Longitudinal binomial data are frequently generated from multiple questionnaires and assessments in various scientific settings for which the binomial data are often over-dispersed. The standard generalized linear mixed effects model (GLMM) may result in severe underestimation of standard errors of estimated regression parameters in such cases and hence potentially bias the statistical inference. In this paper, we propose a longitudinal beta-binomial model for over-dispersed binomial data and estimate the regression parameters under a probit model using the Generalized Estimating Equation (GEE) method. A hybrid algorithm of the Fisher Scoring and the Method of Moments is implemented for computing the method. Extensive simulation studies are conducted to justify the validity of the proposed method. Finally the proposed method is applied to analyze functional impairment in subjects who are at-risk of Huntington disease (HD) from a multi-site observational study of prodromal HD. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Beta-binomial model; Generalized estimating equation; Generalized linear mixed-effects model; Over-dispersion; Probit model.

1. Introduction

In many medical and biological research areas, longitudinal data composed of repeated binary or binomial responses and a set of exploratory variables are commonly generated. One standard approach to deal with such data is the generalized linear mixed effects model (GLMM), which accommodates response variables that follow distributions other than a normal distribution and contains random effects in the linear predictor but assumes the binomial model for the binomial response made from independent and homogenous binary outcomes. However, such responses are very likely to be subject to excess variability when the binary outcomes are either dependent or nonhomogeneous. For example, in an international multi-site observational study, PREDICT-HD [1], subjects at-risk to develop a neurodegenerative disease, Huntington disease (HD), were followed annually to ascertain disease progression markers that are associated with diagnosis. One of the important affected domains in HD is daily function, whose impairment is highly associated with HD progression. A common measure of daily functioning is the functional assessment scale (FAS) of the Unified Huntington’s Disease Rating Scale (UHDRS) [2]. The FAS consists of 25 items with yes/no responses. The items purport to measure the same...
construct, which suggests they are correlated and cannot be treated as independent Bernoulli outcomes. Rather, the items of the FAS constitute correlated binary data. Binomial data arising from an aggregate of correlated binary outcomes can have excess variability beyond the binomial distribution, a property known as over-dispersion. It has been recognized that over-dispersion can cause underestimation of the standard error of regression parameter estimates in GLMM and, therefore, potentially bias the inference of these parameters [3].

The analysis of longitudinal binomial data requires special consideration. An appropriate model should account for the correlation due to repeated measurements. If the data are generated by a binomial process, then the model should also account for potential over-dispersion. In this paper, we propose a model that addresses both issues.

The beta-binomial distribution [4], the double exponential distribution [5], and a multiplicative generalization of the binomial distribution [6] have all been proposed to handle binomial over-dispersion. Among these, the beta binomial model is most commonly used. It was first proposed to describe variation in the probability of success between sets of trials by Skellam [4]. Various researchers including Griffiths, Williams and Crowder [7, 8, 9] promoted the use of the beta-binomial model for over-dispersed proportions. Among them, Griffiths’s reparameterization has been widely adopted since its introduction. Assorted model estimation methods have been developed as well. Williams [10] proposed the maximum likelihood (ML) estimation approach with iterated reweighted least squares for a beta-binomial model using the logistic link function. Nelder and Pregibon [11] used an extended quasi-likelihood method that requires a much weaker distributional assumption. A mixed strategy of the maximum likelihood estimation and quasi-likelihood estimation was introduced by Brooks [12] whereas Carroll and Ruppert [13] suggested the pseudo-likelihood approach.

Limited research has been done to extend the models addressing the issue of over-dispersion in binomial data for longitudinal studies. Molenberghs et al. [14] proposed models with normal and conjugate random effects including Bernoulli-type models for binary data with the logit link and also with the probit link. The former model computed the success rate of the Bernoulli trial as a product of a beta-binomial distributed random factor and another factor of normal random effects in a logistic form, whereas the latter employed an approximation formula to link the logistic densities to the normal densities. The authors pointed out that the ML estimation for the marginalized probit model faces computational challenges stemming from a multivariate normal integral in the marginal likelihood. The authors suggested several estimating methods including quasi-likelihood, pseudo-likelihood, EM algorithm, Bayesian methods, and a technique to transform the beta random effect to a normal random effect with implementation using the SAS NLMIXED procedure. However, the authors did not actually implement the numerical methods, nor evaluate the performance of the estimating methods. Kassahun et al. [15] adopted the model with logit link from Molenberghs et al. [14], and applied the frequentist maximization approach through SAS NLMIXED to two real datasets with binary outcomes. They also alternatively employed the Markov Chain Monte Carlo (MCMC) technique for estimation and compared inference from both methods. It is well known that the logit link does not combine smoothly with normal random effects [14], and the accuracy and stability of the estimator from such a procedure is not always guaranteed. Neither study provided theoretical validation nor any simulation studies to demonstrate the validity of these methods. In addition, only the binary case and not the general binomial case was analyzed in both studies.

In this article, we propose generalized estimating equations (GEE) as an alternative to ML to analyze longitudinal binomial data. GEE not only avoids the numerical challenges in finding the ML estimator in a specific distributional model for over-dispersed longitudinal binomial data, but it also allows a relaxation of the distributional assumptions required for ML. It will be demonstrated that our proposed GEE approach successfully accounts for the over-dispersion caused by heterogeneity in binary data as well as the correlation between measures on the same subject and, therefore, provides more reliable inferences.

