

Modern methods for longitudinal data analysis, capabilities, caveats and cautions

Lin GE¹, Justin X. TU², Hui ZHANG³, Hongyue WANG¹, Hua HE⁴, Douglas GUNZLER⁵

Summary: Longitudinal studies are used in mental health research and services studies. The dominant approaches for longitudinal data analysis are the generalized linear mixed-effects models (GLMM) and the weighted generalized estimating equations (WGEE). Although both classes of models have been extensively published and widely applied, differences between and limitations about these methods are not clearly delineated and well documented. Unfortunately, some of the differences and limitations carry significant implications for reporting, comparing and interpreting research findings.

In this report, we review both major approaches for longitudinal data analysis and highlight their similarities and major differences. We focus on comparison of the two classes of models in terms of model assumptions, model parameter interpretation, applicability and limitations, using both real and simulated data. We discuss caveats and cautions when applying the two different approaches to real study data.

Key words: binary variables, correlated outcomes, generalized linear mixed-effects models, weighted generalized estimating equations, latent variable models, R, SAS

[*Shanghai Arch Psychiatry*. 2016; **28**(5): 293-300. doi: <http://dx.doi.org/10.11919/j.issn.1002-0829.216081>]

1. Introduction

Longitudinal study designs have become increasingly popular in research and practice across all disciplines. Such designs capture both between-individual differences and within-subject dynamics, providing opportunities to study complex biological, psychological and behavioral changes over time such as causal treatment effects and mechanisms of change^[1,2]. Since longitudinal study designs create serial correlations over repeated assessments from same subjects, traditional statistical methods for cross-sectional data analysis such as linear and logistic regression do not apply. In addition, since longitudinal studies are typically of long

duration, missing data is common. Specialized models and methods must be used to address the two major issues.

The two dominant approaches for longitudinal data analysis are the generalized linear mixed-effects model (GLMM) and weighted generalized estimating equations (WGEE)^[1]. Both methods are derived from the same class of models for cross-sectional data, the generalized linear models (GLM). Because different techniques are used to extend the GLM to longitudinal data, the GLMM and WGEE are quite different and some of the differences carry significant implications for their applicability and interpretation of study findings.

¹Department of Biostatistics and Computational Biology, University of Rochester, Rochester, NY, USA.

²SUNY Upstate Medical University, Syracuse, NY, USA.

³Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN, USA.

⁴Department of Epidemiology, Tulane University, New Orleans, LA, USA

⁵Case Western Reserve University at MetroHealth Medical Center, Cleveland, OH, USA.

* correspondence: Dr. Douglas Gunzler. Mailing address: Case Western Reserve University at MetroHealth Medical Center, Center for Health Care Research & Policy, 2500 MetroHealth Drive, Cleveland, OH, USA. Postcode: 44109-1998. E-Mail: dgunzler@metrohealth.org

Despite the existence of an extremely large body of literature discussing the development and application of the two approaches, practitioners still often face many difficult questions when choosing and applying such models to real study data.

For example, which approach is applied given data from a study? Do the two approaches yield identical estimates and/or inferences? If not, how should one approach and interpret such differences? What are the pros and cons associated with each approach? Although some questions have well-documented answers in the literatures, others have only been approached recently and still await answers.

In this report, we first give an overview of the approaches and then discuss major differences between the two classes of models. Unlike the literature on the discussion of the two methods, we focus on their practical implications, which we think provide useful guidance for practitioners for selecting right approaches for their studies and effectively addressing their study questions.

2. Models for Longitudinal Data

Since both the GLMM and WGEE are extensions of the GLM, we start with a brief overview of the latter.

2.1 Generalized Linear Models (GLM)

Consider a sample of n subjects and let $Y_i(X_i)$ denote a continuous response (a vector of explanatory variables). The classic linear model is given by:

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad 1 \leq i \leq n, \quad (1)$$

where $N(\mu, \sigma^2)$ denotes a normal distribution with mean μ and σ^2 variance. The linear model is widely used in research and practice. One major limitation is that it only applies to continuous response Y_i . The generalized linear models (GLM) extend the classic linear model to non-continuous response such as binary.

To express the GLM, we first rewrite the linear regression in (1) as

$$y_i | \mathbf{x}_i \sim N(\mu_i, \sigma^2), \quad \mu_i = E(y_i | \mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n, \quad (2)$$

where $y_i | \mathbf{x}_i$ denotes the conditional distribution of Y_i given X_i and $E(y_i | \mathbf{x}_i)$ denotes the conditional mean of Y_i given X_i . By replacing the normal in (2) with other distributions appropriate for the type of response, we obtain the class of GLM:

$$y_i | \mathbf{x}_i \sim f(\mu_i), \quad g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n, \quad (3)$$

where $f(\mu)$ denotes some distribution with mean μ and $g(\mu)$ is a function of μ . Since $g(\mu)$ links the mean to the explanatory variables, $g(\mu)$ is called the link function.

