Modeling Correlated Binary Outcomes with Time-Dependent Covariates

Trent L. Lalonde[1*], Anh Q. Nguyen[2], Jianqiong Yin[3],
Kyle Irimata[3] and Jeffrey R. Wilson[3]
[1] *University of Northern Colorado*, [2] *PayPal Seller Risk Decision Analytics and* [3] *Arizona State University*

Abstract: We group approaches to modeling correlated binary data according to data recorded cross-sectionally as opposed to data recorded longitudinally; according to models that are population-averaged as opposed to subject-specific; and according to data with time-dependent covariates as opposed to time-independent covariates. Standard logistic regression models are appropriate for cross-sectional data. However, for longitudinal data, methods such as generalized estimating equations (GEE) and generalized method of moments (GMM) are commonly used to fit population-averaged models, while random-effects models such as generalized linear mixed models (GLMM) are used to fit subject-specific models. Some of these methods account for time-dependence in covariates while others do not. This paper addressed these approaches with an illustration using a Medicare dataset as it relates to rehospitalization. In particular, we compared results from standard logistic models, GEE models, GMM models, and random-effects models by analyzing a binary outcome for four successive hospitalizations. We found that these procedures address differently the correlation among responses and the feedback from response to covariate. We found marginal GMM logistic regression models to be more appropriate when covariates are classified as time-dependent in comparison to GEE models. We also found conditional random-intercept models with time-dependent covariates decomposed into components to be more appropriate when time-dependent covariates are present in comparison to ordinary random-effects models. We used the SAS procedures GLIMMIX, NLMIXED, IML, GENMOD, and LOGISTIC to analyze the illustrative dataset, as well as unique programs written using the R language.

Key words: Conditional models, estimating equations, longitudinal logistic regression, marginal models, time-dependence.

---

*Corresponding author.

1. Introduction

There are not many suitable models for binary responses taken over time, when the data include correlation between responses and covariates that are time-dependent. Such longitudinal data are useful as they allow the researchers to study the time course of change and the long-term effects of the covariates. They also offer increased statistical power and robustness for model selection (Zeger and Liang, 1992). Hu *et al.* (1998) summarized the modeling of binary outcomes that arise from repeated measures. Most of the models can be grouped into two classes (Zeger *et al.*, 1988; Neuhaus *et al.*, 1991): the "subject-specific" approaches and the "population-averaged" approaches.

Random-effects logistic models (Stiratelli *et al.*, 1984; Wong and Mason, 1985; Lee and Nelder, 1996) are commonly used to estimate subject-specific effects, while the generalized estimating equations (GEE) method of Liang and Zeger (1986) is often used to provide population-averaged effects. Hu *et al.* (1998) claimed that while both the GEE and random effects approaches are extensions of models for independent observations to time-dependent data, they addressed the problem of time-dependency differently. Also, the regression coefficients or odds ratios obtained from the two approaches are numerically different, as are their interpretations (TenHave *et al.*, 1995; McCulloch *et al.*, 2008; Pendergrast *et al.*, 1991). Lee and Nelder (2004) provide a discussion that favors conditional models, arguing that marginal inferences can be made from conditional models. However, Senn (2004) responds that there are many situations in which a statistician would choose to apply a marginal model, including some situations in which the simplification of a marginal model is necessary.

Population-averaged models allow researchers to make conclusions that compare populations defined by different characteristics according to the covariates in a model. Within such models the response at a given time is often expected to be affected by covariates observed at the same time. For a population-averaged logistic regression model, the interpretations of parameter estimates relate to the odds ratio comparing two populations defined by different covariate values. Subject-specific models allow researchers to make conclusions that compare the effects of successive responses by the same subject. Through random-effects, baseline values are allowed to vary by individual, and distinctions can be made between odds ratios defined across subjects and odds ratios defined within multiple responses by individuals. For a subject-specific logistic regression model, the interpretations of parameter estimates relate to the odds ratio comparing two different covariate values for a single subject, given an individual baseline propensity for the response of interest.

The interpretations provided by these two types of models can be affected by the presence of time-dependent covariates. When a model contains covariates

that change through repeated observations of the same subject, it is possible to evaluate directly the effects of different covariate values on the response for individuals, in addition to making comparisons using odds ratios across different populations. However, when time-dependent covariates are present it is important to account for possible feedback effects between responses and covariates at different times. Some of the models we present account for these types of relationships, while others do not.

Figure 1 provides a summary of the models discussed and used in this paper for the illustrative example. We have constructed this decision tree using methods presented previously in the literature. Based on study intentions, researchers have the choice to select between population-averaged and subject-specific models, while also considering whether time-dependent covariates are present. If time-dependent covariates are present, there are many modeling options.
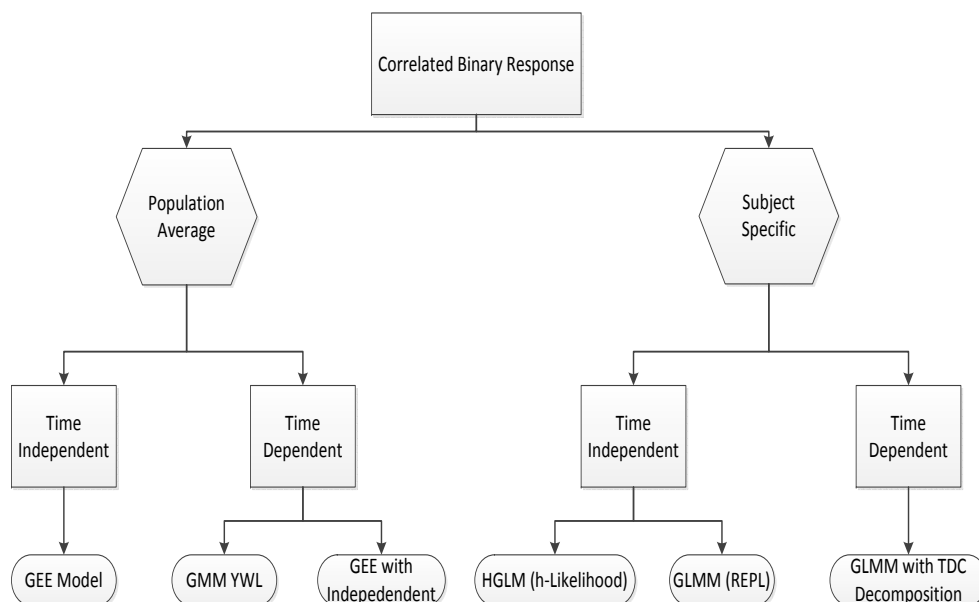