The remainder of this paper is structured as follows. Section 2 introduces the notation and construction of the GEE approach based on the beta-binomial model. We also provide details on the computing procedure. Section 3 discusses extensive simulation studies to compare the performance of the proposed GEE method to the GLMM. In Section 4, we apply the proposed method to analyze the longitudinal FAS data from the PREDICT-HD study for finding the markers
that are responsible for daily function impairment in subjects at-risk for HD. Section 5 provides concluding remarks with potential extensions of the proposed methodology.

2. Model

Suppose we have \( m \) independent subjects and for the \( i \)-th subject \((i = 1, \ldots, m)\), at the \( j \)-th time point \((j = 1, \ldots, n_i)\), we observe a binomial response \( Y_{ij} \) and a \( p \)-dimensional vector of covariates \( X_{ij} \). Let \( Y_i = (Y_{i1}, \ldots, Y_{in_i})^T \) be the \( n_i \times 1 \) response vector and \( X_i = (X_{i1}, \ldots, X_{it}, \ldots, X_{ini})^T \) be the \( n_i \times p \) design matrix which includes a column of 1’s, time-dependent, and time-independent covariates for the \( i \)-th subject. Each subject can have a different number of records either by design or due to missing data. In a longitudinal setting, the \( Y_{ij} \) are usually repeated measures on the same subject and therefore are likely to be correlated. Let \( u_i \) denote the subject-specific random effects for the \( i \)-th subject and \( Z_{ij} \) the \( q \times 1 \) covariate vector for these random effects at the \( j \)-th time point.

2.1. Beta-binomial Model

The beta-binomial model is the most frequently used model to account for the over-dispersion present in binomial data. It is a compound distribution of the beta distribution and the binomial distribution. Suppose that \( Y_{ij} \) follows a beta-binomial distribution. Namely,

\[
Y_{ij}|p_{ij}, u_i \sim \text{Binomial}(K_{ij}, p_{ij})
\]

\[
p_{ij}|u_i \sim \text{Beta}(a_{ij}, b_{ij})
\]

where \( K_{ij} \) is the total number of trials for the \( i \)-th subject at the \( j \)-th time point.

Further assume that for every \( i \) and \( j \), the sum of \( a_{ij} \) and \( b_{ij} \) is fixed. Let \( \mu_{ij} = E(p_{ij}|u_i) = \frac{a_{ij}}{a_{ij}+b_{ij}} \) and \( \tau = \frac{1}{a_{ij}+b_{ij}} \). It can be shown that

\[
E(Y_{ij}|u_i) = K_{ij}E(p_{ij}|u_i) = K_{ij}\mu_{ij}
\]

\[
\text{Var}(Y_{ij}|u_i) = K_{ij}\mu_{ij}(1-\mu_{ij}) + \frac{\tau}{1+\tau}K_{ij}(K_{ij} - 1)\mu_{ij}(1-\mu_{ij})
\]

and

\[
\text{Corr}(Y_{ijk1}, Y_{ijk2}|p_{ij}, u_i) = \frac{\tau}{1+\tau} \text{ for } k_1 \neq k_2
\]

where \( Y_{ijk} \) is the binary component of the binomial distribution \( Y_{ij}|p_{ij}, u_i \), \( k = 1, \ldots, K_{ij} \). This equation characterizes a constant correlation among the binary components of the binomial variable \( Y_{ij} \). The correlation yields some extra amount of variation (the second term of \( \text{Var}(Y_{ij}|u_i) \)) that cannot be modeled by a binomial distribution for binomial data. Let \( \eta = \frac{\tau}{1+\tau} \) denote the correlation coefficient. Then \( \eta(K_{ij} - 1) \) describes the over-dispersion from binomial distribution for the binomial data \( Y_{ij} \).

Consider \( W_{ij} = \frac{Y_{ij}}{K_{ij}} \), the proportion of successful trials for the \( i \)-th subject at the \( j \)-th time. It follows that \( E(W_{ij}|u_i) = \frac{1}{K_{ij}}E(Y_{ij}|u_i) = \mu_{ij} \) and \( \text{Var}(W_{ij}|u_i) = \left\{ \frac{1}{K_{ij}} + \eta(1-\frac{1}{K_{ij}}) \right\} \mu_{ij}(1-\mu_{ij}) \). Let \( W_i = (W_{i1}, \ldots, W_{it}, \ldots, W_{ini})^T \) and \( \mu_i = (E(W_{i1}), \ldots, E(W_{it}), \ldots, E(W_{ini}))^T \). Denote \( g() : (0,1) \rightarrow \mathbb{R} \) as a link function, which will be used to model the marginal mean of the outcome variable \( W_{ij} \). Then

\[
E(W_{ij}|u_i) = g(\mu_{ij}) = X_{ij}^T\beta + Z_{ij}^Tu_i
\]

where \( \beta \) is a \( p \times 1 \) vector of the fixed-effects regression coefficients. Possible choices of the link function include the log function, the logit function, the complementary log-log function, the complementary log-log function, and the inverse function of any strictly increasing cumulative distribution function, including the probit function.
2.2. Review of the Generalized Estimating Equation Method