The specification of $f(\mu)$ and $g(\mu)$ depend on the type of response Y_i . For a binary Y_i , $f(\mu)$ is the Bernoulli distribution and $g(\mu)$ is often set as the logit function $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

The resulting GLM is the logistic regression. For a count Y_i , a popular choice for $f(\mu)$ is the Poisson distribution and $g(\mu)$ is the log function, $g(\mu) = \log(\mu)$. Another choice for $f(\mu)$ is the negative binomial. A major limitation of Poisson is that its variance is the same as its mean. Count responses arising in many real studies often have variances larger than means, a phenomenon known as “overdispersion” [1]. The negative binomial is similar to the Poisson, but unlike the Poisson, allows for overdispersed count responses [1].

Inference for GLM can be based on maximum likelihood (ML) or estimating equations (EE). The classic ML provides most efficient estimates, if the response Y_i follows the specified distribution such as the normal in the linear regression in (1). In many studies, it may be difficult to specify the right distribution, in which case the ML will yield biased estimates if the specified distribution does not match the data distribution. The modern alternative EE uses an approach for inference that does not require specification of a mathematical distribution for Y_i , thereby providing valid inference for a wider class of data distribution. Since no distribution is required under EE, we may also express the GLM in this case as:

$$\mu_i = E(y_i | \mathbf{x}_i), \quad g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n, \quad (4)$$

or simply

$$g(E(y_i | \mathbf{x}_i)) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n, \quad (5)$$

or equivalently

$$E(y_i | \mathbf{x}_i) = h(\mathbf{x}_i^T \boldsymbol{\beta}), \quad 1 \leq i \leq n, \quad (6)$$

where $h = g^{-1}$ is the inverse of $g(\mu)$. When specified without the distribution component, (4), (5) or (6) are also called the semi-parametric GLM. In comparison, (3) is called the parametric GLM.

Example 1. In a suicide study, we may model number of suicide attempts, Y_i , a count response, using a parametric GLM:

$$y_i | \mathbf{x}_i \sim \text{Poisson}(\mu_i), \quad \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n, \quad (7)$$

where X_i denotes a set of explanatory variables such as age, physical problems and history of depression. If concerned about overdispersion, such as indicated empirically by a much larger variance as compared to the mean, the semi-parametric GLM below may be used

$$\mu_i = E(y_i | \mathbf{x}_i), \quad \log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \quad 1 \leq i \leq n.$$

Unlike the parametric model in (7), the semi-parametric GLM above provides valid inference regardless of presence of over-dispersion.

2.2 Extension of GLM to longitudinal data

2.2.1 Weighted Generalized Estimating Equations (WGEE)

Consider a longitudinal study with T time points and let Y_{it} and X_{it} denote the same response and predictors/covariates as in the cross-sectional setting, but with t indicating their dependence on the time of assessment ($1 \leq i \leq n, 1 \leq t \leq T$). For each time t , we can apply the GLM in (3) to model the regression relationship between Y_{it} and X_{it} at each point

$$y_{it} | \mathbf{x}_{it} \sim f(\mu_{it}), \quad \mu_{it} = E(y_{it} | \mathbf{x}_{it}), \quad g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta}_t, \quad 1 \leq t \leq T, \quad 1 \leq i \leq n.$$

We can then get estimates of β_t for each time point t . However, it is difficult to interpret different β_t across the different time points. Moreover, it is technically challenging to combine estimates of β_t to test hypotheses concerning temporal trends because of interdependence between such estimates.

The WGEE addresses the aforementioned difficulties by using a single estimate β to model changes over time based on multiple assessment times^[1,3]. Since the WGEE estimates β using a set of equations that do not rely on assumed distribution $f(\mu_{it})$ in (8), the first part of the GLM can be removed and the resulting model becomes

$$\mu_{it} = E(y_{it} | \mathbf{x}_{it}), \quad g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta}, \quad 1 \leq t \leq T, \quad 1 \leq i \leq n. \quad (8)$$

By comparing the WGEE above with the model in (4), it is seen that the WGEE is an extension of the semi-parametric GLM to longitudinal data. The key difference between (4) and (8) is that (8) is not simply an application of GLM to each of the time points, but rather an extension of the model in (4) to provide a single parameter vector β for easy interpretation and estimate this parameter vector by using data from all time points and accounting for correlations between the repeated assessments. Like the semi-parametric GLM, the WGEE provides valid inference for a wider class of data distributions.

2.3 Generalized Linear Mixed-effects Models (GLMM)

The GLMM extends the GLM to longitudinal data analysis using a completely different approach.

