left(\boldsymbol{\alpha}\right) A_i^{\frac{1}{2}}, \quad A_i = \mathrm{diag}_t\left(v\left(\mu_{it}\right)\right),$$

with $\mathrm{diag}_t\left(a_t\right)$ denoting a diagonal matrix with a_t on the tth diagonal and $R(\boldsymbol{\alpha})$ a working correlation matrix parameterized by vector $\boldsymbol{\alpha}$. Although the GEE is a function of both $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, (3) is solved for $\boldsymbol{\beta}$ after substituting an estimate $\widehat{\boldsymbol{\alpha}}$ in place of $\boldsymbol{\alpha}$ [4, 5]. Thus, strictly speaking, the estimate $\widehat{\boldsymbol{\beta}}$ is a function of $\widehat{\boldsymbol{\alpha}}$, that is, $\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}\right)$, and (3) is satisfied at $\widehat{\boldsymbol{\beta}}\left(\widehat{\boldsymbol{\alpha}}\right)$ and $\widehat{\boldsymbol{\alpha}}$.

The matrix $R\left(\boldsymbol{\alpha}\right)$ is generally not equal to the true within-subject correlation of $\mathbf{y}_i$, and its role is to increase efficiency (or decrease standard errors) [5]. The simplest choice is $R\left(\boldsymbol{\alpha}\right) = I_T$, with I_T denoting the $T \times T$ identity matrix. In this working independence model, $\boldsymbol{\alpha}$ is known, and the correlated components of $\mathbf{y}_i$ are treated as if they were independent. Another common choice is to use a uniform compound symmetry correlation matrix, $R\left(\boldsymbol{\alpha}\right) = C_T\left(\rho\right)$, where $\boldsymbol{\alpha} = \rho$ is a single parameter in this exchangeable correlation model. If unknown, $\boldsymbol{\alpha}$ is readily estimated for most applications [4, 5].

Although primary interest lies in $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$ must be estimated before proceeding with the computation of the GEE estimate of $\boldsymbol{\beta}$ by solving the equations in (3). Although the consistency of the GEE estimate $\widehat{\boldsymbol{\beta}}$ is independent of how $\boldsymbol{\alpha}$ is estimated, judicious choices of the type of estimates of $\boldsymbol{\alpha}$ not only ensure the asymptotic normality but also simplify the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. To this end, we require that $\widehat{\boldsymbol{\alpha}}$ be $\sqrt{n}$-consistent, that is, $\widehat{\boldsymbol{\alpha}}$ converges to some point $\boldsymbol{\alpha}$ and $\sqrt{n}\left(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\right)$ is bounded in probability [4, 5].

Most popular estimates of $\widehat{\boldsymbol{\alpha}}$ such as the moment estimate are asymptotically normal and thus are $\sqrt{n}$-consistent. Given such an estimate of $\widehat{\boldsymbol{\alpha}}$, the estimate $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically normal under some mild regularity conditions, that is,

$$\begin{aligned}\sqrt{n}\mathbf{w}_n(\boldsymbol{\beta}) &\to_d N(0,\Sigma_w), \quad \Sigma_w = E\left(D_i^\top V_i^{-1} S_i S_i^\top V_i^{-1} D_i\right), \\ \sqrt{n}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right) &\to_d N\left(0, \Sigma_{\boldsymbol{\beta}} = B^{-1}\Sigma_w B^{-1}\right), \quad B = E\left(D_i^\top V_i^{-1} D_i^\top\right),\end{aligned} \tag{4}$$

where $\to_d$ denotes convergence in distribution. Thus, for large sample size, $\widehat{\boldsymbol{\beta}}$ has an approximate normal distribution, which is typically used for inference about $\boldsymbol{\beta}$ via the Wald test [4, 5]. The asymptotic normal of $\mathbf{w}_n(\boldsymbol{\beta})$ is also useful for constructing alternative score tests, which in general provide more robust inference than Wald tests. We discuss both tests in detail within our setting in Section 2.3.

The choice of $R(\boldsymbol{\alpha})$ and properties associated with estimates for GEE have been extensively discussed in the literature, which are stated for ease of reference without justifications [3, 5, 10]. In particular, the GEE estimate may not be consistent in the presence of time-varying covariates under working correlation structures other than the working independence model [3]. Thus, the working independence model may be used in general to ensure valid inference. Although this simple working correlation structure may incur a substantial loss of efficiency for time-dependent covariates [10] and thus other models such as the uniform compound symmetry matrix may be used in some specific applications to improve power, it suffices for our purpose in this study when comparing the performance between the proposed and standard GEE-based goodness-of-fit tests. We focus on the simplest working independence model in what follows unless otherwise stated.

2.2. *Missing data and weighted generalized estimating equations*

Missing data are a common issue in longitudinal studies. The biggest problem with GEE is its very limited ability to address missing data. For example, missing data arising from many real studies follow the MAR, rather than the more limited MCAR mechanism. Unfortunately, GEE estimates are generally biased under this popular missing data model, unless the working correlation assumption is correct [2, 11]. Thus, GEE has quite limited applications in practice.

To address the bias, we must model the occurrence of missing data and combine the information with the GEE. For convenience, we assume that all subjects are observed at baseline $t = 1$ and missing responses only occur at post-baseline. Within the longitudinal study setting, define a missing (or rather, observed) data indicator as follows:

$$r_{it} = \begin{cases} 1 & \text{if } y_{it} \text{ is observed} \\ 0 & \text{if otherwise} \end{cases}, \quad \boldsymbol{r}_i = (r_{i1},\ldots,r_{iT},)^\top, \quad 1 \leqslant i \leqslant n,\ 1 \leqslant t \leqslant T. \tag{5}$$

Let $\boldsymbol{\Delta}_i = \text{diag}_t(r_{it})$, a diagonal matrix with r_{it} on the tth diagonal. Then, we may use the following revised GEE for inference about $\boldsymbol{\beta}$, if we want to keep all observed data in the analysis:

$$\mathbf{w}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{ni} = \frac{1}{n}\sum_{i=1}^n D_i V_i^{-1}\boldsymbol{\Delta}_i S_i = \mathbf{0}. \tag{6}$$

Unfortunately, the aforementioned equation generally yields biased estimates, unless missing data follow the more restricted MCAR assumption. In order to obtain consistent estimates of $\boldsymbol{\beta}$ under MAR, we can use the inverse probability weights (IPW) estimation method [2, 4, 12].

To illustrate the idea of IPW, consider the relatively simple pre–post design with a homogeneous sample, and inference about the mean response at pre-assessment and post-assessment, $\boldsymbol{\mu}_i = E(\mathbf{y}_i)$. Because all subjects are observed at pre-assessment, $r_{i1} \equiv 1$ $(1 \leqslant i \leqslant n)$. Let

$$\pi_{i2} = \Pr(r_{i2} = 1 \mid \mathbf{y}_i, \mathbf{x}_i), \quad \Delta_{i1} = 1, \quad \Delta_{i2} = \frac{r_{i2}}{\pi_{i2}}, \quad 1 \leqslant i \leqslant n, \tag{7}$$

where $\mathbf{x}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{iT})^\top$. The weighted GEE (WGEE) below, obtained by substituting $\boldsymbol{\Delta}_i = \text{diag}_t(\Delta_{it})$ or a consistent estimate of this quantity into (6), is unbiased, that is, $E(\mathbf{w}_n(\beta,\pi_{i2})) = 0$, thereby yielding consistent estimates of $\boldsymbol{\beta}$ [2]:

$$\mathbf{w}_n(\boldsymbol{\beta},\pi_{i2}) = \frac{1}{n}\sum_{i=1}^n \mathbf{w}_{ni} = \sum_{i=1}^n D_i V_i^{-1}\boldsymbol{\Delta}_i S_i = \sum_{i=1}^n D_i V_i^{-1}\begin{pmatrix} 1 & 0 \\ 0 & \frac{r_{i2}}{\pi_{i2}} \end{pmatrix}\begin{pmatrix} y_{i1}-\mu_{i1} \\ y_{i2}-\mu_{i2} \end{pmatrix} = \mathbf{0}. \tag{8}$$

Because π_{i2} involves unobserved y_{i2}, it is not estimable. However, under MAR, the missing data mechanism only depends on the observed y_{i1} and $\mathbf{x}_{i1}$, reducing (7) to the following:

$$\pi_{i2} = \Pr(r_{i2} = 1 \mid y_{i1}, \mathbf{x}_{i1}), \quad 1 \leqslant i \leqslant n.$$

We may model $\Pr(r_{i2} = 1 \mid y_{i1}, \mathbf{x}_{i1})$ using logistic regression.