Figure 1: Options for fitting logistic models with correlated binary responses

In this paper we considered the population-averaged and the subject-specific models, as well as time-dependent versus time-independent covariate models. We addressed these modeling decisions in the context of predicting rehospitalization probabilities using a Medicare dataset. For illustrative purposes we utilized data extracted from the Arizona State Inpatient databases pertaining to Medicare beneficiaries admitted to a hospital for a period of four visits. These data contain information on patient discharges from Arizona hospitals between 2003 and 2005. The dataset analyzed includes information on 1,625 patients, each of which has

three responses and complete information on the covariates at each admission time. Our response variable of interest is an indicator of rehospitalization within 30 days of discharge for the same condition for which they were initially hospitalized. Rehospitalizations are relevant because Medicare will pay for all subsequent visits for patients, except in the case where readmission occurs within thirty day for the same procedure (Jencks *et al.*, 2009). In particular, we chose to consider the effects on rehospitalization of total number of diagnoses, length of stay, total number of procedures, and the existence of coronary atherosclerosis. We found these data attractive in that the responses are correlated and the covariates were time-dependent. Yin *et al.* (2013) first examined these data when exemplifying a new procedure to address time-dependent covariates with the generalized method of moments.

In Section 2, we discuss the standard logistic regression model, which ignores correlation of any kind among the observations and the correlation inherent due to the time-dependent covariates. Further, we examine correlated logistic regression models, which model the correlation among the responses but ignore the correlation due to the time-dependent covariates. We present marginal logistic regression models suitable for population-averaged inferences, including standard GEE models and the GMM models which we present for analyzing correlated logistic regression models with time-dependent covariates based on Yin *et al.* (2013). In Section 3, we discuss conditional logistic regression models suitable for subject-specific inferences, such as the random intercept model, the random slopes model, and models including decompositions of time-dependent covariates. In Section 4 we fit correlated logistic regression models to the Medicare data for illustrative purposes. We fit the GEE models using SAS PROC GENMOD, and we fit the GMM models using SAS PROC IML and using programs in R. We fit the random-intercept model and models including decompositions of time-dependent covariates using SAS PROC GLIMMIX. In Section 5, we provide some discussion pertaining to appropriate conclusions, model complexity and convergence, and relative advantages and disadvantages of the options presented.

2. Standard Logistic Regression Model

2.1 Standard Logistic Regression

It is popular as it indirectly models the odds of an outcome through the log-odds. Logistic regression is most often utilized for testing a relationship between a binary response and one or more continuous or categorical predictor variables. It relies on the use of the logit through the natural logarithm of the odds ratio in a linear relation. The logit transformation is applied to the probability of the outcome, making it possible to predict the probability of the response of interest

from a set of independent variables in a linear form such that the population-averaged model is written

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ},$$

where β_j is the coefficient associated with the jth predictor variable, $j = 1, \cdots, J$. This model does not differentiate among the times of the response for each unit, as the standard logistic regression model assumes that all observations are independent. The outcomes Y_i are assumed to follow the Bernoulli distribution with success probability π_i, and as such the variance depends on the mean (Agresti, 2002). Each coefficient β_j can be interpreted as the log-odds ratio for the probability of success associated with two populations that differ by one unit of predictor x_{ij}, with all other predictors constant.

The standard logistic regression model is a member of the class of generalized linear models (GLM) and as such can be modeled as having three components; a random component, a systematic component, and a link component. As a generalized linear model, logistic regression serves as an expansion upon traditional linear models to allow for the analysis of non-normal data. This particular extension is especially useful in analyzing models with binary responses. However, these standard logistic regression models do not maintain precision when the responses are correlated as a result of clustering or repeated measurements taken on a single subject. With correlated observations we need a model with the capability of accounting for the correlation that may be present in the response and possibly in the predictors. As positively correlated binary data show greater variation than independent binary data, such models are often referred to as overdispersed logistic regression models.

2.2 Longitudinal Studies

Longitudinal studies can address how each subject changes over time, and what variables predict differences among subjects and within subjects in their changes over time. In fact, one major advantage of a longitudinal study is its capacity to separate change over time within subjects and differences among subjects (cohort effects) (Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2004). Longitudinal data often contain repeated measurements of each subject at multiple points in time. Such correlated observations are commonly encountered in studies of clinical trials, healthcare surveys, population polling, marketing, educational outcomes effectiveness, and other types of behavioral research. Standard generalized linear models are often inappropriate in analyzing such sample data due to the clustering and hence non-independence, which leads to overdispersion.

For longitudinal data, marginal models are appropriate when inferences about the population average are our primary interest (Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2004) or when we require the expectation of the response variable to be a function of current covariates in order to make future applications of the results (Pepe and Anderson, 1994). Marginal regression models are useful in characterizing the expectation of a response at a specific time as a function of the respective covariates observed at that same time and are useful when the goal of the analysis is to model the population average. In contrast to the population-averaged model, the subject-specific model can distinguish observations belonging to the same or different subjects. Random-effect models are commonly used to estimate subject-specific effects. Two key methods used to estimate the subject-specific effects in the random-effects models: maximum likelihood and conditional likelihood procedures (Diggle *et al.*, 2002; Fitzmaurice *et al.*, 2004).

When dealing with longitudinal data, in addition to the responses changing over time, we can also have covariate values that change over time. Using data with such characteristics, it is possible to model the differing effects of different covariate values on the changing response at an individual level, in addition to making comparisons of effects on the response across populations defined by covariate values. Thus the treatment of time-dependent covariates in the analysis of longitudinal data allows strong statistical inferences about dynamic relationships and provides more efficient estimators than can be obtained using cross-sectional data (Zeger and Liang, 1992; Hedeker and Gibbons, 2006). For this reason we will consider the effectiveness of both marginal and conditional logistic regression models with respect to time-dependent covariates.

