Research Article

Statistics in Medicine

Received XXXX

(www.interscience.wiley.com) DOI: 10.1002/sim.0000

Longitudinal Beta-Binomial Modeling using GEE for Over-Dispersed Binomial Data

Hongqian Wu[a*†], Ying Zhang[b], Jeffrey D.Long[a,c]

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

[a] *Department of Biostatistics, University of Iowa, Iowa City, Iowa, U.S.A.*

[b] *Department of Biostatistics, Indiana University, Indianapolis, Indiana, U.S.A. and Department of Mathematics, Shanghai Jiao Tong University, Shanghai, China*

[c] *Departments of Psychiatry, University of Iowa,Iowa City, Iowa, U.S.A.*

[*] *Correspondence to: Hongqian Wu, Department of Biostatistics, University of Iowa, Iowa City, Iowa, U.S.A.*

[†] E-mail: hongqian-wu@uiowa.edu

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 ith subject ($i = 1, \ldots, m$), at the jth time point ($j = 1, \ldots, n_i$), we observe a binomial response Y_{ij} and a p-dimensional vector of covariates X_{ij}. Let $Y_i = (Y_{i1}, \ldots, Y_{it}, \ldots, Y_{in_i})^T$ be the $n_i \times 1$ response vector and $X_i = (X_{i1}, \ldots, X_{it}, \ldots, X_{in_i})^T$ be the $n_i \times p$ design matrix which includes a column of 1's, time-dependent, and time-independent covariates for the ith 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 ith subject and Z_{ij} the $q \times 1$ covariate vector for these random effects at the jth 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,

$$Y_{ij}|p_{ij}, u_i \sim Binomial(K_{ij}, p_{ij})$$

$$p_{ij}|u_i \sim Beta(a_{ij}, b_{ij})$$

where K_{ij} is the total number of trials for the ith subject at the jth 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

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

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

and

$$Corr(Y_{ijk_1}, Y_{ijk_2}|p_{ij}, u_i) = \frac{\tau}{1+\tau} \quad \text{for} \quad k_1 \neq k_2$$

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 $Var(Y_{ij}|u_i)$) 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 ith subject at the jth time. It follows that $E(W_{ij}|u_i) = \frac{1}{K_{ij}}E(Y_{ij}|u_i) = \mu_{ij}$ and $Var(W_{ij}|u_i) = \left\{\frac{1}{K_{ij}} + \eta(1 - \frac{1}{K_{ij}})\right\}\mu_{ij}(1-\mu_{ij})$. Let $W_i = (W_{i1}, \ldots, W_{it}, \ldots, W_{in_i})^T$ and $\mu_i = (E(W_{i1}), \ldots, E(W_{it}), \ldots, E(W_{in_i}))^T$. Denote $g(\cdot) : (0,1) \to \mathbb{R}$ as a link function, which will be used to model the marginal mean of the outcome variable W_{ij}. Then

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

where β 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 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 $D_i=\dfrac{\partial\mu_i}{\partial\beta}$, μ_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 α. In practice, the set of α parameters is usually replaced by its $\sqrt{m}$-consistent estimator $\hat{\alpha}$, given β. 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 $H=\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}D_i^T V_i^{-1}D_i$ and $\Omega=\lim_{m\to\infty}\frac{1}{m}\sum_{i=1}^{m}D_i^T V_i^{-1}Var(Y_i)V_i^{-1}D_i$.