For general $T \geqslant 3$, (7) becomes the following:

$$\pi_{it} = \Pr(r_{it} = 1 \mid \mathbf{y}_i, \mathbf{x}_i), \quad \Delta_{i1} = 1, \quad \Delta_{it} = \frac{r_{it}}{\pi_{it}}, \quad 1 \leqslant i \leqslant n, \quad 2 \leqslant t \leqslant T. \tag{9}$$

Again, it is not possible to model π_{it}, because $\mathbf{y}_i$ may not be observed under missing data. However, under MAR, $\{\mathbf{y}_i, \mathbf{x}_i\}$ is replaced by all observed data prior to t. Unfortunately, unlike the pre–post design, there could be many different patterns (2^{T-1} total) in the observed data, making modeling π_{it} extremely difficult. The monotone missing data pattern (MMDP) is generally used to facilitate applications of MAR within the current context.

Under the MMDP assumption, a subject with missing y_{it} and $\mathbf{x}_{it}$ implies that all subsequent components, y_{is} and $\mathbf{x}_{is}$ $(t \leqslant s \leqslant T)$, are also missing. With the help of MMDP, we can express (9) as follows:

$$\pi_{it} = \Pr(r_{it} = 1 \mid \mathbf{y}_{it-}, \mathbf{x}_{it-}), \quad \Delta_{it} = \frac{r_{it}}{\pi_{it}}, \quad 1 \leqslant i \leqslant n, \quad 2 \leqslant t \leqslant T, \tag{10}$$

where $\mathbf{y}_{it-} = \left(y_{i1}, \ldots, y_{i(t-1)}\right)^\top$ and $\mathbf{x}_{it-} = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{i(t-1)}\right)^\top$, all observed. Thus, we can model and estimate π_{it} above (Appendix A). By substituting such estimated $\widehat{\boldsymbol{\Delta}}_i = \text{diag}_t\left(\frac{r_{it}}{\widehat{\pi}_{it}}\right)$ into (6), we obtain the WGEE for inference about $\boldsymbol{\beta}$ for general T. As in the case of GEE, both $\mathbf{w}_n(\widehat{\boldsymbol{\beta}})$ and the WGEE estimate $\widehat{\boldsymbol{\beta}}$ are again asymptotically normal. The respective asymptotic distributions follow the same general forms in (4), albeit with more complex Σ_w and $\boldsymbol{\Sigma}_\beta$ (Appendix B). Thus, by estimating Σ_w and $\boldsymbol{\Sigma}_\beta$, we can again use (4) for inference about $\boldsymbol{\beta}$. The asymptotic properties of these quantities provide the basis for the proposed goodness-of-fit tests we discuss next.

2.3. *Goodness-of-fit statistics*

Let $\widetilde{\mathbf{x}}_{it} = \left(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{ip}\right)^\top$ denote the vector of covariates $\mathbf{x}_{it}$ without the variable for the intercept term, and consider partitioning the covariates' space defined by $\widetilde{\mathbf{x}}_{it}$ into M distinct regions in the p-dimensional space. Let $\mathbf{I}_{it} = (I_{it1}, \ldots, I_{itM})^\top$ be a vector of indicators, with $I_{itm} = 1$ if the ith subject is in the mth region at time point t and 0 otherwise. So $\mathbf{I}_i = [\mathbf{I}_{i1}, \ldots, \mathbf{I}_{iT}]^\top$ is a $T \times M$ matrix. Define $\mathbf{Z}_T$ be a $T \times (T-1)$ design matrix with the first row consisting of zeros and the remaining $(T-1) \times (T-1)$ forming an identity matrix.

Consider an expanded model of (2) defined by the following:

$$\boldsymbol{\mu}_i = E(\mathbf{y}_i \mid \mathbf{x}_i) = h(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{Z}_T\boldsymbol{\tau} + \mathbf{I}_i\boldsymbol{\eta} + \mathbf{U}_i\boldsymbol{\varsigma}), \quad v(\boldsymbol{\mu}_i) = \text{Var}(\mathbf{y}_i \mid \mathbf{x}_i), \quad 1 \leqslant i \leqslant n, \tag{11}$$

where $\mathbf{U}_i = [\mathbf{0}, \text{diag}(\mathbf{I}_{i2}, \ldots, \mathbf{I}_{iT})]^\top$ is a $T \times (T-1)M$ matrix, $\mathbf{0}$ is a $(T-1) \times 1$ vector of zeros, $h^{-1}(\cdot)$ is a known link function, $\boldsymbol{\tau}$ is a $(T-1) \times 1$ vector of time effects, $\boldsymbol{\eta}$ is an $M \times 1$ vector of parameters associated with the partitioned regions of the covariates space, and $\boldsymbol{\varsigma}$ is a $(T-1)M \times 1$ vector of parameters for the time by region interaction terms.

Thus, (11) is the model in (1) augmented by $\mathbf{Z}_T$, $\mathbf{I}_i$, and $\mathbf{U}_i$. Let $\boldsymbol{\delta} = \left(\boldsymbol{\tau}^\top, \boldsymbol{\eta}^\top, \boldsymbol{\varsigma}^\top\right)^\top$ be a $J \times 1$ vector with $J = (T-1) + M + (T-1)M$, denoting the collection of parameters from these extra terms in (11). Let $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \boldsymbol{\delta}^\top\right)$ denote the vector of all parameters for the augmented model. If the posited model in (1) is correct, then $\boldsymbol{\delta} = 0$ and vice versa. Thus, the goodness-of-fit test concerns the null hypothesis $H_0 : \boldsymbol{\delta} = 0$. By identifying $\boldsymbol{\theta}$ as $\boldsymbol{\beta}$ and applying the WGEE to (11), we can use Wald statistics to test H_0 [4, 5]. However, Wald tests are typically anticonservative for small to moderate samples [13], and score statistics are often used as an alternative to reduce bias, especially for type I error rates, in such situations [8, 14]. We next describe a score test statistic, which is asymptotically equivalent to the Wald, to extend the test statistic in [8] to the WGEE setting.

Let $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top\right)^\top$, with $\boldsymbol{\theta}_{(1)} = \boldsymbol{\beta}$ and $\boldsymbol{\theta}_{(2)} = \boldsymbol{\delta}$. The null for the goodness-of-fit test is a special case of $H_0 : \boldsymbol{\theta}_{(2)} = \boldsymbol{\theta}_{(20)}$, with $\boldsymbol{\theta}_{(20)}$ denoting a vector of known constants, which, in the case of

goodness-of-fit test, is $\mathbf{0}$. The idea of the score test is to use the distribution of the score $\mathbf{w}_n\left(\widehat{\boldsymbol{\theta}}\right)$ rather than the estimate $\widehat{\boldsymbol{\theta}}$ to form a test statistic. Unfortunately, because only $\boldsymbol{\theta}_{(1)}$ is unknown, we need to estimate $\boldsymbol{\theta}_{(1)}$ before $\mathbf{w}_n(\boldsymbol{\theta})$ can be used as a statistic. In Appendix C, we discuss how to obtain a particular estimate $\widetilde{\boldsymbol{\theta}}_{(1)}$ of $\boldsymbol{\theta}_{(1)}$ so that the score statistic $\widetilde{\mathbf{w}}_{n(2)}$ defined by

$$\widetilde{\mathbf{w}}_{n(2)} = \mathbf{w}_{n(2)}\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) = \frac{1}{n}\sum_{i=1}^{n} D_{i(2)} V_i^{-1} \boldsymbol{\Delta}_i S_i, \quad D_{i(2)} = \frac{\partial \mathbf{h}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{(2)}}, \tag{12}$$

has an asymptotic normal distribution, $\sqrt{n}\,\widetilde{\mathbf{w}}_{n(2)} \to_d N\left(0, \Sigma_{(2)}\right)$, where V_i, $\boldsymbol{\Delta}_i$ and S_i are defined the same way as in WGEE, and $\Sigma_{(2)}$ is given in Appendix C.