3. Marginal Correlated Logistic Regression

In the following sections we present some marginal logistic regression models for correlated data. For the purposes of discussion, we will describe interpretations with respect to longitudinal data over time instead of the more general correlated data situation. For subject i, let $\boldsymbol{y}_i = [y_{i1}, \cdots, y_{iT_i}]^T$ be a $T_i \times 1$ vector of binary responses with associated design matrix

$$\boldsymbol{X}_i = \begin{bmatrix} x_{i11} & \cdots & x_{i1J} \\ & \vdots & \\ x_{iT_i1} & \cdots & x_{iT_iJ} \end{bmatrix},$$

where $t = 1, \cdots, T_i$ denotes different times and $j = 1, \cdots, J$ denotes different covariates. For subject i at time t the row vector $\boldsymbol{x}_{it.} = [x_{it1}, \cdots, x_{itJ}]$ gives the J covariate values, and for the jth covariate for subject i the column vector $\boldsymbol{x}_{i.j} = [x_{i1j}, \cdots, x_{iT_ij}]^T$ gives the T_i covariate values across times. The full response

vector for all N subjects is given by the column vector $\boldsymbol{y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N]^T$, and the full design matrix is similarly given by $\boldsymbol{X} = [\boldsymbol{X}_1^T, \cdots, \boldsymbol{X}_N^T]^T$. Notice that each design matrix $\boldsymbol{X}_i$ may have a unique number of rows, as the number of times observed for each subject may differ.

3.1 Generalized Estimating Equations

The generalized estimating equations (GEE), as presented by Zeger and Liang (1986) and Liang and Zeger (1986), applies to marginal models for longitudinal binary data. An important aspect of this approach is the specification of a working correlation structure by the researcher. The working correlation structure represents the correlation believed to be present among responses within subjects, and as such is incorporated into the random component of the model. For subject i, let the working correlation structure be denoted by $\boldsymbol{R}_i(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ is a $s \times 1$ vector of correlation parameters that fully describes the working correlation; i.e. no other parameters are necessary. When fitting a logistic regression model

$$\ln\left(\frac{\pi_{it}}{1 - \pi_{it}}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ},$$

while accounting for the autocorrelation among responses, the marginal response variance $\boldsymbol{V}_i(\boldsymbol{\alpha})$ for subject i can be defined in terms of the working correlation $\boldsymbol{R}_i(\boldsymbol{\alpha})$,

$$\boldsymbol{V}_i(\boldsymbol{\alpha}, \phi) = \phi \boldsymbol{A}_i^{1/2} \boldsymbol{R}_i(\boldsymbol{\alpha}) \boldsymbol{A}_i^{1/2},$$

where $\boldsymbol{A}_i$ is a diagonal matrix representing the response variance under the assumption of independence and ϕ is the overdispersion factor. Thus the generalized estimating equations for N independent subjects

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(\frac{\partial \pi_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\alpha}}(\boldsymbol{\beta}), \hat{\phi}(\boldsymbol{\beta}))(\boldsymbol{y}_i - \pi_i(\boldsymbol{\beta})) = \boldsymbol{0},$$

with the dispersion parameters $\boldsymbol{\alpha}$ and ϕ. Solving these estimating equations provides parameter estimates $\hat{\boldsymbol{\beta}}$. Each coefficient β_j can be interpreted similarly to those of the standard logistic regression model, with the added condition that the autocorrelation has been accounted for (Zeger and Liang, 1992).

Liang and Zeger (1986) showed that when $\boldsymbol{R}_i(\boldsymbol{\alpha}) = \boldsymbol{I}$, the GEE can be simplified to the score functions as from a likelihood function that assumes independence among repeated observations from a subject. The GEE estimates for $\boldsymbol{\beta}$ are consistent regardless of the choice of working correlation structure for time-independent covariates, although a correct specification of the working correlation structure does enhance efficiency. Liang and Zeger (1986) established

that the vector $\hat{\boldsymbol{\beta}}$ that satisfies $U(\boldsymbol{\beta}) = \mathbf{0}$ is asymptotically unbiased in the sense that $\lim_{N \to \infty} (\mathrm{E}_{\boldsymbol{\beta}} [U(\boldsymbol{\beta})]) = \mathbf{0}$, under suitable regularity conditions. Diggle *et al.* (2002) showed that the GEE approach is usually satisfactory when the data consist of short, essentially complete, sequences of measurements observed at a common set of times on many subjects, and a conservative selection in the choice of a working correlation matrix is applied.

In using GEE, the specified covariance structure takes into consideration the correlation that arises as a result of repeated measurements on the same subject being related to each other, or due to clustering (Zeger and Liang, 1986; Liang and Zeger, 1986; Diggle *et al.*, 2002; Smith and Smith, 2006). In fact, the GEE estimates with time-independent covariates produce efficient estimates if the working correlation structure is correctly specified, and remain consistent as well as providing correct standard errors if the working correlation structure is incorrectly specified. Thus it is known that GEE can be used to appropriately model correlated binary outcomes when there are time-independent covariates. However, when there are time-dependent covariates, Hu *et al.* (1998) and Pepe and Anderson (1994) have pointed out that the consistency of parameter estimates using GEE is not assured with arbitrary working correlation structures unless that a subject repeated measurements are independent, i.e., the independent working correlation is satisfied as employed. Pepe and Anderson (1994) suggested the use of the independent working correlation structure when using GEE with time-dependent covariates as a "safe" choice of analysis. However, Fitzmaurice (1995) discussed the losses in efficiency that arise from using the independent working correlation structure with GEE when the data are, in fact, not independent.

3.2 Generalized Method of Moments

The generalized method of moments (GMM), when time-dependent covariates are involved, can provide more efficient estimates than using the GEE estimates based on the independent working correlation (Lai and Small, 2007). Lai and Small (2007) maintained that the GEE approach with time-independent covariates is an attractive approach as it provides consistent estimates under all correlation structures for subjects repeated measurements. However, they showed through a simulation study that when there are time-dependent covariates, some of the estimating equations applied by using the GEE method with an arbitrary working correlation structure are not valid. In other words, some estimating equations will not have zero expected value. The "safe" choice of independent working correlation structure with GEE has been shown to produce inefficient estimates when time-dependent covariates are present (Fitzmaurice, 1995). Therefore, GMM has become the preferred method of estimation for

marginal correlated logistic regression models when time-dependent covariates are present.