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

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

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

$$(V_i)_{jk}=\begin{cases}\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\left\{1-\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\right\}-\frac{(K_{ij}-1)(1-\eta)}{K_{ij}}\left\{\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)-A_{ijj}\right\} & \text{if } j=k\\ A_{ijk}-\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\Phi\left(\frac{X_{ik}^T\beta}{\sqrt{1+\sigma^2}}\right) & \text{if } j\neq k\end{cases}$$

where $(V_i)_{jk}$ refers to the component in the jth row and kth column of the variance matrix V_i, u_i is the random intercept, assumed to follow a normal distribution $N(0,\sigma^2)$, and $A_{ijk}=E\left\{\Phi(X_{ij}^T\beta+u_i)\Phi(X_{ik}^T\beta+u_i)\right\}$. The variance σ^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 β with this working covariance matrix. Specifically, the GEE for this setting is given by

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

where X_i, V_i and W_i are as defined above,

$$\Delta_i(\beta,\sigma^2)=\frac{1}{\sqrt{1+\sigma^2}}Diag\left(\phi\left(\frac{X_{i1}^T\beta}{\sqrt{1+\sigma^2}}\right),\ldots,\phi\left(\frac{X_{in_i}^T\beta}{\sqrt{1+\sigma^2}}\right)\right)$$

and

$$\mu_i(\beta,\sigma^2) = \left(\Phi\left(\frac{X_{i1}^T\beta}{\sqrt{1+\sigma^2}}\right),\ldots,\Phi\left(\frac{X_{in_i}^T\beta}{\sqrt{1+\sigma^2}}\right)\right)^T$$

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 (σ^2,η), are given by

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

and

$$\frac{1}{\sum_{i=1}^{m} n_i} \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{\left\{W_{ij}-\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\right\}^2}{\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\left\{1-\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\right\}-\frac{(K_{ij}-1)(1-\eta)}{K_{ij}}\left\{\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)-A_{ijj}\right\}} = 1, \qquad (3)$$

respectively. In these equations, although $A_{ijk}=E\left\{\Phi(X_{ij}^T\beta+u_i)\Phi(X_{ik}^T\beta+u_i)\right\}$ cannot be explicitly evaluated, it can be easily approximated by Gauess-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 α.

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, β. 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

$$\beta^{(s+1)} = \beta^{(s)} + \left\{\sum_{i=1}^{N} X_i^T\Delta_i\left(\beta^{(s)},\sigma^{2(s)}\right)V_i^{-1}\left(\beta^{(s)},\sigma^{2(s)},\eta^{(s)}\right)\Delta_i\left(\beta^{(s)},\sigma^{2(s)}\right)X_i\right\}^{-1}\times$$
$$\left[\sum_{i=1}^{N} X_i^T\Delta_i\left(\beta^{(s)},\sigma^{2(s)}\right)V_i^{-1}\left(\beta^{(s)},\sigma^{2(s)},\eta^{(s)}\right)\left\{W_i-\Phi(\beta^{(s)},\sigma^{2(s)})\right\}\right] \qquad (4)$$

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 β. 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 sth 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 $[\sigma_L^2,\sigma_U^2]$ containing the true value σ^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 β, to be less than 10^{-5}. Then a crude search algorithm with increment of 10^{-7} for σ^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 β with $\eta^{(s+1)}$ and $\sigma^{2(s+1)}$ by (4) and obtain $\beta^{(s+1)}$

Step 4. Check if $max|\left(\beta^{(s+1)},\sigma^{2(s+1)},\eta^{(s+1)}\right)^T - \left(\beta^{(s)},\sigma^{2(s)},\eta^{(s)}\right)^T| \le \epsilon$, where ϵ is a small constant. If this condition is satisfied, $\left(\beta^{(s+1)},\sigma^{2(s+1)},\eta^{(s+1)}\right)^T$ 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 $\hat{\theta} = (\hat{\beta}_{GEE}, \hat{\sigma}^2, \hat{\eta})^T$ is obtained, the estimated variance of the estimate $\hat{\beta}_{GEE}$ is then given by $\hat{\Sigma}_\beta = \hat{H}_2^{-1}\hat{H}_1\hat{H}_2^{-1}$ where

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

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 exam 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 and 200), and different degree of over-dispersion. In each sample, the data constitute $\left(\underline{Y_i}, X_i = (\underline{1}, \underline{T_i})\right) : i = 1, 2, \ldots, m$. For each subject, $\underline{T_i}$ 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 $\underline{T_i}$ 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 η 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 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, $\underline{Y_i}$ 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 ith subject at the jth time point, p_{ij}, follows the beta distribution $Beta\left(\frac{\mu_{ij}(1-\eta)}{\eta}, \frac{(1-\mu_{ij})(1-\eta)}{\eta}\right)$. Finally Y_{ij} is generated according to $Binomial(K, p_{ij})$.