Consider again a longitudinal study with T time points and let Y_{it} and X_{it} denote the same response and predictors/covariates as in the WGEE above. The GLMM is specified by:

$$y_{it} | \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{b}_i \sim \tilde{f}(\tilde{\mu}_{it}), \\ g(\tilde{\mu}_{it}) = \mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}} + \mathbf{z}_{it}^T \mathbf{b}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \Sigma_b), \quad 1 \leq i \leq n, \quad 1 \leq t \leq T, \quad (9)$$

where $N(\boldsymbol{\mu}, \Sigma)$ denotes a multivariate normal with mean $\boldsymbol{\mu}$ and variance Σ , \mathbf{z}_{it} is a vector of predictors/covariates (often set equal to X_{it}), and $g(\boldsymbol{\mu})$ is the appropriate link function for the type of response Y_{it} . The vector of latent variables, \mathbf{b}_i , is called the random effects, denoting individual differences from the

population mean \mathbf{b}_i , which is known as the fixed effects. Although $\tilde{\boldsymbol{\beta}}$ is typically assumed to follow a multivariate normal as in (3), other types of distributions may also be considered.

Unlike the WGEE, the GLMM accommodates correlated responses Y_{it} by directly modeling their joint distribution. Since multivariate distributions are extremely complex except for the multivariate normal, latent variables \mathbf{b}_i are generally employed to model the correlated responses. Thus, although Y_{it} is still modeled for each time point t , by including the random effect \mathbf{b}_i in the specification of the conditional distribution of Y_{it} given \mathbf{b}_i (X_{it} and \mathbf{z}_{it}), the GLMM in (9) allows the resulting Y_{it} 's to be correlated (conditional on X_{it} and \mathbf{z}_{it} only). This approach allows one to specify multivariate distributions using familiar univariate distributions such as the Bernoulli (for binary responses) and the Poisson (for count responses).

2.4 Key differences between GLMM and WGEE

Although both WGEE and GLMM are extensions of the GLM, the two approaches are quite different because of the way the extensions are accomplished. In this section, we discuss such key differences.

2.4.1 Interpretation of Model Parameters

A fundamental difference between the WGEE and GLMM is in interpretation of model parameters. As noted earlier, the WGEE is an extension of the semi-parametric GLM, while the GLMM is an extension of the parametric GLM. Although the parameter vector β has the same interpretation between the parametric (1) and semi-parametric (4) GLM, the parameter vectors β in the WGEE (8) and β in the GLMM (9) are generally different except for the linear regression.

Example 2. Consider the model in Example 1, but now assume a longitudinal study with T assessment times. Let Y_{it} and X_{it} denote the longitudinal versions of Y_i and X_i . Again, if Y_{it} at each point is modeled to follow a Poisson, the GLMM has the form:

$$y_{it} | \mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{b}_i \sim \text{Poisson}(\tilde{\mu}_{it}), \\ \log(\tilde{\mu}_{it}) = \mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}} + \mathbf{z}_{it}^T \mathbf{b}_i, \quad \mathbf{b}_i \sim N(\mathbf{0}, \Sigma_b), \\ 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (10)$$

On the other hand, the WGEE for a count response has the form

$$\log(E(y_{it} | \mathbf{x}_{it})) = \mathbf{x}_{it}^T \boldsymbol{\beta}, \quad \text{or} \\ E(y_{it} | \mathbf{x}_{it}) = \exp(\mathbf{x}_{it}^T \boldsymbol{\beta}), \quad 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (11)$$

The two models look quite the same, except for the additional random effect in (10). So, to compare the two models, we integrate out \mathbf{b}_i in (10) and calculate

the conditional expectation $E(y_{it} | \mathbf{x}_{it})$ to obtain (see Zhang et al., 2012 for derivation)^[4]

$$E(y_{it} | \mathbf{x}_{it}) = E \left[\exp \left(\frac{1}{2} \mathbf{z}_{it}^T \Sigma_b \mathbf{z}_{it} \right) | \mathbf{x}_{it} \right] \exp(\mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}}), \quad 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (12)$$

This conditional expectation is different from the expression in (11) for the WGEE unless

$$E \left[\exp \left(\frac{1}{2} \mathbf{z}_{it}^T \Sigma_b \mathbf{z}_{it} \right) | \mathbf{x}_{it} \right] = 1$$

In some special cases, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ may be identical except for the intercept term. For example, if \mathbf{z}_{it} is independent of \mathbf{x}_{it} , then

$$E \left[\exp \left(\frac{1}{2} \mathbf{z}_{it}^T \Sigma_b \mathbf{z}_{it} \right) | \mathbf{x}_{it} \right] = E \left[\exp \left(\frac{1}{2} \mathbf{z}_{it}^T \Sigma_b \mathbf{z}_{it} \right) \right] = \exp(\gamma_0)$$

Thus (12) reduces to

$$E(y_{it} | \mathbf{x}_{it}) = \exp(\gamma_0 + \mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}}), \quad 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (13)$$

Except for the intercept term, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ have the same interpretation. The next Example illustrates this special case with data from a real study.