To handle longitudinal outcomes of various distributions, Liang and Zeger [16] developed a moment-based GEE method that only requires specification of the first two moments of the outcome vector for each subject. Generally, the regression parameters are obtained by solving the GEE,

\[ U(\beta, \alpha) = \sum_{i=1}^{m} D_i^T V_i^{-1} (W_i - \mu_i) = 0 \]

where \( D_i \) is \( \frac{\partial \mu_i}{\partial \beta} \), \( \mu_i \) is the vector of mean responses, and \( V_i \) is the “working” covariance matrix of \( W_i \), which depends on a set of variance-covariance parameters \( \alpha \). In practice, the set of \( \alpha \) parameters is usually replaced by its \( \sqrt{m} \)-consistent estimator \( \hat{\alpha} \), given \( \beta \). Liang and Zeger [16] proved that under mild regularity conditions \( \hat{\beta}_{GEE} \), defined as the solution of \( U(\beta, \hat{\alpha}) = 0 \), is consistent with an asymptotic multivariate normal distribution \( N(0, H^{-1}\Omega H^{-1}) \), where

\[ H = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} D_i^T V_i^{-1} D_i \text{ and } \Omega = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} D_i^T V_i^{-1} \text{Var}(Y_i) V_i^{-1} D_i. \]

2.3. The GEE Method under the Beta-Binomial Model

To construct the GEE for longitudinal beta-binomial data, the first two moments need to be specified based on the distributional assumptions. In order to account for the correlation due to repeated measures and for simplicity of computation, the mixed effects model with random intercept is considered in this article. As for the choice of the link function, the probit function is preferred here because (1) it facilitates the formulation and computation of the mean structure as well as the variance structure, and (2) it is commonly used as the link function for binomial type data. Under this model setting with binomial data, it can be shown that

\[ E(W_{ij}) = E(E(W_{ij}|u_i)) = E\{\Phi(X_{ij}^T \beta + u_i)\} = \Phi\left(\frac{X_{ij}^T \beta}{\sqrt{1+\sigma^2}}\right) \]

where \( \Phi(\cdot) \) denotes the cumulative distribution function of the standard normal distribution, and

\[ (V_i)_{jk} = \begin{cases} \Phi\left(\frac{X_{ij}^T \beta}{\sqrt{1+\sigma^2}}\right), & \text{if } j = k \\ \Phi\left(\frac{X_{ij}^T \beta}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{X_{ik}^T \beta}{\sqrt{1+\sigma^2}}\right), & \text{if } j \neq k \end{cases} \]

where \( (V_i)_{jk} \) refers to the component in the \( j \)th row and \( k \)th column of the variance matrix \( V_i \), \( u_i \) is the random intercept, assumed to follow a normal distribution \( N(0, \sigma^2) \), and \( A_{ijk} = E\{\Phi(X_{ij}^T \beta + u_i)\Phi(X_{ik}^T \beta + u_i)\} \). The variance \( \sigma^2 \) is an unknown nuisance parameter. The variance matrix \( V_i = cov(W_i) \) is employed as the “working” covariance matrix for constructing the GEE. If the data truly follow a beta-binomial distribution, the proposed GEE approach will be efficient in estimating \( \beta \) with this working covariance matrix. Specifically, the GEE for this setting is given by

\[ \sum_{i=1}^{m} X_i^T \Delta_i (\beta, \sigma^2) V_i^{-1} (\beta, \sigma^2, \eta) \{W_i - \mu_i(\beta, \sigma^2)\} = 0 \]

where \( X_i, V_i \) and \( W_i \) are as defined above,

\[ \Delta_i(\beta, \sigma^2) = \frac{1}{\sqrt{1+\sigma^2}} \text{Diag} \left( \phi\left(\frac{X_{i1}^T \beta}{\sqrt{1+\sigma^2}}\right), \ldots, \phi\left(\frac{X_{im}^T \beta}{\sqrt{1+\sigma^2}}\right) \right) \]
and
\[
\mu_i(\beta, \sigma^2) = \left( \Phi \left( \frac{X_i^T \beta}{\sqrt{1 + \sigma^2}} \right), \ldots, \Phi \left( \frac{X_{ni}^T \beta}{\sqrt{1 + \sigma^2}} \right) \right)^T
\]
Here \(\phi(\cdot)\) refers to the probability density function of a standard normal distribution.

2.4. Numerical Algorithm

As stated in the previous section, solving the GEE of \(U(\beta, \hat{\alpha}) = 0\) for \(\hat{\beta}_{GEE}\) requires finding a \(\sqrt{m}\)-consistent estimator of \(\alpha = (\sigma^2, \eta)\). In this article, we adopt the Method of Moments in the spirit of Zeger’s approach [17], which is often used in practice for count data. Two extra estimating equations for \((\sigma^2, \eta)\), are given by

\[
\frac{1}{\sum_{i=1}^{m} n_i(n_i - 1)} \sum_{i=1}^{m} \sum_{j \neq k} \left\{ W_{ij} - \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \right\} \left\{ W_{ik} - \Phi \left( \frac{X_{ik}^T \beta}{\sqrt{1 + \sigma^2}} \right) \right\} = 1
\]

(2)

and

\[
\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \left\{ 1 - \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \right\}^2 \left\{ (K_{ij} - 1)(1 - \eta) \right\} \left\{ \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) - A_{ij} \right\} = 1,
\]

(3)

respectively. In these equations, although \(A_{ijk} = E \{ \Phi(X_{ij}^T \beta + u_i)\Phi(X_{ik}^T \beta + u_i) \}\) cannot be explicitly evaluated, it can be easily approximated by Gauss-Hermite Quadrature. Note, however, that as long as \(\hat{\alpha} = (\hat{\sigma}^2, \hat{\eta})\) is \(\sqrt{m}\)-consistent, the asymptotic normality of \(\hat{\beta}_{GEE}\) does not depend on the choice of the estimator. The above Method of Moments estimates can be shown to be \(\sqrt{m}\)-consistent using the same arguments as given in Hua, Zhang and Tu [18] and hence, they were used for estimating the nuisance parameter \(\alpha\).