Mechanism 2 : Using a procedure proposed by Ahn and Chen [20], we simulate binary data $Y_{ijk}, i = 1, \ldots, m, j = 1, \ldots, n_i, k = 1, \ldots, K$ such that, for any $1 < l \le n_i,\ k < l$, $Corr(Y_{ijk}, Y_{ijl})$ is set to be a constant randomly drawn from the uniform distribution with length 0.1 and centered at η, i.e. $Corr(Y_{ijk}, Y_{ijl}) \sim \mathcal{U}(\eta - 0.05,\ \eta + 0.05)$. The sum $\sum_{k=1}^{K} Y_{ijk}$ 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

	η	$\hat{\beta}_0$				$\hat{\beta}_1$				$\hat{\sigma}$	$\hat{\eta}$
		Bias %	ASE	ESD	CP	Bias %	ASE	ESD	CP		
						GLMM					
K = 10	0	0.47	0.0685	0.0705	93.8%	0.80	0.0073	0.0077	94.1%	0.4931	-
m = 100	0.10	3.21	0.0733	0.0857	85.9%	-2.60	0.0073	0.0106	81.2%	0.5576	-
	0.25	7.67	0.0839	0.1105	67.9%	-5.20	0.0076	0.0138	71.0%	0.6591	-
	0.50	17.49	0.1062	0.1524	35.7%	-9.60	0.0080	0.0192	54.2%	0.8451	-
						GEE					
K = 10	0	0.12	0.0709	0.0767	92.6%	1.00	0.0074	0.0077	93.9%	0.4946	-0.0004
m = 100	0.10	0.12	0.0831	0.0865	94.3%	0.00	0.0101	0.0103	94.4%	0.4918	0.1000
	0.25	0.13	0.0989	0.1048	92.9%	0.40	0.0132	0.0132	94.3%	0.4935	0.2494
	0.50	0.16	0.1201	0.1316	93.1%	1.60	0.0171	0.0176	94.0%	0.4872	0.4979
						GLMM					
K = 10	0	0.42	0.0489	0.0483	94.1%	-0.60	0.0052	0.0052	95.8%	0.4939	-
m = 200	0.10	3.30	0.0535	0.0618	82.8%	-2.40	0.0053	0.0074	83.7%	0.5628	-
	0.25	7.68	0.0605	0.0779	53.7%	-5.80	0.0054	0.0100	70.0%	0.6615	-
	0.50	17.28	0.0751	0.1093	12.9%	-11.20	0.0056	0.0141	54.1%	0.8474	-
						GEE					
K = 10	0	-0.06	0.0500	0.0520	93.7%	-0.40	0.0052	0.0052	95.7%	0.4928	-0.0001
m = 200	0.10	0.21	0.0591	0.0631	92.9%	0.00	0.0072	0.0073	94.0%	0.4979	0.0996
	0.25	0.20	0.0702	0.0767	92.2%	-0.20	0.0094	0.0097	94.8%	0.4989	0.2491
	0.50	0.03	0.0854	0.0942	92.3%	0.40	0.0121	0.0127	94.3%	0.4940	0.5004
						GLMM					
K = 25	0	0.48	0.0582	0.0626	92.6%	-0.20	0.0047	0.0049	94.1%	0.4963	-
m = 100	0.10	3.41	0.0659	0.0780	84.5%	-1.60	0.0048	0.0087	71.4%	0.5786	-
	0.25	8.81	0.0659	0.1091	56.9%	-5.20	0.0042	0.0131	47.1%	0.7005	-
	0.50	21.15	0.1006	0.1525	20.5%	-12.60	0.0051	0.0184	41.1%	0.9280	-
						GEE					
K = 25	0	0.19	0.0612	0.0670	93.5%	0.20	0.0048	0.0049	93.8%	0.4927	-0.0003
m = 100	0.10	-0.06	0.0760	0.0808	93.5%	1.00	0.0086	0.0086	94.6%	0.4930	0.0995
	0.25	-0.05	0.0938	0.1046	93.4%	0.60	0.0123	0.0126	94.2%	0.4923	0.2494
	0.50	0.23	0.1172	0.1258	93.6%	-0.20	0.0165	0.0167	94.0%	0.4884	0.4981
						GLMM					
K = 25	0	0.27	0.0411	0.0413	94.9%	-0.00	0.0033	0.0033	94.6%	0.4966	-
m = 200	0.10	3.59	0.0470	0.0564	74.0%	-2.80	0.0034	0.0060	72.1%	0.5807	-
	0.25	8.70	0.0550	0.0743	36.5%	-5.40	0.0034	0.0092	52.0%	0.7019	-
	0.50	20.69	0.0710	0.1068	4.8%	-12.60	0.0036	0.0134	37.6%	0.9264	-
						GEE					
K = 25	0	0.11	0.0436	0.0466	94.3%	0.00	0.0034	0.0034	94.1%	0.4969	-0.0001
m = 200	0.10	0.22	0.0543	0.0596	93.1%	-0.40	0.0061	0.0060	95.6%	0.4991	0.0998
	0.25	-0.10	0.0666	0.0714	93.1%	0.20	0.0087	0.0089	93.8%	0.4968	0.2499
	0.50	-0.04	0.0833	0.0905	92.9%	-0.20	0.0118	0.0121	94.6%	0.4928	0.4999