Example 3. The COMBINE study was a multi-site randomized clinical trial with longitudinal follow-up to compare intervention effects between nine pharmacological and/or psychosocial treatments. For illustration purposes, we combined all the 9 treatment conditions and focused on the two drinking outcomes, days of any drinking and days of heavy drinking over the past month, at three visits during treatment at weeks 8, 16 and 26.

Denote the three visits by $1 \leq t \leq 3$. To see whether the mean days of any drinking (heavy drinking) varied in the three time periods, we created two dummy variables indicating weeks 16 and 26 as the fixed effect X_i and a random intercept to account for correlations among the repeated assessments for the GLMM

$$y_{it} | \mathbf{x}_{it}, b_i \sim \text{Poisson}(\tilde{\mu}_{it}),$$

$$\log(\tilde{\mu}_{it}) = x_{i1t} \tilde{\beta}_0 + x_{i2t} \tilde{\beta}_1 + x_{i3t} \tilde{\beta}_2 + z_i b_i, \quad b_i \sim N(0, \tau^2),$$

$$x_{i1t} = 1, \quad x_{ikt} = \begin{cases} 1 & \text{if } t = k \\ 0 & \text{if otherwise} \end{cases}$$

$$k = 2, 3, \quad z_i = 1, \quad 1 \leq i \leq n, \quad 1 \leq t \leq 3.$$

Since $Z_i=1$ and $X_{i1t}=1$ are both constants, they are independent. We fit a WGEE with the same predictors $X_{it}=(X_{i1t}, X_{i2t}, X_{i3t})^T$ as in the fixed effect of GLMM above. As indicated in Example 2, $\tilde{\beta}_0$ differs from its WGEE counterpart β_0 by a constant $\gamma_0 = \ln(E(\exp(\frac{1}{2} z_i^T \tau^2)))$, whereas $\tilde{\beta}_1$ and $\tilde{\beta}_2$ retain the same scales as the corresponding β_1 and β_2 in the WGEE model.

Shown in Table 1 are the estimates of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ and associated standard errors, and test statistics and associated p-values for both models. As expected, estimates of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ were quite close to the respective WGEE estimates β_1 and β_2 , but those of were much smaller (in magnitude) than their WGEE counterparts (0.694 vs. 1.908 for Days of Any Drinking and -0.30 vs. 1.307 for Days of Heavy Drinking).

Since $\tilde{\beta}_1$ and $\tilde{\beta}_2$ (β_1 and β_2) represent changes relative to the reference level $\tilde{\beta}_0$ (β_0) from Visit 1 to Visit 2 and from Visit 1 to Visit 3 under GLMM (WGEE), the difference between the estimated $\tilde{\beta}_0$ and β_0 implies not only different means at visit 1, but also at visits 2 and 3 for each drinking outcome. For example, for the outcome of Days of Any Drinking, the WGEE estimates indicate a mean of 6.7, 7.7 and 8.9 days of any drink over the three visits, much lower than 1.9, 2.3 and 2.7 days of any drink at the corresponding visits estimated by the GLMM. The WGEE estimates are actually identical to sample means of this drinking outcome at each visit, but the GLMM estimates are not, making the latter difficult to interpret.

Table 1. Estimates of reference level at Visit 1 (β^*_0 for WGEE and β_0 for GLMM) and change from reference level at Visit 2 (β_1 and β^*_1) and at Visit 3 (β_2 and β^*_2), along with associated standard errors, for Days of Any Drinking and Days of Heavy Drinking for the real COMBINE Study in Example 3.

Comparison of Estimates (Standard Errors) between GEE and GLMM			
	Visit 1 (β_0 or β^*_0)	Visit 2 (β_1 or β^*_1)	Visit 3 (β_2 or β^*_2)
Days of Any Drinking			
Model Fit			
GEE($\boldsymbol{\beta}$)	1.908 (3.6)	0.144 (2.2)	0.144 (2.8)
GLMM($\boldsymbol{\beta}^*$)	0.694 (6.2)	0.144 (1.5)	0.144 (1.5)
Days of Heavy Drinking			
GEE($\boldsymbol{\beta}$)	1.307 (4.7)	0.224 (3.3)	0.216 (4.1)
GLMM($\boldsymbol{\beta}^*$)	-0.30 (7.3)	0.224 (1.9)	0.216 (1.9)

Example 4. In Example 2, now suppose that y_{it} is binary and is modeled by a GLMM for binary response as follows

$$y_{it} | \mathbf{x}_{it}, \mathbf{z}_{it}, b_i \sim \text{Bernoulli}(\tilde{\mu}_{it}),$$

$$\text{logit}(\tilde{\mu}_{it}) = \mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}} + \mathbf{z}_{it}^T b_i \sim N(0, \Sigma_b), \quad 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (14)$$