Our score test statistic is define by $T_S\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right) = n\widetilde{\mathbf{w}}_{n(2)}^{\top}\widetilde{\Sigma}_{(2)}^{-1}\widetilde{\mathbf{w}}_{n(2)}^{\top}$, where $\widetilde{\Sigma}_{(2)}$ is an estimate of $\Sigma_{(2)}$ obtained by substituting $\widetilde{\boldsymbol{\theta}}_{(1)}$ in place of $\boldsymbol{\theta}_{(1)}$. Because $\widetilde{\mathbf{w}}_{n(2)}$ is asymptotically normal, $T_s\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right)$ has an asymptotic χ_J^2 distribution with J degrees of freedom χ_q^2. As in the case of the Wald test, this asymptotic distribution holds, if the right π_{it} in $\boldsymbol{\Delta}_i$ is used. For example, if π_{it} is modeled by (10), T_S is asymptotically χ_J^2, regardless of whether the missing data follow MCAR or MAR. However, if π_{it} is set equal to a constant 1, a special case considered by Barnhart and Williamson [8], T_S generally yields incorrect inference under MAR (Section 3.1).

Although asymptotically equivalent, this score test generally provides improved accuracy over its Wald counterpart. However, as the asymptotic distribution of $T_S\left(\widetilde{\boldsymbol{\theta}}_{(1)}, \boldsymbol{\theta}_{(20)}\right)$ completely ignores the variability in the estimated $\widetilde{\Sigma}_{(2)}$, we may further improve its performance by accounting for this extra variation. If $\widetilde{\Sigma}_{(2)}^{-1}$ is assumed to follow a Wishart, then the normalized Hotelling T^2-like statistic, $T_S^H = \frac{(n-1)-(J-1)}{J(n-1)}T_S$, has an $F_{J,n-J}$ distribution with numerator J and denominator $n-J$ degrees of freedom. To account for larger sampling variability for small samples, we also estimated Σ_w for the WGEE discussed in 2.2 differently (Appendix D). As shown in Section 3, this Hotelling T^2-like score statistic does seem to outperform its asymptotic counterpart T_S in small sample cases.

Note that the degree-of-freedom for the score statistic with an asymptotic chi-square distribution does not always equal the number of parameters in $\boldsymbol{\theta}_{(2)}$ under the null hypothesis, such as J within the current context, because of linear dependence between the covariates $\mathbf{x}_i$ in the posited model in (1) and the extra variables $(\mathbf{Z}_T, \mathbf{I}_i, \mathbf{U}_i)$ from the augmented model in (11). Because the two models in (1) and (11) involve the design matrices of the following form:

$$H_1 = \left[\mathbf{x}_1^{\top}, \ldots, \mathbf{x}_n^{\top}\right]^{\top}, \quad H_2 = [H_1, \mathbf{Z}, \mathbf{I}, \mathbf{U}],$$
$$\mathbf{I} = \left[\mathbf{I}_1^{\top}, \ldots, \mathbf{I}_n^{\top}\right]^{\top}, \quad \mathbf{U} = \left[\mathbf{U}_1^{\top}, \ldots, \mathbf{U}_n^{\top}\right]^{\top}, \quad \mathbf{Z} = \mathbf{Z}_T \otimes \mathbf{1}_n,$$

the degrees of freedom of the asymptotic χ^2 for the score statistic T_S is actually given by $J_c = \text{rank}(H_2) - \text{rank}(H_1)$, where $\mathbf{1}_n$ denotes an $n \times 1$ vector of 1's and $\text{rank}(A)$ the rank of matrix A. Similarly, the F distribution for the Hotelling T^2 version of the score $\widetilde{T}_S^H$ has J_c and $n - J_c$ as its numerator and denominator degrees of freedom, respectively.

3. Application

We demonstrate our considerations with both simulated and real data. We first investigate the performance of the proposed goodness-of-fit statistics with small ($n = 50$) to moderate ($n = 200$) to large ($n = 500$) sample sizes by simulation and then present an application to a real study on geriatric depression and associated medical comorbidities. In all the examples, we set the statistical significance at $\alpha = 0.05$. All analyses were carried out using codes developed by the authors for implementing the models considered using the R software platform.

3.1. Simulation studies

We carried out a series of simulation studies to assess the performance of the goodness-of-fit tests discussed in Section 2 for models with a continuous or binary longitudinal response under missing data following MAR. For brevity and without the loss of generality, we focused on a pre–post longitudinal design. In each case, data were generated from models containing both time-independent and time-varying covariates, and the proposed goodness-of-fit statistics were applied to assess model fit. The

performances of test statistics is assessed by their ability to confirm the adequacy of a model fit to the data generated by the same model (type I error) as well as to data generated under an alternative specification, involving either an interaction term or an extra covariate in the fitted model (type II error or power).

In the case of continuous response, the original model (under the null H_0), upon which model fit test statistics are based, and an expanded version containing extra terms (under the alternative H_a) are specified as follows for a longitudinal design with two assessments:

$$\begin{aligned}
\text{Model I} \quad &: \quad H_0 : y_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_i + \epsilon_{it}, \\
&: \quad H_a : y_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta^* x_{i1} x_{i2} + b_i + \epsilon_{it}, \\
\text{Model II} \quad &: \quad H_0 : y_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_i + \epsilon_{it}, \\
&: \quad H_a : y_{it} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta^* x_{i3t} + b_i + \epsilon_{it}, \\
\text{Model III} \quad &: \quad H_0 : y_{it} = \beta_0 + \beta_1 x_{i4} + b_i + \epsilon_{it}, \\
&: \quad H_a : y_{it} = \beta_0 + \beta_1 x_{i4} + \beta^* x_{i4}^2 + b_i + \epsilon_{it},
\end{aligned} \tag{13}$$

$$\epsilon_{it} \sim \left(\chi_1^2 - 1\right) \sqrt{\frac{\sigma_x^2}{2}}, \quad b_i \sim \left(\chi_1^2 - 1\right) \sqrt{\frac{\sigma_b^2}{2}}, \quad x_{i4} \sim \text{Uniform}(-1.5, 1.5);$$

$$x_{i1} \sim \text{Bernoulli}(0.5), \quad x_{i2} \sim \text{Uniform}(-1, 1), \quad x_{i3} \sim \text{Uniform}(0, 1), \quad x_{i3t} = t \cdot x_{i3}$$

where $t = 1, 2$ denotes the time point in this pre–post study design. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ under H_0 and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta^*)^\top$ under H_a. For our simulations, we fixed $\boldsymbol{\beta}$ so that $\boldsymbol{\beta}$ has the same value under H_0, but we varied β^* for the interaction or the extra term under H_a for all three models to facilitate evaluation of power:

$$\begin{aligned}
H_0 \quad &: \quad \beta_0 = \beta_1 = \beta_2 = 1, \\
H_a \quad &: \quad \beta_0 = \beta_1 = \beta_2 = 1; \\
&: \quad \beta^* = 0,\ 0.4,\ 0.8,\ 1.2,\ 1.6, \quad \text{for Models I and II;} \\
&: \quad \beta^* = 0,\ 0.25,\ 0.5,\ 0.75,\ 1, \quad \text{for Model III.}
\end{aligned}$$

Note that power for model III grew quite rapidly as β^* increased, and thus, a different of set of β^* was used to slow its growth.

The covariates, x_{i1} (categorical) and x_{i2} (continuous), were both time-invariant variables, while x_{i3t} (continuous) was a time-dependent variable, but following the same distribution as x_{i2}. In models I and II, β^* was associated with the interaction $x_{i1}x_{i2}$ (x_{i3t}) under H_a, while ϵ_{it} and b_i both followed a rescaled chi-square. We varied σ_b^2 and σ^2 to control for the within-subject correlation, $\rho = \sigma_b^2 / \left(\sigma_b^2 + \sigma^2\right)$, and for the simulation study, we set $\sigma_b^2 = 1$ and $\sigma^2 = 1$ so that $\rho = 0.5$. In model III, we also evaluated the performance of the proposed tests in highly correlated responses with $\rho = 0.8$ by setting $\sigma_b^2 = 4$ and $\sigma^2 = 1$.