The GMM method generalizes the standard method of moments, which involves constructing estimating equations by setting the expectation of known functions of observable random variables equal to known functions of unknown parameters Hansen (1982). The GMM estimators have become very popular because they have properties that are easy to characterize given a large sample, which makes comparisons relatively easy. Further, these methods can be applied without indicating the full data generating process. Therefore, GMM methods of estimation can be adapted to a wide variety of applications (Hansen, 2007).

The general process of GMM estimation involves forming estimating equations as weighted linear combinations of "valid" moment conditions with zero expected value (Hansen, 1982, 2007). It utilizes a positive definite weight matrix that assigns different levels of importance to the moment conditions based upon how informative each moment is with respect to the parameters $\boldsymbol{\beta}$. These moment conditions consist of products of residual terms $(y_{it} - \mu_{it})$ at time t and covariate terms $\partial \mu_{is}/\partial \beta_j$ at time s. Parameter estimates $\hat{\boldsymbol{\beta}}_{GMM}$ are obtained by minimizing the quadratic form $Q(\boldsymbol{\beta}) = \boldsymbol{G}^T \boldsymbol{W}^{-1} \boldsymbol{G}$, where $\boldsymbol{G}$ is a vector of valid moment conditions $(\partial \mu_{is}(y_{it} - \mu_{it}))/\partial \beta_j$ and $\boldsymbol{W}$ is a weight matrix typically chosen to be $\text{Cov}(\boldsymbol{G})$. The resulting parameter estimates $\hat{\beta}_j$ have the same interpretations as with the GEE method.

Lai and Small (2007) used GMM estimators with time-dependent covariates by selecting the linear combinations of moment conditions according to the nature of the time-dependence. They defined Type I time-dependent covariates as those covariates that do not involve any effects between covariate process and response process, and consequently products of residual and covariate terms at all times t and s are valid. Type II time-dependent covariates are defined as those that may involve feedback from covariates to future responses, and so only residual and covariate products with $s \geq t$ are valid. Type III time-dependent covariates are defined as those that may involve feedback from responses to future covariate values, and so only residual and covariate products with $s = t$ are valid. Lai and Small (2007) showed gains in efficiency over the "safe" choice of independent GEE.

Yin *et al.* (2013) presented an extension of the approach of Lai and Small (2007), first defining a Type IV time-dependent covariate in which responses may affect future covariate values (but not the converse), so only products with $s \leq t$ are valid. Additionally, Yin *et al.* (2013) proposed an extended classification method in which the data determine individual combinations of residual and covariate terms that form valid moment conditions instead of researcher-selected classifications. Their process follows. In their process, it is of interest to fit a

logistic regression model

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ},$$

while accounting for the autocorrelation among responses. They presented a method of selecting valid moment conditions for time-dependent covariates based on the observed correlation between residuals and covariate values. Let e_{it} denote the residual for subject i from a preliminary model fit using only the data from time t, and let $d_{isj} = \partial\mu_{is}/\partial\beta_j$ represent the covariate term evaluated using the parameter estimates from the same preliminary model, so that T different preliminary models are fit, one for each time observed. Define ρ_{sjt} to be the linear correlation between the standardized errors $\boldsymbol{e}_t = (e_{1t}, \cdots, e_{Nt})^T$ and the standardized covariate values $\boldsymbol{d}_{sj} = (d_{1sj}, \cdots, d_{Nsj})^T$ for the jth covariate. Under certain regularity conditions, Yin *et al.* (2013) showed that

$$\frac{\hat{\rho}_{sjt}}{\sqrt{\hat{\mu}_{22}/N}} \sim \mathcal{N}(0,1),$$

where N is the number of subjects, and $\hat{\mu}_{22}$ is the estimated mixed fourth moment of $\boldsymbol{e}_t$ and $\boldsymbol{d}_s$ for the jth covariate,

$$\hat{\mu}_{22} = (1/N)\sum_i (\tilde{d}_{sji})^2 (\tilde{e}_{ti})^2.$$

In this way the correlation between residual terms and covariate terms can be evaluated directly. Any pair with non-significant correlation will be treated as forming a valid moment condition within the GMM process. One the valid moment conditions have been selected, estimation can proceed using either two-step GMM (2SGMM) or continuously updating GMM (CUGMM) (Hansen *et al.*, 1996). 2SGMM proceeds by estimating the GMM weight matrix and parameter estimates in two separate steps. CUGMM proceeds by maximizing a single expression for both the weight matrix and the parameters of interest simultaneously.

3.3 Model Fitting

The marginal models presented in this section can be fit in some statistical software, including SAS and R. The GEE models can be fit in SAS using PROC GENMOD (Smith and Smith, 2006) and in R using the GEE function within the packages GEE or GEEPACK. Our GMM models were fit using programs written using SAS PROC IML and also using programs in R.

4. Conditional Correlated Logistic Regression

4.1 Random Intercept Models

Consider the logistic regression model for binary responses with a random intercept,

$$\ln\left(\frac{\pi_{it}}{1 - \pi_{it}}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_J x_{iJ} + \gamma_i,$$

where γ_i is a random effect associated with the clustering by subjects. It is customary to think of these random effects as distributed normally with mean 0 and variance σ_γ^2. Because this random effect is additive within the model and is not associated with any covariates, it is often referred to as a random intercept. When we include a random intercept in the model, the interdependencies among the repeated observations within subjects are fully taken into account. The variance-covariance structure is analogous to the "compound symmetry" form assumed in a generalized linear mixed model, and also the "exchangeable working correlation" in the GEE model. Hu *et al.* (1998) pointed out that with the standard logistic model, the baseline risk is simply the proportion of positive responses in the control group at baseline, while in the random-intercept logistic model, the baseline risk is assumed to follow a distribution. Therefore, the corresponding change in absolute risk with and without the covariate varies from one subject to another, depending on the baseline rate. In this sense conditional models that include random subject terms are referred to as "subject-specific" models and lend themselves to such interpretations.