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 ($\eta = 0$ 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 β_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 η 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 (σ^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 η 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

	η	$\hat{\beta}_0$				$\hat{\beta}_1$				$\hat{\sigma}$	$\hat{\eta}$
		Bias %	ASE	ESD	CP	Bias %	ASE	ESD	CP		
						GLMM					
K = 10	0.10	3.26	0.0730	0.0892	83.8%	-3.00	0.0073	0.0109	79.4%	0.5600	-
m = 100	0.25	7.61	0.0834	0.1088	67.0%	-5.26	0.0076	0.0138	70.7%	0.6548	-
	0.50	16.84	0.1044	0.1492	36.0%	-10.40	0.0079	0.0188	59.2%	0.8371	-
						GEE					
K = 10	0.10	0.23	0.0834	0.0916	92.5%	-0.60	0.0102	0.0107	93.7%	0.4952	0.0997
m = 100	0.25	0.09	0.0980	0.1064	93.1%	0.60	0.0131	0.0133	95.2%	0.4872	0.2484
	0.50	0.05	0.1202	0.1325	92.4%	1.00	0.0170	0.0170	94.2%	0.4906	0.4993
						GLMM					
K = 10	0.10	3.14	0.0523	0.0604	82.9%	-2.20	0.0053	0.0072	84.4%	0.5592	-
m = 200	0.25	7.61	0.0584	0.0741	50.7%	-5.80	0.0054	0.0097	71.8%	0.6549	-
	0.50	16.71	0.0744	0.1033	13.7%	-10.00	0.0056	0.0132	56.6%	0.8395	-
						GEE					
K = 10	0.10	0.08	0.0591	0.0625	93.7%	0.20	0.0072	0.0072	94.6%	0.4951	0.0996
m = 200	0.25	0.19	0.0700	0.0729	93.6%	-0.00	0.0094	0.0093	95.3%	0.4927	0.2493
	0.50	0.04	0.0854	0.0915	93.3%	1.00	0.0121	0.0121	94.5%	0.4985	0.4989
						GLMM					
K = 25	0.10	3.07	0.0644	0.0765	85.3%	-1.20	0.0047	0.0086	70.3%	0.5723	-
m = 100	0.25	8.71	0.0668	0.1056	58.0%	-6.20	0.0043	0.0133	45.6%	0.6894	-
	0.50	20.47	0.0624	0.1492	18.0%	-12.00	0.0033	0.0187	26.8%	0.9080	-
						GEE					
K = 25	0.10	-0.19	0.0759	0.0786	93.6%	1.00	0.0086	0.0085	94.1%	0.4928	0.0987
m = 100	0.25	0.21	0.0939	0.1016	93.3%	-0.20	0.0123	0.0128	92.8%	0.4904	0.2493
	0.50	0.63	0.1177	0.1270	92.4%	-0.60	0.0166	0.0170	94.1%	0.4916	0.4979
						GLMM					
K = 25	0.10	3.39	0.0464	0.0554	77.1%	-2.60	0.0034	0.0063	68.8%	0.5761	-
m = 200	0.25	8.57	0.0482	0.0728	38.4%	-5.40	0.0031	0.0092	47.7%	0.6904	-
	0.50	20.15	0.0292	0.1007	3.7%	-12.80	0.0016	0.0132	18.0%	0.9128	-
						GEE					
K = 25	0.10	0.12	0.0540	0.0584	93.3%	-0.20	0.0061	0.0062	94.1%	0.4970	0.0995
m = 200	0.25	0.13	0.0667	0.0710	93.1%	0.40	0.0087	0.0089	94.2%	0.4955	0.2496
	0.50	0.24	0.0837	0.0863	93.9%	-0.80	0.0118	0.0119	94.9%	0.4970	0.5000