To simulate missing data, we assumed no missingness at pre-treatment $t = 1$ and generated missing responses at post-treatment $t = 2$ under MAR according to the following logistic regression:

$$\text{logit}\left(\pi_{i2}\right) = \text{logit}\left(\Pr\left(r_{i2} = 1 \mid y_{i1}\right)\right) = \eta_0 + \eta_1 y_{i1}, \quad 1 \leqslant i \leqslant n. \tag{14}$$

We set $\eta_0 = -0.4$ for models I and II and $\eta_0 = 0.1$ for model III, and $\eta_1 = 1.0$ to create about 25% missing response rate at $t = 2$. A different η_0 was necessary to ensure 25% missing for model III because of different effects of $\eta_1 = 1.0$ on the missing data probability $\Pr\left(r_{i2} = 1 \mid y_{i1}\right)$ between model III and the other two models.

To assess the performance of the goodness-of-fit statistics proposed, the covariate's space was partitioned as follows for the three models:

$$\begin{aligned}
\text{Models I and II} \quad &: \quad \{x_1 = 0, x_2 \leqslant 0\} \cup \{x_1 = 0, x_2 > 0\} \cup \{x_1 = 1, x_2 \leqslant 0\} \cup \{x_1 = 1, x_2 > 0\}, \\
\text{Model III} \quad &: \quad \{-1.5 \leqslant x_4 \leqslant -0.5\} \cup \{-0.5 < x_4 < 0.5\} \cup \{0.5 \leqslant x_4 \leqslant 1.5\}.
\end{aligned}$$

Thus, for our simulation study, τ was a scalar, while $\boldsymbol{\eta}$ and $\boldsymbol{\rho}$ were both a 4×1 vector, giving rise a 9×1 vector $\boldsymbol{\delta}$ of parameters.

In the second case with the binary response, we considered a special type of mixed-effects logistic regression with a 'bridge' random effect to yield the same marginal models:

$$\begin{aligned}
\text{Model I} \quad &: \quad H_0 : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_i, \\
&: \quad H_a : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta^* x_{i1} x_{i2} + b_i, \\
\text{Model II} \quad &: \quad H_0 : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_2 + b_i, \\
&: \quad H_a : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_2 + \beta^* x_{i3t} + b_i, \\
\text{Model III} \quad &: \quad H_0 : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i4} + b_i, \\
&: \quad H_a : \text{logit}(E\,(y_{it} \mid \mathbf{x}_{it}, b_i)) = \beta_0 + \beta_1 x_{i4} + \beta^* x_{i4}^2 + b_i, \\
b_i \sim BR(x, \phi) &: \text{Density } f(x \mid \phi) = \frac{1}{2\pi} \frac{\sin(\phi\pi)}{\cosh(\phi x) + \cos(\phi\pi)}, \quad 0 < \phi < 1, -\infty < x < \infty,
\end{aligned} \tag{15}$$

where $\mu_{it} = E(y_{it} \mid x_{it})$ and b_i was a random intercept following a distribution $BR(x)$ defined by the density $f(x)$. Because of this particular random effect used, the marginal of each of the mixed-effect models in (15) retains the same form as the respective model in (15) without the random effect b_i, despite he overdispersed Bernoulli response [15]. For example, for model I under H_0, we have the following:

$$E\,(y_{it} \mid \mathbf{x}_{it}) = \mu_{it}, \quad \text{Var}\,(y_{it} \mid \mathbf{x}_{it}) > \mu_{it}\,(1 - \mu_{it})\,, \quad \text{logit}\,(\mu_{it}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}.$$

We set $\phi = 0.6$ and used the resulting overdispersed Bernoulli response to investigate the robustness of the proposed tests.

The simulation settings for the continuous response above were carried over to the logistic models, with the exceptions for the values of the parameters $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and ϕ:

$$\begin{aligned}
\eta_0 \quad &= \quad 0.6, \quad \eta_1 = 1.0, \quad \phi = 0.6, \\
H_0 \quad &: \quad \beta_0 = 0,\ \beta_1 = -0.5,\ \beta_2 = 0.2; \\
H_a \quad &: \quad \beta_0 = 0,\ \beta_1 = -0.5,\ \beta_2 = 0.2; \\
&: \quad \beta^* = 0,\ 0.5,\ 1,\ 2, \qquad \text{for Models I and II;} \\
&: \quad \beta^* = 0,\ 0.1,\ 0.3,\ 0.5, \quad \text{for Model III.}
\end{aligned}$$

These changes are necessary to ensure that there is about 25% missing responses y_{i2} at post-treatment, and μ_{it} has a reasonable range. Also, we changed the distribution of x_{i4} from the Uniform$(-1.5, 1.5)$ to Uniform$(-3, 3)$ to help compare our results with those from [8]. The latter led to a different partitioning of the covariate's space:

$$\{-3 \leqslant x_4 \leqslant -1\} \cup \{-1 < x_4 < 1\} \cup \{1 \leqslant x_4 \leqslant 3\}$$

For both continuous and binary responses, the proposed goodness-of-fit test statistics were calculated under the working independence model. The power of each test statistics (Score T_S and Hotelling T_S^H) was estimated by the frequency of ejecting the model under the null H_0, when fit to the data generated by the model under the alternative H_a by Monte Carlo (MC) simulation with the MC size set to 1000. For example, if $T_S^{(m)}$ denotes the score statistic at the mth MC replication, the type I error rate for testing the fit of model under H_0 is estimated by $\widehat{\alpha} = \frac{1}{1000} \sum_{m=1}^{1000} I_{\left\{T_S^{(m)} \geqslant q_{0.95}\right\}}$, where $q_{0.95}$ is the 95th percentile of the asymptotic χ^2 distribution of T_S. The same quantity was used for evaluating power for rejecting this model when fit to the data simulated from the model under H_a. Finally, by replacing $q_{0.95}$ with the 95th percentile of the F corresponding to the Hotelling T^2-like statistic, we obtained type I and power estimates to assess the performance of this 'small' sample statistic.

Shown in Table I are the estimates of type I error (power) under the null $H_0 : \beta^* = 0$ (alternative $H_a : \beta^* > 0$) by the proposed score and Hotelling T^2-like alternative T_S^H for both models I and II in (13), with π_{i2} modeled by (10) (labeled WGEE) and $\pi_{i2} = 1$ (labeled GEE). For comparison purposes, the table also shows such estimates obtained from the two statistics when applied to the complete data (labeled Complete data). Because of the fast convergence of the score statistics to asymptotic normality in the linear model setting, $n = 200$ suffices to characterize the asymptotic behavior of the tests.

The results show that there was upward bias in the estimated type I error for both models across all three cases (WGEE, complete data, and GEE) under the smaller sample size $n = 50$. However, the proposed WGEE-based goodness-of-fit statistics yielded type I error estimates closer to the nominal value

Table I. Type I error ($\beta^* = 0$) and power ($\beta^* > 0$) estimates for testing the null H_0 versus the alternative model H_a for two linear regression models, models I and II, by the proposed Score (T_S) and Hotelling (T_S^H) score statistics for both the missing data (WGEE and GEE) and complete data (Complete data); model I involves a missing interaction, while model II entails a missing time-varying covariate, both defined in (13).

		$n = 50$ / $n = 200$				
Sample size		$\beta^* = 0.0$	$\beta^* = 0.4$	$\beta^* = 0.8$	$\beta^* = 1.2$	$\beta^* = 1.6$
Model I						
T_S	WGEE	0.10/0.05	0.16/0.14	0.27/0.47	0.40/0.80	0.57/0.96
	Complete data	0.09/0.05	0.14/0.16	0.29/0.56	0.47/0.88	0.65/0.99
	GEE	0.16/0.18	0.18/0.27	0.29/0.55	0.41/0.86	0.58/0.97
T_S^H	WGEE	0.04/0.05	0.07/0.14	0.16/0.47	0.27/0.81	0.42/0.97
	Complete data	0.03/0.05	0.06/0.16	0.18/0.56	0.33/0.89	0.49/0.99
	GEE	0.07/0.19	0.09/0.27	0.18/0.56	0.28/0.86	0.43/0.98
Model II						
T_S	WGEE	0.11/0.04	0.12/0.17	0.31/0.71	0.59/0.98	0.79/1.00
	Complete data	0.10/0.04	0.16/0.27	0.42/0.86	0.71/0.99	0.90/1.00
	GEE	0.15/0.18	0.30/0.65	0.55/0.95	0.79/1.00	0.91/1.00
T_S^H	WGEE	0.05/0.04	0.06/0.17	0.18/0.72	0.42/0.98	0.66/1.00
	Complete data	0.05/0.05	0.08/0.28	0.27/0.86	0.58/0.99	0.82/1.00
	GEE	0.07/0.19	0.19/0.65	0.41/0.95	0.67/1.00	0.83/1.00

WGEE, weighted generalized estimating equations; GEE, generalized estimating equations.