Consequently, the odds ratios estimated from a random-effect logistic model are adjusted for the heterogeneity of the subjects, which can be considered to be due to unmeasured variables. In the illustrative Medicare example we can think of the random intercept as a patient's constant propensity to be rehospitalized across the four time points of the study, which is independent of the effects of the time-dependent covariates. If we were to include additional random subject effects into the model, then this propensity will be allowed to vary across time and any other factors included, Hu *et al.* (1998). As a consequence of such properties, the random effects are sometimes thought of as omitted subject-dependent covariates, Longford (1994).

4.2 Random Slopes Models

The random slopes model can be thought of as including an additional random error term for the intercept of the model. As such it is also reasonable to include random error terms associated with the coefficients of each of the predictors. In

this case the model is referred to as a "random slopes" model,

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = (\beta_0 + \gamma_{0i}) + (\beta_1 + \gamma_{1i})x_{i1} + \cdots + (\beta_J + \gamma_{Ji})x_{iJ},$$

where each γ_{ji} represents an "error" associated with a model coefficient. This type of model is also commonly referred to as a hierarchical logistic regression model and may be written using multiple equations,

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i} + \beta_{1i}x_{i1} + \cdots + \beta_{Ji}x_{iJ},$$
$$\beta_{ji} = \delta_{j0} + \gamma_{ji}.$$

In this formulation δ_{j0} can be thought of as an intercept term in the model for β_{ji}, with random error γ_{ji}. Using this notation it is clear that additional predictors can be included to model each regression coefficient. In this way the regression coefficient can be modeled over time, thus allowing for a subject's propensity of success to change over time with a time-dependent covariate. Specifically, addition of a time covariate into the regression model would allow for a covariate-averaged time adjustment,

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i} + \beta_{1i}x_{i1} + \cdots + \beta_{Ji}x_{iJ} + \beta_t t,$$
$$\beta_{ji} = \delta_{j0} + \gamma_{ji},$$

while inclusion of a time covariate into the coefficient models would allow for a covariate-specific time adjustment,

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_{0i} + \beta_{1i}x_{i1} + \cdots + \beta_{Ji}x_{iJ},$$
$$\beta_{ji} = \delta_{j0} + \delta_t t + \gamma_{ji}.$$

In this way a random slopes model can account for time-dependent covariates within a longitudinal logistic regression model.

4.3 Decomposition of Time-Dependent Covariates

While the random intercept and random slopes models can be used to effectively account for interdependencies among responses within subjects, and also to account for subject-to-subject heterogeneity due to potential latent variables, the models may not properly account for the different effects of time-dependent covariates. Specifically, Neuhaus and Kalbfleisch (1998) argued that the standard subject-specific model will produce an odds ratio for time-dependent covariates

that is challenging to interpret. Its effect will be an unknown combination of the effect of varying covariate values within subjects, and the effect of varying covariate values between subjects. To identify these different time-dependent covariate effects, Neuhaus and Kalbfleisch proposed a decomposition of any time-dependent covariate factor in a model into two terms; one term accounting for the "within" variation and the other term accounting for the "between" variation. Then the random intercept logistic regression model is

$$\ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + (\beta_{1B}\bar{x}_{i.1} + \beta_{1W}(x_{it1} - \bar{x}_{i.1})) + \cdots$$
$$+ (\beta_{JB}\bar{x}_{i.J} + \beta_{JW}(x_{itJ} - \bar{x}_{i.J})) + \gamma_i,$$

where each β_{jB} corresponds to the "between" contribution of the time-dependent covariate, and β_{jW} corresponds to the "within" contribution. Time-independent covariates are analyzed without any change, as there will be no variation "within". The coefficient β_{jB} is associated with the change in log-odds for subjects from different populations defined by x_{itj}, while β_{jW} is associated with the change in log-odds for a single subject at different values of x_{itj}. For example, consider the illustrative Medicare example, in which we model the probability of rehospitalization within 30 days, using the time-dependent cocovariate number of procedures. The associated between-subjects coefficient β_{jB} would represent the expected change in the log-odds of rehospitalization for different populations of subjects, for a unit increase in the average number of procedures. The associated within-subjects coefficient β_{jW} would represent the expected change in the log-odds of rehospitalization, for a single subject, for a unit increase in the number of procedures on different hospital visits.

Neuhaus and Kalbfleisch argued that, without this decomposition, the single parameter associated with any time-dependent covariate will be a biased combination of the within and between parameters. Neuhaus and Kalbfleisch (1998), and also Scott and Holt (1982), have shown that for time-dependent covariates with equivalent subject averages ($\bar{x}_{i.j} = \bar{x}_{..j}$ for all i, no between-subject variation), the effect of the covariate will be completely described by β_{jW}. Similarly, for time-dependent covariates with individual subject values equivalent to the subject average ($x_{itj} = \bar{x}_{i.j}$ for all i, no within-subject variation), the effect of the covariate will be completely described by β_{jB}.

4.4 Model Fitting

Conditional logistic regression models can be fitted using SAS and R. In SAS, PROC GLIMMIX can be used to fit the conditional models, using the restricted pseudo-likelihood (REPL) (Wolfinger and O'Connell, 1993; Dai *et al.*, 2006). Alternatively, PROC NLMIXED can be used directly to fit a nonlinear mixed

model (Wolfinger, 2000). In R, the GLMER function within the LME4 package can be used to fit the models using REPL, which implicitly assumes normal random effects. Additionally, the packages HGLMMM, HGLM, and DHGLM can be used to apply the h-Likelihood of Lee and Nelder (1996) and Lee *et al.* (2006). The function HGLM can be applied from the package HGLM, and the function DHGLMMODELING can be applied from the package DHGLM.

5. Illustrative Example: Rehospitalization

5.1 Modeling Rehospitalization Data

For illustrative purposes, we revisited data from the Arizona State Inpatient Database (SID), Yin *et al.* (2013). The dataset contained patient information from Arizona hospital discharges for 3-year period from 2003 through 2005, of those who were admitted to a hospital exactly 4 times. There were 1625 patients in the dataset with complete information; each has three observations indicating three different times to rehospitalizations. We classified those who returned to the hospital within 30-days as "one" opposed to "zero" for those who did not return within 30 days. Table 1 provides the percentage of the patients who were readmitted to the hospital within 30 days of discharge against the percentages of the patients who were not readmitted for each of their first three hospitalizations.