The corresponding WGEE has the form

$$\text{logit}(E(y_{it} | \mathbf{x}_{it})) = \mathbf{x}_{it}^T \boldsymbol{\beta}, \text{ or}$$

$$E(y_{it} | \mathbf{x}_{it}) = \frac{\exp(\mathbf{x}_{it}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_{it}^T \boldsymbol{\beta})}, \quad 1 \leq i \leq n, \quad 1 \leq t \leq T. \quad (15)$$

As in Example 2, the two models look the same, except for the additional random effect in (14). To compare the two, we again need to integrate out b_i in (14). In general, it is difficult to obtain a closed-form expression for the resulting integral. But, in the special case when Z_{it} is a subset of X_{it} and $Z_{it} = Z_t$, the integral can be expressed as (see Zhang, Xia et al., 2011 for derivation)^[5]

$$\text{logit}[E(\tilde{\mu}_{it} | \mathbf{x}_{it})] \approx (1 + c^{-2} \mathbf{z}_{it}^T \Sigma_b \mathbf{z}_{it})^{-1} \mathbf{x}_{it}^T \tilde{\boldsymbol{\beta}}, \quad c = \frac{15\pi}{16\sqrt{3}}. \quad (16)$$

Thus $\tilde{\boldsymbol{\beta}}$ is a rescaled $\boldsymbol{\beta}$. Even in this special case, it may not be easy to interpret $\tilde{\boldsymbol{\beta}}$, as demonstrated by a real study example next.

Example 5. In a recent study on smoking cessation, 276 subjects participated in multi-component program adapted to seriously mentally ill patients within an outpatient mental health clinic. Out of these subjects, 99 also participated in a formal evaluation, in which interviews were conducted at the point of enrollment (baseline) and again at 3, 6 and 12 months. A primary outcome of the study is the 7-day point prevalence abstinence (defined as no smoking at all in the previous 7 days). We modeled changes of this longitudinal binary abstinence outcome using data from the 99 subjects.

We applied both longitudinal approaches and used three dummy variables to model the rate of 7-day point prevalence abstinence at each of the follow-up times for the WGEE and also included a random intercept for the GLMM model. Thus, the GLMM has the form

$$y_{it} | \mathbf{x}_{it}, b_i \sim \text{Bernoulli}(\tilde{\mu}_{it}),$$

$$\log(\tilde{\mu}_{it}) = x_{i1t} \tilde{\beta}_0 + x_{i2t} \tilde{\beta}_1 + x_{i3t} \tilde{\beta}_2 + x_{i4t} \tilde{\beta}_3 + z_i b_i, \quad b_i \sim N(0, \tau^2),$$

$$x_{i1t} = 1, \quad x_{i2t} = \begin{cases} 1 & \text{if } t = k \\ 0 & \text{if otherwise} \end{cases}$$

$$k = 2, 3, 4, \quad z_i = 1, \quad 1 \leq i \leq n, \quad 1 \leq t \leq 4,$$

while the WGEE is given by:

$$\text{logit}[E(y_{it} | \mathbf{x}_{it})] = x_{i1t} \beta_0 + x_{i2t} \beta_1 + x_{i3t} \beta_2 + x_{i4t} \beta_3,$$

$$x_{i1t} = 1, \quad x_{i2t} = \begin{cases} 1 & \text{if } t = k \\ 0 & \text{if otherwise} \end{cases}, \quad k = 2, 3, 4, \quad 1 \leq i \leq n, \quad 1 \leq t \leq 4,$$

Shown in Table 2 are the estimates of $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\beta}}$ and associated standard errors, and test statistics and associated p-values for both models. The estimates were quite different between the two models, although

the ratio $\frac{\tilde{\beta}_k}{\beta_k} \approx 0.7$, a constant for all $1 < k < 3$. This is

not a coincidence, but expected, because of (16) and the fact that Z_{it} in the GLMM model is a subset of X_{it} .

Also, since $H_0 : \beta_k = 0$ under one model implies that H_0 holds true for the corresponding coefficient under the other model, same conclusions were obtained regarding statistical significance of the parameters, although the p-values under the two models were slightly different.

The scale difference in the parameters between the two models did have serious implications for the interpretation of estimates from the GLMM. Shown in the last two columns of Table 2 are the rates of 7-day point prevalence abstinence over time estimated by each of the two models. Since the WGEE becomes the logistic regression at each assessment time, it yields estimates identical to the observed rates of 7-day point prevalence abstinence. Estimates from the GLMM were quite different, underestimating the observed rates by over 50%. Even in this simplest case where estimates of GLMM are a scale shift of their GEE counterparts, GLMM estimates are difficult to interpret.

2.4.2 Computational Issues for GLMM

Inference for the WGEE is based on a set of equations. Although it is generally not possible to express solutions (estimates) in closed-form, the equations are readily solved numerically^[1]. Inference for the GLMM, however, is much more challenging. Since the likelihood function arising from the GLMM is generally quite complex, involving multidimensional integrals, it is difficult to directly maximize this function except for the special linear models case for continuous outcomes. Different approaches have been proposed to address the computational issues when modeling non-continuous responses using the GLMM.