$\alpha = 0.05$ than their GEE-based counterparts, regardless of the sample sizes. Although the two had similar p-values in some cases for $n = 50$, the differences widened when n increased to $n = 200$. The larger differences for the latter sample indicate the GEE-based test statistics are biased under MAR.

The WGEE-based T_S did show some upward bias for the smaller sample $n = 50$. However, at $n = 50$, type I error estimates were biased even under complete data. In comparison, the Hotelling T^2-like T_S^H performed much better, substantially reducing the amount of bias across the board and yielding type I error estimates even closer to the nominal value than those from the standard score test T_S when applied to the complete data. Thus, the WGEE-based tests performed remarkably well, which is specially true for the Hotelling T^2-like test T_S^H.

Because WGEE/GEE cannot have larger power than the complete-data case, the results presented indicate a small amount of upward bias in power estimates by both test statistics for model I with $n = 50$ when $\beta^* = 0.4$. However, the bias seemed to have diminished quickly as β^* increased.

Type I error rates and power estimates are shown in Table II for model III in (13) for detecting the missing quadratic term in the linear regression models. Compared with the simulation results in Table I for detecting the missing linear interaction for models I and II in (13), we also obtained quite accurate type I error rates $\alpha = 0.05$, especially under the moderate within-subject correlation $\rho = 0.5$. When the correlation increased to $\rho = 0.8$, the type I error rate became a bit conservative. When the sample size was small $n = 50$, the Hotelling T^2 -like statistic T_S^H not only beat the regular score statistic T_S but even showed better type I error rates than the complete data case. The power estimates showed more power for detecting the missing quadratic term than for detecting the missing interaction, consistent with the results observed in the goodness-of-fit tests for longitudinal GEE model with binary response in [8]. As expected, power reduced as the correlation ρ increased from 0.5 to 0.8.

Shown in Table III are the same type of estimates as in Table I but obtained from the models for the overdispersed binary response defined in (15). Because of the slower convergence for binary responses, we have included results from a large sample size $n = 500$ to help evaluate bias associated with GEE. The results show quite similar patterns as observed for the models with the continuous response. For example, bias in type I error estimates (under $\beta^* = 0$) diminished for WGEE as well as complete data as n increased. Also, WGEE seemed to be sensitive to model misspecification (under $\beta^* > 0$), albeit with some upward bias for the small ($n = 50$) and moderate ($n = 200$) sizes.

Type I error estimates from the large sample size $n = 500$ confirm that WGEE is unbiased. This proposed approach also yielded quite accurate type I error estimates for the moderate sample size $n = 200$. For $n = 50$, the score-based WGEE did not perform well, but neither did the same test based on the

Table II. Type I error ($\beta^* = 0$) and power ($\beta^* > 0$) estimates for testing the null H_0 versus the alternative model H_a for model III in linear regression by the proposed Score (T_S) and Hotelling (T_S^H) score statistics for both the missing data (WGEE and GEE) and complete data (Complete data); model III involves a missing quadratic term defined in (13).

		$n = 50$ / $n = 200$				
Sample size		$\beta^* = 0.0$	$\beta^* = 0.25$	$\beta^* = 0.5$	$\beta^* = 0.75$	$\beta^* = 1.0$
$\rho = 0.5$						
T_S	WGEE	0.09/0.05	0.15/0.21	0.30/0.66	0.52/0.96	0.72/0.99
	Complete data	0.09/0.05	0.15/0.24	0.32/0.71	0.57/0.97	0.79/1.00
	GEE	0.13/0.14	0.17/0.31	0.31/0.69	0.51/0.94	0.72/0.99
T_S^H	WGEE	0.05/0.05	0.08/0.22	0.20/0.66	0.41/0.96	0.62/0.99
	Complete data	0.04/0.05	0.08/0.25	0.21/0.72	0.47/0.97	0.70/1.00
	GEE	0.08/0.14	0.10/0.32	0.20/0.70	0.41/0.95	0.62/0.99
$\rho = 0.8$						
T_S	WGEE	0.09/0.04	0.11/0.09	0.18/0.28	0.27/0.60	0.43/0.84
	Complete data	0.07/0.04	0.13/0.12	0.21/0.34	0.34/0.68	0.49/0.90
	GEE	0.36/0.80	0.35/0.78	0.37/0.80	0.42/0.88	0.51/0.95
T_S^H	WGEE	0.05/0.04	0.07/0.10	0.12/0.29	0.18/0.60	0.30/0.85
	Complete data	0.04/0.04	0.07/0.12	0.12/0.34	0.23/0.69	0.37/0.90
	GEE	0.28/0.81	0.26/0.79	0.25/0.80	0.31/0.88	0.40/0.95

WGEE, weighted generalized estimating equations; GEE, generalized estimating equations.

Table III. Type I error ($\beta^* = 0$) and power ($\beta^* > 0$) estimates for testing the null H_0 versus the alternative model H_a for two logistic regression models, models I and II, by the proposed Score (T_S) and Hotelling (T_S^H) score statistics for both the missing data (WGEE and GEE) and complete data (Complete data); model I involves a missing interaction, while model II entails a missing time-varying covariate, both defined in (15).

		$n = 50$ / $n = 200$ / $n = 500$			
Sample size		$\beta^* = 0.0$	$\beta^* = 0.5$	$\beta^* = 1.0$	$\beta^* = 2.0$
Model I					
T_S	WGEE	0.16 /0.06 /0.05	0.17 /0.07 /0.10	0.20 /0.12 /0.23	0.26 /0.35 /0.76
	Complete data	0.12 /0.05 /0.05	0.12 /0.07 /0.09	0.17 /0.11 /0.27	0.20 /0.39 /0.83
	GEE	0.13 /0.07 /0.10	0.12 /0.09 /0.14	0.16 /0.13 /0.30	0.21 /0.36 /0.78
T_S^H	WGEE	0.09 /0.06 /0.05	0.11 /0.08 /0.10	0.13 /0.12 /0.24	0.16 /0.36 /0.76
	Complete data	0.05 /0.05 /0.05	0.05 /0.07 /0.09	0.09 /0.11 /0.27	0.11 /0.40 /0.83
	GEE	0.06 /0.07 /0.10	0.05 /0.09 /0.15	0.09 /0.13 /0.30	0.11 /0.37 /0.79
Model II					
T_S	WGEE	0.18 /0.06 /0.05	0.17 /0.07 /0.12	0.21 /0.16 /0.43	0.30 /0.57 /0.98
	Complete data	0.11 /0.05 /0.05	0.13 /0.09 /0.14	0.16 /0.21 /0.53	0.29 /0.69 /0.99
	GEE	0.16 /0.07 /0.11	0.14 /0.14 /0.36	0.19 /0.30 /0.77	0.33 /0.76 /1.00
T_S^H	WGEE	0.11 /0.06 /0.05	0.10 /0.08 /0.13	0.12 /0.17 /0.44	0.17 /0.58 /0.98
	Complete data	0.05 /0.05 /0.05	0.06 /0.09 /0.15	0.09 /0.21 /0.53	0.17 /0.69 /0.99
	GEE	0.06 /0.07 /0.11	0.08 /0.15 /0.37	0.11 /0.31 /0.77	0.19 /0.76 /1.00

WGEE, weighted generalized estimating equations; GEE, generalized estimating equations.

complete data, indicating inadequacy in approximating the sampling distribution by the asymptotic normal model. In all cases, the Hotelling version of the score test T_S^H seemed working really well in reducing such small sample bias in its unadjusted counterpart T_S.