Plugging estimates of nuisance parameters \(\hat{\alpha} = (\hat{\sigma}^2, \hat{\eta})\) into Equation (1), we can solve it to obtain the estimates of the parameters of interest, \(\beta\). We adopt the Fisher scoring algorithm to compute the proposed GEE model for updating \(\hat{\beta}_{GEE}\) with the parameter estimates \((\beta(s), \sigma^2(s), \eta(s))\) from the previous step. That is

\[
\beta(s+1) = \beta(s) + \left\{ \sum_{i=1}^{N} X_i^T \Delta_i \left( \beta(s), \sigma^2(s) \right) V_i^{-1} \left( \beta(s), \sigma^2(s), \eta(s) \right) \Delta_i \left( \beta(s), \sigma^2(s) \right) X_i \right\}^{-1} \times \left\{ \sum_{i=1}^{N} X_i^T \Delta_i \left( \beta(s), \sigma^2(s) \right) V_i^{-1} \left( \beta(s), \sigma^2(s), \eta(s) \right) \left\{ W_i - \Phi(\beta(s), \sigma^2(s)) \right\} \right\}
\]

(4)

The algorithm for computing the proposed GEE is summarized in the following steps:

**Step 1.** Choose an initial value \(\beta(0)\) for the regression parameter \(\beta\). Here, we use the estimate from probit regression for binomial GLMM model, which includes the same fixed effects and a random intercept.

**Step 2.** At the \(s\)th iteration, use the current regression parameter estimate \(\beta(s)\) to update the estimate of \(\alpha = (\sigma^2, \eta)\). Given an educated guess of an admissible interval \([\sigma^2_L, \sigma^2_U]\) containing the true value \(\sigma^2\), perform bisection method described in [19] to reduce the length of the admissible interval, which contains the root of Equation (2) with \(\beta(s)\) plugged in for \(\beta\), to be less than \(10^{-5}\). Then a crude search algorithm with increment of \(10^{-7}\) for \(\sigma^2\) within this updated admissible interval is implemented to identify \(\sigma^2(s+1)\) that warrants its accuracy to the solution of Equation (2) being at least \(10^{-7}\). Using the same numerical methods, \(\eta(s+1)\) can be solved from equation (3) with \(\beta(s)\) and \(\sigma^2(s+1)\) plugged into it.

**Step 3.** Update the estimate of \(\beta\) with \(\eta(s+1)\) and \(\sigma^2(s+1)\) by (4) and obtain \(\beta(s+1)\).
Step 4. Check if \( \max |(\beta^{(s+1)}, \sigma^2(s+1), \eta^{(s+1)}^T - (\beta^{(s)}, \sigma^2(s), \eta^{(s)})^T | \leq \epsilon \), where \( \epsilon \) is a small constant. If this condition is satisfied, \( (\beta^{(s+1)}, \sigma^2(s+1), \eta^{(s+1)})^T \) is the final parameter estimate. Otherwise, repeat Step 2 to 4.

Once the convergence criterion is reached and the final estimate of all unknown parameters \( \hat{\beta} = (\hat{\beta}_{GEE}, \hat{\sigma}^2, \hat{\eta})^T \) is obtained, the estimated variance of the estimate \( \hat{\beta}_{GEE} \) is then given by \( \hat{Σ} = \hat{H}_2^{-1} \hat{H}_1 \hat{H}_2^{-1} \) where

\[
\hat{H}_1 = \sum_{i=1}^m X_i^T \Delta_i \left( \hat{\beta}_{GEE}, \hat{\sigma}^2 \right) \hat{V}_i^{-1} \left( \hat{\beta}_{GEE}, \hat{\sigma}^2, \hat{\eta} \right) \left\{ W_i - \Phi(\hat{\beta}_{GEE}, \hat{\sigma}^2) \right\} \left\{ W_i - \Phi(\hat{\beta}_{GEE}, \hat{\sigma}^2) \right\}^T \times \hat{V}_i^{-1} \left( \hat{\beta}_{GEE}, \hat{\sigma}^2, \hat{\eta} \right) \Delta_i \left( \hat{\beta}_{GEE}, \hat{\sigma}^2 \right) X_i
\]

and

\[
\hat{H}_2 = \sum_{i=1}^m X_i^T \Delta_i \left( \hat{\beta}_{GEE}, \hat{\sigma}^2 \right) \hat{V}_i^{-1} \left( \hat{\beta}_{GEE}, \hat{\sigma}^2, \hat{\eta} \right) \Delta_i \left( \hat{\beta}_{GEE}, \hat{\sigma}^2 \right) X_i
\]

3. Simulation

The program written in R to compute the proposed model is available from the first author upon request. To examine the validity of the proposed GEE approach and its performance relative to the GLMM, simulation studies were conducted. The outcome \( Y_{ij} \) either followed the regular beta-binomial distribution or was generated as the sum of heterogeneous binary data that produce the over-dispersion of binomial data. The two types of mechanism of over-dispersion were used here because (1) over-dispersed data are frequently modeled by the beta-binomial distribution in practice, which is used to develop our proposed GEE model; (2) the robustness of the beta-binomial model for over-dispersed binomial data against the underlying constant correlation assumption between the binary components needs to be evaluated in order to promote the use of the proposed method.