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 β. 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 β_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, β_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

	Corr.Coef.	$\hat{\beta}_0$				$\hat{\beta}_1$				$\hat{\sigma}$	$\hat{\eta}$
		Bias %	ASE	ESD	CP	Bias %	ASE	ESD	CP		
					GLMM						
K = 10	[0.05,0.45]	5.22	0.0796	0.1033	74.8%	-2.40	0.0073	0.0136	69.2%	0.6373	-
m = 100	[0.30,0.70]	13.48	0.0995	0.1445	47.7%	-8.60	0.0077	0.0184	58.7%	0.8075	-
					GEE						
K = 10	[0.05,0.45]	-1.57	0.0961	0.1032	92.1%	2.60	0.0129	0.0131	94.0%	0.4822	0.2392
m = 100	[0.30,0.70]	-1.60	0.1158	0.1279	90.8%	1.80	0.0164	0.0169	93.8%	0.4861	0.4633
					GLMM						
K = 10	[0.05,0.45]	5.11	0.0583	0.0723	70.2%	-3.00	0.0054	0.0095	71.9%	0.6363	-
m = 200	[0.30,0.70]	13.25	0.0711	0.1034	27.2%	-8.00	0.0055	0.0132	58.0%	0.8036	-
					GEE						
K = 10	[0.05,0.45]	-1.65	0.0682	0.0694	92.8%	2.20	0.0092	0.0092	94.1%	0.4841	0.2391
m = 200	[0.30,0.70]	-1.76	0.0820	0.0897	91.5%	2.20	0.0116	0.0122	93.0%	0.4856	0.4650
					GLMM						
K = 25	[0.05,0.45]	-0.41	0.0685	0.0947	84.0%	3.00	0.0046	0.0118	57.8%	0.6150	-
m = 100	[0.30,0.70]	10.05	0.0734	0.1278	54.1%	-4.80	0.0042	0.0160	37.9%	0.7989	-
					GEE						
K = 25	[0.05,0.45]	-6.30	0.0845	0.0936	75.5%	7.00	0.0111	0.0113	92.5%	0.4567	0.2118
m = 100	[0.30,0.70]	-4.29	0.1066	0.1115	88.5%	4.80	0.0149	0.0147	94.7%	0.4662	0.4168
					GLMM						
K = 25	[0.05,0.45]	-0.55	0.0488	0.0629	86.5%	3.00	0.0033	0.0082	57.9%	0.6172	-
m = 200	[0.30,0.70]	9.77	0.0528	0.0880	39.6%	-5.40	0.0030	0.0114	35.9%	0.8053	-
					GEE						
K = 25	[0.05,0.45]	-6.41	0.0599	0.0618	61.2%	7.00	0.0078	0.0080	91.1%	0.4646	0.2118
m = 200	[0.30,0.70]	-4.47	0.0759	0.0796	81.7%	3.60	0.0106	0.0105	95.1%	0.4773	0.4180