Shown in Table IV are the type I error rates and power estimates for detecting the missing quadratic term for model III in (15) for binary responses. All estimates behaved in the same manner as we expected from the logistic regression model and simulation results for the linear regression models for continuous responses discussed earlier. For example, the Hotelling T^2 -like statistic T_S^H again outperformed the regular score statistic T_S, and there was more power for detecting departure from a missing quadratic than an interaction term.

Table IV. Type I error ($\beta^* = 0$) and power ($\beta^* > 0$) estimates for testing the null H_0 versus the alternative model H_a for model III in logistic regression by the proposed Score (T_S) and Hotelling (T_S^H) score statistics for both the missing data (WGEE and GEE) and complete data (Complete data); model III involves a missing quadratic term defined in (15).

		$n = 50$ / $n = 200$ / $n = 500$			
Sample size		$\beta^* = 0.0$	$\beta^* = 0.1$	$\beta^* = 0.3$	$\beta^* = 0.5$
T_S	WGEE	0.14/0.04/0.05	0.16/0.08/0.13	0.25/0.34/0.77	0.35/0.64/0.98
	Complete data	0.12/0.05/0.05	0.14/0.08/0.15	0.20/0.39/0.81	0.32/0.67/0.99
	GEE	0.15/0.10/0.14	0.18/0.16/0.30	0.27/0.45/0.87	0.37/0.70/0.99
T_S^H	WGEE	0.06/0.05/0.05	0.07/0.08/0.13	0.11/0.35/0.77	0.17/0.64/0.98
	Complete data	0.05/0.06/0.05	0.05/0.08/0.15	0.08/0.39/0.81	0.13/0.68/0.99
	GEE	0.07/0.10/0.15	0.08/0.16/0.30	0.12/0.45/0.87	0.19/0.71/0.99

WGEE, weighted generalized estimating equations; GEE, generalized estimating equations.

Table V. Estimates (Estimate) of parameters of logistic regression, along with standard errors (SE) and p-values (p-value) for modeling missingness at 1-year follow-up for the real study on geriatric research.

Predictors	Estimate	SE	p-value
Intercept	−0.645	0.086	<0.001
Baseline IADL	0.049	0.018	0.007

IADL, instrumental activities of daily living.

Note that we only compared type I errors and power estimates in the simulation study. GEE estimates of $\boldsymbol{\beta}$ in general are also biased when missing data follows MAR. We did not discuss such bias within the current context because of space consideration and published studies on this topic (e.g., [2]).

3.2. A case study of geriatric depression

In geriatric research, overall functional disability is of particular importance, as it reflects both the mental and physical health conditions of an individual. A primary measure of overall functional status is the instrumental activities of daily living (IADL, range 0–24), with higher scores representing greater functional disability [16]. In this real data example, we use data from a study on geriatric depression and associated medical comorbidities to model the change of IADL over time [17].

In this study, 744 primary care patients 65 years and older who presented for care and were capable of giving informed consent were enrolled from private practices and University-affiliated clinics in general internal medicine, geriatrics, and family medicine in Monroe County, New York [17]. All subjects underwent semi-structured interviews for functional status and medical comorbidities and Structured Clinical Interview for DSM-IV for major depression [18]. Assessments took place at study intake and again every year for 5 years. For illustration purposes, we analyzed the IADL y_{it} from baseline ($t = 1$) to the first year follow-up ($t = 2$).

Of the 744 enrolled, 468 completed the IADL at the 1-year follow-up. We modeled the missingness at the follow-up using a logistic model of the form (14), with y_{i1} as the predictor controlling for some demographic covariates available from the study. Because only y_{i1} was significant, the reduced model with only this variable was used for the WGEE analysis. Shown in Table V are the estimates of the intercept and coefficient of this baseline IADL from the fitted logistic regression. The significance of IADL at baseline indicates a MAR mechanism for the missing data at the 1-year follow-up, with the positive sign of the estimate implicating that subjects with higher baseline IADL were more likely to come for assessment at the follow-up. As higher IADL scores are associated with poorer functioning status, the observed sample at the follow-up visit seems biased towards those with more severe overall functional disability at baseline.

Because IADL is a function of mental and physical health, we started a model with $\mathbf{x}_i$ consisting of depression diagnosis, a binary indicator of major depression diagnosis based on Structured Clinical

Interview for DSM-IV, and medical burdens assessed using the cumulative illness rating scale, which basically counts the total number of major medical problems [19], in addition to an intercept term:

$$E(\mathbf{y}_i) = \mathbf{x}_i^\top \boldsymbol{\beta} + W_i \boldsymbol{\delta}, \quad \mathbf{y}_i = (y_{i1}, y_{i2})^\top, \tag{16}$$

where $\mathbf{W}_i$ is the design matrix based on some partition of the space formed by the covariates $\mathbf{x}_i$, and $\boldsymbol{\rho}$ is the corresponding vector of parameters. We assessed the adequacy of the model by testing the null hypothesis, $H_0 : \boldsymbol{\delta} = 0$, as discussed in Section 2.3.

Shown in Table VI are the estimates of $\boldsymbol{\beta}$ and standard errors, along with the goodness-of-fit statistics under GEE and WGEE. As seen, GEE and WGEE estimates were generally different, although they were quite close and were all significant. The high levels of significance for both Score and Hotelling T^2-like tests under WGEE indicate severe lack of fit. To improve model fit, we added additional predictors, one at a time, and repeated the aforementioned steps of assessment until the two model fit test statistics became insignificant.

Shown in Table VII are the estimates of $\boldsymbol{\beta}$ from the final model resulting from this model building process. Because both Score and Hotelling T^2-like statistics under WGEE were all larger than 0.2, this

Table VI. Estimates (Estimate) of parameters, standard errors (SE), and p-values (p-value) for a model with depression and medical burdens as predictors for the IADL response in the real study on geriatric research, along with goodness-of-fit test statistics (Statistic) from Score (T_S) and Hotelling $\left(T_S^H\right)$ score statistics and associated degree of freedom (DF) and p-values (p-value) for the adequacy of the initial model under the proposed (WGEE) and standard (GEE) methods.

	WGEE			GEE		
Predictors	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	−2.156	0.349	<0.001	−2.144	0.371	<0.001
Major depression (1 = yes, 0 = no)	1.557	0.447	<0.001	1.434	0.440	0.0011
Medical burdens	0.478	0.049	<0.001	0.503	0.053	<0.001
Goodness-of-fit test						
	Statistic	DF	p-value	Statistic	DF	p-value
T_S (χ^2_J test)	31.9	7	<0.001	32.1	7	<0.001
T_S^H ($F_{J,(n-1)-(J-1)}$ test)	4.73	7, 737	<0.001	4.75	7, 737	<0.001

IADL, instrumental activities of daily living; WGEE, weighted generalized estimating equations; GEE, generalized estimating equations.

Table VII. Estimates (Estimate) of parameters, standard errors (SE), and p-values (p-value) for a model with depression and medical burdens as predictors for the IADL response in the real study on geriatric research, along with goodness-of-fit test statistics (Statistic) from Score (T_S) and Hotelling $\left(T_S^H\right)$ score statistics and associated degree of freedom (DF) and p-values (p-value) for the adequacy of the final model under the proposed (WGEE) and standard (GEE) methods.

	WGEE			GEE		
Predictors	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	−6.165	2.018	0.002	−5.804	2.186	0.008
Major depression (1 = yes, 0 = no)	1.083	0.474	0.022	0.931	0.485	0.055
Medical burdens	0.342	0.050	<0.001	0.360	0.054	<0.001
Age (years)	0.095	0.023	<0.001	0.097	0.025	<0.001
Education (years)	−0.203	0.063	0.001	−0.241	0.067	<0.001
Ham-D	0.109	0.027	<0.001	0.109	0.027	<0.001
Goodness-of-fit test						
	Statistic	DF	p-value	Statistic	DF	p-value
T_S (χ^2_J test)	63.1	57	0.270	93.9	57	0.002
T_S^H ($F_{J,(n-1)-(J-1)}$ test)	1.12	57, 687	0.265	1.75	57, 687	<0.001

IADL, instrumental activities of daily living; WGEE, weighted generalized estimating equations; GEE, generalized estimating equations; Ham-D, Hamilton Rating Scale for Depression.

model seemed to fit the data quite well. Further, all the predictors included in the model were highly significant, suggesting that all had a significant impact on the subject's functional status in this particular population of old primary care patients. The model indicates that depression severity, as measured by the Hamilton Rating Scale for Depression [20], age, and education contributed to greater functional disability above and beyond major depression diagnosis and medical burdens. Patients who were older and less educated had poorer functioning status. Education was assessed using an ordinal scale from 0 to 17 based on years of education from primary school to college (0–17) to any postgraduate work (17) but treated as a continuous outcome in the model.