For each simulation setting, we generated 1,000 Monte Carlo samples of different sample size \((m = 100 \text{ and } 200)\), and different degree of over-dispersion. In each sample, the data constitute \((Y_i, X_i = (1, \bar{T}_i)) : i = 1, 2, \ldots, m\). For each subject, \( \bar{T}_i \) is a \( n_i \times 1 \) vector of time variables, \( t_{ij}, j = 1, \ldots, n_i \) where \( n_i \) is an integer randomly drawn from a discrete uniform distribution \([1, 10]\). The first element of \( \bar{T}_i \) is 1 and the other elements are simulated by a two-stage procedure. First, \( n_i - 1 \) integers are drawn without replacement from \([2, 10]\). Then \( t_{ij} \) is drawn from the uniform distribution with length 1 and centered at each of these integers, \( j = 2, \ldots, n_i \). The parameter \( \eta \) takes three values: 0.10, 0.25, and 0.50 (one at a time) and the number of trials \( K_{ij} \) is assumed to be the same \((K = 10 \text{ or } 25)\) for all \( i \) and \( j \). Note that these two quantities determine the degree of over-dispersion. The greater these two quantities, the more over-dispersion. The true values for other parameters are set to be: \( (\beta_0, \beta_1)^T = (1.5, -0.05)^T \), and \( \sigma = 0.5 \). Using these values, \( Y_i \) are generated using the aforementioned mechanisms.

Mechanism 1 : The random intercept \( u_i \) is drawn from the normal distribution \( N(0, 0.25) \) which results in \( \mu_{ij} = \Phi(\beta_0 + \beta_1 t_{ij} + u_i) \). The probability for the \( j \)th subject at the \( j \)th time point, \( p_{ij} \), follows the beta distribution \( \text{Beta} \left( \frac{\mu_{ij}(1-\eta)}{\eta}, \frac{1-\mu_{ij}(1-\eta)}{\eta} \right) \). Finally \( Y_{ij} \) is generated according to \( \text{Binomial}(K, p_{ij}) \).

Mechanism 2 : Using a procedure proposed by Ahn and Chen [20], we simulate binary data \( Y_{ijkl}, i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, K \) such that, for any \( 1 < l \leq n_i, k < l, \text{Corr}(Y_{ijkl}, Y_{ijl}) \) is set to be a constant randomly drawn from the uniform distribution with length 0.1 and centered at \( \eta \), i.e. \( \text{Corr}(Y_{ijkl}, Y_{ijl}) \sim U(\eta - 0.05, \eta + 0.05) \). The sum \( \sum_{k=1}^K Y_{ijkl} \) constitutes an over-dispersed binomial outcome, which does not follow a beta-binomial distribution.

In addition to the over-dispersed data, standard binomial data \((\eta = 0)\) are also generated and included in the simulation. The binomial GLMM with the probit link function and the correctly specified subject random intercept was fitted to every simulated dataset, in addition to the proposed GEE model. Note that the only misspecification in the GLMM is the
Table 1. Comparison of estimation results from the GLMM and the proposed GEE model on simulated beta-binomial data

<table>
<thead>
<tr>
<th>η</th>
<th>GLMM</th>
<th>GEE</th>
<th>GLMM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β₀</td>
<td>β₁</td>
<td>σ</td>
<td>η</td>
</tr>
<tr>
<td>K = 10</td>
<td>0.47</td>
<td>0.0705</td>
<td>93.8%</td>
<td>0.80</td>
</tr>
<tr>
<td>m = 100</td>
<td>0.10</td>
<td>0.0857</td>
<td>85.9%</td>
<td>-2.60</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.1105</td>
<td>67.9%</td>
<td>-5.20</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.1524</td>
<td>35.7%</td>
<td>-9.60</td>
</tr>
<tr>
<td></td>
<td>0.12</td>
<td>0.0767</td>
<td>92.6%</td>
<td>1.00</td>
</tr>
<tr>
<td>m = 100</td>
<td>0.10</td>
<td>0.0865</td>
<td>94.3%</td>
<td>-2.40</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.1048</td>
<td>92.9%</td>
<td>0.40</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.1316</td>
<td>93.1%</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>0.42</td>
<td>0.0483</td>
<td>94.1%</td>
<td>-0.60</td>
</tr>
<tr>
<td>m = 200</td>
<td>0.10</td>
<td>0.0561</td>
<td>82.8%</td>
<td>-2.40</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.0779</td>
<td>53.7%</td>
<td>-5.80</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.1093</td>
<td>12.9%</td>
<td>-11.20</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0520</td>
<td>92.7%</td>
<td>-0.40</td>
</tr>
<tr>
<td>m = 200</td>
<td>0.10</td>
<td>0.0618</td>
<td>92.9%</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.0865</td>
<td>82.8%</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.0942</td>
<td>92.3%</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td>0.48</td>
<td>0.0626</td>
<td>92.6%</td>
<td>-0.20</td>
</tr>
<tr>
<td>m = 250</td>
<td>0.10</td>
<td>0.0779</td>
<td>53.7%</td>
<td>-5.80</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.1093</td>
<td>12.9%</td>
<td>-11.20</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.0520</td>
<td>92.7%</td>
<td>-0.40</td>
</tr>
<tr>
<td>m = 250</td>
<td>0.10</td>
<td>0.0618</td>
<td>92.9%</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.0865</td>
<td>82.8%</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.0942</td>
<td>92.3%</td>
<td>1.60</td>
</tr>
</tbody>
</table>