The approach implemented in most major software such as SAS and R is integral approximation. This approach first approximates the log-likelihood function and then maximizes the approximated function using the Newton and Gauss-Hermite quadratures. However, studies have shown that the approach does not work well as one would expect^[6], especially for modeling binary responses. Below, we highlight some key findings from these studies.

Example 6. Zhang, Lu and Feng et al. (2011)^[6] examined the computational issues by fitting GLMM for simulated binary responses using both SAS and R. They simulated an explanatory variable X_i and the response Y_{it} from the GLMM:

$$y_{it} | x_i, b_i \sim \text{Bernoulli}(\mu_{it}),$$

$$\text{logit}(\mu_{it}) = \beta_0 + b_{i0} + x_i (\beta_1 + b_{i1}),$$

$$\mathbf{b}_i = \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0.25 \\ 0.25 & 1 \end{pmatrix}\right), \quad x_{it} \sim N(0, 1), \quad t = 1, 2, 3, \quad (17)$$

where $\beta_0 = \beta_1 = 1$ and $\tau = 0.001$ and 2. For $\tau = 0.001$, the within-subject correlation was very small and thus negligible, making the Y_{it} 's almost independent. For $\tau = 2$, the within-subject correlations was about 0.5.

Table 2. Estimates of parameters, standard errors, p-values and rates of 7-day point prevalence abstinence at baseline and 3 follow-up visits from the WGEE and GLMM models for the real Smoking Cessation Study in Example 5.

Parameters	Estimates (Standard Errors)		p-values		Rates of 7-day point prevalence abstinence	
	WGEE	GLMM	GEE	GLMM	GEE	GLMM
Baseline (β or β^*)	4.05 (0.92)	5.47 (1.31)	<.001	<.001	0.076	0.021
Month 3 (β_0 or β^*_0)	0.87 (0.42)	1.27 (0.56)	0.036	0.026	0.164	0.070
Month 6 (β_1 or β^*_1)	1.01 (0.42)	1.49 (0.56)	0.016	0.009	0.184	0.086
Month 12 (β_2 or β^*_2)	0.55 (0.45)	0.79 (0.57)	0.217	0.173	0.124	0.044

They simulated data from the GLMM in (17) with a sample size $n=500$, fit the same model to the simulated data using R and SAS, and repeated the process 1,000 times. Shown in Table 3 are the estimates of parameter vector β (under “ $\beta_0=1$ ” and “ $\beta_1=1$ ” based on averaging over 1,000 estimates from fitting the model to each simulated data) and associated standard errors (under $S.E. \times 10$ based on sample standard deviations from the Monte Carlo 1,000 estimates of β) from fitting SAS NLMIXED procedure and R lme4 function. For $\tau=0.001$, the estimates were quite similar between the two, but for $\tau=2$, the SAS NLMIXED procedure provided more accurate estimates. There were more pronounced differences in the standard errors between the two, with the R lme4 consistently yielding lower standard errors than the SAS NLMIXED. Such differences play a significant role in hypothesis testing.

Shown in Table 3 under “Type I error” are the type I error rates for testing the hypothesis, $H_0: \beta_1=1$ vs. $H_a: \beta_1 \neq 1$ based on the Wald statistic from the SAS and R procedures. The Wald statistic is $T_W = n(\hat{\sigma}_{\beta_1}^2)^{-1}(\hat{\beta}_1 - 1)^2$ where $\hat{\sigma}_{\beta_1}^2$ is the standard error for the estimate. The type I error rate was calculated as:

$$\hat{\alpha} = \frac{1}{1000} \sum_{m=1}^{1000} I_{\{T_W^{(m)} \geq q_{1,0.95}\}}$$

where $q_{1,0.95}$ is the 95th percentile of χ^2_1 and $T_W^{(m)}$ denotes the Wald statistic for testing the hypothesis based on m th simulated data ($1 \leq m \leq 1,000$). Since the null hypothesis $H_0: \beta_1=1$ is true, the type I error rate $\hat{\alpha}$ should be close to the nominal value $\alpha=0.05$, if the SAS and R procedures provided correct inference. Although the SAS NLMIXED did yield type I error rates close to the nominal value, R lme4’s estimates were inflated, especially for the case with higher within-subject correlation $\tau=2$.

In addition to the SAS and R procedures, Zhang, Lu and Feng et al. (2011)^[6] also considered other procedures in SAS and R and found that none provided correct estimates. Their conclusion was that the SAS

NLMIXED procedure provided more accurate estimates and type I error rates. However, more recent studies by Chen, Knox, Arora et al. (2016)^[7] and Chen, Lu, Arora et al. (2016)^[8] show that this SAS procedure did not provide correct estimates either.