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

	GLMM			GEE		
Effect	Estimate	SE	p-value	Estimate	SE	p-value
Intercept	-0.4489	0.3679	0.2225	0.4909	0.4681	0.2943
SEX	0.2546	0.1134	0.0248	0.2070	0.1126	0.0659
CAP	0.0005	0.0007	0.4660	-0.0016	0.0007	0.0125
NEUROTOT	-0.0438	0.0028	$< 10^{-6}$	-0.0427	0.0039	$< 10^{-6}$
SYDIGTOT	0.0244	0.0039	$< 10^{-6}$	0.0196	0.0043	$5{\cdot}10^{-6}$
STROOPCO	0.0120	0.0033	0.0003	0.0128	0.0044	0.0036
STROOPWO	0.0111	0.0027	0.0001	0.0100	0.0033	0.0024

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

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

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

$$\begin{aligned}
Var(W_{ij}) &= E\left\{Var(W_{ij}|u_i)\right\} + Var\left\{E(W_{ij}|u_i)\right\} \\
&= \frac{1}{K_{ij}^2}\left[E\left\{Var(Y_{ij}|u_i)\right\} + Var\left\{E(Y_{ij}|u_i)\right\}\right] \\
&= \frac{1}{K_{ij}^2}\left[\left\{K_{ij} + \eta K_{ij}(K_{ij}-1)\right\} E\left\{\mu_{ij}(1-\mu_{ij})\right\} + K_{ij}^2\left\{E\mu_{ij}^2 - (E\mu_{ij})^2\right\}\right] \\
&= \frac{1}{K_{ij}^2}\left[\left\{K_{ij} + \eta K_{ij}(K_{ij}-1)\right\} E\mu_{ij} + K_{ij}(K_{ij}-1)(1-\eta)E\mu_{ij}^2 - K_{ij}^2(E\mu_{ij})^2\right] \\
&= \Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\left\{1 - \Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\right\} - \\
&\frac{(K_{ij}-1)(1-\eta)}{K_{ij}}\left\{\Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right) - E\Phi^2\left(u_i + X_{ij}^T\beta\right)\right\}
\end{aligned}$$

Also,

$$\begin{aligned}
cov(W_{ij}, W_{ik}) &= E(W_{ij}W_{ik}) - E(W_{ij})E(W_{ik}) \\
&= E\left\{E(W_{ij}W_{ik}|u_i)\right\} - E(W_{ij})E(W_{ik}) \\
&= \frac{1}{K_{ij}K_{ik}} E\left\{E(Y_{ij}|u_i)E(Y_{ik}|u_i)\right\} - E(W_{ij})E(W_{ik}) \\
&= E\left\{\Phi(u_i + X_{ij}^T\beta)\Phi(u_i + X_{ik}^T\beta)\right\} - \Phi\left(\frac{X_{ij}^T\beta}{\sqrt{1+\sigma^2}}\right)\Phi\left(\frac{X_{ik}^T\beta}{\sqrt{1+\sigma^2}}\right)
\end{aligned}$$

References

1. Paulsen J, Long J, Ross C, Harrington D, Erwin C, Williams J, Westervelt H, Johnson H, Aylward E, Zhang Y, *et al.*. Prediction of manifest Huntington's disease with clinical and imaging measures: a prospective observational study. *Lancet Neurology* 2014; **13**(12):1193–1201.
2. Huntington Study Group. Unified huntington's disease rating scale: Reliability and consistency. *Movement Disorders* 1996; **11**:136–142.
3. Demidenko E. *Mixed Models: Theory and Applications*. John Wiley & Sons, Inc.: Hoboken, NJ, 2005.
4. Skellam J. A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets for trials. *Journal of the Royal Statistical Society. Series B* 1948; **10**(2):257–261.
5. Efron B. Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association* 1986; **81**(395):709–721.
6. Altham P. Two generalisations of the binomial distribution. *Journal of the Royal Statistical Society. Series C* 1978; **27**(2):162–167.
7. Griffiths D. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics* 1973; **29**(4):736–648.
8. Williams D. The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* 1975; **31**(4):949–952.
9. Crowder M. Beta-binomial anova for proportions. *Journal of the Royal Statistical Society. Series C* 1978; **27**(1):34–37.