underlying binomial distribution when over-dispersion is indeed present. The GLMM was fitted using the “glmer” function in the R package “lme4” with the default method of Laplace approximation to compute the integral. The algorithm with the default Laplace approximation method worked very well in our simulation studies and converged successfully in every case we studied. In each setting of the simulation, the percent bias (bias %), average standard error (ASE), empirical standard deviation (ESD) of the estimates of β, coverage probability of the 95% Wald confidence intervals of β, and the average estimate of the variance parameter σ were calculated, for both the GLMM and the proposed GEE method. For our proposed GEE method, the average estimate of the correlation coefficient η was also provided. With data generated by Mechanisms 1 and 2, simulation results are summarized in Tables 1 and 2, respectively.

When binomial data are generated from the standard binomial distribution (η = 0 in Table 1), the GLMM is actually...
the ML estimation method and hence, is asymptotically efficient. As can be seen from Table 1, the GLMM method performs very well in terms of unbiasedness of estimation of the regression parameters, standard errors, and coverage probabilities. For this case, our proposed GEE method also gives unbiased inference and its performance is comparable to the GLMM method, particularly for the regression parameter $\beta_1$. However, when binomial data are generated from the beta-binomial distribution ($\eta > 0$ in Table 1), the GLMM method does not work properly. Not only does it underestimate the standard errors by comparing ASE to ESD that yields undercoverage of 95% confidence intervals, it also results in biased estimation of regression parameters. The inference becomes more biased when data are more over-dispersed as induced by $\eta$ increasing from 0.10 to 0.50, or $K$ increasing from 10 to 25. For this case, our proposed GEE method still works very well with unbiased inferences regardless of the extent of over-dispersion. In addition, we note that for the over-dispersed binomial data, the standard GLMM procedure also badly overestimates the variability of the random intercept ($\sigma^2$) and the overestimation worsens as over-dispersion increases; whereas the proposed GEE approach estimates the variability of the random intercept and the over-dispersion parameter $\eta$ unbiasedly.

Table 2. Comparison of estimation results from the GLMM and the proposed GEE model on over-dispersed binomial data generated from binary data with small heterogeneity

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>$\beta_0$ Bias</th>
<th>$\beta_0$ ASE</th>
<th>$\beta_0$ ESD</th>
<th>$\beta_0$ CP</th>
<th>$\beta_1$ Bias</th>
<th>$\beta_1$ ASE</th>
<th>$\beta_1$ ESD</th>
<th>$\beta_1$ CP</th>
<th>$\hat{\sigma}$</th>
<th>$\eta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GLMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 10</td>
<td>0.10</td>
<td>3.26</td>
<td>0.0730</td>
<td>0.0892</td>
<td>83.8%</td>
<td>-3.00</td>
<td>0.0073</td>
<td>0.0109</td>
<td>79.4%</td>
<td>0.5600</td>
</tr>
<tr>
<td>m = 100</td>
<td>0.25</td>
<td>7.61</td>
<td>0.0834</td>
<td>0.1088</td>
<td>67.0%</td>
<td>-5.26</td>
<td>0.0076</td>
<td>0.0138</td>
<td>70.7%</td>
<td>0.6548</td>
</tr>
<tr>
<td>K = 10</td>
<td>0.10</td>
<td>0.23</td>
<td>0.0834</td>
<td>0.0916</td>
<td>92.5%</td>
<td>0.80</td>
<td>0.0102</td>
<td>0.0120</td>
<td>79.4%</td>
<td>0.4952</td>
</tr>
<tr>
<td>m = 100</td>
<td>0.25</td>
<td>0.09</td>
<td>0.0980</td>
<td>0.1064</td>
<td>93.1%</td>
<td>0.60</td>
<td>0.0131</td>
<td>0.0133</td>
<td>95.2%</td>
<td>0.4872</td>
</tr>
<tr>
<td>GEE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 10</td>
<td>0.10</td>
<td>0.50</td>
<td>0.0104</td>
<td>0.1492</td>
<td>36.0%</td>
<td>-10.40</td>
<td>0.0079</td>
<td>0.0188</td>
<td>95.2%</td>
<td>0.8371</td>
</tr>
<tr>
<td>m = 100</td>
<td>0.25</td>
<td>0.50</td>
<td>0.1202</td>
<td>0.1325</td>
<td>92.4%</td>
<td>0.00</td>
<td>0.0170</td>
<td>0.0170</td>
<td>94.2%</td>
<td>0.4906</td>
</tr>
</tbody>
</table>

When the over-dispersed binomial data are generated from binary data with moderate heterogeneity (Mechanism 2), the results are similar as in the beta-binomial data. As expected, Table 2 shows the GLMM underestimates the standard...
errors in estimating the regression parameters, which directly causes the shrinkage of the coverage probability and also potentially inflates Type 1 error in making inference regarding $\beta$. In contrast, the inference based on the proposed GEE method remains asymptotically unbiased with $\hat{\eta}$ estimating the mean of correlations among individual binary outcomes.