Table 1: Cross-classification of re-admit by time

		Time			Total
		1	2	3	
Re-Admit	**No**	231 46.48%	272 54.73%	253 50.91%	756
	Yes	266 53.52%	225 45.27%	244 49.09%	735

For the sake of simplicity, we ignored the intraclass correlations due to hospitals. In particular, in our models we chose to consider the following as predictors of the probability of rehospitalization within 30 days: total number of diagnoses, total number of procedures performed, length of patient hospitalization, the existence of coronary atherosclerosis, and indicators for time 2 and time 3.

5.2 Marginal Logistic Regression Models for Rehospitalization

In this section we present marginal models for the probability of rehospitalization within 30 days. Some of the models account only for the correlation inherent in repeated observation of individuals, while others additionally included the

feedback due to the time varying aspect of some covariates. We present a standard logistic regression model, a correlated model fit using GEE with compound symmetry, unstructured, and independent working correlation structures, and a correlated model fit using GMM as presented by Yin *et al.* (2013), using both continuously updating GMM (CUGMM) as well as two-step GMM (2SGMM).

We first utilized a standard logistic regression model to analyze the data. This model assumes that the information used was taken from 4,875 independent observations. However, we have 1,625 independent sampling units, each measured three times. Our results are provided in Table 2 in the column labeled "Standard" (* indicates significance at the 0.05 level, ** indicates significance at the 0.01 level, and *** indicates significance at the 0.001 level). We found that length of hospitalization and number of diagnoses each have a significant increasing impact on rehospitalization. Specifically, a population with an increase of one night of length of hospitalization over another population gives an odds ratio of $\exp(0.0344) \approx 1.067$ for probability of rehospitalization, while a population with an increase of one diagnosis gives an odds ratio of $\exp(0.0648) \approx 1.035$. We see significant differences between the first and second times, and the first and third times. Number of procedures showed marginal significance at the 0.10 level, but this result is not compelling as these results provided standard errors that were smaller than expected because of the independence model that was used.

Table 2: Parameter estimates and standard errors. Marginal logistic regression models

	Parameter Estimate (Standard Error)					
	Standard	**UGEE**	**CSGEE**	**IGEE**	**CUGMM**	**2SGMM**
Number of Diagnoses	0.0648 (0.0154)***	0.0686 (0.0160)***	0.0664 (0.0160)***	0.0648 (0.0160)***	0.0543 (0.0154)***	0.0642 (0.0151)***
Number of Procedures	-0.0306 (0.0186)	-0.0272 (0.0190)	-0.0268 (0.0190)	-0.0306 (0.0192)	-0.0453 (0.0187)*	-0.0315 (0.0187)
Length of Hospitalization	0.0344 (0.0056)***	0.0314 (0.0075)***	0.0314 (0.0075)***	0.0344 (0.0077)***	0.0531 (0.0058)***	0.0396 (0.0049)***
Existence of C. A.	-0.1143 (0.0913)	-0.1260 (0.0934)	-0.1327 (0.0934)	-0.1143 (0.0937)	0.0133 (0.0942)	-0.0517 (0.0929)
Time 2	-0.3876 (0.0716)***	-0.3868 (0.0710)***	-0.3859 (0.0710)***	-0.3876 (0.0711)***	-0.4419 (0.0695)***	-0.3840 (0.0672)***
Time 3	-0.2412 (0.0721)***	-0.2390 (0.0688)***	-0.2390 (0.0688)***	-0.2413 (0.0688)***	-0.2674 (0.0683)***	-0.2686 (0.0672)***
QIC		6648.73	6648.75	6648.52		
QICu		6646.87	6646.87	6646.56		

To address the non-independence of the observations, we allowed the correlation in the response to be accounted for by applying marginal logistic regression models with different working correlation structures fit using GEE. The unstructured and compound symmetry working correlation structures were applied. Our findings are given in Table 2 in the columns labeled "UGEE" and "CSGEE". The signs of parameter estimates and their interpretations remain consistent, as does the significance of number of diagnoses and length of hospitalization. We continue to see differences in the probability of rehospitalization among the three times. The marginal significance of number of procedures is no longer present, as the standard errors have been corrected to account for repeated observation of patients. However, these estimates and standard errors do not account for possible feedback due to the time-dependent covariates.

To account for the time-dependent covariates and the associated feedback we fitted a marginal GEE logistic regression model with independent working correlation structure, presented in Table 2 in the column labeled "IGEE". While this method is an appropriate choice for logistic regression with time-dependent covariates, independent GEE does not take advantage of all possible estimating equations and as a consequence can lack efficiency as compared to the GMM approach. Therefore we also fitted a marginal correlated logistic regression model using GMM methods, Lai and Small (2007), also based on the recent work of Yin *et al.* (2013). The GMM estimates were obtained using both CUGMM and 2SGMM, presented in Table 2 in the columns labeled "CUGMM" and "2SGMM". The parameter estimates and interpretations remain consistent with other models, with the exception of a change in sign on the coefficient for existence of coronary atherosclerosis using CUGMM, which was not found significant. There is a significant increase in the probability of rehospitalization with increased number of diagnoses and length of hospitalization. There are also differences across the three times. Using the GMM approach, there is a significant decrease in probability of rehospitalization associated with the number of procedures using CUGMM, and marginal significance using 2SGMM. The method of Yin *et al.* (2013) applies a greater number of valid moment conditions with respect to number of procedures than when using independent GEE.

Comparison of GEE models is often made using information criteria such as QIC and QICu (Panm, 2001). In this case both QIC and QICu are nearly identical for all three correlation structures, with the lowest values associated with the independent GEE model. This suggests the independent GEE model is most appropriate among the GEE models, as expected for data with time-dependent covariates. The QIC summary measure is not appropriate for GMM estimation, and in fact there is no current fit statistic appropriate for both GEE and GMM estimation. But Lai and Small (2007) have shown GMM methods may gain effi-

ciency over GEE in the presence of time-dependent covariates, if the correlation within subjects is sufficiently large or if the working correlation structure has been misspecified. Therefore we prefer GMM as a marginal estimation method for data with strong within-subject correlation, with CUGMM slightly preferred over 2SGMM due to more consistent convergence properties. When there is not strong within-subject correlation, we prefer independent GEE.