The Score T_S and Hotelling T^2-like T_S^H test from the final model were almost identical under WGEE, confirming the asymptotic equivalence between the two. However, when applied under the wrong assumption, the GEE versions of the tests were highly biased, continuing to indicate severe lack of fit for the final model. The difference between WGEE and GEE also showed up in the estimates of $\boldsymbol{\beta}$, associated standard errors, and p-values. As noted earlier, GEE estimates of $\boldsymbol{\beta}$ are biased when missing values follow MAR, and as such, the difference in the estimates of $\boldsymbol{\beta}$ between WGEE and GEE is expected.

4. Discussion

We extended a GEE-based goodness-of-fit statistic to address its limitations when used for longitudinal data with missing values following MAR, a popular mechanism arising from most real research and clinical studies. In particular, we proposed a WGEE-based score test and a revised version to provide more accurate inference for small samples based on the construction of Hotelling's T^2 statistic in classic multivariate models. We examined the performance of the proposed tests through simulated data.

For linear models, both tests performed well; although small bias was present in the score test for the small sample size $n = 50$, the small-sample-adjusted Hotelling T^2-like test eliminated such bias. There also does not seem to be much power loss in these tests in rejecting a poorly fit model. For logistic regression, both tests exhibit some bias for the small sample size $n = 50$, which may be in part due to unstable estimates for binary data, as evidenced by the bias in the tests even under complete data. The Hotelling T^2-like score test worked extremely well in reducing the bias in the standard score test under both WGEE and complete data. The inflated type I error estimates from the results of the simulation study show that the existing GEE-based goodness-of-fit statistic is biased under MAR.

Like the GEE-based score statistic, the proposed tests require that the partitioned regions of the covariates' space contain a sufficient number of observations, such as 10 or more observations in each partitioned region [8]. With a large number of covariates, we may first fit a series of simple models, each with a couple of covariates, and then build more complex models by selecting significant variables from the initial simpler models.

A main drawback of approaches based on partitioning of covariates' space such as the proposed tests is the dependence on the partition. For categorical covariates, the distinct values of the covariate serves as the natural boundaries of partitioned regions. For continuous covariates, one may decide a priori a set of cut-points before looking at the data to avoid data dredging.

Appendix A. Estimates of π_{it}

To estimate the π_{it}'s in (10), note first that

$$\pi_{it} = \prod_{s=2}^{t} \Pr(r_{is} = 1 \mid \mathbf{y}_{is-}, \mathbf{x}_{is-}, r_{i(s-1)} = 1). \tag{A.1}$$

By modeling each transition probability, $p_{it} = \Pr(r_{it} = 1 \mid \mathbf{y}_{it-}, \mathbf{x}_{it-}, r_{i(t-1)} = 1)$, using a logistic regression model [4], we can use estimates of p_{it} to estimate the π_{it}'s in (A.1). The transition probability p_{it} must be specified correctly to ensure consistent estimates of β, unless some additional models are considered for modeling the missing response, for example, in the augmented WGEE [2, 21].

Let η_t denote the vector of parameters for modeling p_{it} and $\eta = \left(\eta_2^\top, \ldots, \eta_T^\top\right)^\top$. Then, $p_{it}(\eta)$, $\pi_{it}(\eta)$ and $\Delta_{it}(\eta)$ are all determined by η. We can readily estimate the parameter of each η_t using either maximum likelihood or estimating equations [4]. For instance, under maximum likelihood, we can combine

the score equations across all η_t's to obtain a set of estimating equations for joint η:

$$\begin{aligned}\mathbf{Q}_{ni} &= (\mathbf{Q}_{i2}, \ldots, \mathbf{Q}_{iT}), \\ \mathbf{Q}_{it} &= \frac{\partial}{\partial \eta_t}\left\{r_{i(t-1)}\left[r_{it}\log(p_{it}) + (1-r_{it})\log(1-p_{it})\right]\right\},\end{aligned} \tag{A.2}$$

where $\mathbf{Q}_{it}$ is the score vector for η_t from each ith subject ($i = 1, 2, \ldots, n$, $2 \leqslant t \leqslant T$).

Appendix B. Asymptotic variances Σ_w and Σ_β for WGEE

Under suitable regularity conditions, both $\mathbf{w}_n(\beta)$ and the WGEE estimate $\widehat{\beta}$ are asymptotically normal, with the asymptotic variances Σ_w and Σ_β given by

$$\begin{aligned}&\Sigma_w = \Sigma_U + \Phi, \quad \Sigma_\beta = B^{-1}\Sigma_w B^{-1}, \\ &\Sigma_U = E\left(D_i V_i^{-1}\Delta_i S_i S_i^\top \Delta_i V_i^{-1} D_i^\top\right), \quad B = E\left(D_i V_i^{-1}\Delta_i D_i^\top\right), \\ &\Phi = -CH^{-T}C^\top - F - F^\top, \quad C = E\left[\frac{\partial}{\partial \eta}\left(D_i V_i^{-1}\Delta_i S_i\right)\right]^\top, \\ &H = E\left(\frac{\partial}{\partial \eta}\mathbf{Q}_{ni}\right)^\top, \quad F = E\left(D_i V_i^{-1}\Delta_i S_i \mathbf{Q}_{ni}^\top H^{-\top} C^\top\right),\end{aligned} \tag{B.1}$$

where $A^{-\top}$ denotes the inverse of the transposed matrix A. The first term of Σ_w (Σ_β) is identical to its GEE counterpart in (4), while the second term accounts for the additional variability in an estimated $\widehat{\eta}$. A consistent estimate of Σ_w (Σ_β) is given by

$$\begin{aligned}&\widehat{\Sigma}_\beta = \widehat{B}^{-1}\widehat{\Sigma}_U\widehat{B}^{-\top}, \quad \widehat{B} = \frac{1}{n}\sum_{i=1}^n \widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{D}_i^\top, \quad \widehat{C} = \frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial \eta}\left(\widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{S}_i\right), \\ &\widehat{\Sigma}_U = \frac{1}{n}\sum_{i=1}^n \widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{S}_i\widehat{S}_i^\top\widehat{\Delta}_i\widehat{V}_i^{-1}\widehat{D}_i^\top, \quad \widehat{F} = \left(\frac{1}{n}\sum_{i=1}^n \widehat{D}_i\widehat{V}_i^{-1}\widehat{\Delta}_i\widehat{S}_i\widehat{\mathbf{Q}}_{ni}^\top\right)\widehat{H}^{-\top}\widehat{C}^\top,\end{aligned} \tag{B.2}$$

where $\widehat{A}$ denotes a quantity A in (B.1) with β and η replaced by their respective estimates $\widehat{\beta}$ and $\widehat{\eta}$.

Appendix C. Asymptotic distribution of the score statistic $\tilde{\mathbf{w}}_{n(2)}$

We develop the asymptotic normal distribution of $\widetilde{\mathbf{w}}_{n(2)}$ through (a) finding an appropriate estimate $\widetilde{\theta}_{(1)}$ of $\theta_{(1)}$ and (b) proving that the resulting $\widetilde{\mathbf{w}}_{n(2)}$ has an asymptotic normal distribution.