Table 3. Estimates of parameters, standard errors, Type I error rates (for testing null: $H_0: \beta_1=1$ from SAS NLMIXED and R lme4 procedures for two within-subject correlation cases ($\tau=0.001$ and $\tau=0.0001$)) for the simulation study in Example 6.

τ	Software	Type I error for testing null $H_0: \beta_1=1$	$\beta_0=1$	S.E.	$\beta_1=1$	S.E.
0.001	SAS	0.046	1.025	0.075	1.029	0.085
	NLMIXED	0.098	1.022	0.067	1.029	0.075
2	SAS	0.066	0.992	0.125	0.976	0.166
	NLMIXED	0.256	0.983	0.123	0.912	0.141

Example 7. Chen, Knox, Arora et al. (2016)^[7] considered clustered binary responses arising from multi-center studies. They modeled and simulated clustered binary responses from the following GLMM

$$y_{ki} | x_{ki}, \lambda_k \stackrel{i.d.}{\sim} \text{Bernoulli}(\mu_{ki}), \quad \mu_{ki} = \frac{\exp(\beta_0 + \beta_1 x_{ki} + \lambda_k)}{1 + \exp(\beta_0 + \beta_1 x_{ki} + \lambda_k)},$$

$$x_{ki} \sim \text{Bernoulli}\left(\frac{1}{2}\right), \quad \lambda_k \sim N(0, \sigma_\lambda^2), \quad 1 \leq i \leq n, \quad 1 \leq k \leq K. \tag{18}$$

where K denotes the number of study sites, n number of subjects within each site (cluster size), x_{ki} is a binary variable indicating treatment condition assigned to the i th subject within the k th site, and λ_k is the random effect accounting for correlations between responses y_{ki} from subjects with the k th site. They considered testing the hypothesis:

$$H_0 : p_0 = 0.10 \quad \text{vs.} \quad H_a : p_1 = 0.05, \quad (19)$$

where $p_j = E(y_{ki} | x_{ki} = j)$ is the mean response y_{ki} (or proportion of $y_{ki} = 1$) for the treatment group. The parameters β_0 and β_1 in (18) are related to p_0 and p_1 in (19) as follows:

$$p_k = \Phi \left(\frac{\beta_0 + \beta_1 x_{ki}}{\sqrt{c^2 + \sigma_\lambda^2}} \right), \quad c = \frac{15\pi}{16\sqrt{3}}.$$

Thus given (19) and σ_λ^2 , we can solve the above for β_0 and β_1 (numerically).

Chen, Knox, Arora et al. (2016)^[7] considered $\sigma_\lambda^2 = 0$, $\sigma_\lambda^2 = 0.1$ and $\sigma_\lambda^2 = 1$, with $K = 20$ clusters and $n = 25$ within each cluster. For each σ_λ^2 , they obtained β_0 and β_1 and simulated data from the GLMM in (18) under $H_a : p_1 = 0.05$. They fit the model (18) to the data generated using SAS NLMIXED, tested the null $H_0 : p_0 = 0.10$ and rejected the null if the p-value is larger than $\alpha = 0.05$. The process was repeated 1,000 times and power was estimated by the percent of times the null was rejected.

Shown in Table 4 are power estimates under “NLMIXED” for the three cases of σ_λ^2 , along with true power values under (“True”) obtained using a different approach developed in Chen, Knox, Arora et al. (2016)^[7]. As seen, power estimates from obtained from the SAS NLMIXED were quite different from the true power for each case of σ_λ^2 , all underestimating the true power. Note that when $\sigma_\lambda^2 = 0$, there is no data cluster and power only depends on the total sample size Kn . In this case, we can also obtain power estimates by using instead the SAS LOGISTIC NLMIXED for non-clustered data. This is indeed the case, since Chen, Knox, Arora et al. (2016)^[7] reported that they obtained power estimates similar to 0.561 when running the Monte Carlo substitution to estimate power using the SAS LOGISTIC.

Table 4. Power estimates from SAS NLMIXED along with true power values for two data clustering cases ($\sigma_\lambda^2=0.1$ and $\sigma_\lambda^2=1$) for the simulation study in Example 7.

K	n	$\sigma_\lambda^2=0$		$\sigma_\lambda^2=0.1$		$\sigma_\lambda^2=1$	
		True	NL-MIXED	True	NL-MIXED	True	NL-MIXED
50	10	0.561		0.565	0.546	0.598	0.473
20	25	0.561	0.275	0.561	0.336	0.590	0.455

3. Discussion

In this report, we discussed the two most popular regression models for longitudinal data. We focused on interpretation and computation of model parameters. For parameter interpretation, we discussed differences between the GLMM and WGEE when applied to model

binary and count responses. Since parameters from the two longitudinal models are generally quite different, we should not expect similar estimates when applying the two models to real study data. Moreover, except for some special cases, it is generally not possible to find a relationship between estimates from the two models. Our analysis indicates that GLMM estimates can be quite difficult to interpret, while WGEE estimates afford straightforward interpretation.