In addition to the analyses above, we conducted a simulation study for over-dispersed data generated using Mechanism 2, but with large heterogeneity in correlations among individual binary components. Specifically, we considered two scenarios: the correlation coefficient of any pair of binary components is randomly drawn from $[0.05, 0.45]$ and from $[0.30, 0.70]$. The results are summarized in Table 3. For this case, the proposed GEE method does appear to have noticeable bias in estimating the regression parameters as well as the variance components. For estimating $\beta_0$, the standard errors are also slightly underestimated and the bias becomes more noticeable when $K$ increases from 10 to 25, which contributes to the undercoverage of the confidence intervals. However the standard errors for estimating the more practically meaningful regression parameter, $\beta_1$, are still estimated unbiasedly, which results in the confidence intervals having a coverage probability around the nominal value.

In summary, the simulation studies provide evidence that the proposed beta-binomial based GEE method has a robustness property for the underlying distribution with longitudinal over-dispersed binomial data. The robustness pertains to inferences of the regression parameters that are related to the effects of covariates on the binomial outcome, particularly if the heterogeneity of the correlation coefficients is not substantial. This evidence strengthens applicability of this proposed GEE method in analyzing longitudinal binomial data subject to potential over-dispersion.

Table 3. Comparison of estimation results from the GLMM and the proposed GEE model on over-dispersed binomial data generated from binary data with large heterogeneity

<table>
<thead>
<tr>
<th>Corr.Coeff.</th>
<th>$\hat{\beta}_0$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\sigma}$</th>
<th>$\hat{\eta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bias %</td>
<td>ASE</td>
<td>ESD</td>
<td>CP</td>
</tr>
<tr>
<td>GLMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLMM</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GEE</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K = 25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>m = 200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prepared using simauth.cls
4. Application to the PREDICT-HD Study

The PREDICT-HD study [21] is an international multi-site longitudinal observational study of prodromal HD. One of the study goals is to assess clinical markers that are associated with the loss of daily functioning for prodromal HD individuals who are genetically at-risk for HD, but who have not yet received a diagnosis. An HD diagnosis is made when a trained examiner (e.g., a neurologist) indicates in the UHDRS Diagnostic Confidence Level that they are at least 99% confident the examinee is presenting unequivocal motor signs of HD, based on the standard motor examination. Individuals not yet diagnosed are referred to as prodromal because they could be displaying “soft” motor signs. In this study, 1,021 gene-expanded prodromal participants were evaluated annually with the 25-item yes/no FAS questionnaire. The FAS purports to measure everyday function, with the total score computed as the count of “yes” responses. There were on average 4.53 visits per individual with a total number of 4,628 observations. Some individuals had intermittent missing values that can be reasonably assumed missing completely at random (MCAR) as the missing values were due to administrative reasons (e.g., scheduling conflicts). Initial analysis showed that answers to 20 out of the 25 FAS questions had no variability (all “no” response). Therefore, only the five questions with possible “yes” responses were included for the analysis. At the baseline, the pairwise phi coefficient [22] for the five binary outcomes ranged from 0.1773 to 0.5857 showing good reproducibility of the five outcomes and hence, suggesting the presence of over-dispersion in the binomial FAS data.

The clinical markers other than the FAS in PREDICT-HD include measures of motor abnormality, cognition, psychiatric symptoms and brain imaging. An important research question is whether FAS is related to these other clinical measures, especially the cognitive and motor variables. To address this question, the following variables were selected based on previous research [23, 24]. The cognitive domain was represented by SYDIGTOT, which is a score from the symbol digit modalities test (SDMT) that measures the number of correct responses on a timed task of symbol to digit transcription; STROOPCO and STROOPWO are two scores from the Stroop Color and Word Test [25] that measure the number of correct matches of color and word, respectively. Representing the motor domain was NEUROTOT, which is the total motor score of multiple individual motor signs (each rating 0 = normal to 4 = greatest impairment) based on the standard motor examine (NEUROTOT may influence the diagnosis decision but it does not define diagnosis). Control variables included in the analysis were sex and CAG-Age product (CAP), the latter being a commonly used index of the cumulative toxicity of the mutant protein characteristic of HD [26, 27].

For the data analysis, both the binomial GLMM with the probit link and the proposed GEE model were fitted to the FAS data. The results from both models were summarized in Table 4. The correlation coefficient estimate from the GEE model was \( \hat{\eta} = 0.1474 \). As shown in Table 4, the fixed effects estimates for the four clinical markers were similar in both models. However, the standard errors of the estimated regression parameters were consistently larger in the GEE approach compared to the GLMM. Given the estimated correlation for the data (\( \hat{\eta} = 0.1474 \)), the results are perhaps not surprising in light of what was found for the over-dispersed longitudinal binomial data in the simulation study. Interestingly, inference for the CAP control variable was different between the two models. For the GEE method, CAP was statistically significant at the 0.05 level with a negative coefficient indicating prodromal HD subjects with more genetic toxicity had less probability to maintain unimpaired daily functioning. CAP is a very important variable in HD research because it indexes both the genetic loading of the disease and the length of exposure to the toxic protein. The finding regarding CAP is consistent with many other data analysis results in PREDICT-HD and other studies [23, 24]. The GLMM results showed an insignificant positive coefficient. The inferential discrepancy between the two models may be explained by the finding from the simulation studies that the GLMM generally yields a biased inference about regression parameters for over-dispersed binomial data.
Table 4. The inference from the GLMM and the proposed GEE model for the PREDICT-HD FAS data