To achieve (a), consider the quantities

$$\mathbf{w}_{n(1)}(\theta) = \frac{1}{n}\sum_{i=1}^n D_{i(1)}V_i^{-1}\Delta_i S_i, \quad \mathbf{w}_{n(2)}(\theta) = \frac{1}{n}\sum_{i=1}^n D_{i(2)}V_i^{-1}\Delta_i S_i, \tag{C.1}$$

where

$$D_i = \begin{pmatrix} \frac{\partial \mathbf{h}(\theta)}{\partial \theta_{(1)}} \\ \frac{\partial \mathbf{h}(\theta)}{\partial \theta_{(2)}} \end{pmatrix} = \begin{pmatrix} D_{i(1)} \\ D_{i(2)} \end{pmatrix}, \quad D_i V_i^{-1}\Delta_i S_i = \begin{pmatrix} D_{i(1)}V_i^{-1}\Delta_i S_i \\ D_{i(2)}V_i^{-1}\Delta_i S_i \end{pmatrix}.$$

Denote by $\widetilde{\theta}_{(1)}$ an estimate of $\theta_{(1)}$ from solving the WGEE defined by the first quantity in (C.1):

$$\mathbf{w}_{n(1)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) = \frac{1}{n}\sum_{i=1}^n D_{i(1)}V_i^{-1}\Delta_i S_i = \mathbf{0}. \tag{C.2}$$

To prove (b), let

$$\begin{aligned}&\widetilde{\theta} = \begin{pmatrix} \widetilde{\theta}_{(1)} \\ \theta_{(20)} \end{pmatrix}, \quad B = E\left(D_i V_i^{-1}\Delta_i D_i\right) = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}, \\ &G = \begin{pmatrix} -B_{21}B_{11}^{-1} & \mathbf{I}_J \end{pmatrix}, \quad \Sigma_{(2)} = G\Sigma_w G^\top, \quad \widetilde{\Sigma}_{(2)} = \Sigma_{(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right),\end{aligned} \tag{C.3}$$

where $\mathbf{I}_J$ designates the $J \times J$ identity matrix, B_{11} denotes the $P \times P$ submatrix, B_{12} the $P \times J$ submatrix, and B_{22} the $J \times J$ submatrix from the partitioned $(P + J) \times (P + J)$ matrix B defined in (C.3), and Σ_w is defined in (B.1). First, assume no missing data. Then, $B = E\left(D_i V_i^{-1} D_i\right)$. By applying the law of large numbers,

$$\frac{\partial}{\partial\theta}\mathbf{w}_n(\theta) = \begin{pmatrix} \frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(1)}(\theta) & \frac{\partial}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}(\theta) \\ \frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}(\theta) & \frac{\partial}{\partial\theta_{(2)}}\mathbf{w}_{n(2)}(\theta) \end{pmatrix} \to_p B = E\left(D_i V_i^{-1} D_i\right) = \begin{pmatrix} B_{11} & B_{12} \\ B_{12}^\top & B_{22} \end{pmatrix}, \quad \text{(C.4)}$$

where $\to_p$ denotes convergence in probability. It follows from a Taylor's series expansion and (C.4) that

$$\mathbf{0} = \mathbf{w}_{n(1)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) = \mathbf{w}_{n(1)}(\theta) - B_{11}^{-\top}\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right),$$

where $\mathbf{o}_p(1)$ denotes the stochastic $\mathbf{o}(1)$ [4]. Thus,

$$\widetilde{\theta}_{(1)} - \theta_{(1)} = B_{11}^{-1}\mathbf{w}_{n(1)}(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \quad \text{(C.5)}$$

Similarly, because $B_{12}^\top = B_{21}$, we have

$$\begin{aligned} \mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) &= \mathbf{w}_{n(2)}(\theta) - \left(\frac{\partial^\top}{\partial\theta_{(1)}}\mathbf{w}_{n(2)}\right)\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \quad \text{(C.6)} \\ &= \mathbf{w}_{n(2)}(\theta) - B_{21}\left(\widetilde{\theta}_{(1)} - \theta_{(1)}\right) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \end{aligned}$$

It follows from (C.5) and (C.6) that

$$\begin{aligned} \mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) &= \mathbf{w}_{n(2)}(\theta) - B_{21}\left[B_{11}^{-1}\mathbf{w}_{n(1)}(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right)\right] + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right) \\ &= G\mathbf{w}_n(\theta) + \mathbf{o}_p\left(n^{-\frac{1}{2}}\right). \end{aligned}$$

By the central limit theorem,

$$\sqrt{n}\mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right) = \sqrt{n}G\mathbf{w}_n(\theta) + \mathbf{o}_p(1) \to_d N\left(\mathbf{0}, \Sigma_{(2)} = G\Sigma_\theta G^\top\right), \quad \text{(C.7)}$$

where G is defined in (C.3) and Σ_θ in (B.1).

In the presence of missing data, $B = E\left(D_i V_i^{-1} \Delta_i D_i\right)$ as defined in (B.1). By a similar argument, $\mathbf{w}_{n(2)}\left(\widetilde{\theta}_{(1)}, \theta_{(20)}\right)$ has an asymptotic normal distribution.

Appendix D. An estimate of Σ_w for the Hotelling T^2-like statistic T_S^H

For the Hotelling T^2-like statistic T_S^H discussed in Section 2.3, we propose to use a different estimate of $\Sigma_w = \Sigma_U + \Phi$ than the one in (B.1) to further improve the performance of the score test statistic for small samples. This new estimate of Σ_w is again obtained by substituting estimates of Σ_U and Φ. For Φ, the same estimate in (B.2) is used, but for Σ_U, an different estimate is constructed, $\widehat{\Sigma}_U = \frac{1}{n-1}\sum_{i=1}^n (\mathbf{w}_{ni} - \mathbf{w}_n)(\mathbf{w}_{ni} - \mathbf{w}_n)^\top$.

Acknowledgements

We like to thank Dr. Tang of the Department of Biostatistics and Computational Biology at the University of Rochester for the fruitful discussions about goodness-of-fit statistics for semi-parametric regression models.

References

1. Demidenko E. *Mixed Models: Theory and Applications*. Wiley: New York, 2004.
2. Lu N, Tang W, He H, Yu Q, Crits-Christoph P, Zhang H, Tu XM. On the impact of parametric assumptions and robust alternatives for longitudinal data analysis. *Biometrical Journal* 2009; **51**:627–643.
3. Pepe MS, Anderson GL. A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communication in Statistics-Simulation* 1994; **23**:939–951.

4. Kowalski J, Tu XM. *Modern Applied U Statistics*. Wiley: New York, 2007.
5. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**:13–22.
6. Evans S, Li L. A comparison of goodness of fit tests for the logistic GEE model. *Statistics in Medicine* 2005; **24**:1245–1261.
7. Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* 1980; **67**:250–251.
8. Barnhart HX, Williamson JM. Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics* 1998; **54**:720–729.
9. Rubin DB. Inference and missing data (with discussion). *Biometrika* 1976; **63**:581–592.
10. Fitzmaurice GM. A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics* 1995; **51**:309–317.
11. Kenward MG, Molenberghs G. Likelihood based frequentist inference when data are missing at random. *Statistical Science* 1998; **13**:236–247.
12. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**:106–121.
13. Pan W. On the robust variance estimator in generalized estimating equations. *Biometrika* 2001; **88**:901–906.
14. Zhang H, Xia Y, Chen R, Lu N, Tang W, Tu XM. On modeling longitudinal binomial responses — implications from two dueling paradigms. *Applied Statistics* 2011; **38**:2373–2390.
15. Wang ZR, Louis TA. Mathing conditional and marginal shapes in binary random intercept models using a bridge distribution function. *Biometrika* 2003; **90**(4):765–775.
16. Lawton MP, Brody EM. Assessment of older people; self-maintaining and instrumental activities of daily living. *Gerontologist* 1969; **9**:179–186.
17. Lyness JM, Niculescu A, Tu XM, Reynolds CF, Caine ED. The relationship of medical comorbidity to depression in older primary care patients. *Psychosomatics* 2007; **47**:435–439.
18. Spitzer RL, Williams JBW. *Structural Clinical Interview for DSM-III*. State Psychiatric Institute, Biometrics Research Department: New York, 1986.
19. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. *Journal of the American Geriatrics Society* 1968; **16**:622–626.
20. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 1988; **45**:742–747.
21. Rotnitzky A, Robins JM, Scharfsteinc DO. Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association* 1998; **93**:1321–1339.