Another major issue with the GLMM is to obtain reliable estimates using existing software. It seems that even software giants like SAS cannot provide correct estimates when applying GLMM to model binary responses. Until the computational problem is resolved, one may want to consider applying WGEE when modeling longitudinal binary responses.

We focused on the binary and count response when discussing interpretation of model parameters in this report. When modeling continuous responses, the two longitudinal models have the same interpretation for their model parameters and thus the interpretational issue does not arise. The computational problem seems only relevant to binary responses. For continuous responses, the log-likelihood function can be solved accurately^[1]. For count responses, major software such as R and SAS seem to provide quite reliable estimates.

Funding

The work was supported in part by a grant from NIH/NCRR CTSA KL2TR000440. The funding agreement ensured the authors’ independence in designing the study, interpreting the data, writing, and publishing the report.

Conflict of interest statement

The authors report no conflict of interest related to this manuscript.

Authors’ contributions

All authors worked together on the manuscript and contributed their expertise to the relevant sections.

Mr. Lin and Dr. Gunzler outlined the structure of the manuscript and integrated sections from the other authors.

Mr. Lin also performed some of the simulations under the supervision of Dr. Gunzler and Dr. Zhang.

Mr. Tu helped develop the real data examples and interpretation of results with Dr. Zhang, Dr. Wang and Dr. He helped develop the sections on longitudinal models and interpretation of simulation study results.

纵向数据的现代分析方法及其功能、说明和注意事项

GE L, TU JX, ZHANG H, WANG H, HE H, GUNZLER D

概述：纵向研究可用于精神卫生及其服务领域的科研中。纵向数据分析的主要方法是广义线性混合效应模型 (GLMM) 和加权广义估计方程 (WGEE)。虽然这两个模型已被广泛应用，也有大量文献发表，但是人们并没有清晰地描述这些方法间的差别以及方法本身的局限性，缺少相关的文献记录。遗憾的是，有些差别和局限性会明显影响对研究结果的报告、比较和解释。

本文回顾了纵向数据分析的两种主要方法，强调两者的相似之处和主要差别。我们使用真实数据和模拟数据着重比较这两类模型的假设、对参数的解释、适用性和局限性，并讨论了将这两种不同的方法用于真实数据研究时的注意事项，提出了相关的警示。

关键词：二分类变量，相关结果，广义线性混合效应模型，加权广义估计方程，潜变量模型，R，SAS

References

1. Tang W, He H, Tu XM. *Applied Categorical and Count Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC; 2012
2. Gunzler D, Lu N, Tang W, Wu P, Tu XM. A Class of Distribution-free Models for Longitudinal Mediation Analysis. *Psychometrika*. 2014; **17**(4): 543-568. doi: <http://dx.doi.org/10.1007/s11336-013-9355-z>
3. Kowalski J, Tu XM. *Modern Applied U Statistics*. New York: Wiley; 2007
4. Zhang H, Yu Q, Feng C, Gunzler D, Wu P, Tu XM. A new look at the difference between GEE and GLMM when modeling longitudinal count responses. 2012; *J Appl Stat*. 2012; **39**(9): 2067-2079. doi: <http://dx.doi.org/10.1080/02664763.2012.700452>
5. Zhang H, Xia Y, Chen R, Gunzler D, Tang W, Tu XM. On modeling longitudinal binomial responses --- Implications from two dueling paradigms. *Appl Stat*. 2011; **38**: 2373-2390. doi: <http://dx.doi.org/10.1080/02664763.2010.550038>
6. Zhang H, Lu N, Feng C, Thurston S, Xia Y, Tu XM. On fitting generalized linear mixed-effects models for binary responses using different statistical packages. *Stat Med*. 2011; **30**: 2562-2572. doi: <http://dx.doi.org/10.1002/sim.4265>
7. Chen T, Knox K, Arora J, Tang W, Tu XM. Power analysis for clustered non-continuous responses in multicenter trials. *Appl Stat*. 2016; **43**(6): 979-995. doi: <http://dx.doi.org/10.1080/02664763.2015.1089218>
8. Chen T, Lu N, Arora J, Katz I, Bossarte R, He H, et al. Power analysis for cluster randomized trials with binary outcomes modeled by Generalized Linear Mixed-effects Models. *Appl Stat*. 2016; **43**(6): 1104-1118. doi: <http://dx.doi.org/10.1080/02664763.2015.1092109>



Lin Ge obtained his bachelor's of science degree in Mathematics from Southwest University, PRC, in 2015. He is currently a Master's student in the Department of Biostatistics and Computational Biology at the University of Rochester in New York, USA. His research interests include categorical data analysis, longitudinal data analysis, social networks and machine learning.