<table>
<thead>
<tr>
<th>Effect</th>
<th>GLMM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.4489</td>
<td>0.3679</td>
</tr>
<tr>
<td>SEX</td>
<td>0.2546</td>
<td>0.1134</td>
</tr>
<tr>
<td>CAP</td>
<td>0.0005</td>
<td>0.0007</td>
</tr>
<tr>
<td>NEUROTOT</td>
<td>-0.0438</td>
<td>0.0028</td>
</tr>
<tr>
<td>SYDIGTOT</td>
<td>0.0244</td>
<td>0.0039</td>
</tr>
<tr>
<td>STROOPCO</td>
<td>0.0120</td>
<td>0.0033</td>
</tr>
<tr>
<td>STROOPWO</td>
<td>0.0111</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

5. Final Remarks

Longitudinal over-dispersed binomial data can arise in numerous applications where binary components are correlated. The GLMM is widely used as a default method for modeling binomial data because it is a natural extension of the generalized linear model and various statistical software are available for computation. However, the GLMM does not account for the over-dispersion that often exists in longitudinal binomial data. We showed the GLMM tends to underestimate the standard errors of regression parameter estimates and it may also result in biased estimation of regression parameters with the presence of over-dispersion. Inflation of Type I error for inference is a direct consequence of the biased estimation of those quantities. Therefore, a special treatment must be considered in practice when over-dispersion is a potential factor for longitudinal binomial data.

In this article, we developed a beta-binomial based GEE model to analyze longitudinal binomial data that accounts for both the over-dispersion in binomial data and correlations among the repeatedly measured binomial data. Our proposed beta-binomial GEE model requires the sum of parameters in the Beta distribution to be constant, which is indicative of constant correlation among the binary components and is overly restrictive in view of practical applications. We conducted extensive simulation studies to demonstrate that the proposed GEE method is valid and provides unbiased inference for the regression parameters if the correlations among the binary components do not vary too much. This result implies a desired inference property of robustness for the proposed beta-binomial GEE model in analyzing longitudinal over-dispersed binomial data. Hence the proposed method is expected to have broad practical application.

In this article, only the random intercept was included in the linear predictor for ease of model presentation. A more general approach should include random slopes with additional effort to deal with computation complexity. An ML approach has been theoretically proposed [14] and it will be of interest to implement the ML estimator and compare its inference with the proposed GEE model. Besides probit regression, one could investigate other types of regression models such as the traditional logistic regression model. However, it is anticipated that the logistic regression model will require more computing effort than the probit regression with the proposed GEE method.

Acknowledgement

Drs Zhang and Long’s research is partly supported by NIH grant 2 UL1 TR000442-06. The authors are grateful to the associate editor and two anonymous reviewers for their insightful comments which largely help improve the presentation of this work.
Appendix A  Derivation of the variance matrix $V_i$

First, the mean of $W_{ij}$ is given by

$$E(W_{ij}) = \frac{1}{K_{ij}} E \{ E(Y_{ij}|u_i) \} = \frac{K_{ij} E(\mu_{ij})}{K_{ij}} = E(\Phi(u_i + X_{ij}^T \beta)) = \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right)$$

Then the variance of $W_{ij}$ can be derived.

$$\text{Var}(W_{ij}) = E \{ \text{Var}(W_{ij}|u_i) \} + \text{Var} \{ E(W_{ij}|u_i) \}$$

$$= \frac{1}{K_{ij}^2} \left[ E \{ \text{Var}(Y_{ij}|u_i) \} + \text{Var} \{ E(Y_{ij}|u_i) \} \right]$$

$$= \frac{1}{K_{ij}^2} \left[ (K_{ij} + \eta K_{ij}(K_{ij} - 1)) E \{ \mu_{ij} (1 - \mu_{ij}) \} + K_{ij}^2 \{ E \mu_{ij}^2 - (E \mu_{ij})^2 \} \right]$$

$$= \frac{1}{K_{ij}^2} \left[ (K_{ij} + \eta K_{ij}(K_{ij} - 1)) E \mu_{ij} + K_{ij}(K_{ij} - 1)(1 - \eta) E \mu_{ij}^2 - K_{ij}^2 (E \mu_{ij})^2 \right]$$

$$= \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \left( 1 - \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \right) - \Phi^2 (u_i + X_{ij}^T \beta)$$

Also,

$$\text{cov}(W_{ij}, W_{ik}) = E(W_{ij}W_{ik}) - E(W_{ij})E(W_{ik})$$

$$= E \{ E(W_{ij}W_{ik}|u_i) \} - E(W_{ij})E(W_{ik})$$

$$= \frac{1}{K_{ij} K_{ik}} E \{ E(Y_{ij}|u_i)E(Y_{ik}|u_i) \} - E(W_{ij})E(W_{ik})$$

$$= E \Phi(u_i + X_{ij}^T \beta) \Phi(u_i + X_{ik}^T \beta) - \Phi \left( \frac{X_{ij}^T \beta}{\sqrt{1 + \sigma^2}} \right) \Phi \left( \frac{X_{ik}^T \beta}{\sqrt{1 + \sigma^2}} \right)$$

References