5.3 Conditional Logistic Regression Models for Rehospitalization

In this section we present the results of fitting conditional logistic regression models to the Medicare rehospitalization data. We present models with a random intercept, with a random intercept and a random slope, and with a random intercept that includes a decomposition of the time-dependent covariates.

First consider a subject-specific logistic regression model with a random intercept term. In fitting this model we allowed each patient to have his/her own level of propensity but with common correlation. The results of this analysis are given in Table 3 in the column labeled "Random Intercept". We found that number of diagnoses and length of hospitalization were both significant and associated with an increase in probability of rehospitalization. Accounting for different baseline propensity for rehospitalization, a unit increase in length of hospitalization for an individual gives an odds ratio of $\exp(0.0327) \approx 1.033$ for probability of rehospitalization, while a unit increase in diagnoses for an individual gives an odds ratio of $\exp(0.0706) \approx 1.073$. There are significant differences across different times. Further we found that the random intercept variation was also significant, indicating that it is necessary to allow the patients to have different baseline propensities for rehospitalization.

In addition we wanted the model to allow the rate of change associated with length of hospitalization to be allowed to vary among patients, thus we included a random slope for this variable. The results of this analysis are given in Table 3 in the column labeled "Random Slope". When accounting for both the random slope as well as the random intercept we found that the number of diagnoses and length of hospitalization were both significant. Both times remain significant. We also found that the intercept variation and the length of hospitalization slope variation were significant at the 0.05 level in this model, indicating that it is necessary to allow the patients to have different baseline propensities for rehospitalization and to allow patients to have different impacts of length of hospitalization.

In order to further investigate the dynamic relationships between time-dependent covariates and the probability of rehospitalization, a random intercept model was fit including a decomposition of each time-dependent covariate into "within" and "between" components. Results of this analysis are presented in Table 3 in the columns labeled "TDC (Within)" and "TDC (Between)". The

Table 3: Parameter estimates and standard errors. Conditional logistic regression models

	Parameter Estimate (Standard Error)				
	Standard	**Random Intercept**	**Random Slope**	**TDC (Within)**	**TDC (Between)**
Number of Diagnoses	0.0648 (0.0154)***	0.0663 (0.0159)***	0.0589 (0.0167)***	0.0780 (0.0220)***	0.0444 (0.0229)
Number of Procedures	-0.0306 (0.0186)	-0.0279 (0.0191)	-0.0354 (0.0201)	0.0188 (0.0251)	-0.0824 (0.0302)**
Length of Hospitalization	0.0344 (0.0056)***	0.0322 (0.0056)***	0.0498 (0.0082)***	0.0008 (0.0074)	0.0736 (0.0085)***
Existence of C. A.	-0.1143 (0.0913)	-0.1278 (0.0937)	-0.0992 (0.0981)	-0.2607 (0.1270)*	0.2223 (0.1435)
Time 2	-0.3876 (0.0716)***	-0.3881 (0.0718)***	-0.4206 (0.0748)***	-0.3730 (0.0727)***	
Time 3	-0.2412 (0.0721)***	-0.2405 (0.0723)***	-2588 (0.0750)***	-0.2130 (0.0740)**	
Intercept Variance	–	0.1578 (0.0532)**	0.1748 (0.0826)*	0.1472 (0.0541)**	
Slope Variance	–	–	0.0025 (0.0013)*	–	
Generalized χ^2/DF	1.19*	0.98	0.96	0.98	

number of diagnoses remains significant, but only within patients. This suggests that a change in the number of diagnoses on successive visits has a significant and increasing impact on probability of rehospitalization for an individual patient, but two individuals from populations with different mean numbers of diagnoses are not expected to differ significantly in probability of rehospitalization. On the other hand, the significance of length of hospitalization remains, but only between subjects. This suggests that populations with different mean lengths of hospitalization are expected to differ significantly in probability of rehospitalization, but for an individual patient an increase in length of hospitalization on successive visits does not have an impact on rehospitalization. Additionally, significance of number of procedures is found between subjects, implying that populations with different mean numbers of procedures will differ significantly in probability of rehospitalization. We found significance for existence of coronary

atherosclerosis, within subjects only. For an individual patient, development of coronary atherosclerosis on a follow-up visit is associated with a significant decrease in probability of rehospitalization. This combination of subject-specific and population-averaged conclusions was not possible using previous fitting methods.

Comparisons of conditional models using likelihood calculations such as restricted log-likelihood or various information criteria should not be made because the estimates and consequently the likelihood values are based on different pseudo-likelihoods (Wolfinger and O'Connell, 1993). For this reason we prefer Generalized χ^2/DF as a fit statistic for comparing these mixed logistic models. Each of the three mixed logistic models presented has similar quality of fit based on generalized chi-square divided by degrees of freedom. The value reported for the standard logistic regression model is simply the Pearson χ^2/DF, not the Generalized χ^2/DF, and shows evidence of overdispersion. The choice between a model with decomposed time-dependent covariates and one without depends on the relative trade-off between model parsimony and flexibility of available conclusions. We prefer to use the random-intercept model with decomposed time-dependent covariates because both subject-specific and population-averaged conclusions are readily made. In this specific case, the decomposed time-dependent covariate model shows that the significance of number of diagnoses is only within subjects, while the significance of length of hospitalization is only between subjects. Additionally, the decomposition allows us to see the significance of number of procedures between subjects only, and also the significance of existence of coronary atherosclerosis within subjects only. These specific conclusions cannot be reached using the models without decomposed of time-dependent covariates, and may be worth the cost of four additional parameters in the model.

6. Conclusions

We presented a survey of methods for fitting models to longitudinal data with time-dependent covariates. Longitudinal models were classified according to marginal models and conditional models, which differ with respect to parameter interpretations. These models also differ with respect to handling time-dependent covariates.

As an informative example, we analyzed data consisting of repeated measurements giving rise to correlated responses with covariates that are time-dependent. We found that the standard logistic regression model was inadequate for modeling this type of data as it is not able to account for repeated measures on a single subject, resulting in standard errors that were smaller than expected. Through our illustrative example we showed how the parameters of both population-averaged logistic regression models and subject-specific logistic regression models can be

interpreted. Consistent with the statistical literature our results show that regression models ignoring the time-dependence of predictors tend to overestimate the standard errors of time-dependent covariates.