10. Williams D. Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C* 1982; **31**(2):144–148.
11. Nelder J, Pregibon D. An extended quasi-likelihood function. *Biometrika* 1987; **74**(2):221–232.
12. Brooks R. Approximate likelihood ratio tests in the analysis of beta-binomial data. *Journal of the Royal Statistical Society. Series C* 1984; **33**(3):285–289.
13. Carroll R, Ruppert D. *Transformation and Weighting in Regression*. Chapman & Hall: London, 1988.
14. Molenberghs G, Verbeke G, Demtrio C, Vieira A. A family of generalized linear models for repeated measures with normal and conjugate random effects. *Statistical Science* 2010; **25**(3):325–347.
15. Kassahun W, Neyens T, Molenberghs G, Faes C, Verbeke G. Modeling overdispersed longitudinal binary data using a combined beta and normal random-effects model. *Archives of Public Health* 2012; **70**(7):1–13.
16. Liang K, Zeger S. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**(1):13–22.
17. Zeger S. A regression model for time series of counts. *Biometrika* 1988; **75**(4):621–629.
18. Hua L, Zhang Y, Tu W. A spline-based semiparametric sieve likelihood method for over-dispersed panel count data. *The Canadian Journal of Statistics* 2014; **42**(2):217–245.
19. Burden R, Faires J. *Numerical Analysis*. Available Titles CengageNOW Series, Cengage Learning, 2004.
20. Ahn H, Chen J. Generation of over-dispersed and under-dispersed binomial variates. *Journal of Computational and Graphical Statistics* 1994; **4**(1):55–64.
21. Tabrizi SJ, Reilmann R, Roos RAC, Dur A, Leavitt B, Owen G, Jones R, Johnson H, Crauford D, Hicks SL, *et al.*. Potential endpoints for clinical trials in premanifest and early Huntington’s disease in the TRACK-HD study: analysis of 24 month observational data. *The Lancet Neurology* 2012; **11**(1):42–53.
22. Cramér H. *Mathematical Methods of Statistics*. Princeton University Press: Princeton, NJ, 1946.
23. Paulsen J, Long J, Johnson H, Aylward E, Ross C, Williams J, Nance M, Erwin C, Westervelt H, Harrington D, *et al.*. Clinical and biomarker changes in premanifest Huntington disease show trial feasibility: a decade of the PREDICT-HD study. *Frontiers in Aging Neuroscience* 2014; **6**(78):1–11.
24. Paulsen J, Long J, Ross C, Harrington D, Erwin C, Williams J, Westervelt H, Johnson H, Aylward E, Zhang Y, *et al.*. Prediction of manifest Huntington’s disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology* 2014; **13**(12):1193–1201.
25. Golden C. *Stroop color and word test*. Stoelting Company: Illinois, 1978.
26. Zhang Y, Long J, Mills J, Warner J, Lu W, Paulsen J, Researchers of the PREDICT-HD Huntington’s Study Group. Indexing disease progression at study entry with individuals at-risk for Huntington disease. *American Journal of Medical Genetics Part B-Neuropsychiatric Genetics* 2011; **156**(7):751–763.
27. Long J, Paulsen J, Marder K, Zhang Y, Kim J, Mills J, Researchers of the PREDICT-HD Huntington’s Study Group. Tracking motor impairments in the progression of Huntington’s disease. *Movement Disorders* 2014; **29**(3):311–319.