The population-averaged approaches discussed include the moment-based methods GEE and GMM with extended classification. These models allow researchers to make interpretations based on comparisons of populations similar to the interpretations of standard logistic regression coefficients. These methods have the added benefit of accounting for the correlation in the response, which does not affect the interpretations. Consistent with previous literature, the GMM fitting methods showed evidence of smaller standard errors of parameter estimates as compared to other methods that accounted for the time-dependent nature of the data.

The subject-specific approaches discussed include the generalized linear mixed models with random-intercept, random slopes, and decomposition of time-dependent covariates. These models allow researchers to make interpretations based on comparisons of multiple responses of the same individual. The resulting odds ratios are averaged over all subjects observed, but are presented as conditional on random-effects representing varying baseline values. These methods account for the correlation in the response, but the assumed random-effect distributional properties affect parameter estimates and thus interpretations, and these distributional assumptions are difficult to assess. The random-intercept model with decomposed time-dependent covariates included the largest number of parameters, but also provided the most informative conclusions about varying relationships between time-dependent covariates and response. Using this decomposition, significant covariates from other subject-specific models were shown to have significance either within subjects or between subjects, but not both within and between. Additionally, significance of some variables was detected using the decomposition where it was not detected using other subject-specific models.

The choice between population-averaged and subject-specific models is often driven by desired interpretations. Both types of models can be appropriate for longitudinal data with time-dependent covariates. For our example data set, the conditional model with decomposed time-dependent covariates revealed the nature of the dynamic relationships between different time-dependent covariates and response.

Acknowledgements

References

Agresti, A. (2002). *Categorical Data Analysis*, 1st edition. Wiley, Hoboken, New Jersey.

Dai, J., Li, Z. and Rocke, D. (2006). Hierarchical logistic regression modeling with SAS GLIMMIX. In *Proceedings of Western Users of SAS Software Conference.* Cary, North Carolina.

Diggle, P. J., Heagerty, P., Liang, K. Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd edition. Oxford University Press, New York.

Fitzmaurice, G. M. (1995). A caveat concerning independence estimating equations with multiple multivariate binary data. *Biometrics* **51**, 309-317.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis*, 1st edition. Wiley, Hoboken, New Jersey.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50**, 1029-1054.

Hansen, L. P. (2007). Generalized method of moments estimation. Department of Economics, University of Chicago, 1-14. http://home.uchicago.edu/llian/paper/GMM_estimation.pdf.

Hansen, L. P., Heaton, J. and Yaron, A. (1996). Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics* **14**, 262-280.

Hedeker, D. and Gibbons, R. D. (2006). *Longitudinal Data Analysis.* Wiley, New York.

Hu, F. B., Goldberg, J., Hedeker, D., Flay, B. R. and Pentz, M. A. (1998). Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. *American Journal of Epidemiology* **147**, 694-703.

Jencks, S., WIlliams, M. and Coleman, E. (2009). Rehospitalization among patients in the medicare fee-for-service program. *New England Journal of Medicine* **360**, 1418-1428.

Lai, T. L. and Small, D. (2007). Marginal regression analysis of longitudinal data with time-dependent covariates: a generalized method-of-moments approach. *Journal of the Royal Statistical Society, Series B* **69**, 79-99.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B* **58**, 619-678.

Lee, Y. and Nelder, J. A. (2004). Conditional and marginal models: another view. *Statistical Science* **19**, 219-228.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*, 1st edition. Chapman & Hall/CRC, Boca Raton, Florida.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.

Longford, N. T. (1994). Logistic regression with random coefficients. *Computational Statistics and Data Analysis* **17**, 1-15.

McCulloch, C. E., Searle, S. R. and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd edition. Wiley Series in Probability and Statistics, Hoboken, New Jersey.

Neuhaus, J. M. and Kalbfleisch, J. D. (1998). Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* **54**, 638-645.

Neuhaus, J. M., Kalbfleisch, J. D. and Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review* **59**, 25-35.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics* **57**, 120-125.

Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M. and Fisher, M. R. (1991). A survey of methods for analyzing clustered binary response data. *International Statistical Review* **64**, 89-118.

Pepe, M. S. and Anderson, G. L. (1994). A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data. *Communications in Statistics - Simulation and Computation* **23**, 939-951.

Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848-854.

Senn, S. (2004). Comment of "Conditional and marginal models: another view". *Statistical Science* **19**, 228-230.

Smith, T. and Smith, B. (2006). PROC GENMOD with GEE to analyze correlated outcomes data using SAS. Department of Defense Center for Deployment Health Research, California.

Stiratelli, R., Laird, N. and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics* **40**, 961-971.

Ten Have, T. R., Landis, J. R. and Weaver, S. L. (1995). Association models for periodontal disease progression: a comparison of methods for clustered binary data. *Statistics in Medicine* **14**, 413-429.

Wolfinger, R. D. (2000). Fitting nonlinear mixed models with the new NLMIXED procedure. In *Proceedings of the 24th Annual SAS Users Group International Conference.* Cary, North Carolina.

Wolfinger, R. D. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood approach. *Journal of Statistical Computation and Simulation* **48**, 233-243.

Wong, G. Y. and Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association* **80**, 513-524.

Yin, J., Wilson, J. R. and Lalonde, T. L. (2013). Correlated GMM logistic regression models with time-dependent covariates and valid estimating equations. Submitted to *Statistics in Medicine.*

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121-130.

Zeger, S. L. and Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine* **11**, 1825-1839.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics* **44**, 1049-1060.

Trent L. Lalonde
Department of Applied Statistics and Research Methods
University of Northern Colorado
McKee 520, Campus Box 124, Greeley, CO 80639, USA
trent.lalonde@unco.edu

Anh Q. Nguyen
Seller Risk Decision Analytics
PayPal
9999 North 90th Street, Scottsdale, AZ 85258, USA
anguyen2@paypal.com

Jianqiong Yin
Program of Statistics
Arizona State University
Tempe, AZ 85287, USA
Jianqiong.Yin@azahcccs.gov

Kyle Irimata
Department of Economics
Arizona State University
Tempe, AZ 85287, USA
Kyle.Irimata@asu.edu

Jeffrey R. Wilson
Department of Economics
Arizona State University
Tempe, AZ 85287, USA
jeffrey.wilson@asu.edu